

Machine Learning Model in Diagnosing Diabetes Mellitus among Women of Reproductive Age Using Kaggle Datasets

K.V. Odunugu¹, C.I. Osuji², F. Andibanbang³, S. Ariyo⁴, P. O. Amosu⁵, T.G. Agbejaye⁶, O.E. Owoicho⁷, O.A. Bamgbose⁸, O.E. Oke⁹, S.I. Okogu¹⁰, O. Anipole¹¹, M.A. Sunmola¹², L.E. Aminu-Ayinde¹³, E. Ademola¹⁴, J. Idowu¹⁵, T. Ademola¹⁶, O. Oyewole¹⁷, and S.E. Paul¹⁸

^{1, 6, 9, 13}Olabisi Onabanjo University Teaching Hospital Sagamu, Nigeria.

²Federal University, Wukari, Nigeria.

³TICEHURST Priory Mental Health Hospital, United Kingdom.

⁴Mountain Top University, Ibafo, Nigeria.

⁵University of Ilorin, Ilorin, Nigeria.

^{7, 12}Corporation for Africa and Overseas (CFAO) group

^{8, 9, 12, 15, 16, 17} Olabisi Onabanjo University, Ago-iwoye, Nigeria.

¹⁰University of Uyo Teaching Hospital, Uyo, Nigeria

¹⁸Thomas Adewumi University, Oko, Nigeria.

Received 8 January 2025; Acceptance 3 February 2025; Published 28 February 2025.

Abstract

The diagnosis of diabetes mellitus using machine learning (ML) offers a modern, efficient, and non-invasive approach to identifying the disease. ML algorithms analyze complex patterns in patient data, such as blood glucose levels, demographic factors, lifestyle habits, and genetic information, to make accurate predictions. This study aimed to develop and implement a machine-learning model for predicting diabetes based on clinical and demographic data such as glucose levels, blood pressure, BMI, and Diabetic predegree Function. The data consists of 768 females at least 21 years old. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies, BMI, blood glucose, skin thickness, Diabetes Predegree Function, insulin level and age. The study adopted the support vector machines (SVM) for

Correspondence addressed to Kayode Valentine Odunugu, email: kayodeodunugu94@gmail.com

Open Access This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License. Any other distribution of this content must maintain attribution to the author (s) and the title of the work, citation of the journal and DOI. Published under licence by Scholar J Publishing Ltd

classification. The data was divided into training sets (70%) and testing (30%), both of which were trained. The model was evaluated for performance using precision and accuracy scores. Finding reveal that there are 500 (65%) diabetics and 268 (35%) non-diabetics. The mean values are Pregnancies: 3.85 ± 3.37 . Glucose: 69.11 ± 19.36 mg/dL. Blood Pressure: 20.54 ± 15.95 mmHg. Skin Thickness: 20.54 ± 15.95 mm. Insulin: 79.80 ± 115.24 μ U/mL. BMI: 31.99 ± 7.88 . Diabetes Pedigree Function: 0.47 ± 0.33 . Age: 33.24 ± 11.76 years. The accuracy of this model in predicting diabetes for the training data was 88.66 % and 87.27% for the test data set. Overall, the performance of the machine learning model presented in this study indicates that this approach can be effective for diabetes prediction in clinical practices.

Keywords: Machine learning, Model, Diabetes, Women reproduction.

Introduction

Diabetes mellitus (DM) is a chronic metabolic disease marked by raised sugar levels in the blood with absolute or relative deficiencies in carbohydrate, protein, and lipid metabolism because of insulin insufficiency, with the dysfunction of the organs [1-2]. Thus, proper diagnosis and treatment should be done at an early stage to avoid complications such as cardiovascular disease, kidney failure and neuropathy [3].

Traditional diagnostic methods such as fasting blood glucose, oral glucose tolerance test and HbA1c have been useful [4]. They are time consuming, invasive and are not sensitive enough to diagnose early stages of Diabetes [5]. In recent years, concerns have arisen regarding the potential use of Machine Learning (ML) to transform the healthcare industry, offering innovative solutions to complex problems [6].

Machine learning algorithms can process vast amounts of data, uncover patterns that are not easily noticeable, and predict results with remarkable accuracy, making them ideal for transforming the approach to diabetes diagnosis [4,7]. Through an efficient, non-invasive, and more comprehensive collection of patient features, including their demographic profile, lifestyle, genetic profile, and clinical history, more accurate and quicker diagnostic models based on the patient data can be attained using the ML models [8].

Keshtkar, et al. in [9] stated that diagnostic accuracies of SVMs and RFs are very high for differentiating various types of Diabetes. Singh and Tiwari in [10] revealed that AI usage has unlimited opportunities to transform diabetes diagnosis into a simple process of risk assessment and early detection, increase the efficiency of diabetes therapy, and identify severe complications.

Khalifa and Albadawy in [11] revealed that the integration of artificial intelligence in diabetes care offers a personalized, efficient, and proactive model. It improves patient care and provides health education to patients on their management and treatment plans [12]. AI in diabetes care entails anticipatory and patient-centric care and leads to elevated average life expectancy and a better standard of living [4].

This study seeks to address the gap in the literature by not only creating a very robust and accurate model but also emphasizing this model's applicability, adaptability, usefulness in real clinical settings, and relevance.

Research Method

The dataset is originally from the Kaggle Dataset. The purpose of the data set is to determine whether or not a patient has diabetes based on certain potential diagnostic measurements incorporated within the data set [5]. The data contains 768 females at least 21 years old. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include:

Pregnancies: The number of times the individual has been pregnant. This value can indicate the woman's reproductive history.

Glucose: The plasma glucose concentration in mg/dL during a 2-hour oral glucose tolerance test. This measure is crucial for diagnosing diabetes and pre-diabetes.

Blood Pressure: The diastolic blood pressure (in mmHg). High blood pressure can be a risk factor for various health issues, including gestational hypertension.

Skin Thickness: The thickness of the skin fold at the triceps (in mm), which can be an indicator of body fat distribution and insulin resistance.

Insulin: The serum insulin level (in $\mu\text{U/mL}$). This is a critical hormone that regulates blood glucose levels. High levels may indicate insulin resistance.

Body Mass Index (BMI): a measure of body fat based on height and weight (kg/m^2). A higher BMI can be associated with an increased risk of diabetes.

Diabetes Pedigree Function: A function that represents the genetic predisposition to diabetes. This value ranges between 0 and 2.5, where higher values indicate a greater genetic risk.

Age: The Age of the individual (in years). Age can influence the risk of developing diabetes.

Outcome: A binary variable indicating whether the individual has diabetes (1) or not (0). This is the target variable for classification models aimed at predicting diabetes.

There were no missing values, particularly in fields like Skin Thickness and Insulin. This is because, in medical datasets, zeros can sometimes indicate missing data rather than a value of zero. The target variable, Outcome, suggests the prevalence of diabetes in the dataset. A balance between 0s and 1s was ideal for training classification models.

The data was normalized or standardized to numerical variables to ensure they were on a similar scale.

The study adopted the support vector machines (SVM) for classification.

Model Training. The data was divided into training (70%) and testing (30%) sets, both of which were trained. The model was evaluated for performance using precision and accuracy scores.

Results and Discussion

Descriptive Statistics

The mean includes: Pregnancies: 3.85 ± 3.3 Glucose: 69.11 ± 19.36 mg/dL

Blood Pressure: Average diastolic blood pressure is 20.54 ± 15.95 mmHg.

Skin Thickness: 20.54 ± 15.95 mm. Insulin: 79.80 ± 115.24 μ U/mL.

BMI: 31.99 ± 7.88 , which is above the normal BMI range (18.5–24.9).

Diabetes Pedigree Function: 0.47 ± 0.33 Age: 33.24 ± 11.76 years.

The findings revealed that out of 768 individuals involved in the datasets, total observations, 500 individuals (approximately 65%) do not have diabetes. This represents the majority of the dataset, indicating that two-thirds of the subjects are diabetes-free. Thus 268 individuals (about 35%) are diagnosed with diabetes. This is a significant portion of the dataset, which highlights that over one-third of the individuals are living with diabetes.

The dataset printed contains 768 rows and 8 columns. It seems to include demographic and health-related data points, particularly related to diabetes.

Pregnancies: This variable is a factor to consider for women, as multiple pregnancies may impact their health and risk for diabetes. The values range from 0 to 10+.

Glucose: A high value indicates potential diabetes or pre-diabetes. Some individuals have extremely high glucose levels, which could indicate poor glucose regulation.

Blood Pressure: High blood pressure is a common risk factor for diabetes and cardiovascular diseases. Normal diastolic pressure is usually around 80 mmHg, but many observations are lower, indicating potential hypotension.

Skin Thickness: A value of "0" indicates missing or unmeasured data in this context.

Insulin: Many entries have insulin levels recorded as "0," likely indicating missing data rather than an actual measurement of zero.

BMI: BMI is a strong indicator of obesity and diabetes risk. Several values are high, indicating that many individuals may be overweight or obese, a significant risk factor for diabetes.

Diabetes Pedigree Function: Higher values indicate a stronger family history of diabetes.

Age: The age of the individual. Age is a known risk factor for diabetes, especially for those above 45.

The data was split into two parts: Training Set: 614 samples (80% of the data).

Test Set: 154 samples (20% of the data).

Training the model

Training Accuracy 88.66% (0.8866449511400652). This means the model is correctly predicting the labels for 88.66% of the training data.

Finding the Accuracy Score of test data

The accuracy score of the model's test data was 0.8727 or 87.27%. This indicates that for 87.27% of the instances in the test data set the outcome as predicted by the model was right.

Discussion

The information from this dataset is relevant to understanding health implications linked to diabetes, especially among pregnant individuals. It provides a background for developing a risk prediction model and for subsequent studies of risk factors and incidence rates of diabetes in this group [1].

Most insulin levels were extremely low, and the spread was seen to be very large due to the presence of some very high values, which ranged up to 846 $\mu\text{U/mL}$. The average BMI is above the healthy imperative, implying that many people in this dataset may be overweight or obese, a triggering factor for diabetes. The study revealed that 35% of patients were identified to have diabetes, which is a large proportion. The distribution of the cases in the dataset is dominated by non-diabetic participants rather than participants with diabetes. This is a frequently seen problem in medical

databases, as symptomatic states are often more frequent than pathological ones in many populations [2]. Thus, 35% of people with diabetes is high. This finding can be worrisome for diabetes in this population and highlights the need for initiatives, changes in life, and necessary healthcare assistance. This distribution of diabetic data reinforces the potential areas of Diabetes within the sample population and the potential that the healthcare research field can harness to develop solutions for the prevention or improved handling of Diabetes. Non-diabetic pregnancy average was 3.30, whereas Pregnancy average of diabetic was 4.87. The present study identified that, on average, women who had diabetes had a higher number of pregnancies than women who did not have diabetes. It may also mean that multiple pregnancies are linked to diabetes, maybe because multiple pregnancies place a strain on the body's metabolism, such as gestation-induced diabetes.

The study revealed that non-diabetic had an average glucose of 109.98 mg/dL, and diabetic had an average glucose of 141.26 mg/dL. Glucose levels are significantly higher in diabetic individuals. Elevated glucose is a primary indicator of diabetes and is expected in this case, as it is a key diagnostic criterion [9].

The non-diabetic had an average blood pressure of 68.18 mmHg, and diabetic 70.82 mmHg. Diabetic individuals have slightly higher average blood pressure. Though the difference isn't large, it indicates a tendency for increased blood pressure among diabetic individuals, which could contribute to comorbid conditions such as hypertension.

The average skin thickness of 19.66 mm among non-diabetics and 22.16 mm among diabetic. Diabetic individuals tend to have higher skin thickness, which may be linked to body fat distribution and insulin resistance, factors related to the development of type 2 diabetes. The average non-diabetic insulin is 68.79 $\mu\text{U/mL}$, and diabetic is 100.34 $\mu\text{U/mL}$. Diabetic individuals have higher average insulin levels, which may suggest that many of them are insulin-resistant. Higher insulin levels indicate the body's effort to manage high blood sugar, common in type 2 diabetes.

The average Body Mass Index (BMI) of non-diabetics is 30.30 kg/m^2 , and diabetic 35.14 kg/m^2 . Diabetic individuals tend to have a higher BMI, indicating a higher prevalence of overweight or obesity in this group. This aligns with the known association between obesity and the risk of developing type 2 diabetes.

The Diabetes Pedigree Function of non-diabetic is 0.43, and diabetic is 0.55. Diabetic individuals have a higher diabetes pedigree function, suggesting a greater genetic predisposition to diabetes. This indicates that those with a family history of diabetes are more likely to develop the condition [11].

The average age of non-Diabetic is 31.19 years and diabetic is 37.07 years. Diabetic individuals are older on average, which is consistent with the fact that age is a risk factor for developing diabetes. As people age, the likelihood of insulin resistance and other metabolic issues increases, contributing to the higher prevalence of diabetes in older individuals [2].

Many individuals have elevated glucose and BMI values, which are two key factors associated with a higher risk of diabetes.

The standardized data represents the features after applying the transformation. Each entry in ``standardized_data`` is a transformed version of the corresponding feature in ``X``, standardized according to the formula: The first row starts as ``[0.63994726, 0.84832379, 0.14964075, ..., 1.4259954]``: Each number represents the standardized value of a corresponding feature (like glucose, insulin, BMI, etc.). A value of ``0.63994726`` means that the first feature (e.g., glucose) is slightly above the mean. A ``-1.12339636`` value indicates the second feature (e.g., blood pressure) is significantly below the mean.

Standardization is particularly important in models that rely on distance measures (like K-Nearest Neighbors or Support Vector Machines) or models sensitive to input scale (like neural networks). It prevents features with larger ranges (e.g., glucose, BMI) from dominating those with smaller ranges (e.g., diabetes pedigree function) [4]. Standardization makes the training process more efficient and improves model accuracy, especially for gradient-based algorithms, where unscaled features can slow convergence or cause poor model performance [14]. Thus, while previous studies may have focused on traditional machine learning models (e.g., logistic regression, decision trees), this study adopts the use of a Support Vector Machine (SVM), which has shown improved performance across various healthcare applications [12,15]. A Support Vector Machine (SVM) classifier trained with a linear kernel was being used. This implies that the SVM map also uses a linear model to create the decision functions or just the decision interface to segregate the classes [7,16]. The model is being trained based on the training datasets, which are “X_train”; this contains 614 samples, while “Y_train” contains the corresponding labels for the samples.

SVMs are one of the most often-used algorithms for classification when working with large datasets in specified large-dimensional space [10,17].

SVMs are efficient even when the worlds are in high dimensions and can easily remove noise using the support vectors that define the boundary. The chosen SVM model balances the degree of accuracy with computation time, which will prove useful for this research's diabetes prediction task. Nonetheless, in the case of the availability of computational resources, models like Random Forests

or even Neural Networks can have slightly better predictive accuracy, especially when working with large datasets [17].

The model achieves high accuracy on the training set, indicating that it has learned well the relationships between features and the target labels in the training data [6]. However, checking the test accuracy is important to ensure the model is not overfitting (i.e., performing well on the training data but poorly on unseen test data).

The accuracy score of 87.27% indicates that the model is performing well on the test data. It suggests that the classifier can predict the correct outcomes for most instances in the test set, but about 12.73% of the predictions are incorrect. An accuracy score of 87.27% is generally considered strong in many machine learning tasks, especially if the test dataset is representative of real-world data. However, it is important to note that accuracy alone does not always provide the full picture of model performance, particularly if the dataset is imbalanced (i.e., one class is more frequent than another [10]).

This accuracy score reflects the model's ability to generalize to unseen data (test set). A score above 85% typically indicates a well-trained model that can make reliable predictions. If 87.27% accuracy is insufficient, model tuning (e.g., hyperparameter optimization) or trying different algorithms might improve performance. Thus, while the prediction result is binary (diabetic or not diabetic), this doesn't convey the model's confidence. Exploring other model outputs, like probabilities (using ``predict_proba()``), would be beneficial to assess how confident the model is in its prediction.

Conclusion

This dataset highlights key risk factors associated with diabetes, including high glucose levels, elevated BMI, and genetic predisposition. The results show that the SVM classifier performs well on the training data with an accuracy of 88.66%. The model's accuracy on the test data is 87.27%, which shows that it is effective regarding the data. The predictive system is rather accurate at using the trained model to predict whether someone is in a favourable position based on medical data. In addition, the study proposed more complicated ML models, such as ensemble models, Gradient Boosting, and Deep learning, to take the prediction accuracy to another level. The results of machine learning models should then be compared to clinical judgment to determine how good the machine learning models are.

References

1. Dankwa-Mullan, I., Rivo, M., Sepulveda, M., Park, Y., Snowdon, J., & Rhee, K. (2019). Transforming diabetes care through artificial intelligence: the future is here. *Population health management*, 22(3), 229-242.
2. Guan, Z., Li, H., Liu, R., Cai, C., Liu, Y., Li, J., Wang, X., Huang, S., Wu, L., Liu, D., Yu, S., Wang, Z., Shu, J., Hou, X., Yang, X., Jia, W., & Sheng, B. (2023). Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. *Cell Reports Medicine*, 4(10), 101213. <https://doi.org/10.1016/j.xcrm.2023.101213>
3. Ngugi, M. P., Njagi, J., Kibiti, C., & Ngeranwa, J. (2012). Diagnosis of diabetes mellitus. *International Journal of Experimental Diabetes Research*, 1(2), 24-27. <https://doi.org/10.5923/j.diabetes.20120102.01>
4. Tiwari, K., & Kumar, D. (2018). Recent methods for diagnosis of diabetes mellitus. *International Journal of Innovative Science and Technology*, 3(4), 58-63. IJIST Publishing Group. Available at <http://www.ijistweb.com>
5. Binhowemel, S. F., Alfakhri, M. O., AlReshaid, K. M., & Alyani, A. M. (2023). The role of artificial intelligence in diabetes research, diagnosis, and prognosis: A narrative review. *Journal of Health Informatics in Developing Countries*, 17(2). <http://www.jhdc.org/>
6. Shaukat, Z., Zafar, W., Ahmad, W., Haq, I. U., Husnain, G., Al-Adhaileh, M. H., Ghadi, Y. Y., & Algarni, A. (2023). Revolutionizing diabetes diagnosis: Machine learning techniques unleashed. *Healthcare*, 11(21), 2864. <https://doi.org/10.3390/healthcare11212864>
7. Farajollahi, B., Mehmanavaz, M., Mehrjoo, H., Moghbeli, F., & Sayadi, M. J. (2021). Diabetes diagnosis using machine learning. *Frontiers in Health Informatics*, 10, 65. <https://www.ijmi.ir>
8. Mohsen, F., Al-Absi, H. R. H., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. *npj Digital Medicine*, 6(1), 197. <https://doi.org/10.1038/s41746-023-00933-5>
9. Keshtkar, A., Ayareh, N., Atighi, F., & Hamidi, R. (2024). Artificial intelligence in diabetes management: Revolutionizing the diagnosis of diabetes mellitus; A literature review. *Shiraz E-Medical Journal*, 25(7). <https://doi.org/10.5812/semj-146903>
10. Singh, Y., & Tiwari, M. (2024). Revolutionizing diabetes disease prediction through novel machine learning techniques. *Nano*, 19(4), 2350056. <https://doi.org/10.1142/S179329202350056X>
11. Khalifa, M., & Albadawy, M. (2024). Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management. *Computer Methods and Programs in Biomedicine Update*, 5, 100141. <https://doi.org/10.1016/j.cmpbup.2024.100141>
12. Guan, H., Wang, Y., Niu, P., Zhang, Y., Zhang, Y., Miao, R., Fang, X., Yin, R., Zhao, S., Liu, J., and Tian, J. (2024) The role of machine learning in advancing diabetic foot: a review. *Front. Endocrinol.* 15:1325434. doi: 10.3389/fendo.2024.1325434
13. Ngugi, M. P., Njagi, J., Kibiti, C., & Ngeranwa, J. (2012). Diagnosis of diabetes mellitus. *International Journal of Experimental Diabetes Research*, 1(2), 24-27. <https://doi.org/10.5923/j.diabetes.20120102.01>

14. Buchanan, C., Howitt, M. L., Wilson, R., Booth, R. G., Risling, T., & Bamford, M. (2020). Predicted influences of artificial intelligence on the domains of nursing: Scoping review. *JMIR Nursing*, 3(1), e23939. <https://doi.org/10.2196/23939>
15. Gautier, T., Ziegler, L. B., Gerber, M. S., Campos-Náñez, E., & Patek, S. D. (2021). Artificial intelligence and diabetes technology: A review. *Metabolism*, 124, 154872. <https://doi.org/10.1016/j.metabol.2021.154872>
16. Kaul, S., & Kumar, Y. (2020). Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. *SN Computer Science*, 1(6), 322.
17. Mackenzie, S. C., Sainsbury, C. A. R., & Wake, D. J. (2024). Diabetes and artificial intelligence beyond the closed loop: A review of the landscape, promise, and challenges. *Diabetologia*, 67(2), 223–235. <https://doi.org/10.1007/s00125-023-06038-8>

Publisher's Note Scholar J Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.