Proceedings of ICEGEE 2025 Symposium: Sensor Technology and Multimodal Data Analysis 2025
DOI: 10.54254/2753-8818/2025.AU24408

Comparative Analysis of Machine Learning Models in Early-Stage Diabetes Prediction

Fangting Zhang

School of Nursing and Rehabilitation, Shandong University, Jinan, China
202200250070@mail.sdu.edu.cn

Abstract. As the implementation of machine learning (ML) techniques in the medical industry is growing, utilizing ML models for prediction is crucial for diabetes prevention and treatment. Employing data extracted from the early-stage diabetes risk prediction dataset on the Kaggle platform, consisting of 520 samples, this study established the confusion matrix, applied four ML techniques: logistic regression, decision trees, random forests, and k-nearest neighbors, and compared the application outcomes of several models to assess the effect. According to the study, the four models performed well in predicting diabetes in its early stages. Decision trees and random forests achieved 99% and 98% accuracy, respectively. The interpretation of results provides favorable evidence supporting the application of ML in the early diabetes diagnosis. It manifests that ML has great potential in the early forecasts of diabetes. The research, however, still has the problem of a tiny sample size and insufficient model training. Increasing the incorporation of real-time data in order to strengthen the model's flexibility and long-term prediction capabilities is an excellent way to tackle the matter.

Keywords: Early-stage diabetes, Logistic regression, Decision trees, Random forests, K-nearest neighbors

1. Introduction

Diabetes, a metabolic disease characterized by chronic hyperglycemia, is caused by problems in the activity or secretion of insulin [1]. From 1990 to 2021, the global age-standardized prevalence of diabetes surged by 90.5%. In 2021, 529 million people globally were thought to be diagnosed with diabetes. Alarmingly, projections indicate that the number of diabetics will exceed 1.31 billion by 2050, highlighting an urgent need for targeted prevention and intervention strategies. With its increasing prevalence, diabetes has emerged as a significant global public health issue [2]. Diabetes may lead to various complications, such as microvascular and macrovascular complications [1]. The harm brought on by advanced diabetes worsens with time [3]. Implementing timely management and prevention measures can diminish the complications, slow down its onset, and effectively lower the incidence of diabetes [4]. Emerging evidence indicates that proactive identification and comprehensive management during the initial disease phase may enable therapeutic reversal in select cases [2]. Using early diagnosis and forecasting of diabetes occurrence through measuring

DOI: 10.54254/2753-8818/2025.AU24408

several risk factors may detect individuals at high risk of diabetes as early as possible, which is of great significance in the field of public health.

Currently, amid the digital transformation of healthcare, machine learning (ML) is widely applied in diabetes prediction and performs exceptionally well in diabetes prediction, with the benefits of affordability, high efficiency, and practicality. Most researchers evaluated models with accuracy, sensitivity, specificity, AUC, and F1 scores [4, 5].

Al-Zebari and Sengur came up with a method for diagnosing type 2 diabetes using decision trees (DT) and discovered that coarse tree (CT) was the most effective, which served as an auxiliary function in the subsequent medication treatment [6]. DT is frequently utilized because it has excellent releasing properties while not being the most advanced and comprehensive [7]. Based on the Kuwait Health Network dataset, Farran et al. constructed a model that used ML techniques such as k-nearest neighbors (KNN) and logistic regression (LR) to identify people at high risk of type 2 diabetes and their prognosis in the early stages [8]. Xu and Wang (2019) and Wang (2019) studied a composite model that used the random forest method to optimize feature selection and employed XGBoost for classification to predict the probability of having type 2 diabetes [9]. This model was validated through the Pima Indians diabetes dataset and assessed in terms of accuracy, specificity, and sensitivity, all of which surpassed 90%.

Recent research has shown that LR, DT, random forest (RF), and KNN are widely used methods for diabetes-related research using ML [10]. Therefore, this research explores the application of the above four methods to predict early-stage diabetes using a specialized dataset. The primary objective is to compare the effect of various techniques in predicting diabetes in its early phase and provide valuable insights into the effectiveness of these algorithms.

2. Methods

2.1. Data source and description

The data utilized in this study was sourced from the Kaggle platform, which the UCSI Machine Learning Repository supplied, a trusted repository for machine learning datasets. Comprising 520 case studies, this dataset contains essential data about signs and symptoms of individuals who either exhibit early signs of diabetes or are at risk of developing diabetes, ranging from demographic details to specific symptoms associated with diabetes [8]. There are 17 variables, which are age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class.

2.2. Variables and data pre-processing

In the original dataset, there were 17 variables, consisting of 16 categorical variables and 1 numerical variable. The paper encodes the binary variables by 0 and 1 (0 = No; 1 = Yes, 0 = Male; 1 = Female, and 0 = Negative; 1 = Positive).

The dataset is devoid of any missing data. Meanwhile, since the data set is composed of only binary attributes, 0 and 1, except for the 'age' attribute, it is not necessary to remove outliers [11]. However, the "age" ranges from 16 to 90. The paper uses Min-Max Scaling to perform data normalization.

DOI: 10.54254/2753-8818/2025.AU24408

2.3. Technical route

Focusing on the dataset, firstly, this research processes data by normalization of numerical variables and assignment of categorical variables. Secondly, it is vital to use the independent samples t-test and the chi-square test to proceed with feature selection. Thirdly, the datasets are split 8:2 between the training and test phases. The machine learning algorithm is trained by using the sample features as input, and the model is evaluated by the test set. Ultimately, the performance of the model is evaluated by the index of precision, accuracy, recall, and F1-score. The technical route of the paper is depicted in Figure 1.

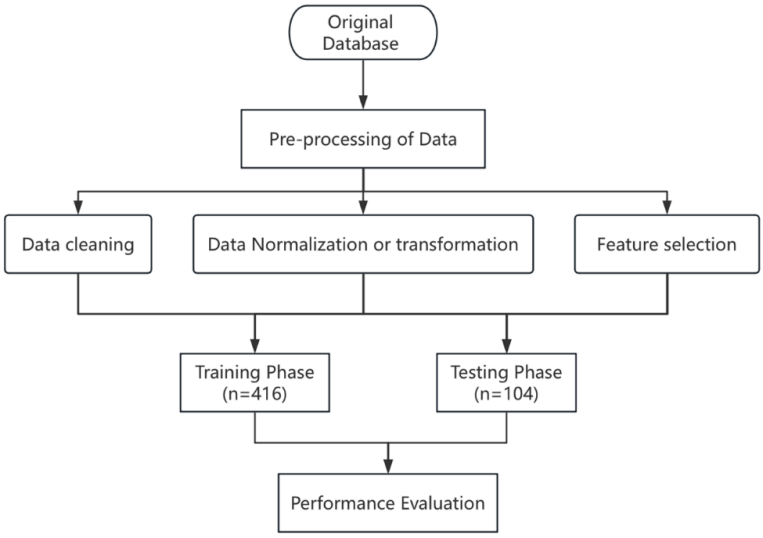


Figure 1. Technical Route (Picture credit: Original)

2.4. Feature selection

In this study, an independent samples t-test is employed to analyze age, and a chi-square test is used to analyze 15 binary classification variables to determine the significance of the relationship between the disease and the characteristics. It can also obtain the degree of their correlation. The test's significance level is set at 0.05.

2.5. Machine evaluation

This study compares the results of four ML models: KNN, LR, DT, and RF. All models presented in this research are split at an 8:2 ratio between the training phase and the test phase. To evaluate the LR models, this research decides to use a 2x2 confusion matrix. Accuracy, precision, recall, and F1-score are exploited to evaluate the performance of ML.

3. Results and discussion

3.1. Evaluation of feature scaling

The cross-tabulation is used to investigate the relationship among the 15 classes. According to Table 1, the class does not display significance for any of the three things: itching, delayed healing, and obesity. Exclude these three non-significant variables and select the remaining 12 features for

DOI: 10.54254/2753-8818/2025.AU24408

research. Given that the independent-sample t-test result is P=0.03<0.05 (Table 2), age and the class of diabetes were revealed to be correlated.

In conclusion, 13 features were selected for further model prediction in this study, comprising 12 categorical variables (et al., gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, irritability, partial paresis, muscle stiffness, and alopecia) and the numerical variable age.

Table 1. Cross Tabulation Number (Percentage).

Item	Option	class		Total	χ2	p Value
		Negative	Positive			
Gender	0	90.50%	45.94%	63.08%	104.942	0.000**
	1	9.50%	54.06%	36.92%		
Polyuria	0	92.50%	24.06%	50.38%	230.595	0.000**
	1	7.50%	75.94%	49.62%		
Polydipsia	0	96.00%	29.69%	55.19%	218.845	0.000**
	1	4.00%	70.31%	44.81%		
Sudden weight loss	0	85.50%	41.25%	58.27%	99.108	0.000**
	1	14.50%	58.75%	41.73%		
Weakness	0	56.50%	31.87%	41.35%	30.775	0.000**
	1	43.50%	68.13%	58.65%		
Polyphagia	0	76.00%	40.94%	54.42%	61.001	0.000**
	1	24.00%	59.06%	45.58%		
Genital thrush	0	83.50%	74.06%	77.69%	6.325	0.012*
	1	16.50%	25.94%	22.31%		
Visual blurring	0	71.00%	45.31%	55.19%	32.839	0.000**
	1	29.00%	54.69%	44.81%		
Itching	0	50.50%	51.88%	51.35%	0.093	0.76
	1	49.50%	48.13%	48.65%		
Irritability	0	92.00%	65.63%	75.77%	46.634	0.000**
	1	8.00%	34.38%	24.23%		
Delayed healing	0	57.00%	52.19%	54.04%	1.148	0.284
	1	43.00%	47.81%	45.96%		
Partial paresis	0	84.00%	40.00%	56.92%	97.174	0.000**
	1	16.00%	60.00%	43.08%		
Muscle stiffness	0	70.00%	57.81%	62.50%	7.8	0.005**
	1	30.00%	42.19%	37.50%		
Alopecia	0	49.50%	75.63%	65.58%	37.212	0.000**
	1	50.50%	24.38%	34.42%		
Obesity	0	86.50%	80.94%	83.08%	2.709	0.1
	1	13.50%	19.06%	16.92%		

* $p < 0.05$ ** $p < 0.01$

DOI: 10.54254/2753-8818/2025.AU24408

Table 2. Independent-Samples t-Test.

	class (Mean ± S.D.)		t	p Value
	Negative (n = 200)	Positive (n = 320)		
Age	0.41 ± 0.16	0.45 ± 0.16	-2.488	0.013*

* $p < 0.05$ ** $p < 0.01$

3.2. Models results

3.2.1. Logistic regression

From Figure 2, it appears that the disease status of 92 participants was correctly predicted, 8 participants were wrongly predicted to have diabetes, and 4 participants were omitted as not having the disease, using the LBFGS optimization algorithm. In summary, LR's result on the test phase achieves an F1-score of 0.89, 88.46% accuracy, 88.95% precision, and 88.46% recall, as presented in Table 3.

3.2.2. K-nearest neighbors

In Figure 2, there are 97 participants whose predictions were consistent with the actual situation, while 5 participants were incorrectly predicted to have diabetes, and 2 patients were incorrectly predicted to be optimistic.

The number of neighbors (K) is set to 5, and the sample voting weights are set to uniform voting weights; the neighbor search used is automatic, and the distance calculation uses Euclidean distance calculation. KNN on the test set accomplishes an F1-score of 0.93, 93.27% accuracy, 93.53% precision, and 93.27% recall, presented in Table 3.

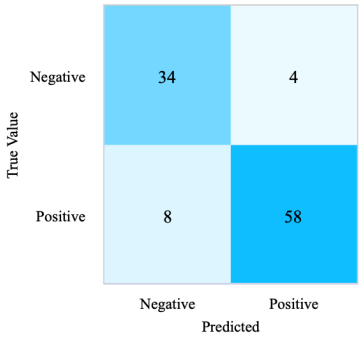
3.2.3. Random forests

The number of decision trees is set to 100, and the node splitting criterion is gini with no restriction on the maximum tree depth. From Figure 2, it finds that 102 participants were correctly predicted to have diabetes, 2 participants without the disease were predicted to have diabetes, and no participant was wrongly predicted to have the disease. It can be seen that RF achieves an F1-score of 0.98, 98.08% accuracy, 98.13% precision, and 98.08% recall, presented in Table 3.

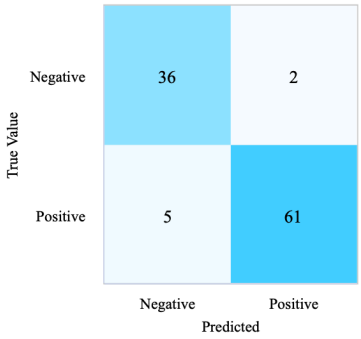
3.2.4. Decision trees

When the training set ratio is adjusted to 0.8, the node splitting criterion is gini, and the node splitting is best. The confusion matrix in Table 3 indicates that the disease status of 103 participants was correctly predicted, and 1 participant was wrongly predicted to not have diabetes. The results shown in Table 3 can demonstrate that DT achieves a 99.04% accuracy rate, 99.05% precision rate, 99.04% recall rate, and 0.99 F1-score.

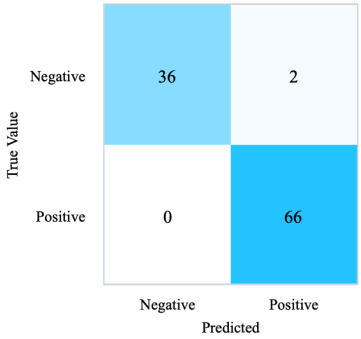
DOI: 10.54254/2753-8818/2025.AU24408



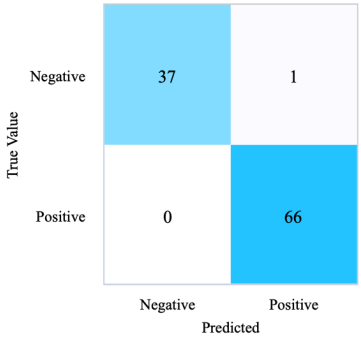
(a) Confusion Matrix of LR (Test Set)



(b) Confusion Matrix of KNN (Test Set)



(c) Confusion Matrix of RF (Test Set)



(d) Confusion Matrix of DT (Test Set)

Figure 2. Confusion matrices of four models (Photo/Picture credit: Original).

Table 3. Independent-Samples t-Test.

Model	Accuracy	Precision	Recall	F1-score
LR	88.46%	88.95%	88.46%	0.886
DT	99.04%	99.05%	99.04%	0.99
KNN	93.27%	93.53%	93.27%	0.933
RF	98.08%	98.13%	98.08%	0.981

3.3. Discussion

Although the ultimate results were positive, there are still certain shortcomings. The results of the research demonstrate that random forests and decision trees provide excellent feedback. But throughout the training process, decision trees could overfit the training data, particularly if there is little data or if the tree's depth is not adequately managed. However, random forests might not work as well with unbalanced datasets. This research's sample size is tiny in comparison to certain huge data sets, so the specificity of the data set or model parameters may have an impact on the performance [12]. For instance, in Varun Gulshan's study, an extensive data set provided the benefit of being able to explain consistency [13]. Real-time data can be incorporated into studies, while the data set can be enlarged to better train as the healthcare environment changes, which can increase the model's adaptability and long-term prediction capacity [8].

DOI: 10.54254/2753-8818/2025.AU24408

4. Conclusion

Based on the Kaggle database, the performance of the following ML algorithms is compared in the early phase of diabetes prediction: (a) LR, (b) RF, (c) KNN, and (d) DT. Gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, visual blurring, irritability, partial paresis, and alopecia were all found to be strongly associated with diabetes in the data set variable screening. Researchers can concentrate on these associated symptoms in subsequent investigations. The results show that overall, the four methods have achieved good results on this dataset. In summary, the most comfortable is decision trees since it has a performance of 99% or more in each aspect. Meanwhile, random forest is second. The good feedback of the machine model provides favorable evidence for using ML to forecast diabetes in its early stages. With continuous model optimization and adjustment and the verification of a large number of data sets, the application of ML in diabetes prediction can increase illness prediction efficiency and aims to promote medical progress in diabetes prevention and intervention.

References

[1] Ahmad, E., Lim, S., Lamptey, R., Webb, D. R., & Davies, M. J. Type 2 diabetes. The Lancet, 400(10365), 1803–1820.

[2] Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., Hagins, H., Lindstedt, P. A., Aali, A., Abate, Y. H., Abate, M. D., Abbasian, M., Abbasi-Kangevari, Z., Abbasi-Kangevari, M., Abd ElHafeez, S., Abd-Rabu, R., … Vos, T. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: A systematic analysis for the Global Burden of Disease Study 2021. The Lancet, 402(10397), 203–234.

[3] Battineni, G., Sagaro, G. G., Nalini, C., Amenta, F., & Tayebati, S. K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Machines, 7(4), 74.

[4] Ahsan, M. M., Luna, S. A., & Siddique, Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare, 10(3), 541.

[5] Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review. Diabetology & Metabolic Syndrome, 14(1), 196.

[6] Chaki, J., Thillai Ganesh, S., Cidham, S. K., & Ananda Theertan, S. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University - Computer and Information Sciences, 34(6), 3204–3225.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116.

[8] Nimmagadda, S. M., Suryanarayana, G., Kumar, G. B., Anudeep, G., & Sai, G. V. A Comprehensive Survey on Diabetes Type-2 (T2D) Forecast Using Machine Learning. Archives of Computational Methods in Engineering, 31(5), 2905–2923.

[9] Xu, Z., & Wang, Z. A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), 278–283.

[10] Suyanto, S., Meliana, S., Wahyuningrum, T., & Khomsah, S. A new nearest neighbor-based framework for diabetes detection. Expert Systems with Applications, 199, 116857. https://doi.org/10.1016/j.eswa.2022.116857

[11] Saxena, S., Mohapatra, D., Padhee, S., & Sahoo, G. K. Machine learning algorithms for diabetes detection: A comparative evaluation of performance of algorithms. Evolutionary Intelligence, 16(2), 587–603.

[12] Oikonomou, E. K., & Khera, R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovascular Diabetology, 22(1), 259.

[13] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402.