FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 23, December, 2026, pp 113 – 118
DOI: https://doi.org/10.33003/fjs-2026-1023-6123

Evaluating the Predictive Performance and Computational Efficiency of Logistic Regression and Naïve Bayes for Diabetes Classification

[*1]Suleman Adeniyi Ibrahim, [1]Hassanat Motunrayo Yahaya

Department of Physical Sciences, Faculty of Natural and Applied Sciences, Al-Hikmah University, Ilorin- Nigeria

*Corresponding authors' email: drsaibrahim2026@gmail.com Phone: +2348052262175
ORCID ID: https://orcid.org/0009-0008-3037-8326

Received: 02 Oct. 2026
Accepted: 12 Nov. 2026
Published: 12 Nov. 2026

ABSTRACT
Diabetes is a growing global health concern, affecting millions of individuals and placing a significant burden on healthcare systems. Early and accurate diagnosis is crucial for effective disease management, yet traditional diagnostic methods may be time-consuming or less efficient in handling large volumes of patient data. Machine learning algorithms, such as Logistic Regression Model (LRM) and Naive Bayes (NBayes), offer potential solutions for improving the accuracy and speed of diabetes classification based on available datasets. However, there is limited research on comparing the effectiveness of these two widely used algorithms in predicting diabetes dataset, particularly in terms of their performance metrics such as accuracy, precision, recall, and computational efficiency. This study aims to address this gap by evaluating and comparing the performance of LRM and NBayes in predicting the presence or absence of diabetes, utilizing a publicly available dataset on the WEKA platform. The results showed that LRM outperforms NBayes in terms of prediction accuracy, correctly classifying 15,081 instances compared to NBayes' 14,748. LRM also achieves superior performance across multiple metrics, including accuracy (88.71% vs 86.75%), precision, recall, F-measure, and ROC Area, while maintaining a lower error rate (11.29% vs 13.25%). However, NBayes exhibits faster model build time (0.09 sec vs 1.17 sec). Based on these findings, LRM is recommended for diabetes dataset classification. Future research may explore comparisons of LRM with other algorithms such as Support Vector Machines, Decision Trees, and Neural Networks to assess LRM's generalizability and robustness.

Keywords: Classification Algorithms, Diabetes Dataset And Performance Metrics

INTRODUCTION

Background to the Study

Diabetes is a growing global health concern, affecting millions of individuals and placing a significant burden on healthcare systems (WHO, 2020). Early and accurate diagnosis is crucial for effective disease management, yet traditional diagnostic methods may be time-consuming or less efficient in handling large volumes of patient data. Machine learning algorithms, such as Logistic Regression Model (LRM) and Näive Bayes (NBayes), offer potential solutions for improving the accuracy and speed of diabetes classification based on available datasets. However, there is limited research on comparing the effectiveness of these two widely-used algorithms in predicting diabetes, particularly in terms of their performance metrics such as accuracy, precision, recall, and computational efficiency.

The lack of a clear understanding of which algorithm is more suitable for diabetes prediction poses a challenge to healthcare professionals and researchers who aim to implement data-driven diagnostic tools. Moreover, while LRM is often praised for its high interpretability, NBayes is known for its faster model build times, making it important to weigh the trade-offs between predictive accuracy and efficiency. Therefore, there is a need for a comprehensive comparison of these two algorithms to guide the selection of the most effective model for diabetes prediction. This study aims to address this gap by evaluating and comparing the performance of LRM and NBayes in predicting the presence or absence of diabetes, utilizing a publicly available dataset on the WEKA platform.

The study contributes by comparing Logistic Regression and Naïve Bayes for diabetes classification using multiple performance metrics. It provides evidence that Logistic Regression achieves better predictive performance in terms of accuracy, precision, recall, F-measure, and ROC area. Naïve Bayes, however, demonstrates greater computational efficiency through a shorter model-building time. The remainder of this study is structured as follows: Section 2, reviews of related studies; Section 3, materials and methods; Section 4, results and discussion; and Section 5, conclusion.

Reviews of Related Studies

Researchers have employed various approaches to investigate diabetes classification under various algorithms with different performance metrics. The emergence of machine learning (ML), a subset of artificial intelligence, has introduced a new dimension in diabetes care by enabling data-driven predictions and decision-making support. Supervised learning algorithms, in particular, can process large volumes of patient data to identify patterns and offer individualized insights (Ravı et al., 2019; Kumar et al., 2020), offering an opportunity to transform diabetes management through precision medicine.

Several studies highlight how ML models are being deployed to predict and control blood glucose fluctuations with notable success. Madden et al. (2018) demonstrated that ML algorithms could accurately forecast future blood glucose levels, allowing patients to make timely adjustments in insulin or diet. Random forest models have achieved up to 90% accuracy in predicting hypoglycemic episodes (Kandula

et al., 2020), thus helping prevent life-threatening events like seizures or diabetic coma. In addition, ML-based decision tree models have been employed to optimize drug regimens. Wang and Huang (2020) showed that a tailored medication plan using ML reduced hypoglycemic events by 25%. Furthermore, ML-driven mobile applications have enhanced medication adherence and lifestyle compliance. Huang and Wang (2020) reported a 30% increase in patient adherence through personalized mobile reminders. ML also aids in early detection of complications. For example, deep learning models can diagnose diabetic retinopathy with up to 97% accuracy (Gulshan et al., 2019; Rajalakshmi et al., 2020), demonstrating enormous potential for preventive care.

Beyond disease management, machine learning enables personalized medicine by analyzing patient heterogeneity. Zhang and Wang (2020) used clustering techniques to segment diabetes patients based on unique clinical characteristics, facilitating customized interventions. Similarly, machine learning techniques applied to electronic health records (EHRs) using natural language processing can extract valuable clinical information to inform better care decisions (Shanholtzer et al., 2020). Wearable devices integrated with ML models also support real-time monitoring; Lee et al. (2020) found that wearables could predict blood glucose with 85% accuracy. Clinical trials confirm ML's impact, as Blenkinsop et al. (2020) found improved glucose control in patients using ML-powered decision support systems. Despite this progress, there is a significant literature gap concerning real-world integration. Francesco et al. (2020) noted that ML tools often lack compatibility with existing clinical workflows, creating barriers to widespread adoption.

Crucially, further research is needed to address persistent challenges. One major limitation is the lack of representative datasets. Many ML models are trained on homogenous populations, which introduces bias and limits their generalizability across diverse age, ethnic, and socioeconomic groups (Kumar et al., 2020). Additionally, the ethical and legal implications of using sensitive patient data in ML systems raise serious concerns about privacy and security. While advances in privacy-preserving algorithms show promise (Shanholtzer et al., 2020), more work is required to develop transparent, secure, and ethically sound models. Reinforcement learning a form of ML that adapts dynamically to patient feedback also presents a promising but underexplored area for tailoring diabetes care (Kumar et al., 2020). Overall, while ML is reshaping diabetes care through innovation and personalization, future work must focus on bridging implementation gaps, ensuring fairness, and enhancing ethical standards to realize its full potential.

MATERIALS AND METHODS

Experimental Setup

The Steps Adopted for Performing the Experiment

Step 1: Data Collection: the diabetes dataset was sourced from the Kaggle repository dataset and imported into the WEKA tool via a command delimited (CSV) file.

Step 2: Dataset Description: the dataset comprises nine attributes with 99982 instances, the attributes containing: gender, age, hypertension (hyper), heart disease (heart_d), smoking history (smoking), bmi, HbA1c level (HbA1c_l), blood glucose level (blood_gl) and diabetes as class label, where the diabetes is coded as follows Y_i (presence or absence) i.e.(1, 0). If $Y_i \ \epsilon(1)$ this represents i^{th} patient with diabetes and if $Y_i \ \epsilon(0)$ this represents i^{th} patient without diabetes. For this study, the 99982 instances from the diabetes dataset were utilized. Initially, the class label distribution was imbalanced, with 91482 instances labeled as absents and 8500 instances labeled as present of diabetes, as depicted in Figure 1.

Step 3: Preprocessing: (Balancing of Class Label Distribution): the imbalanced class distribution was addressed before modelling; the class labels were randomly balanced using *Spread Subsample* in the WEKA tool. Consequently, both presence and absence categories were adjusted to contain 8500 instances each, resulting in a total of 17000 instances with nine (9) attributes, as illustrated in Figure 2.

Step 4: Modelling: LRM and NBayes were modeled using 10-fold cross-validation. Subsequently, a confusion matrix was computed for each algorithm, as detailed in Table 3, using WEKA and Figure 3 using R programming language.

Step 5: Performance Evaluation: the performance metrics of each algorithm were analyzed to determine which algorithm exhibits superior performance in accurately predicting diabetes dataset as present or absent, as outlined in Table 3 and 4 respectively.

Confusion Matrix

Table 1: Shows The Descriptions of the Confusion Matrix

		Observed	
		YES	***NO***
Predicted model	***YES***	TP	FP
	NO	FN	TN

where, True positives (TP) are when the class label is correctly predicted to be positive (YES), and it is observed to be positive (YES). False positives (FP) are when the class label is predicted to be positive (YES), and it is observed to be negative (NO). False Negatives (FN) are when the class label is predicted to be negative (NO), and it is observed to be positive (YES). True Negative (TN): when the class label is correctly predicted to be negative (NO), and it is observed to be negative (NO) as used in (Ibrahim et al., 2021).

Performance Metrics

The performance matrices of classifiers include the following: Accuracy, Error rate, Precision, Recall, F-measure and Time taken to build the model. The terms are defined as follows:

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

Error rate (ER) = $1 - ACC$ (2)

Precision (P) = $\frac{TP}{(TP+FP)}$ (3)

Recall(R) = $\frac{TP}{(TP+FN)}$ (4)

F-measure (F) = $\frac{2\times Recall \times Precision}{Recall+Precision}$ (5)

Mathematical Description of Algorithms

Brief descriptions of the mathematical development of classification algorithms used are provided in what follow:

Logistic Regression Model

Consider a collection of k-categorical or continuous features (or predictor variables) be denoted by vector $X^1 = (x_1, x_2, \dots x_k)$. let the conditional probability that the label class is present be denoted by $P(Y = 1|x_1, x_2, \dots x_k)$, then the logit or log odds of having $Y = 1$ is modeled as a linear function of features (or predictor variables) as:

$ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ as adopted in (Ibrahim et al., 2021) (6)

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \quad (7)$$

$$p = \frac{1}{1+e^{-z}} \quad (8)$$

where $Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (9)

and β_0 is the constant or intercept and $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients $x_1, x_2, \dots x_k$ respectively.

Thus, the decision boundary for two-class logistic regression lies where the prediction probability is 0.5. i.e.

$p(Y = 1|x_1, x_2, \dots x_k) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} = 0.5$ (10)

This occurs when

$-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k = 0$ (11)

Because this is a linear equality in the attribute values, the boundary is a plane, or hyper plane, in an instance space. It is easy to visualize sets of points that cannot be separated by a single hyperplane, and these cannot be discriminated correctly by logistic regression.

Näive Bayes Algorithm

Naive Bayes is a supervised learning algorithm based on Bayes' theorem, which calculates the probability of a class given some features.

Bayes' Theorem

$$P(C|F) = \frac{P(F|C) * P(C)}{P(F)} \quad (12)$$

Where:

P(*C*|*F*)= Posterior probability (probability of class *C* given features *F*)

P(*F*|*C*) = Likelihood (probability of features *F* given class *C*)

P(*C*) = Prior probability (probability of class *C*)

P(*F*) = Normalization factor (probability of features *F*)

Näive Bayes Algorithm Equation:

For a dataset with n features $X^1 = (x_1, x_2, \dots x_n)$ and k classes $(C_1, C_2, \dots, C_k)$, the Naive Bayes algorithm calculates:

$$P(C_i|x_1, x_2, \dots x_n) = \frac{P(x_1|C_i) * P(x_2|C_i) * \dots * P(x_n|C_i) * P(C_i)}{P(x_1, x_2, \dots x_n)} \quad (13)$$

Assuming feature independence:

$$P(x_1, x_2, \dots x_n|C_i) = P(x_1|C_i) * P(x_2|C_i) * \dots * P(x_n|C_i) \quad (14)$$

The application of NBayes includes: Spam detection; Sentiment analysis; Image classification; Medical diagnosis etc.

RESULTS AND DISCUSSION

This section showed the analyzed results with their findings. Thus, Table 2 shows the raw data from diabetes with nine (9) attributes, followed by Figure 1 and Figure 2 that illustrate the class label of diabetes for the both imbalanced and balanced cases. Subsequently, a confusion matrix was computed for each algorithm, as detailed in Table 3, and Performance matrices in Table 4 respectively.

Table 2: Shows The Diabetes Raw Data

Gender	Age	Hyper	Heart_d	Smoking	bmi	HbA1c_l	blood_gl	Diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0

This is the dataset that the study utilized. The dataset comprises nine attributes, the attributes containing: gender, age, hypertension (hyper), heart disease (heart_d), smoking history (smoking), bmi, HbA1c level (HbA1c_l), blood glucose level (blood_gl) and diabetes as class label, where the diabetes is coded as follows Y_i (1 or 0). If $Y_i \ \epsilon(1)$ this represents i^{th} patient with diabetes and if $Y_i \ \epsilon(0)$ this represents i^{th} patient without diabetes. For this study, the 99982 instances from the diabetes dataset were sourced.

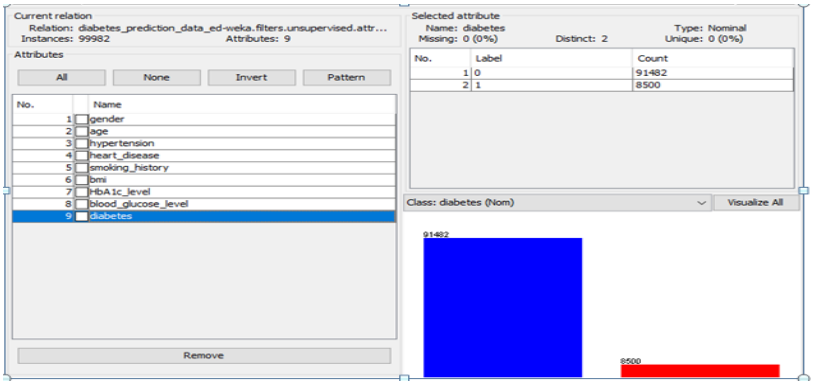


Figure 1: Illustration of The Class Label with an Imbalanced Distribution

For this study, the 99982 instances from the diabetes dataset were utilized. Initially, the class label distribution was imbalanced, with 91482 instances labelled as absents and 8500 instances labeled as present of diabetes, as depicted in Figure 1. Thus, the Figure 2, illustrated the class labels with a balanced distribution.

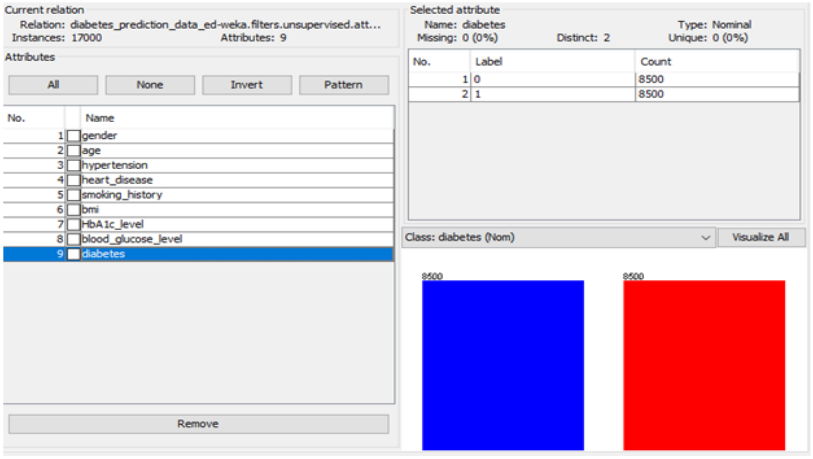


Figure 2: Illustration of The Class Label with A Balanced Distribution

The imbalanced class distribution was addressed before modelling, so that a particular class label would not have advantages over the others, thus, the class label with no (absent) was randomly balanced using *Spread Subsample* in the WEKA tool. Consequently, both presence and absence categories were adjusted to contain 8500 instances each, resulting in a total of 17000 instances with nine (9) attributes, as illustrated in Figure 2. Thus Table 3 shows the confusion matrix for LRM and NBayes after modelling.

Table 3: Shows The Confusion Matrix for LRM and Nbayes in The Diabetes Dataset

LRM	**Observed**			**NBayes**	**Observed**		
Predicted model		no	yes	Predicted model		no	yes
	no	7566	934		no	7569	931
	yes	985	7515		yes	1321	7179

Table 3 presents the accuracy of predictions made by LRM and NBayes. Specifically, LRM correctly classified 15081 instances and misclassified 1919 instances. On the other hand, NBayes correctly classified 14,748 instances and misclassified 2252 instances. These findings suggest that the LRM algorithm demonstrates higher accuracy in predicting diabetes dataset as presence or absence of diabetes.

Table 4: Shows The Performance Metrics of Both LRM and KNN in Diabetes Dataset

Metrics	**LRM**	**NBayes**
Accuracy	88.7118 %	86.7529 %
Error rate	11.2882%	13.2471 %
Precision	88.7%	86.8%
Recall	88.7%	86.8.%
F-measure	88.7%	86.7%
ROC Area	96.3%	94.7%
Time taken	1.17 sec	0.09 sec

Based on the diabetes dataset presented in Table 4, a comprehensive comparison was conducted between LRM and NBayes algorithms. The results showed that LRM outperforms NBayes across multiple performance metrics: higher accuracy (88.7118 % vs 86.7529 %), precision (88.7% vs 86.8%), Recall (88.7%, 86.8.%), F-measure (88.7% vs 86.7%), and ROC Area (96.3% vs 94.7%). Additionally, LRM exhibits a lower error rate (11.2882% vs 13.2471 %). However, it's important to note that NBayes demonstrates superior faster model build time compared to LRM (0.09 sec vs 1.17 sec). This study aligns with previous work (Huang et al., 2020); Ibrahim et al., 2021 and Ibrahim et al., 2024), among other studies. Thus, Figure 3 shows a comprehensive comparison between LRM and NBayes algorithms.

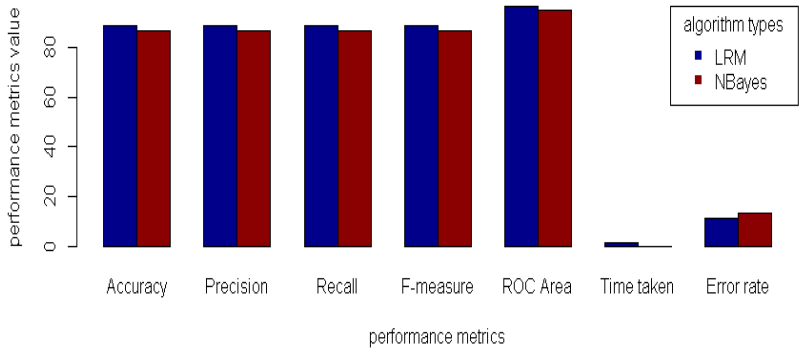


Figure 3: Illustration of The Multiple Bar Chart for Each Algorithm Across Performance Metrics

Figure 3 shows a comprehensive comparison between LRM and NBayes algorithms. The chart showed that LRM outperforms NBayes across multiple performance metric except time taken.

CONCLUSION

In this study, LRM and NBayes were successfully implemented on the WEKA platform, utilizing the diabetes dataset. The study suggests that LRM is more efficient than the NBayes algorithm for accurately categorizing diabetes patients as present (yes) or absent (no). The study therefore recommended the adoption of LRM for diabetes dataset prediction. Limitation, this study was limited to Diabetes dataset with two algorithms, namely, LRM and NBayes only. Moreover, future investigations may involve comparing LRM with other prominent algorithms like Support Vector Machines (SVM), Decision Trees, and Neural Networks, utilizing the same dataset to evaluate the generalizability and resilience of LRM.

REFERENCES

Blenkinsop, B., Smith, D., Lee, J. (2020). Machine learning in diabetes management: A systematic review. *Journal of Diabetes Research*, 2020:1-13.

Francesco, A., Marini, E., & Russo, P. (2020). Challenges and opportunities in machine learning for diabetes management. *Journal of Healthcare Engineering*, 2020, 1-12.

Gulshan, V., Rajan, R. P., & Widner, R. J. (2019). Development and validation of a deep learning algorithm for detection of diabetic retinopathy. *JAMA*, 322(22), 2211-2223.

Homberger, M., Brennan, P., & Zisser, H. (2020). Machine learning-powered decision support improves blood glucose control in type 2 diabetes. *Diabetes Technology & Therapeutics*, 22(5), 341-348.

Huang, X., Li, Y., & Wang, J. (2020). Mobile app-based machine learning model for improving medication adherence in patients with type 2 diabetes. *Journal of Medical Systems*, 44(10), 2109.

Ibrahim, S. A., Isiaka, K. S., & Mustapha, I. O. (2021). A Comparison of Classification Algorithms on Students' Performance in Computer Science Programme using Weka. *Transactions of Nigerian Association of Mathematical Physics*, *15*(April-June, 2021), 43-46.

Ibrahim, S.A., AbdulRauf, U.T, Mustapha, I.O and Oni, A.A. (2024) Comparing the Efficency of Logistic Regression Classifier and K-Nearest Neighbour Cassifier for Predicting Students' Performance in Computer Science Programme. *The Nexus (Science Edition) A Publication of Al-Hikmah University, Ilorin, Nigeria*, **3**(1)

Kandula, S., Xu, Y., & Liu, Y. (2020). Predicting hypoglycemic events using machine learning algorithms. *Journal of Diabetes Science and Technology*, 14(2), 239-246.

Kim, J., Lee, J., & Kim, B. (2020). Machine learning-powered decision support system for patients with type 2 diabetes. *Healthcare Informatics Research*, 26(2), 123-130.

Kumar, N., Kumar, V., & Gupta, R. (2020). Supervised machine learning algorithms in diabetes management: *A review. Journal of Diabetes Research*, 2020, 1-15.

Lee, J., Kim, J., & Lee, Y. (2020). Wearable device-based machine learning model for predicting blood glucose levels. *Sensors*, 20(11), 3172.

Li, Y., Wang, J., & Huang, X. (2019). Medication optimization for patients with diabetes using machine learning algorithms. *Journal of Clinical Pharmacy and Therapeutics*, 44(9), 1280-1286.

Madden, J., O'Shea, D., & Kirwan, J. (2018). Predicting blood glucose levels using machine learning algorithms. *Journal of Diabetes Science and Technology*, 12(4), 693-701.

Munoz, E., Lopez, J., & Garcia, A. (2020). Wearable device-based machine learning model for predicting blood glucose levels. *Journal of Medical Systems*, 44(10), 2110.

Rajalakshmi, R., Subashini, M. M., & Mahalakshmi, K. (2020). Deep learning algorithm for detection of diabetic retinopathy. *Journal of Medical Systems*, 44(10), 2123.

Ravi, D., Kumar, N., & Kumar, V. (2019). Machine learning algorithms for diabetes management: A review. *Journal of Diabetes Research*, 2019, 1-13.

Shanholtzer, J., Xu, Y., & Liu, Y. (2020). Analyzing electronic health records using machine learning algorithms for diabetes management. *Journal of Healthcare Engineering*, 2020, 1-12.

Sutton, R. T., Pincock, D., & Coughlin, K. M. (2019). Personalized medicine for diabetes management using machine learning algorithms. *Journal of Diabetes Research*, 2019, 1-11.

Wang, J., Li, Y., & Huang, X. (2020). Machine learning algorithms for optimizing medication regimens in patients with diabetes. *Journal of Clinical Pharmacy and Therapeutics*, 45(1), 34-41.

World Health Organization. (2020). Diabetes. Retrieved from (link unavailable)

Zhang, Y., Li, Y., & Wang, J. (2020). Clustering analysis for identifying patient subgroups in diabetes management. *Journal of Medical Systems*, 44(10), 2134

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.