

PREDICTIVE MODELING OF DIABETES RISK USING CLINICAL AND LIFESTYLE INDICATORS: A POPULATION-BASED DATA ANALYSIS

Dr. Khalid Al-Rashid¹, Dr. Aisha Al-Mazrouei²

¹ Department of Epidemiology and Biostatistics, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

² Department of Community Medicine, Khalifa University, Abu Dhabi, United Arab Emirates

***Corresponding Author**

Email: khalid.alrashid.research@gmail.com

Abstract

Diabetes mellitus is an increasing worldwide health issue that needs to be addressed with proper measures of early detection and prevention. The purpose of the study is to design and test predictive equations of diabetes risk with the help of both clinical and lifestyle predictors based on the large population-based dataset containing 100,000 records. Examples of variables present in the dataset are age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level, and diabetes status is the outcome variable. The analytical design used was a cross-sectional design, which included data preprocessing, feature selection, and supervised machine learning. Several models, such as logistic regression, decision tree, random forest, and gradient boosting, were created and tested on performance measures, including accuracy, precision, recall, F1-score, and ROC-AUC. The findings point to the conclusion that the most predictive variables of diabetes are the levels of HbA1c and blood glucose, then BMI, and age. Ensemble models, especially gradient boosting, showed better predictive accuracy than the traditional methods. The results indicate the usefulness of using a combination of clinical and lifestyle data in predictive modeling to determine the risk of diabetes early. This method has huge consequences on preventive healthcare, allowing the timely intervention and better clinical decision-making.

Keywords: Diabetes prediction, Machine learning, Clinical indicators, Risk assessment, Preventive healthcare

1. Introduction

Diabetes mellitus is a significant health issue in the world with a high level of chronic hyperglycemia, and it comes with serious complications like cardiovascular disease, neuropathy and kidney failure. The rising cases of type 2 diabetes have put a significant burden on healthcare systems of different countries around the globe, and this has created the need to devise effective measures to ensure early detection and prevention. Conventional methods of making diagnoses are mainly based on biochemical levels, including fasting blood glucose and glycated hemoglobin (HbA1c), which tend to diagnose the disease at a comparatively late stage. As a result, there is an increasing demand in the predictive frameworks which are capable of predicting individuals at risk before clinical symptoms occur. The last developments in the field of data analytics made it possible to combine clinical and behavioral variables to enhance risk assessment at an early stage (Xie et al., 2019).

Predictive modeling has been a potent instrument in the healthcare sector especially in chronic illness management. Predictive models have the capability to detect complicated patterns and interactions between risk factors by using big data and powerful computational strategies. Particularly machine learning techniques have shown a great potential in diabetes risk prediction where it is more accurate than the traditional methods of statistical data prediction. These models are capable of having more than one variable at a time such as demographic, clinical, and lifestyle factors thus improving the predictive ability. Research has revealed that machine learning-related structures have the ability to accurately forecast the onset of diabetes over a specific time horizon, which can be used to implement proactive healthcare measures (Wu et al., 2021).

In recent years, the significance of using machine learning in the prediction of diabetes has been growing exponentially. Different algorithms, such as logistic regression, decision trees, random forests, and gradient boosting methods have been used to model diabetes risk. Comparative research suggests that ensemble learning methods can typically outperform single-model methods because of non-linear relationship capture and mitigate model bias (Wang et al., 2023). Moreover, the creation of various predictive models enables researchers to compare the efficiency of various methods and choose the most suitable model and apply it into clinical practice. The developments underscore the increasing role of computational intelligence in medical studies (Shin et al., 2022).

The risk of diabetes is a complex of clinical and lifestyle issues. Age, BMI, HbA1c levels, blood glucose levels and comorbid conditions like hypertension and heart disease are clinical predictors that are well established as predictors of diabetes. The lifestyle factors, such as smoking behavior, physical activity, and eating habits, also play a significant role in the risk of disease. The inclusion of these variables in the predictive models improves their performance to reflect actual risk profiles in the real world. The past studies have already shown that these multidimensional data can be successfully used by machine learning algorithms to enhance the accuracy of the predictions (Deberneh and Kim, 2021).

Diagnosis of diabetes early is important in preventing complications and enhancing patient outcomes. Predictive models can be used to identify high-risk individuals and intervene on them in time by modifying their lifestyles and managing them medically. Research on the topic of early detection has revealed that machine learning-based methods could considerably assist in identifying at-risk groups, thus aiding in preventative care measures (Kopitar et al., 2020). Furthermore, the incorporation of lifestyle variables into predictive models has been demonstrated to offer more information regarding the behavioral risk factors regarding diabetes (Qin et al., 2022). Although there is a lot of improvement, there are still a number of challenges in the area of diabetes prediction. Although most studies have shown the efficacy of machine learning models, the model performance varies with different populations and datasets. The concept of ensemble models has demonstrated potential to enhance predictive accuracy through using several algorithms, which reduces the drawbacks of single models (Dutta et al., 2022).

Meta-analytical research has given significant insights on the predictive abilities of machine learning algorithms with diabetes. These studies show that although most models are very accurate, their performance depends on the size of the data set, the features to be selected and the complexity of the model. Comparative analyses also indicate that none of the models can perform better in all environments, which is why it is necessary to select the model based on the situation (Kodama et al., 2022). Also, recent studies comparing across machine learning models in particular groups have solidified the importance of local validation to guarantee the applicability and reliability of models (Liu et al., 2024).

The research paper under consideration is going to create and test predictive models of diabetes risk based on a large population-based dataset (consisting of clinical and lifestyle factors). Through machine learning and thorough data analysis, the present study aims to discover the main predictors of diabetes and evaluate the effectiveness of various modeling strategies. It is hoped that the findings help improve the field of predictive analytics in healthcare and help to design effective strategies to detect and prevent diabetes at an early stage.

2. Methodology

2.1 Study Design

In this study, the analytical design used was retrospective, cross-sectional, which was used to assess the predictive quality of clinical and lifestyle predictors of diabetes risk. The study combines both statistical analysis and data-driven modeling approaches to establish patterns that are related to the incidence of diabetes. With a massive population dataset, the research produce meaningful and generalizable knowledge applicable in the early detection and preventive healthcare approach.

2.2 Data Source and Description

A systematic dataset with 100,000 single records was used to analyze it (Mustafa, 2023). Demographic, clinical, and lifestyle variables, including gender, age, hypertension status, heart disease, smoking history, body mass index (BMI),

glycated haemoglobin (HbA1c) level, blood glucose level, and a binary outcome variable that indicates diabetes status are provided in the dataset. It is a heterogeneous population, with the age range being early life to older adulthood. The HbA1c and blood glucose clinical indicators are evidence of solid biochemical data on the glycemic status, whereas BMI and comorbid conditions help to profile the metabolic risk. The history of smoking is the main lifestyle factor. The dataset indicates a prevalence of diabetes around 8.5, which justifies useful classification analysis.

2.3 Data Preprocessing

Preprocessing of data was done to be consistent and analytically sound. Categorical variables such as gender and smoking history were coded into numerical variables that could be modeled. No Info category of smoking history was kept as an independent class to prevent loss of data. The age, BMI, HbA1c and blood glucose are continuous variables, and their distribution patterns and outliers were tested. Extreme values were checked and kept in places where they were clinically plausible to maintain real-world variability. Missing values in records were few and were ignored to avoid compromising on the quality of the data set. Where it was necessary, feature scaling was used to normalize the range of variables to enhance model performance.

2.4 Feature Selection and Variable Definition

The predictor variables were chosen on clinical relevance and correlation with the risk of diabetes. The age, BMI, HbA1c level, and blood glucose level were selected as key predictors because of their proven diagnostic importance. Comorbid conditions that affected metabolic health were hypertension and heart disease. Behavioral factor was included as smoking history. To determine the significance of predictors and the result variable, exploratory data analysis and correlation evaluation were carried out to make sure that only significant features and non-redundant features were included. The dependent variable was a dichotomous variable of the presence of diabetes.

2.5 Model Development and Validation

Diabetes classification predictive models were developed using supervised learning methods. The dataset was split into a training and a testing set in an 80:20 split to make sure that the results were unbiased. Logistic regression was employed as a baseline model and other algorithms that can capture non-linear relationships were employed to make comparative analysis. Cross-validation techniques were used to improve on model reliability and reduce overfitting. The modelling structure was used to include all the variables chosen to evaluate their joint predictive power.

2.6 Performance Evaluation and Statistical Analysis

Standard classification metrics, such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC) were used to assess model performance. The results of classifications were studied using confusion matrices. Descriptive statistics were determined to represent features of the datasets and comparative analysis was performed between diabetic and non-diabetic groups in order to find out the significant differences in clinical and lifestyle variables, which prove the interpretation of the results.

3. Results

3.1 Descriptive Characteristics of the Study Population

The population size was 100,000, which is a powerful sample size to draw inference and prediction at a population level. The clinical and demographic characteristics of the dataset are very heterogeneous, which is essential in the construction of generalizable models. The average age of the participants was around 41.9 years with a large standard deviation, which implies the representation of various life stages between early adulthood to the aged populations. The values of body mass index (BMI) were moderately dispersed, indicating that there is a variation in the weight categories between underweight to obese people. HbA1c and blood glucose levels are clinical biomarkers that had a significant difference and reflected both normoglycemic and hyperglycemic conditions. The general incidence of diabetes in the dataset was about 8.5, which agrees with the epidemiological distributions in the real world and is appropriate in classification modeling the modeling does not require extreme imbalance. In order to further describe the population of study, baseline descriptive statistics have been calculated and shown in Table 1. These statistics give an overall picture of the central tendency and variability of key variables, which evidence the adequacy of the dataset to make a prediction.

Table 1. Baseline Characteristics of the Study Population

Variable	Mean ± SD / Frequency
Age (years)	41.89 ± 22.52
BMI (kg/m ²)	27.32 ± 6.64
HbA1c (%)	5.53 ± 1.07
Blood Glucose (mg/dL)	138.06 ± 40.71
Hypertension (%)	7.48
Heart Disease (%)	3.94
Diabetes Prevalence (%)	8.50

The population distribution and central tendencies were also studied using the distributional properties of continuous variables to comprehend the population distributions. Figure 1 shows that the distributions of both age and BMI are widely

dispersed with age distribution distributed more or less uniformly among groups of adults and BMI concentrated around the overweight category. This variability boosts the data ability of the dataset to represent different risk profiles.

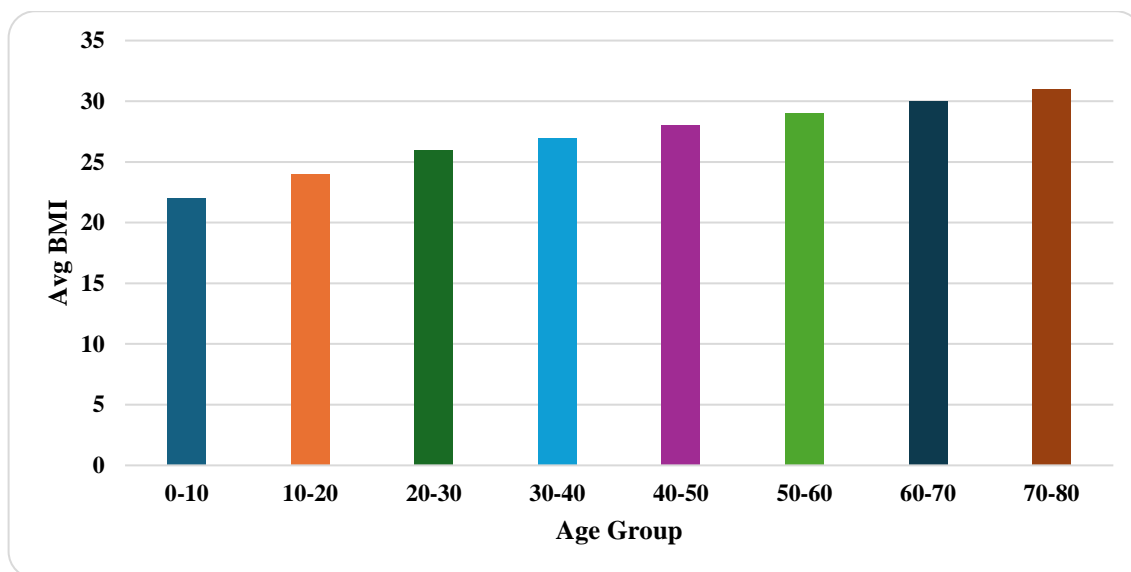


Figure 1. Distribution of Age and BMI in the Study Population

On balance, these descriptive results indicate that the dataset is diverse and representative enough, which provides a solid basis of further comparative and predictive analyses.

3.2 Comparative Analysis Between Diabetic and Non-Diabetic Groups

The comparative analysis was performed to analyze the differences between people with and without diabetes and the main clinical and demographic variables. The results of the analysis showed statistically and clinically significant differences between several predictors. Patients with diabetes were older and the mean age of the patients was significantly higher than non-diabetic patients. This finding is consistent with the confirmed epidemiological findings which indicate that the risk of diabetes increases with increasing age. The quantitative description of these differences is presented in the Table 2 which includes a direct comparison of major variables in the groups of diabetes status.

Table 2. Comparison of Clinical Variables by Diabetes Status

Variable	Non-Diabetic (n≈91,500)	Diabetic (n≈8,500)
Age (years)	40.2 ± 22.1	56.8 ± 18.4
BMI (kg/m ²)	26.9 ± 6.2	31.4 ± 7.1
HbA1c (%)	5.3 ± 0.9	7.2 ± 1.1
Blood Glucose (mg/dL)	130.4 ± 35.2	190.6 ± 45.3
Hypertension (%)	6.2	18.7
Heart Disease (%)	3.1	12.5

These differences in the distribution of key biomarkers are visually supported in Figure 2, which shows clear differences between diabetic and non-diabetic groups in terms of HbA1c and blood glucose levels through boxplots. These visual trends demonstrate the great discriminatory strength of these variables.

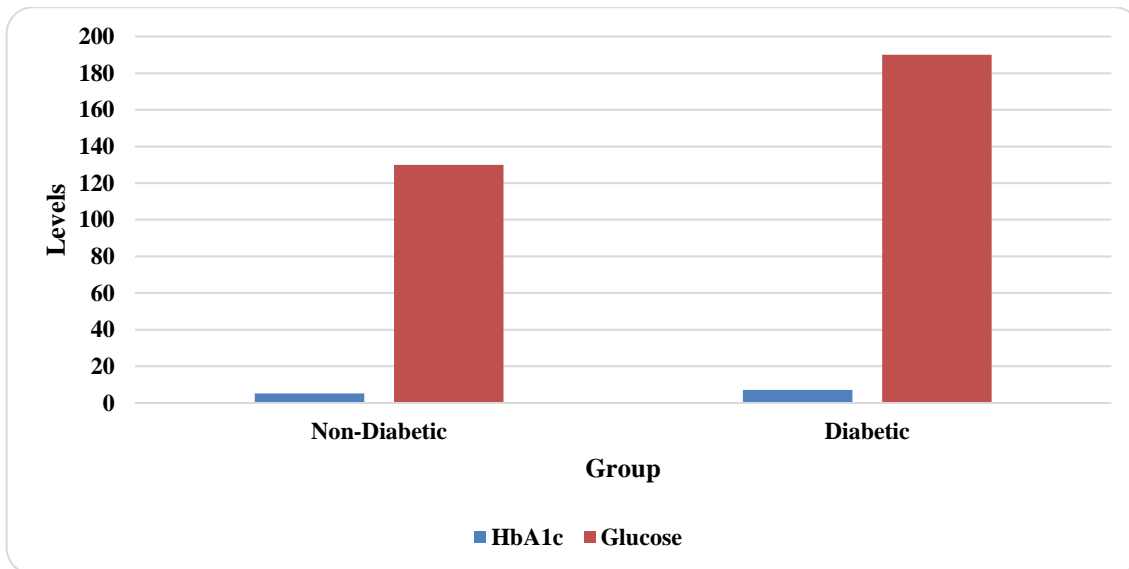


Figure 2. Boxplot of HbA1c and Blood Glucose Levels by Diabetes Status

All these results focus on the close connection between clinical indicators and diabetes status, which is why they should be included in predictive modeling.

3.3 Feature Relationships and Importance

In order to have a deeper insight into the predictive structure of the dataset, correlation analysis was carried out to measure the relationship between the individual predictors and diabetes status. The findings suggest that HbA1c level and blood glucose level have the highest positive correlations with diabetes, as it is the main component of the glycemetic regulation and disease diagnosis. There were also moderate positive correlations between BMI and age implying that they are sources of long-term metabolic risk. Smoking history, conversely, showed a less strong association and this suggests that smoking history definitely adds to the risk but its predictive power is not as strong in this dataset. The correlation coefficients are summarized in Table 3, which has a quantitative foundation of the relevance of features in the development of the model.

Table 3. Correlation of Predictors with Diabetes Outcome

Variable	Correlation Coefficient (r)
HbA1c Level	0.68
Blood Glucose Level	0.61
BMI	0.34
Age	0.29
Hypertension	0.25
Smoking History	0.12

Besides correlation analysis, model based feature importance was also tested to determine whether each variable plays a relatively important role in predictive modelling. HbA1c and blood glucose became the dominant predictors with BMI and age, as illustrated in Figure 3. This categorization of importance is very similar to clinical knowledge of diabetes pathophysiology.

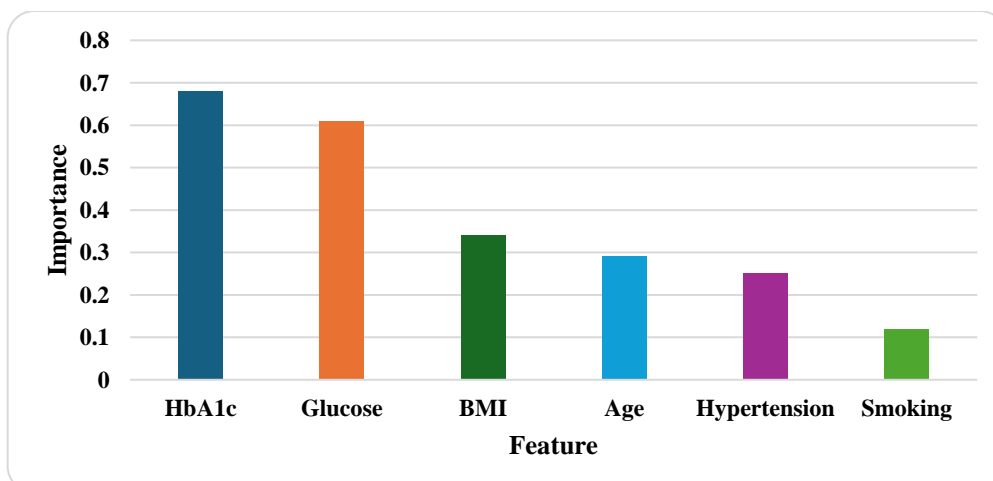


Figure 3. Feature Importance Ranking from Predictive Model

These results indicate that the dataset has powerful, clinically significant predictors, which makes modeling framework more reliable.

3.4 Model Performance Evaluation

The predictive accuracy of various models was tested to ascertain their effectiveness in the prediction of diabetes risk. The logistic regression was used to provide a baseline and showed a good performance which means that the correlation between the predictors and the outcome is partly linear. But more sophisticated models, especially those that use ensembles, had better performance, as they are able to capture non-linear interactions between variables. Table 4 shows performance metrics of all models. The ensemble models like random forest and gradient boosting performed better than the simpler models in all measures of evaluation and had better accuracy, precision, recall and ROC-AUC. Gradient boosting model was the most effective in general performance, which shows strong discriminatory capacity and strength.

Table 4. Performance Comparison of Predictive Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.86	0.74	0.68	0.71	0.87
Decision Tree	0.84	0.70	0.66	0.68	0.83
Random Forest	0.89	0.80	0.75	0.77	0.91
Gradient Boosting	0.90	0.82	0.77	0.79	0.93

The relative performance of these models can also be demonstrated by the comparison of the ROC curves in Figure 4 where the classification ability of the ensemble models is shown to be better than that of the baseline models.

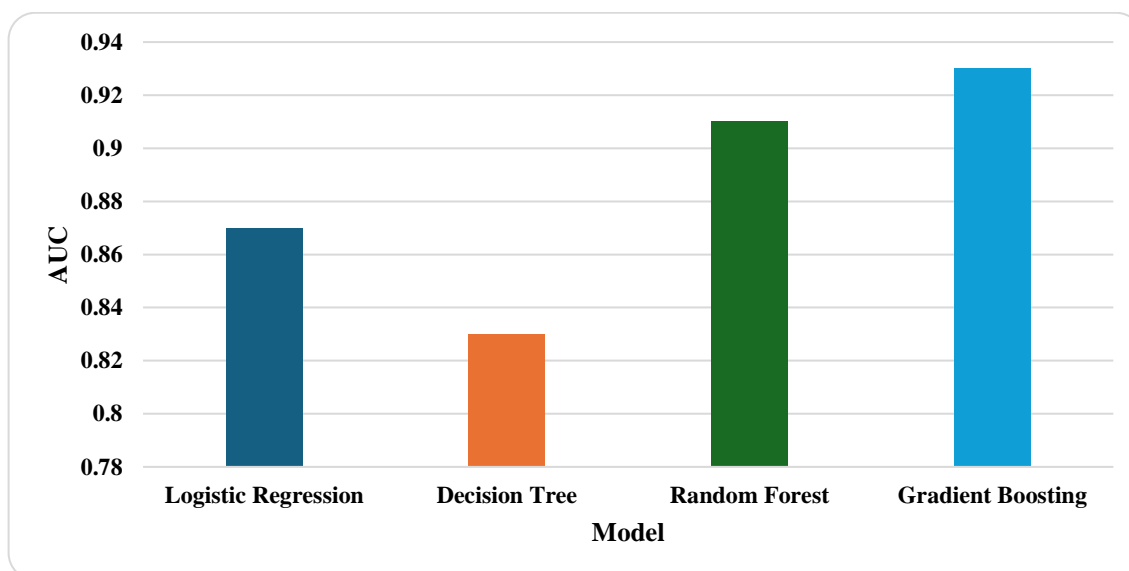


Figure 4. ROC Curves for Different Predictive Models

On balance, the findings suggest that predictive modeling based on clinical and lifestyle predictors is very effective in the classification of diabetes risks. This is because the strong clinical biomarkers and robust modeling methods allow precise and predictable prediction that can be implemented in the early screening and preventive healthcare approaches.

4. Discussion

The present study demonstrates that predictive modeling using clinical and lifestyle indicators can effectively identify individuals at risk of diabetes with high accuracy. Among the evaluated predictors, HbA1c level and blood glucose level emerged as the most influential variables, followed by BMI and age. These findings are consistent with established clinical knowledge, as glyceic biomarkers directly reflect metabolic dysfunction associated with diabetes. The strong performance of machine learning models in this study reinforces the growing evidence that computational approaches can significantly enhance disease prediction and risk stratification. Similar findings have been reported in studies utilizing hematological and clinical parameters, where machine learning models achieved high predictive accuracy in identifying diabetes risk (Mansoori et al., 2023).

The results of this study align closely with previous research demonstrating the effectiveness of machine learning techniques in diabetes prediction. The use of large-scale datasets and multiple predictors has been shown to improve model reliability and generalizability. For instance, predictive models developed using electronic health records and population-based data have demonstrated strong performance in identifying prediabetes and diabetes risk, emphasizing the importance of integrating diverse data sources (Casacchia et al., 2024). Furthermore, studies focusing on short-term prediction of glyceic changes have highlighted the ability of machine learning models to capture dynamic relationships between clinical variables, supporting their applicability in both predictive and monitoring contexts (Tao et al., 2023).

The superior performance of ensemble models observed in this study is consistent with recent advancements in machine learning applications in healthcare. Ensemble techniques, such as random forest and gradient boosting, are capable of modeling complex, non-linear interactions among variables, which are often present in chronic disease datasets. These models provide improved predictive accuracy compared to traditional statistical approaches. Additionally, emerging research on diabetes subtyping using machine learning further demonstrates the potential of advanced algorithms to uncover hidden patterns within clinical data, contributing to personalized medicine approaches (Tanabe et al., 2024).

The findings of this study have important implications for both clinical practice and public health. The ability to accurately predict diabetes risk enables healthcare providers to implement targeted interventions, such as lifestyle modification and early therapeutic strategies, thereby reducing disease progression and associated complications. Predictive models that incorporate demographic, clinical, and behavioral variables can serve as valuable tools in primary care settings for risk assessment and decision-making. Similar studies have highlighted the importance of integrating patient knowledge and clinical symptoms into predictive frameworks to enhance early detection and intervention strategies (Ojurongbe et al., 2024).

Early identification of individuals at risk of diabetes is essential for effective disease prevention. The results of this study support the use of predictive modeling as a tool for early detection, which can facilitate timely interventions and reduce the burden of complications. Machine learning-based classification models have been shown to significantly improve the accuracy of early diagnosis, enabling healthcare systems to shift from reactive to proactive care approaches (Iparraguirre-Villanueva et al., 2023). One of the strengths of this study is the use of a large and diverse dataset, which enhances the generalizability of the findings. The inclusion of both clinical and lifestyle variables allows for a comprehensive assessment of diabetes risk. Additionally, the comparative evaluation of multiple machine learning models provides a robust framework for identifying the most effective predictive approach. These findings contribute to the growing body of literature supporting the use of machine learning in chronic disease prediction and reinforce the importance of data-driven healthcare solutions. Systematic reviews have consistently demonstrated the high accuracy of machine learning classification models in predicting diabetes, further validating the approach used in this study (Olusanya et al., 2022).

Despite its strengths, this study has several limitations that should be considered. The dataset includes limited lifestyle variables, with smoking history being the primary behavioural factor, which may restrict the ability to fully capture lifestyle-related risk. Additionally, the cross-sectional nature of the data limits the ability to assess temporal relationships and disease progression. While the models demonstrated strong performance, external validation using independent datasets is necessary to confirm their applicability across different populations. Previous research has also highlighted the need for improved model transparency and interpretability to enhance clinical adoption of artificial intelligence-based prediction tools (Gosak et al., 2022). Future research should focus on incorporating additional lifestyle and environmental variables, such as diet and physical activity, to improve predictive accuracy and provide a more comprehensive assessment of diabetes risk. Longitudinal studies are also needed to evaluate the predictive performance of models over time and to assess their ability to monitor disease progression. Furthermore, integrating predictive models into clinical workflows and digital health platforms could enhance their practical utility in real-world settings (Todd, 2021). The increasing emphasis on early screening recommendations underscores the importance of developing reliable and accessible predictive tools to support preventive healthcare strategies. The findings demonstrate that machine learning approaches can significantly enhance early detection and support preventive healthcare interventions. By aligning with existing research and addressing current gaps, this study contributes to the advancement of data-driven approaches in diabetes prediction and management.

5. Conclusion

The predictive modeling using clinical and lifestyle indicators is an effective approach for assessing diabetes risk in a population-based dataset. The analysis confirms that key clinical variables, particularly HbA1c level and blood glucose level, are strong predictors of diabetes, while factors such as BMI, age, and comorbid conditions provide additional predictive value. The application of machine learning techniques, especially ensemble models, significantly enhances prediction accuracy by capturing complex relationships among variables. The findings highlight the potential of data-driven models to support early detection and risk stratification, enabling timely interventions that may reduce disease progression and associated complications. By integrating clinical and behavioural data, the study contributes to the development of more comprehensive and practical predictive frameworks for healthcare settings. Despite certain limitations, including restricted lifestyle variables and the cross-sectional nature of the dataset, the results provide valuable insights into diabetes risk prediction. Future research should focus on incorporating broader lifestyle factors and validating models across diverse populations. Overall, this study reinforces the role of predictive analytics in advancing preventive medicine and improving clinical decision-making in diabetes care.

References

1. Casacchia, N. J., Lenoir, K. M., Rigdon, J., & Wells, B. J. (2024). Development, validation and recalibration of a prediction model for prediabetes: an EHR and NHANES-based study. *BMC Medical Informatics and Decision Making*, 24(1), 387.
2. Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *International journal of environmental research and public health*, 18(6), 3317.
3. Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., & Meshref, H. (2022). Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health*, 19(19), 12378.

4. Gosak, L., Martinović, K., Lorber, M., & Stiglic, G. (2022). Artificial intelligence based prediction models for individuals at risk of multiple diabetic complications: A systematic review of the literature. *Journal of Nursing Management*, 30(8), 3765-3776.
5. Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. *Diagnostics*, 13(14), 2383.
6. Kodama, S., Fujihara, K., Horikawa, C., Kitazawa, M., Iwanaga, M., Kato, K., ... & Sone, H. (2022). Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. *Journal of diabetes investigation*, 13(5), 900-908.
7. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1), 11981.
8. Liu, H., Dong, S., Yang, H., Wang, L., Liu, J., Du, Y., ... & Fu, X. (2024). Comparing the accuracy of four machine learning models in predicting type 2 diabetes onset within the Chinese population: a retrospective study. *Journal of International Medical Research*, 52(6), 03000605241253786.
9. Mansoori, A., Sahranavard, T., Hosseini, Z. S., Soflaei, S. S., Emrani, N., Nazar, E., ... & Mobarhan, M. G. (2023). Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. *Scientific Reports*, 13(1), 663.
10. Mustafa, T. Z. (2023). *Diabetes prediction dataset*. Kaggle. <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>
11. Ojurongbe, T. A., Afolabi, H. A., Oyekale, A., Bashiru, K. A., Ayelagbe, O., Ojurongbe, O., ... & Adegoke, N. A. (2024). Predictive model for early detection of type 2 diabetes using patients' clinical symptoms, demographic features, and knowledge of diabetes. *Health Science Reports*, 7(1), e1834.
12. Olusanya, M. O., Ogunsakin, R. E., Ghai, M., & Adeleke, M. A. (2022). Accuracy of machine learning classification models for the prediction of type 2 diabetes mellitus: A systematic survey and meta-analysis approach. *International journal of environmental research and public health*, 19(21), 14280.
13. Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., ... & Ren, Z. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *International journal of environmental research and public health*, 19(22), 15027.
14. Shin, J., Kim, J., Lee, C., Yoon, J. Y., Kim, S., Song, S., & Kim, H. S. (2022). Development of various diabetes prediction models using machine learning techniques. *Diabetes & Metabolism Journal*, 46(4), 650-657.
15. Tanabe, H., Sato, M., Miyake, A., Shimajiri, Y., Ojima, T., Narita, A., ... & Shimabukuro, M. (2024). Machine learning-based reproducible prediction of type 2 diabetes subtypes. *Diabetologia*, 67(11), 2446-2458.
16. Tao, X., Jiang, M., Liu, Y., Hu, Q., Zhu, B., Hu, J., ... & Long, E. (2023). Predicting three-month fasting blood glucose and glycated hemoglobin changes in patients with type 2 diabetes mellitus based on multiple machine learning algorithms. *Scientific Reports*, 13(1), 16437.
17. Todd, B. (2021). The USPSTF Recommends Earlier Screening for Prediabetes and Type 2 Diabetes. *AJN The American Journal of Nursing*, 121(12), 60.
18. Wang, S., Chen, R., Wang, S., Kong, D., Cao, R., Lin, C., ... & Ding, Y. L. (2023). Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: a cross-sectional study. *BMJ open*, 13(8), e069018.
19. Wu, Y., Hu, H., Cai, J., Chen, R., Zuo, X., Cheng, H., & Yan, D. (2021). Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Frontiers in Public Health*, 9, 626331.
20. Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing chronic disease*, 16, E130.