Enhancing Early Detection of Type II Diabetes with Machine Learning: A Performance Evaluation

Fheng Sin Ern[1], Norziha Binti Megat Mohd Zainuddin*[2], Nurulhuda Firdaus Binti Mohd. Azmi[3], Nurazean Binti Maarop[4] & Wan Azlan Bin Wan Hassan[5]

[1,2,3,4]*Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*

[5]*Faculty of Communication, Visual Art & Computing, Universiti Selangor, Malaysia*

[1]*fheng@graduate.utm.my*, [2]*norziha.kl@utm.my*, [3]*huda@utm.my*
[4]*nurazean.kl@utm.my*, [5]*wan.azlan@unisel.edu.my*

Article history

Received:
4 May 2025

Received in revised form:
10 May 2025

Accepted:
20 May 2025

Published online:
27 June 2025

*Corresponding author
norziha.kl@utm.my

Abstract

Type II diabetes is a common issue nowadays and takes a longer time to detect. Detection of diabetes greatly relies on the clinical results from medical professionals, which require a significant amount of time, manpower, and expenses. Machine learning findings may be used as the reference in gaining preliminary understanding about the disease. It is crucial to achieve early detection of type II diabetes in a feasible and efficient manner for broader populations. This study aims to evaluate the performance of selected machine learning models for type II diabetes. The dataset of Behavioral Risk Factor Surveillance System from 2021 was used in this study. Five attributes of high blood pressure, high cholesterol, BMI, general health, and walking difficulty with the highest Cramer's V correlation were selected. Four machine learning models were identified through a literature review, including: (i) Decision Tree, (ii) Neural Network, (iii) Random Forest, (iv) Logistic Regression, and (v) AdaBoost, and were analyzed in the study. The performance of each machine learning model was evaluated based on accuracy, precision, sensitivity, and F1-score. All the algorithms showed acceptable performance, ranging from 68.8% to 74.7%. Neural Network showed the highest accuracy and F1-score of 71.0% and 71.9%, respectively. Decision Tree had the highest sensitivity of 74.7% among all the algorithms. This project suggests Neural Network as the algorithm with the best overall performance in the diabetes prediction model and suggests Decision Tree as the most suitable algorithm specifically for screening diabetes. Preliminary diagnosis based on the interpretation of risk factors may greatly reduce the workload of clinical professionals in identifying the high risk group for type II diabetes to proceed further clinical diagnostic.

Keywords: *Diabetes, Machine Learning, Disease Prediction, Feature Selection, Risk Factors.*

1. Introduction

Diabetes Mellitus, commonly referred to as diabetes, is a disease that involves the condition of hyperglycemia. There are three main types of diabetes: type I diabetes, type II diabetes, and gestational diabetes [1, 2]. As of today, diabetes is a chronic disease with no definite cure for either type I or type II diabetes [3]. Type I diabetes is less common and is a genetic-related disease. Type II diabetes is getting more attention from the public due to its high prevalence. The occurrence of type II

*Corresponding author. norziha.kl@utm.my
DOI: 10.11113/oiji2025.13n1.328

diabetes involves a combination of lifestyle and environmental variables. The accumulation of environmental pressures like stress, unhealthy diet, and insufficient physical exercise may lead to the expression of the diabetic phenotype [4, 5].

This article aims to evaluate the performance of selected machine learning models in predicting type II diabetes based on several evaluation methods which are Decision Tree, Neural Network, Random Forest, Logistic Regression, and AdaBoost. Implementing machine learning may enhance the scalability and accuracy of diagnostics. A preliminary diagnostic solution from prediction models is demanded to provide insights without any lab tests. Therefore, it is crucial to improve early detection of diabetes in a feasible and efficient approach to manage type II diabetes in broader populations.

2. Literature Review

Type II diabetes shows a slow onset of symptoms that are different from the acute presentation of pathological symptoms from type I diabetes. This leads to difficulties in diagnosing type II diabetes due to the mild symptoms at the beginning of diabetes, which may remain undiagnosed until the onset of severe complications [6, 7]. Table 1 provides the summary of the differences between type I diabetes and type II diabetes.

Table 1. Summary of the Differences between Type I Diabetes and Type II Diabetes

Diabetes	Type I	Type II
Related to	Genetic related disease	Lifestyle and environmental variables related disease
Onset of symptoms	Acute onset of symptoms	Slow onset of symptoms

Type II diabetes could be carefully managed or even reversed for pre-diabetes by coping with and reducing environmental pressures from lifestyle factors [5]. Lifestyle interventions with modifications on the risk factors for pre-diabetes, such as obesity, diet, and increasing physical activities, have been shown to have beneficial effects in reducing the risk of diabetes [8]. Pre-diabetes is the condition of having elevated blood glucose levels higher than normal but below the threshold for a diabetes diagnosis. The risk factors for diabetes or complications of diabetes could be controlled before getting worse [9]. Identifying pre-diabetes or mild diabetes subjects would be beneficial in starting lifestyle treatment at the earliest possible stage.

According to the report from the National Diabetes Registry (NDR) for 2020, published by the Disease Control Division, Ministry of Health Malaysia, 902,991 active diabetes patients were reported in 2020, and 99% of them had type II diabetes. The prevalence of diabetes is increasing, with type II diabetes accounting for the majority. The Ministry of Health Malaysia emphasizes the cruciality of identifying prediabetes and type II diabetes mellitus within the population, particularly among those at high risk. According to the 6th edition of the Clinical Practice Guidelines (CPG) on the Management of Type II Diabetes Mellitus issued in 2020, diagnosis tests for type II diabetes involve clinical assessments and three laboratory tests to

be performed at medical facilities. The laboratory tests of (i) fasting plasma glucose, (ii) oral glucose tolerance test, or (iii) HbA1c are performed for the diagnosis of type II diabetes. In symptomatic individuals, a single abnormal test result is sufficient for diagnosis. However, two abnormal test results, including plasma glucose and HbA1c, are required as confirmatory diagnoses for asymptomatic individuals. All laboratory tests for diabetes diagnosis involve the collection of blood samples at least once or more. Both fasting plasma glucose and oral glucose tolerance tests mandate eight hours of overnight fasting. HbA1c testing does not require fasting, but the test is subject to certain factors and conditions affecting its accuracy. These include the presence of hemoglobin variants and certain medications in individuals [10]. An asymptomatic individual may need to repeat HbA1c testing four weeks after the first positive result for diabetes. These diagnostic tests involve complex procedures that are time consuming and shall only be carried out by medical professionals along with specialized equipment [11].

The application of comprehensive machine learning models is useful in tracking the health condition of patients or identifying high-risk groups for related diseases [12]. In disease prediction, the symptoms and medical records of a patient are used to predict the occurrence of specific diseases using machine learning models [7]. Diabetes prediction can also be performed by constructing predictive machine learning models to achieve greater accuracy in identifying early stages of pre-diabetes and diabetes according to diabetes indicators [13]. Machine learning findings may be used as the reference in gaining preliminary understanding about the disease and reducing the workload of healthcare professionals [14].

A risk factor or determinant for a particular disease is reported to be correlated or have a correlation with the disease but not proven to be causal. Lifestyle interventions with modifications on the risk factors have been shown to have beneficial effects in reducing the risk of diabetes [15]. Based on previous studies reviewed, the attributes that are risk factors of depression, age, high body mass index (BMI), cardiovascular disease, hypertension, and income are found to be highly correlated with diabetes. From previous studies, these factors could be further added with sleep quantity, physical inactivity, and smoking [2, 16].

Many studies have been done in predicting diabetes in applying predictive machine learning models. A part of those studies involves the using of clinical diagnostic measures to be carried out by professional clinicians or in hospitals for example HbA1c value, insulin level, glucose level or diabetes pedigree function. However, most of the clinical diagnostic measures involves the using of invasive collection methods and aim in confirming the diagnosis of diabetes [17]. Some of the studies involve the using of dataset with selection bias. A few open data related to diabetes are showing less than 1000 subjects. These may be causing sampling bias or under coverage. Besides that, it is preferable to discover more on type II diabetes that is related to environmental pressures. This is due to type I diabetes and gestational diabetes are incident related diseases and may not be demonstrating any sign or symptoms for prediction before disease. Apart for that, open data related diabetes is limited and leads to the constraint of diabetes prediction study.

A total of 11 studies were reviewed in this project to gain an understanding of similar topics and approaches [13, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26]. All the studies focused on diabetes prediction using machine learning. The Behavioral Risk Factor Surveillance System (BRFSS) dataset from different years is extensively

used in health research. It contains a comprehensive scope of health-related topics and a large sample size with over 400,000 responses each year. The data collected through BRFSS is publicly available and can be accessed on the website of the Centers for Disease Control and Prevention in a de-identified format.

The machine learning classifiers of Decision Tree and Neural Network are frequently employed in disease prediction, including diabetes prediction. Decision Tree was applied and proposed as a suitable model for diabetes prediction in both BRFSS and Pima Indian datasets. The Decision Tree model showed the highest sensitivity of 51.6% using the BRFSS 2014 dataset, which was preferable for type II diabetes screening with the highest detection rate [16]. Using the BRFSS 2015 dataset, Decision Tree achieved a high accuracy of 92% after balancing the classes with the application of the up-sampling technique [20]. The Decision Tree model was an excellent option for type II diabetes screening with high sensitivity and precision in detecting false positive and negative cases after balancing the classes in BRFSS 2020 [23]. A modified version of Decision Tree achieved an accuracy of 99.9% in predicting diabetes when trained using the Pima Indian dataset [21]. Two-class boosted Decision Tree was found to have an AUC score of 99.1% in diabetes prediction using a private dataset from Taipei Municipal Medical Centre [18]. Meanwhile, Neural Network was recommended as a suitable model with the highest specificity (90.2%), AUC (0.7949), and accuracy (82.4%) for diabetes prediction [16]. The experimental results demonstrated that Neural Network had the highest accuracy for the Pima Indian dataset using all available attributes [26].

For ensemble classifiers, Random Forest and AdaBoost were found to be commonly used in diabetes prediction. Eight studies implemented Random Forest for predicting diabetes, demonstrating reliable performance in five studies [13, 19, 22, 23, 26]. According to this study, the Random Forest machine learning model had the highest accuracy and F1-score of 82% for the BRFSS 2015 dataset [13]. Random Forest was highlighted for achieving the highest accuracy in diabetes prediction using the dataset from Sylhet Diabetes Hospital [19]. When applied to the Pima Indian dataset, Random Forest achieved the highest accuracy of 91% with the ADASYN oversampling method [22]. Both Random Forest and Decision Tree were suggested as excellent options for type II diabetes screening using BRFSS 2020 [23]. Random Forest also demonstrated the highest accuracy in diabetic prediction using a private dataset from Luzhou Hospital [26]. AdaBoost was employed and proposed as the classifier with the best performance when using the transformed Pima Indian dataset [24].

Attribute selection was applied in studies using the BRFSS dataset [13, 16, 23, 25]. The study that did not involve attribute selection used a cleaned version of the BRFSS dataset available on Kaggle with fewer attributes [20]. Attribute selection aims to eliminate insignificant attributes and enhance prediction performance [27]. The raw BRFSS survey dataset typically contains around 300 attributes, making attribute selection necessary to remove unrelated attributes.

Oversampling techniques were applied to address the imbalance between majority and minority classes [13, 16, 19, 20, 22, 23, 25]. In diabetes prediction, the minority class of "yes diabetes" was of interest, and balancing classes were crucial for enhancing prediction accuracy for this class [28]. Synthetic Minority Over-sampling Technique (SMOTE) was used to increase instances of the minority class

by generating synthetic instances through random selection and interpolation of nearby instances [29].

According to studies related to diabetes prediction, machine learning models such as Neural Network, Decision Tree, Logistic Regression, Random Forest, and K-Nearest Neighbours were popular and frequently used [13, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26]. These models have shown high potential for diabetes prediction and are supervised machine learning models.

3. Methodology

In this study, there were three phases involved; (i) Phase 1: Study Initiation and Data Preparation, describing the sections of literature review, data acquisition, exploratory data analysis (EDA), and data preparation, (ii) Phase 2: Model Selection, and (iii) Phase 3: Model Evaluation. However, this article focuses only on evaluating the performance of the selected machine learning model based on several evaluation methods.

In Phase 1, the initial step for data investigation which was the exploratory data analysis, was conducted to discover trends, and spot outliers and anomalies. Python was used for data cleaning to filter unrelated attributes and remove instances. SPSS was used for descriptive analysis and correlation studies of the attributes. Figure 1 shows a summarizing flowchart of the data preparation process.

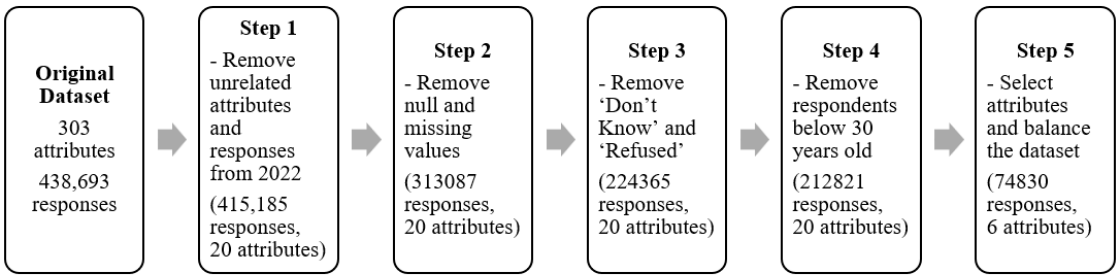

Figure 1. Summary Flow Chart of Data Preparation

The raw dataset used was cleaned to remove attributes not related to health status and diabetes using Python. The responses collected from the year 2022 were excluded. Twenty attributes, including diabetes, were selected based on insights from previous studies, and 283 attributes not related to health status or diabetes were removed. Additionally, 155 attributes were identified as personal information of the respondents rather than health-related data. Sixty-nine attributes detailed other health conditions and diseases including cancers, human immunodeficiency virus or asthma, while 59 attributes were found to be redundant and contained highly similar information to the selected attributes. The dataset contained six attributes of smoking including 'Have you smoked at least 100 cigarettes in your entire life?', 'Do you now smoke cigarettes every day, some days, or not at all?', 'How long has it been since you last smoked a cigarette, even one or two puffs?', 'During the past 12 months, have you stopped smoking for one day or longer because you were trying to quit smoking?', 'Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker' and 'Adults who are current smokers'. All of these attributes were capturing information from repetitive aspect and only one of the attributes was selected to reduce the redundancy. After filtering and selecting

the 20 attributes, data cleaning was applied to check the entries available. Respondent entries with missing values were directly removed. Responses with 'Don't Know' and 'Refused' in any of the attributes were excluded. Thirteen multi-class attributes, including 'Diabetes,' were transformed into binary attributes of class '0' and class '1' through data cleaning. According to the literature review, any responses from respondents younger than 30 years old were removed due to the high possibility of having type I diabetes [30].

The distribution of a categorical variable between two groups or samples may be compared using chi-square test [31]. The data used was categorical. Hence, the chi-square test function was applied to evaluate the relationship between each attribute and the dependent attribute of 'Diabetes.' Cramer's V correlation of each attribute to the dependent attribute of 'Diabetes' was determined using SPSS. Cramer's V correlation is used to assess the magnitude of association between two categorical variables in contingency tables larger than 2 x 2. It is calculated by dividing the chi-square statistic and ranges from zero to one with no negative values. A Cramer's V value greater than 0.15 was considered a strong association, and a value greater than 0.25 indicated a very strong relationship [32]. Five attributes which were identified as statistically significant, were selected as the independent attributes in the analysis. These five attributes had the highest Cramer's V correlation, ranging from 0.20 to 0.26, with the dependent attribute of diabetes.

The dataset show in Table 2 show was imbalanced, with only a minority of the respondent entries showing type II diabetes after data pre-processing. In 212821 responses, 175406 responses were found to be no diabetes class ('0') and only 37415 responses were yes diabetes class ('1'). Random under sampling was applied to prepare a balanced dataset for model construction. The class distribution was balanced by randomly removing responses from the majority class of no diabetes class ('0'). The data contained enough entries to proceed with under sampling the majority group to avoid model bias. A total of 74,830 responses, with 37,415 in the no diabetes class ('0') and 37,415 in the yes diabetes class ('1'), were finalized for machine learning implementation.

Table 2. Class Distribution Before and After Random Under Sampling

Description	No Diabetes Class ('0')	Yes Diabetes Class ('1')	Total
Original dataset entries	175,406	37,415	212,821
After random under sampling (balanced dataset)	37,415	37,415	74,830

In Phase 2, four machine learning models, including (i) Decision Tree, (ii) Neural Network, (iii) Random Forest, (iv) Logistic Regression and (v) AdaBoost, were analyzed in the study. These models are popular and perform well in handling

healthcare data. A stratified cross-validation of five was performed using the default random seed of one. Stratification ensured that each fold maintained the same class distribution, providing reliable results and reducing the risk of overfitting.

A Neural Network is a supervised machine learning algorithm inspired by the structure and function of human neurons. In the learning process, Neural Networks primarily employ a method of modifying weights and adjusting the connection strength between the nodes based on the error between the predictions and the actual outcomes [33, 34]. The iterative modification of weights and connection strength of the algorithm may identify the nonlinear relationships between diabetes and the independent attributes. It is suitable in addressing classification problem within large and complex dataset [35]. It has diverse applications across the fields of science and technology and is widely used in the diagnosis of diseases including diabetes and tuberculosis, as well as in image classification on radiographs [36].

A Decision Tree is a supervised machine learning algorithm commonly applied in classification and prediction in medical research. The complex relationships between the independent attributes and the dependent attribute are simplified through the segmentation of subgroups [37]. It involves the prediction of the dependent attribute based on decision rules derived from the features in the dataset, organized in a flowchart manner. It may assess the importance level of different dependent attributes and identify the combination of attributes with greatest effect in predicting type II diabetes[37]. It is commonly used in disease prediction and decision-making when handling categorized data [38].

Random Forest is an ensemble machine learning algorithm that involves the construction of multiple decision trees [39]. Each decision tree is built using bagging, which involves a random set of data to reduce overfitting. The average prediction from the multiple trees is taken to enhance the accuracy of the predicted outcome [39]. This helps in the classification of type II diabetes with large dataset and multiple independent attributes as all the independent attributes would not be used at once [39]. It can be used in the prediction of both categorical and quantitative attributes in high-dimensional and complex settings [40].

Logistic Regression is a supervised machine learning algorithm commonly used in the classification of binary attributes. It is a multivariable statistical method that models the relationship between multiple independent attributes and a binary or categorical dependent attribute [41]. The probability of the specific outcome is estimated using the sigmoid function and produces output value between 0 to 1. This allows the algorithm to interpret the results as the likelihood of a binary outcome [42]. The algorithm is commonly employed in clinical studies to assess the association of multiple independent attributes to a single binary outcome, such as the presence and absence of diseases [43]. Logistic regression is an important tool in analyzing complex health survey data to obtain health insights and to identify disease behaviors and patterns [44].

AdaBoostM1 is an ensemble machine learning algorithm commonly used to boost the performance of weak algorithms [45]. It is designed to address binary classification tasks through the application of data entry reweighting [46]. The weight adjustment policy of AdaBoostM1 involves assigning higher weights to

misclassified entries and prioritizing the entries that are challenging to classify. The attempt in reducing error rate is important in the prediction of type II diabetes prediction to improve the accuracy in detecting the responses with yes diabetes [47]. It is widely used in diverse domains including facial detection, image classification, and human detection [48].

In Phase 3, the evaluation methods of accuracy, precision, sensitivity, and F1-score were adopted to evaluate the output of the constructed machine learning models. A confusion matrix was generated to visualize the performances of the classification models.

Accuracy shows the ability of the model in getting true positive and true negative in predicting type II diabetes. It involves the differentiation of yes diabetes patient and no diabetes healthy individuals in the correct manner. Specificity shows the ability of the model in getting true negative in predicting type II diabetes which is detecting nondiabetic individual. Sensitivity shows the ability of the model in getting true positive in predicting type II diabetes which is detecting type II diabetic individual. F1 score shows the ability of the model in discriminating the classes of the dependant attribute due to using of precision and sensitivity. It is holding the mean score in combining precision and sensitivity. The model with high sensitivity score is suitable to be used in screening for type II diabetes.

3.1 Dataset

The dataset of the Behavioral Risk Factor Surveillance System (BRFSS) from 2021 was used in this study. The dataset is available to the public on the CDC website (https://www.cdc.gov/brfss/annual_data/annual_2021.html) and contains 438,693 responses with 303 attributes. The dependent attribute of this study was 'Diabetes,' corresponding to the survey question '(Ever told) (you had) diabetes? It was a multi-class attribute with '1' indicating the respondent had diabetes, '2' indicating the respondent had diabetes but only during pregnancy (for females), '3' indicating the respondent did not have diabetes, '4' indicating the respondent did not have diabetes but had prediabetes, '7' indicating the respondent did not know or was not sure, and '9' indicating the respondent refused to answer.

Annually, the BRFSS dataset is readily available and easily accessed through the CDC website. Each BRFSS dataset contains more than 400,000 responses, allowing for robust analysis and providing a greater chance of obtaining substantial instances of 'yes' diabetes as the class of interest. In addition, the dataset contains diverse health-related attributes that facilitate exploratory data analysis in correlation and association studies.

3.2 Research Procedure

In evaluating the performance of the type II diabetes prediction machine learning models, algorithms such as Decision Tree, Neural Network, Random Forest, Logistic Regression, and AdaBoostM1 were applied using Weka. The cleaned dataset, after exploratory data analysis, consisted of 212,821 entries and 6 attributes, including the dependent variable of diabetes. The dataset was imbalanced, with 175,406 entries (82.42%) indicating no diabetes and 37,415 entries (17.58%) indicating yes diabetes. To balance the class distribution of the dataset, random under sampling was applied. All entries from the class of yes diabetes were retained, and 37,415 entries were randomly selected from the class of no diabetes. This

reduced the dataset to a total of 74,830 entries, with a balanced distribution of 37,415 entries indicating yes diabetes and 37,415 entries indicating no diabetes. The balanced dataset was then used for the application of machine learning models on Weka.

4. Findings and Discussion

The findings will be presented in the discussion in this section.

4.1 Exploratory Data Analysis

The exploratory data analysis was carried out using Python and SPSS. The findings showed that the attributes of high blood pressure, high cholesterol, BMI, general health, and walking difficulty had a strong correlation with the dependent attribute of diabetes, with Cramer's V values greater than 0.2. A Cramer's V value greater than 0.15 was considered a strong association, and a value greater than 0.25 indicated a very strong relationship [32]. The five attributes selected had the highest Cramer's V correlation, ranging from 0.20 to 0.26, with the dependent attribute of diabetes. Between 33.16% and 71.84% of diabetes entries had high blood pressure, high cholesterol, and walking difficulty, as shown in Figure 2.

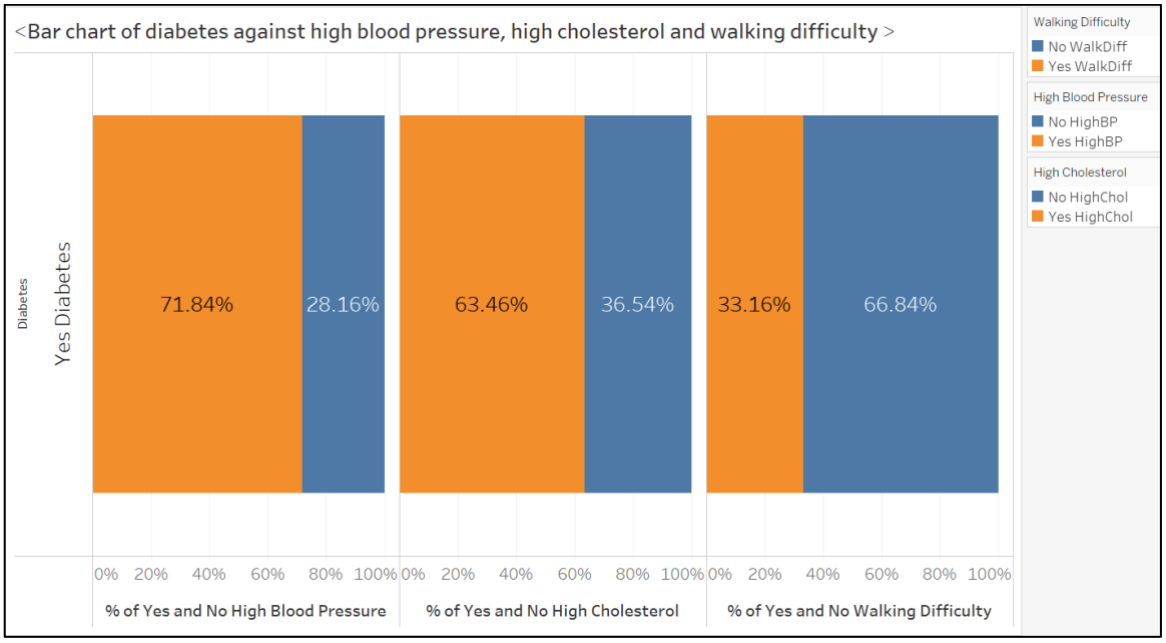

Figure 2. Bar Chart of Diabetes Against High Blood Pressure, High Cholesterol and Walking Difficulty in Percentage

The percentage of yes diabetes increases from 6.19% at BMI of 21 to 41.36% at BMI of 46 in Figure 3.

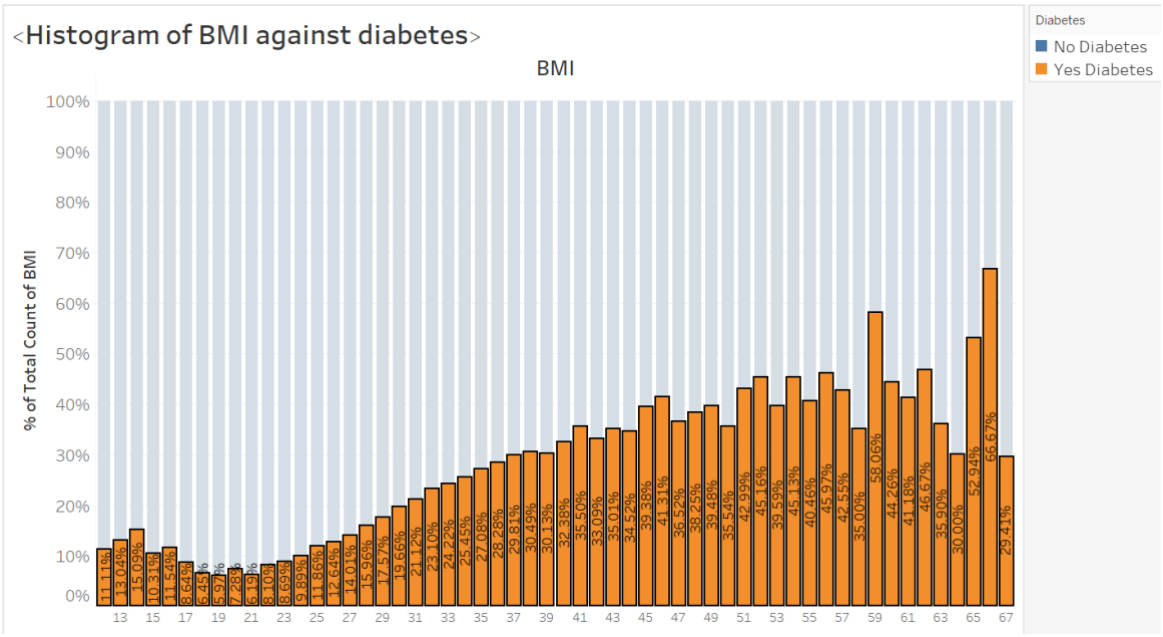

Figure 3. Histogram of BMI Against Diabetes in Percentage

The percentage of diabetes increased as the class of general health increased, as depicted in Figure 4. Class four and class five indicated fair and poor general health, respectively. The higher the class of general health was, the poorer the general health would be.

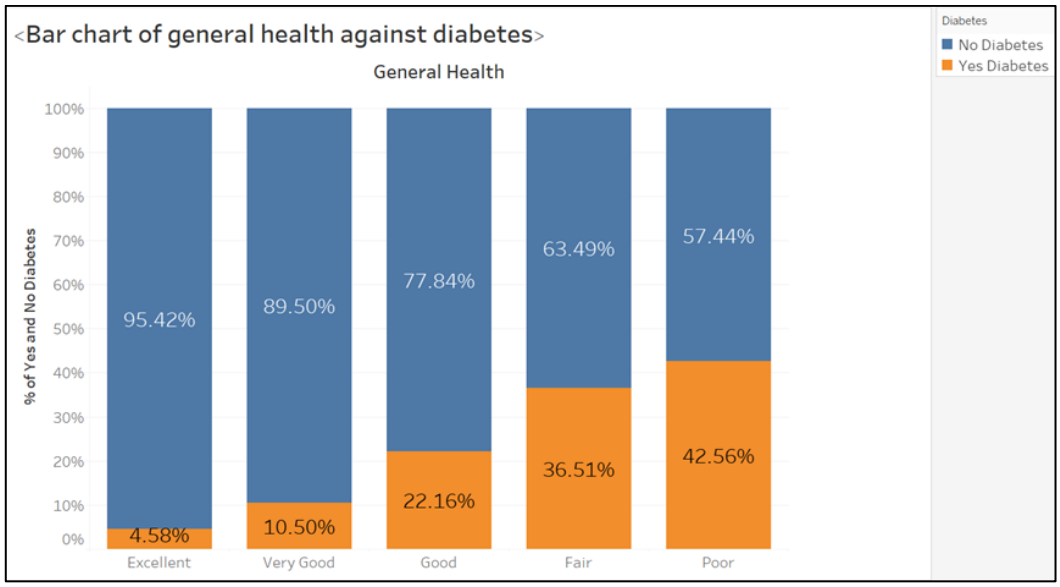

Figure 4. Bar Chart of General Health Against Diabetes in Percentage

The selection of attributes was intended to achieve screening for type II diabetes with lower data input. This approach could enhance feasibility in screening larger populations by reducing data requirements and time constraints.

4.2 Evaluation of Machine Learning Models

Machine learning models including Decision Tree, Neural Network, Random Forest, Logistic Regression, and AdaBoostM1 were applied to predict type II diabetes using WEKA. The classes of diabetes were categorized as '0' for no type II diabetes and '1' for yes type II diabetes. The performance comparison among the four machine learning models was evaluated based on accuracy, precision, sensitivity, and F1-score.

Accuracy indicates the model's ability to correctly predict true positives and true negatives for type II diabetes. It distinguishes between individuals with and without diabetes correctly. Specificity measures the model's ability to correctly predict true negatives for type II diabetes, identifying non-diabetic individuals. Sensitivity gauges the model's ability to correctly predict true positives for type II diabetes, identifying type II diabetic individuals. F1-score assesses the model's capability to discriminate between classes of the dependent attribute by leveraging precision and sensitivity, providing a balanced mean score combining these metrics.

Models with higher sensitivity scores were more suitable for screening type II diabetes. All algorithms demonstrated acceptable performance ranging from 68.8% to 74.7%. Neural Network achieved the highest accuracy (71.0%), and F1-score (71.9%). Decision Tree exhibited the highest sensitivity (74.7%) among all algorithms. Logistic Regression achieved the highest accuracy (71.0%), and Pression (70.5%). Table 2 presents the summary of the performance of the type II diabetes prediction models.

Table 2. Performance of Type II Diabetes Prediction Model

Model	Accuracy	Precision	Sensitivity	F1-score
Neural Network	71.0%	69.7%	74.3%	71.9%
Random Forest	70.6%	69.4%	73.4%	71.4%
AdaBoost	70.0%	69.3%	71.8%	70.6%
Decision Tree	70.4%	68.8%	74.7%	71.7%
Logistic Regression	71.0%	70.5%	72.2%	71.3%

Overall, Neural Network, Decision Tree, and Logistic Regression performed well in predicting type II diabetes based on risk factors such as high cholesterol, high blood pressure, BMI, general health, and walking difficulty. Among the models used for prediction, Neural Network showed the highest accuracy, and F1-score. Decision Tree exhibited the highest sensitivity and comparable performances in accuracy and F1-score. Screening for diseases, including diabetes, was to predict disease occurrence, making sensitivity crucial for minimizing false negative results.

Several studies have recommended the use of Decision Tree models for predicting diabetes, utilizing the BRFSS dataset [16, 20, 23]. These studies utilized BRFSS data from 2014, 2015, and 2020. One study specifically recommended Neural Network algorithms for diabetes prediction using the 2014 BRFSS dataset [16]. This study highlighted Neural Network's superior overall performance in diabetes prediction and suggested Decision Tree as the most suitable algorithm for diabetes screening due to its high sensitivity in minimizing false negatives. Recent research using the BRFSS 2021 dataset has corroborated these findings regarding the evaluation of machine learning models.

This revision clarifies the relationships between different models, emphasizes the importance of sensitivity in disease screening, and improves the overall readability of the paragraph.

5. CONCLUSION

The risk factors of high blood pressure, high cholesterol, BMI, general health, and walking difficulty have demonstrated a high correlation with type II diabetes based on Cramer's V correlation. Both Neural Network and Decision Tree have given a good performance in predicting type II diabetes using the selected attributes. Decision Tree, with its highest sensitivity, has been identified as the most suitable algorithm for type II diabetes prediction, maximizing detection accuracy.

Selecting and utilizing specific risk factors for screening type II diabetes can significantly enhance feasibility by reducing data requirements and time constraints. This approach facilitates early screening and supports the implementation of preventive measures. Preliminary diagnosis based on the interpretation of risk factors may greatly reduce the workload of clinical professionals in identifying the high risk group for type II diabetes to proceed further clinical diagnostic.

However, due to the cross-sectional nature of the BRFSS data used in this study, it is limited in establishing cause-and-effect relationships. The BRFSS dataset relies on survey-based information gathering, which introduces recall bias and confirmation bias. It does not involve any diagnoses or advices from clinical professionals and this raises the concerns on the accuracy of the data collected and the findings derived from this dataset shall be interpreted with caution. Future work may focus on the application of suggested machine learning models on different dataset to enhance the applicability of the findings and asses the performance of the type II diabetes prediction model across diverse populations. An alternative sampling method, for example SMOTE may be considered in generating a balanced dataset. SMOTE technique is commonly used in addressing dataset imbalances by oversampling the minority class on dependent attribute for potential improvements.

Acknowledgments

The authors would like to acknowledge the Universiti Teknologi Malaysia for funding this research. We would also like thank Madam Norazmah Suhailah Binti Abdul Malek for her valuable comments in improving this manuscript.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

[1] ElSayed, N. A., Aleppo, G., Bannuru, R. R., Bruemmer, D., Collins, B. S., Ekhlaspour, L., Gaglia, J. L., Hilliard, M. E., Johnson, E. L., Khunti, K., Lingvay, I., Matfin, G., McCoy, R. G., Perry, M. Lou, Pilla, S. J., Polsky, S., Prahalad, P., Pratley, R. E., Segal, A. R., … Gabbay, R. A. (2024). 2. Diagnosis and Classification of Diabetes: *Standards of Care in Diabetes—2024*. *Diabetes Care*, *47*(Supplement_1), S20–S42. https://doi.org/10.2337/dc24-S002

[2] Ismail, L., Materwala, H., & Al Kaabi, J. (2021). Association of risk factors with type 2 diabetes: A systematic review. *Computational and Structural Biotechnology Journal*, *19*, 1759–1785. https://doi.org/10.1016/j.csbj.2021.03.003

[3] Alam, S., Hasan, Md. K., Neaz, S., Hussain, N., Hossain, Md. F., & Rahman, T. (2021). Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management. *Diabetology*, *2*(2), 36–50. https://doi.org/10.3390/diabetology2020004

[4] Hossain, Md. J., Al-Mamun, Md., & Islam, Md. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Science Reports*, *7*(3). https://doi.org/10.1002/hsr2.2004

[5] Van Ommen, B., Wopereis, S., Van Empelen, P., Van Keulen, H. M., Otten, W., Kasteleyn, M., Molema, J. J. W., De Hoogh, I. M., Chavannes, N. H., Numans, M. E., Evers, A. W. M., & Pijl, H. (2018). From Diabetes Care to Diabetes Cure—The Integration of Systems Biology, eHealth, and Behavioral Change. *Frontiers in Endocrinology*, *8*. https://doi.org/10.3389/fendo.2017.00381

[6] Akhtar, S., Nasir, J. A., Ali, A., Asghar, M., Majeed, R., & Sarwar, A. (2022). Prevalence of type-2 diabetes and prediabetes in Malaysia: A systematic review and meta-analysis. *PLOS ONE*, *17*(1), e0263139. https://doi.org/10.1371/journal.pone.0263139

[7] Al-Behadili, H. N. K., & Ku-Mahamud, K. R. (2021). Fuzzy unordered rule using greedy hill climbing feature selection method: An application to diabetes classification. *Journal of Information and Communication Technology*, *20*(3), 391-422.

[8] Bansal, N. (2015). Prediabetes diagnosis and treatment: A review. *World Journal of Diabetes*, *6*(2), 296. https://doi.org/10.4239/wjd.v6.i2.296

[9] Dwibedi, C., Mellergård, E., Gyllensten, A. C., Nilsson, K., Axelsson, A. S., Bäckman, M., Sahlgren, M., Friend, S. H., Persson, S., Franzén, S., Abrahamsson, B., Carlsson, K. S., & Rosengren, A. H. (2022). Effect of self-managed lifestyle treatment on glycemic control in patients with type 2 diabetes. *Npj Digital Medicine*, *5*(1), 60. https://doi.org/10.1038/s41746-022-00606-9

[10] Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A., & Sakharkar, M. K. (2016). Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. *Biomarker Insights*, *11*, BMI.S38440. https://doi.org/10.4137/BMI.S38440

[11] Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A., & Sakharkar, M. K. (2016). Significance of HbA1c test in diagnosis and prognosis of diabetic patients. *Biomarker Insights, 11*, BMI.S38440.

[12] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8

[13] Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, *2*, 100118.

[14] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), 1-10.

[15] Li, S., Wang, J., Zhang, B., Li, X., & Liu, Y. (2019). Diabetes Mellitus and Cause-Specific Mortality: A Population-Based Study. *Diabetes & Metabolism Journal*, *43*(3), 319. https://doi.org/10.4093/dmj.2018.0060

[16] 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*(9). https://doi.org/10.5888/pcd16.190109

[17] Islam, M. S. (2021). *Machine Learning Approaches for Diabetes Mellitus Prediction and Management* (Doctoral dissertation, Hamad Bin Khalifa University (Qatar)).

[18] Chou, C.-Y., Hsu, D.-Y., & Chou, C.-H. (2023). Predicting the Onset of Diabetes with Machine Learning Methods. *Journal of Personalized Medicine*, *13*(3), 406. https://doi.org/10.3390/jpm13030406

[19] Dritsas, E., & Trigka, M. (2022). Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Sensors*, *22*(14). https://doi.org/10.3390/s22145304

[20] Hama Saeed, M. A. (2023). Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology*, *10*(1). https://doi.org/10.1186/s43067-023-00074-5

[21] Kaur, G., & Chhabra, A. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications*, *98*(22), 975–8887.

[22] Mamandra, E. (2022). *Diabetes Diagnosis Using Machine Learning*. University of Piraeus. Monaghan, M., Helgeson, V., & Wiebe, D. (2015). Type 1 Diabetes in Young Adulthood. *Current Diabetes Reviews*, *11*(4), 239–250. https://doi.org/10.2174/1573399811666150421114957

[23] Mpanga Justin Ngoyi. (2022). *Machine and Deep Learning Approach for Type 2 Diabetes Prediction Using the CDC's BRFSS Dataset: A Retrospective Analysis*. University of Missouri.

[24] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

[25] Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/2557795

[26] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00515

[27] Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

[28] Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), 106. https://doi.org/10.1186/1471-2105-14-106

[29] Hu, F., & Li, H. (2013). A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*, *2013*, 1–10. https://doi.org/10.1155/2013/694809

[30] Monaghan, M., Helgeson, V., & Wiebe, D. (2015). Type 1 Diabetes in Young Adulthood. Current Diabetes Reviews, 11(4), 239–250. https://doi.org/10.2174/1573399811666150421114957

[31] Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics*, *42*(2), 152.

[32] Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

[33] Han, S.-H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dementia and Neurocognitive Disorders*, *17*(3), 83. https://doi.org/10.12779/dnd.2018.17.3.83

[34] Kumar, N., & Kumar, D. (2021). An improved grey wolf optimization-based learning of artificial neural network for medical data classification. *Journal of Information and Communication Technology*, *20*(2), 213-248.

[35] Zhou, Z., Qiu, C., & Zhang, Y. (2023). A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. *Scientific Reports*, *13*(1), 22420.
[36] Goel, A., Goel, A. K., & Kumar, A. (2023). The role of artificial neural network and machine learning in utilizing spatial information. *Spatial Information Research*, *31*(3), 275–285. https://doi.org/10.1007/s41324-022-00494-x
[37] Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
[38] Chern, C.-C., Chen, Y.-J., & Hsiao, B. (2019). Decision tree–based classifier in providing telehealth service. *BMC Medical Informatics and Decision Making*, *19*(1), 104. https://doi.org/10.1186/s12911-019-0825-9
[39] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688
[40] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Frontiers in Aging Neuroscience*, *9*. https://doi.org/10.3389/fnagi.2017.00329
[41] Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. *Anesthesia & Analgesia*, *132*(2), 365-366.
[42] Zaidi, A. (2022). Mathematical justification on the origin of the sigmoid in logistic regression. *Central European Management Journal*, *30*(4), 1327-1337.
[43] Anderson, R. P., Jin, R., & Grunkemeier, G. L. (2003). Understanding logistic regression analysis in clinical reports: an introduction. *The Annals of thoracic surgery*, *75*(3), 753-757.
[44] Dey, D., Haque, M. S., Islam, M. M., Aishi, U. I., Shammy, S. S., Mayen, M. S. A., ... & Uddin, M. J. (2025). The proper application of logistic regression model in complex survey data: a systematic review. *BMC Medical Research Methodology*, *25*(1), 15.
[45] Taha, A. Y., Tiun, S., Abd Rahman, A. H., & Sabah, A. (2021). Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology*, *20*(3), 423-456.
[46] Wang, Y., Feng, L., Zhu, J., Li, Y., & Chen, F. (2022). Improved AdaBoost algorithm using misclassified samples oriented feature selection and weighted non-negative matrix factorization. *Neurocomputing*, *508*, 153–169. https://doi.org/10.1016/j.neucom.2022.08.015
[47] Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An efficient AdaBoost algorithm with the multiple thresholds classification. *Applied sciences*, *12*(12), 5872.
[48] Wu, S., & Nagahashi, H. (2015). Analysis of Generalization Ability for Different AdaBoost Variants Based on Classification and Regression Trees. *Journal of Electrical and Computer Engineering*, *2015*, 1–17. https://doi.org/10.1155/2015/835357