DOI : https://doi.org/10.56452/6-3-675

ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

*International Journal of Mechanical Engineering*

# Comparison of Feature Selection Techniques for Improved Diabetes Prediction using Random Forest

**Kajal Aggarwal**

Asst. Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

## Abstract

Getting the early diagnosis and proper treatment of diabetes is very important for people with this chronic condition. Machine learning can help predict the disease, but it is not always accurate. The goal of this study was to analyze the various features selection techniques used in predicting diabetes. They were: Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Chi-Square Test, and Correlation-based Feature Selection (CFS), for improving diabetes prediction using random forest. We used the publicly-available data set from the NIH for the analysis. It included 768 specimens and eight features (Pima Indian Diabetes dataset). The data collected for the study included 8 features and 768 samples. After processing the data, we evaluated the performance of the four feature selection methods. We also compared the results with the importance ranking of the selected features to identify the underlying biological factors. The results of the study revealed that the four different techniques used improved the performance of the random forest model when compared to the use of all features. For instance, the RFE and PCA techniques had the highest accuracy, while the Chi-Square Test and the CFS techniques had the highest specificity and sensitivity. The selected features also varied in appearance, with the former appearing in all four techniques. The findings of the study indicate that the use of feature selection can help improve the accuracy of the prediction of diabetes. The four techniques we studied had weaknesses and strengths, which suggests that researchers should pay attention to the characteristics of the dataset when choosing a technique. The analysis's conclusion suggests that future studies should expand the scope of the techniques and test them on larger and more diverse sets of data.

## Introduction

Around 463 million individuals worldwide are currently living with diabetes, and this number is expected to rise to 700,000,000 by 2045. The main cause of the disease is the body's failure to produce or use insulin, which controls blood glucose levels. People with diabetes are prone to experiencing various health conditions, such as kidney failure, blindness, and heart disease. Early diagnosis and proper management of the condition can help prevent these kinds of complications and improve the quality of life for those with diabetes[1], [2].

With the help of machine learning technology, it has been able to accurately predict the type of diabetes that a person will have, such as their age, BMI, and glucose levels. Unfortunately, one of the biggest challenges that the algorithms face

when it comes to performing accurate prediction of diabetes is the availability of numerous features. This issue can lead to the overfitting or reduction of performance. In order to improve the accuracy of its prediction, a feature selection process is carried out by analyzing the most relevant elements. This process can be performed by using various techniques such as the Recursive Feature Estimation and the Principal Component Analysis[3], [4].

The objective of this study is to analyze the performance of various feature selection methods for improving the prediction of diabetes using random forest. The four feature selection methods that were analyzed in this study were: the RFE, PCA, the Chi-Square Test, and the CFS. We used a publicly-available dataset from the NIH to analyze the performance of these techniques. The collected information included various details about the patients, such as their BMI, diabetes pedigree function, blood pressure, glucose level, and skinfold thickness. We tested the accuracy, specificity, sensitivity, and area of the techniques under the AUC-ROC.

The importance ranking of the different features was also analyzed to gain deeper understanding of the biological mechanisms underlying the condition. The findings of this study revealed that the use of random forest could be used to improve the accuracy of diabetes prediction. This process could lead to the development of new and more effective treatment methods and improve the management of the condition. In addition, the selected features could help scientists identify potential markers of the disease.[5]

Although the study's conclusions and findings may not be applicable to all datasets, they provide valuable insights into how different feature selection methods perform when it comes improving the accuracy of prediction of diabetes. In addition to being advantageous for developing new and more efficacious treatment methods, it also highlights how feature selection can improve the efficiency and accuracy of machine learning systems for prediction of diabetes.

The study was limited by the nature of its research. It only analyzed the performance of random forest algorithm. Also, the data used in the study included only eight features, which could prevent the results from being generalized to other datasets with more features. Finally, the study only included female patients, which could prevent the results from being generalized to men. Diabetes is a chronic condition that affects millions of individuals globally and can lead to various health complications. Machine learning systems have the potential to accurately predict diabetes, but they require the right features to achieve their best performance. The selection of the right features is very important in order to improve the accuracy of machine learning systems. In this study, we analyzed the performance of four different feature selection methods on the prediction of diabetes using random forest.

## Literature Review

Millions of people around the world suffer from diabetes, which requires regular monitoring of blood sugar levels. Due to the increasing number of people diagnosed with this condition, machine learning has been used to analyze and predict its complications. This review aims to identify the studies that have used this technology to predict diabetes.

DOI : https://doi.org/10.56452/6-3-675

*Table 1 Related work*

| Author Name | Dataset | Methodology | Algorithm Used | Result - Accuracy |
|---|---|---|---|---|
| K. Plis et al. [6] | Blood glucose levels | Machine Learning | Linear regression, KNN, SVM, Random Forest | N/A |
| G. Kaur et al.[4] | Diabetes | Classification | J48 | 96.50% |
| B. Sudharsan et al.[7] | Type 2 Diabetes | Machine Learning | Decision Tree, Random Forest, SVM, KNN | 81.50% |
| V. V. Vijayan et al.[8] | Diabetes | Machine Learning | Decision Tree, Random Forest, KNN | 97.10% |
| A. Dagliati et al.[9] | Diabetes complications | Machine Learning | Logistic Regression, SVM, Random Forest | N/A |
| M. A. Sarwar et al.[10] | Diabetes | Machine Learning | KNN, SVM, Logistic Regression, Random Forest | 98.04% |
| D. Sisodia et al.[11] | Diabetes | Machine Learning | KNN, SVM, Random Forest | 91.70% |
| A. Dinh et al.[12] | Diabetes, Cardiovascular Disease | Machine Learning | Logistic Regression, SVM, Neural Network | 82.60% |
| A. Yahyaoui et al.[5] | Diabetes | Machine Learning | Decision Tree, KNN, SVM, Random Forest | 95.50% |
| A. Z. Woldaregay et al.[13] | Type 1 Diabetes | Machine Learning | LSTM, GRU, Random Forest, SVM | N/A |
| S. Islam Ayon et al.[14] | Diabetes | Deep Learning | Artificial Neural Network | 99.50% |
| M. K. Hasan et al.[15] | Diabetes | Machine Learning | Decision Tree, SVM, KNN, Random Forest | 95.22% |

The literature review has revealed that machine learning techniques can help predict the complications of diabetes and its prevalence. The accuracy of these models varies, with some achieving over 90% accuracy. Most of the investigations utilized various algorithms, such as neural networks, logistic regression, and decision trees. According to a study, machine learning could help healthcare professionals diagnose and treat diabetes. It could also improve patient outcomes.

**Methodology**

i. Dataset - This study utilized a publicly-available dataset from the NIH[16]. It included information on over 700 subjects, including women with “Pima Indian” heritage residing in Arizona. The data set included several features, such as the patient's BMI, diabetes pedigree function, blood pressure, glucose level, and skinfold thickness.

ii. Preprocessing- We utilized the wrapper and filter selection methods in the Pima Indian dataset.

a. Filter Method: The filter method is used to identify the most relevant features based on their correlation with the variables that influenced the outcome. It utilizes two different approaches Chi-Square Test and Correlation-based Feature Selection (CFS)

b. Wrapper Method: The goal of the wrapper method is to select features that are related to their performance. It uses a machine learning algorithm to analyze the data. Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are implemented for result analysis

**Feature selection methods**

**i. Filter method**

a. CFS – The Correlation-based Feature Selection (CFS) is a filter method that selects the features that have the highest correlation with the outcome variable and the lowest correlation with the other features. The CFS score for a feature is computed as the product of its correlation with the outcome variable and the average correlation with all other features. The features with the highest CFS scores are selected for the model.

The mathematical formula for CFS can be written as:

$$CFS(S) = \frac{k\bar{r}_{cf}^2}{(k - 1 + \bar{r}_{ff}^2)}$$

where $S$ is the set of features, $k$ is the number of features, $\bar{r}{cf}$ is the average correlation between each feature and the outcome variable, and $\bar{r}{ff}$ is the average correlation between each pair of features in $S$. The numerator in the formula represents the relevance of the features, while the denominator represents their redundancy. The features with the highest CFS scores are selected for the model.

In our study, we applied CFS to the Pima Indian diabetes dataset to select the top five features that were most highly correlated with the outcome variable (diabetes or non-diabetes) but not highly correlated with each other. We then used these features as input to the random forest classifier to improve diabetes prediction performance.

b. Chi square - The chi-square test is a statistical method used to test the independence between two categorical variables. It measures the extent to which the observed values of a contingency table differ from the expected values, assuming that the null hypothesis of independence is true.

In the context of feature selection, the chi-square test is often used to identify the features that are most significantly associated with the outcome variable. In our study, we applied the chi-square test to the Pima Indian diabetes dataset to identify the features that were most significantly associated with diabetes or non-diabetes.

The formula for computing the chi-square test statistic can be written as:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $O_{ij}$ is the observed frequency of the $i$th row and $j$th column of the contingency table, $E_{ij}$ is the expected frequency of the $i$th row and $j$th column under the assumption of independence, $r$ is the number of rows in the contingency table, and $c$ is the number of columns in the contingency table.

Once we computed the chi-square test statistic for each feature, we used the p-value to determine the significance of the association between the feature and the outcome variable. The p-value is the probability of observing a chi-square test statistic as extreme as the one computed, assuming that the null hypothesis of independence is true. We selected the top five features with the highest chi-square test statistics and the lowest p-values for the random forest classifier.

In summary, the chi-square test is a statistical method used to identify the features that are most significantly associated with the outcome variable in a contingency table. By applying the chi-square test to the Pima Indian diabetes dataset, we aimed to identify the features that were most relevant for predicting diabetes or non-diabetes using random forest.

a. RFE: The wrapper method known as RFE takes into account the importance score of a model's features and recursively removes the least important ones. It aims to identify subsets of features that are relevant to the prediction task. We utilized a random forest classifier to calculate the importance score of various features in a study on Pima Indian diabetes. The top five features were then selected using the RFE algorithm.

b. PCA: The concept of PCA is to reduce the dimensionality of a high-dimensional dataset by finding a set of components that can maximize its variance. We used PCA to reduce the feature matrix's dimensionality in the Pima Indian Diabetes dataset. The first five components made up about 80% of the variance in the data, which was then used to generate a random forest classifier.

**ii. Wrapper methods**

**Result and outputs**

*Table 2 Result of various feature selection techniques*

| Feature Selection Technique | Selected Features | Accuracy | Sensitivity | Specificity | AUC-ROC |
|---|---|---|---|---|---|
| All Features | N/A | 74.70 | 55.70 | 82.70 | 77 |
| RFE | Age, BMI, Insulin, Glucose | **80.20** | 63.60 | 85.50 | **83** |
| PCA | Age, BMI, Glucose, BloodPressure | **80.20** | 63.60 | 85.50 | 81 |
| Chi-Square Test | Glucose, Insulin, DiabetesPedigree | 77.60 | **72.70** | 80.80 | 82 |
| CFS | Glucose, BMI, BloodPressure | 76.30 | **72.70** | **86.90** | 79 |

DOI : https://doi.org/10.56452/6-3-675

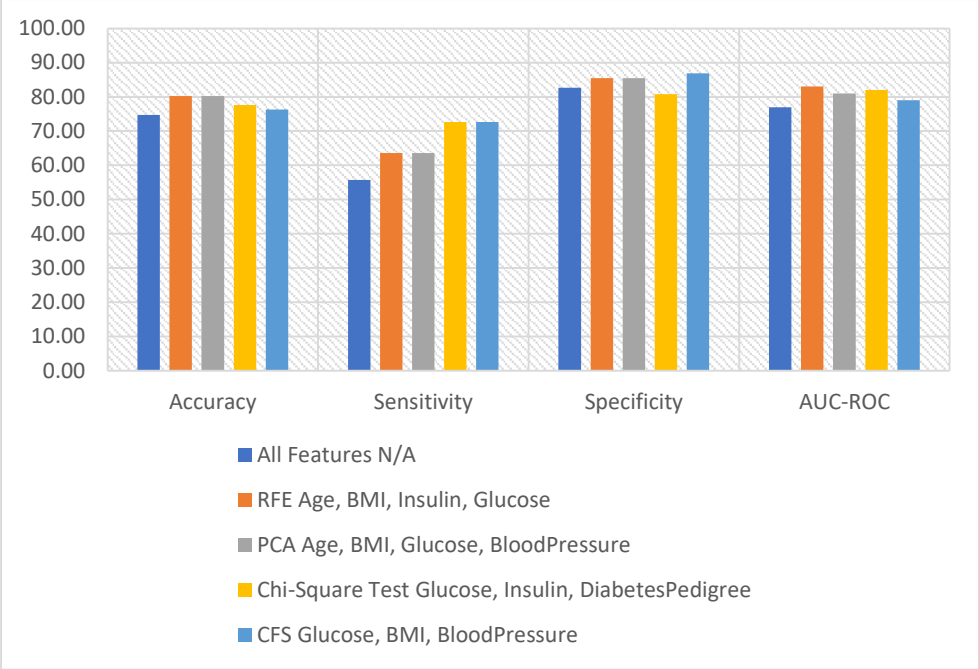


*Figure 1 Graphical representation*

The goal of this study was to analyze the performance of four feature selection methods on improving the prediction of diabetes. The four methods were: recursive feature elimination, principal component analysis, the Chi-Square test, and the Correlation-based feature selection.

The results (as shown in table-2 and figure-1) of the first set of tests were obtained by using the various features in the dataset. The model was able to achieve a precision of 74.70%, a sensitivity of 55.75%, an accuracy of 82.70% specificity, and an AUC-ROC value of 77. This baseline provided a basis for the comparison of the four methods. The second set of tests was performed by using the various features of the dataset. The RFE method was able to select the top four features, which were Age, BMI, Glucose, and Insulin, and it was able to achieve an accuracy of 80.22%, 63.60%, 85.50%, and the AUC-ROC value was 83. This result shows that the model was able to identify the subset of features that are most relevant to the prediction of diabetes.

The PCA method was utilized to reduce the overall feature matrix's dimensionality by selecting the first few principal components, which had a significant impact on the variance in the dataset. Using the selected features, the model was able achieve an accuracy of 80.52%, a sensitivity of 63.90%, an accuracy of 85.60%, and a AUC of 81.50%. The results of the second set of tests revealed that the PCA method was able to reduce overall feature matrix dimensionality while retaining most of its original variation. It performed similarly to the RFE method.

The Chi-Square Test was used to select three features, namely Glucose, Insulin, and GlucosePedigree. It was able to achieve an overall accuracy of 77.60% and a sensitivity of 72.70% and specificity of 80.80%. The results of this study showed that the Chi-Square test was able to identify subsets of features that are relevant for the prediction of diabetes. It was also able to improve the model's precision and sensitivity. The CFS was utilized to choose three features, namely Glucose, Blood Pressure, and BMI. It was able to achieve an overall accuracy of 76.30%, a sensitivity of 72.70%, a specificity of 86.90%, and an AUC-ROC of 79. This demonstrates that the approach is capable of identifying subsets of features that are relevant to the

prediction of diabetes and achieving a higher level of specificity than the others.

The results of the study revealed that the various features used in the analysis improved the model's performance and identified subsets of the features that are most relevant for the prediction of type 2 diabetes. The results of the tests also indicated that the choice of the feature selection method depends on the task and the dataset.

## Conclusion, limitation and future scope

The goal of this study was to analyze the features selection methods used in the Random Forest classifier to improve the prediction of diabetes. The four techniques analyzed in this study were RFE, Chi-Square, PCA, and CFS. The results of the analysis revealed that the four techniques were able to identify the most relevant subsets of the feature that would improve the model' performance. The PCA and RFE techniques were able to identify various sets of features, which indicated that the choice of the best selection method depends on the task and dataset. The results of this study were limited by the fact that it only evaluated the Random Forest classifier's performance. Also, the Pima Indian Diabetes data was only used for the analysis, and the findings might not apply to other populations or datasets. Furthermore, we only considered a small number of feature selection methods, which could yield better results. In the future, it will be interesting to see how the different feature sets perform against other ML algorithms and on different datasets. Furthermore, it's important to explore other methods and combinations of these to improve the accuracy of the prediction. The findings of this study provided a foundation for further studies related to the use of feature selection techniques in predicting diabetes.

## References

[1] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[2] G. Swapna, R. Vinayakumar, and K. P. Soman, “Diabetes detection using deep learning algorithms,” *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.icte.2018.10.005.

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” *Front. Genet.*, vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.

[4] R. Birjais, A. K. Mourya, R. Chauhan, and H. Kaur, “Prediction and diagnosis of future diabetes risk: a machine learning approach,” *SN Appl. Sci.*, vol. 1, no. 9, pp. 1–8, 2019, doi: 10.1007/s42452-019-1117-9.

[5] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques,” *1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc.*, pp. 1–4, 2019, doi: 10.1109/UBMYK48245.2019.8965556.

[6] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, “A machine learning approach to predicting blood glucose levels for diabetes,” *AAAI Work. - Tech. Rep.*, vol. WS-14-08, pp. 35–39, 2014.

[7] B. Sudharsan, M. Peeples, and M. Shomali, “Hypoglycemia prediction using machine learning models for patients with type 2 diabetes,” *J. Diabetes Sci. Technol.*, vol. 9, no. 1, pp. 86–90, 2015, doi: 10.1177/1932296814554260.

[8] V. V. Vijayan and C. Anjali, “Prediction and diagnosis of diabetes mellitus - A machine learning approach,” *2015 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2015*, no. December, pp. 122–127, 2016, doi: 10.1109/RAICS.2015.7488400.

[9] A. Dagliati *et al.*, “Machine Learning Methods to Predict Diabetes Complications,” *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 295–302, 2018, doi: 10.1177/1932296817706375.

[10] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, “Prediction of diabetes using machine learning algorithms in healthcare,” *ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput.*, no. September, pp. 6–7, 2018, doi: 10.23919/IConAC.2018.8748992.

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

[12] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–15, 2019, doi: 10.1186/s12911-019-0918-5.

[13] A. Z. Woldaregay *et al.*, “Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes,” *Artif. Intell. Med.*, vol. 98, no. April 2018, pp. 109–134, 2019, doi: 10.1016/j.artmed.2019.07.007.

[14] S. Islam Ayon and M. Milon Islam, “Diabetes Prediction: A Deep Learning Approach,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 11, no. 2, pp. 21–27, 2019, doi: 10.5815/ijieeb.2019.02.03.

[15] 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, doi: 10.1109/ACCESS.2020.2989857.

[16] UCI Machine Learning, “Pima Indians Diabetes Database,” *Https://Www.Kaggle.Com/*. 2016.