ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

Prediction and Analysis of Diabetes Using Machine Learning

Edelyn A. Bautista[1]

[1]North Eastern Mindanao State University Department of Computer Studies Tandag City, Philippines

Publication Date: 2025/10/14

Abstract: This study focuses on diabetes prediction and analysis using machine learning techniques. Its goal is to develop accurate and reliable models for early detection and better understanding of diabetes. The Diabetes UCI Dataset, containing variables like gender, polyuria, and polydipsia, is used for model training and evaluation. Data preprocessing ensures feature normalization and consistency, while feature selection identifies the most relevant variables. Several classification algorithms, including the Random Tree algorithm, are tested using WEKA. Model performance is evaluated through metrics such as accuracy, precision, and recall. Results show that Random Tree, when combined with other algorithms, achieves high accuracy and robustness in classifying diabetic and non-diabetic individuals. The study highlights the effectiveness of machine learning in early diabetes detection and decision-making support for healthcare professionals. Overall, it demonstrates how computational approaches can enhance diabetes management, improve patient outcomes, and reduce the impact of this chronic disease.

Keywords: *Diabetes Prediction, Dataset, Classification Algorithm, Analysis, Prediction, Healthcare.*

How to Cite: Edelyn A. Bautista (2025) Prediction and Analysis of Diabetes Using Machine Learning. *International Journal of Innovative Science and Research Technology*, 10(10), 650-656. https://doi.org/10.38124/ijisrt/25oct294

I. INTRODUCTION

Diabetes is a long-term metabolic condition marked by high blood glucose also known as blood sugar, which over time can seriously harm the heart, blood vessels, eyes, kidneys, and nerves. More than one in three individuals in the US, or 96 million people have diabetes, and more than 80% are unaware that they possess it [3]. With the rapid technological advancement, machine learning plays a big role in predicting whether a person has diabetes or not. In the medical field, machine learning has been used to create better diagnostic tools, increase the accuracy of diagnoses, and enable earlier disease identification, all of which can improve patient outcomes. In addition, the early detection of diabetes is important because it helps to avoid further complications of the disease. With the use of a dataset, it is feasible to create a machine learning model by using a classification technique that can accurately predict if a patient has diabetes or not, making it simple to utilize data analysis and visualization to derive some conclusions about the information. Manual diagnosis takes a lot of time and usually costs more money, but machine learning makes it more advantageous and accurate to diagnose diabetes early.

Machine learning reduces the cost of care and increases the speed and accuracy of physicians' work. It offers a range of treatment options and specialized care and improves the overall therapeutic efficacy of the hospital and healthcare systems [9]. In addition to enhancing and cutting the cost of medical care, it has the potential to fundamentally alter how systems are designed in ways that will enhance patient flow by reducing queues [18].

Though there were other studies done on diabetes classification and prediction, the innovation continues to fully ensure the precision of the result. The goal of this study is to produce a more accurate result by using the classification technique as it is one of the most commonly used machine learning techniques that examine the training data and creates an inferred result. The classification algorithms have been applied to the Diabetes UCI Dataset and it has been collected from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and approved by a doctor. The dataset contains the attributes to determine whether the patient is diabetic or non-diabetic.

II. LITERATURE REVIEW

Machine learning is a technique for data analysis that uses data to automatically develop analytical models, and utilizes automated optimization techniques to continuously enhance the accuracy of the outputs [11]. It has been a fast-growing trend in the healthcare sector that uses data to analyze a patient's health in real-time. Medical professionals used the technique to analyze data and identify patterns, to better diagnose and treat patients [16]. It was cited by [2] that machine learning algorithms are used in diabetes prediction to predict and detect the disease to prevent further medical complications, as well as to ascertain whether the person is affected and the likelihood that related diseases will occur.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

[1] Implied that, in the healthcare sector, machine learning is essential for discovering hidden information and patterns that may be used to learn from the data and predict results appropriately. The diabetes prediction model contains a few external factors that cause diabetes in addition to standard factors like glucose, BMI, age, insulin, etc. The dataset has been subjected to a variety of machine-learning techniques, with Logistic Regression providing the greatest accuracy of 96%. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%, which can help non-diabetic people to be more aware of what are the symptoms of a diabetic individual.

In connection [8] claimed that, early identification of diabetes is essential because it is a condition with no known cure, which led to the creation of a system that accurately predicts diabetes. The study uses data mining, machine learning, and neural network techniques to predict diabetes. To predict diabetes, nine unique attributes and seven machine learning methods were used. It has been discovered that support vector machine and logistic regression provided approximately 77%–78% accuracy for both train/test split and K-fold cross-validation method that are effective methods for predicting diabetes. In addition, to prevent the sickness through early detection [14] developed diabetes classification using machine learning techniques. The algorithms that were examined include support vector machines, decision tree classifiers, logistic regression, and random forests. The performance assessment of the classifiers was represented using a confusion matrix. The experimental findings indicate that all four machine learning methods work effectively. However, random forest performs better than the other three and has a higher prediction level of 100% compared to other methods and previous studies.

Various machine-learning approaches can be used to analyze the data from different angles and synthesize it into meaningful information. To classify the diabetes dataset effectively and to identify useful patterns, data mining methods and techniques will be considered. By using decision tree, naive Bayes, and logistic regression, the dataset was examined and processed to create a powerful model that predicts and diagnoses diabetes disease, and it shows that logistic regression has the highest accuracy with accuracy value of 90.36% [15]. [4] Affirmed that many disease assessments and predictions are performed using a variety of machine learning algorithms in several analyses, which help to solve larger problems. The viewpoint of classification and prediction is the identification and prediction of diseases and this research analyzes diabetes based on its best attributes. Random forest classifier was found to be a most effective precision with the accuracy of 75.7813 % than support vector machine, which was used to estimate the prevalence of diabetes, and it helps medical professionals decide what kind of care to provide.

III. METHODOLOGY

In view of the problem statement described in the introduction section, this study proposed a classification model with boosted accuracy to predict diabetic patients. In this model, different classifiers are employed like Random Trees, Simple Logistic, Naïve Bayes, and J48. The major focus is to increase the accuracy by resampling the Diabetes UCI Dataset, and changing the parameters of the given technique to avoid over fitting, and preprocess it which includes altering the number of features. The original data set consists of seventeen attributes and is modify to seven through their weight distribution, and from 520 number of sample to 250 by using the stratified sampling.

The dataset is being tested by using WEKA and with the different types of test options to classify data, 10 fold cross-validation has been chosen.

➢ *Random Trees*

Random trees are an ensemble learning technique that can be used for classification, regression, and other tasks. It works by building a large number of decision trees during the training phase, then producing the class that represents the mean of the classes (for classification) or mean prediction (for regression) of the individual trees. Random decision forests are used to correct the decision tree's tendency to overfit the training set. It is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. The three primary hyperparameters of random forest algorithms include the number of trees, node size, and the number of feature samples.

The following stages are necessary for the random forest method to function:

- Step 1: The method will draw samples from the supplied dataset.
- Step 2: For each sample chosen, the algorithm will build a decision tree. After that, each decision tree will yield a forecast result.
- Step 3: Voting will then be conducted for each outcome that was anticipated. It will use mean for a regression problem and mode for a classification task.
- Step 4: The algorithm will then choose the prediction result that received the most votes as the final prediction.
- Step 5: Assess the performance of the Random Trees algorithm using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score).

➢ *Simple Logistic*

Logistic regression is an illustration of supervised learning and it is used to compute or forecast the likelihood of a binary event occurring. When the output or dependent variable is dichotomous or categorical, a categorical variable can be true or false, yes or no, 1 or 0, et cetera, which logistic regression is used to solve the classification problems.

In practice, the logistic regression technique examines variable relationships. It uses the Sigmoid function to assign probabilities to discrete outcomes. For binary predictions, the population will be divided into two groups with a cut-off of 0.5. Everything above 0.5 is classified as belonging to group A, while everything below is classified as belonging to group

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

B. After data points have been allocated to a class, a hyper plane is utilized as a decision line to separate two categories (as far as possible). Using the decision boundary, the class of future data points can then be predicted.

➢ *Naïve Bayes*

Nave Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem that is used in a wide range of classification tasks. It is a classification algorithm for binary (two-class) and multi-class classification problems. Rather than attempting to calculate the values of each attribute value, they are assumed to be conditionally independent given the target value. The representation for naive Bayes is probabilities.

The following process are necessary for the naïve Bayes method to function:

- Step 1: Calculate prior probability for given class labels
- Step 2: Calculate conditional probability with each attribute for each class.
- Step 3: Multiply same class conditional probability.
- Step 4: Multiply prior probability with step 3 probability
- Step 5: See which class has the higher probability, higher probability class belongs to the given input set step.

➢ *J48*

The J48 algorithm is used to classify various applications and produce accurate classification results. It is one of the best machine learning algorithms for categorizing and continuously examining data. It employs a straightforward method to construct a decision tree from training data.

To implement the J48 algorithm on the dataset it should start at the top with the entire training dataset. Choose the attribute to divide on first, and then make a branch for each of its values. Then separate the training data into subsets. Repeat the approach for each branch, selecting an attribute at each node based solely on the cases that reach it. The top-down, recursive, divide-and-conquer technique is used by J48 (aka C4.5), which selects the attribute at each level using a metric called information gain.

Table 1 Dataset Attributes

S No	Attribute	Type
1	Gender	Nominal
2	Polyuria	Nominal
3	Polydipsia	Nominal
4	Sudden weight loss	Nominal
5	Irritability	Nominal
6	Partial paresis	Nominal
7	Class	Categorical

Table 1 above contains the preprocess UCI Diabetes dataset attributes that are used to predict the accuracy of whether an individual is positive or negative from diabetes. The attributes used are polyuria, sudden weight loss, weakness, polyphagia, visual blurring, irritability, muscle stiffness, which are of nominal type and class as categorical.

Table 2 Table of Accuracy Measures

Accuracy Measures	Denotations	Formula
TP Rate	Proportion of correct predictions in predictions of positive class.	TPR = TP/ (TP+ FN)
FP Rate	The probability that an actual negative will test negative.	FPR= FP / (FP + TN)
Precision	Correctness of the classifiers	P = TP/ (TP + FP)
Recall	Measures the correctness or sensitivity of classifiers	R = TP / (TP + FN)
F-Measure	Weighted average of Precision and Recall	F=2*(P*R)/(P+R)
ROC Area	Receiver Operating Characteristic curves which will compare the tests	
PRC Area	The relationship between precision and recall.	

As shown in Table 2, TP Rate, FP Rate, Precision, Recall, F-Measure, Receiver Operating Curve (ROC), and PRC area measures are utilized for the grouping of the outcomes, where, True Positive is meant as TP, True Negative is, indicated as TN, False positive is meant as FP and False Negative is signified as FN.

IV. RESULT

Table 3 Stratified Cross-Validation Summary

Summary	Algorithm Random Tree
Correctly Classified Instances	88.2 %
Incorrectly Classified Instances	11.2 %
Kappa Statistic	0.7479
Mean Absolute Error	0.1214

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

Root Mean Squared Error	0.2657
Relative Absolute Error	28.2416 %
Root Relative Squared Error	57.3343 %
Total Number of Instances	250

Table 4 Stratified Cross-Validation Summary Continuation

Summary	Algorithm		
	Simple Logistic	**Naïve Bayes**	**J48**
Correctly Classified Instances	85.2 %	86.4%	86 %
Incorrectly Classified Instances	14.8 %	13.6 %	14 %
Kappa Statistic	0.6388	0.698	0.6506
Mean Absolute Error	0.1854	0.154	0.1811
Root Mean Squared Error	0.2965	0.3076	0.3259
Relative Absolute Error	43.1083 %	35.8186 %	42.126 %
Root Relative Squared Error	63.9785 %	66.3793 %	70.3273 %
Total Number of Instances	250	250	250

The table 3 shows the performance of different machine learning algorithms in classifying instance of the dataset. The algorithms compared include Random Tree, Simple Logistic, Naïve Bayes and J48. The table shows the various measures of classification accuracy, including the percentage of correctly classified instances, incorrectly classified instance, and the Kappa statistics. The results indicate that due to their differences in functionality, the classifiers offer different resolutions on the datasets. In comparison to the algorithms the Random Tree performs well among the others.

The Random Tree bagged the highest percentage of the correctly classified instances with 88.2%, which means that it gives an efficient classification of accuracy among other algorithms on the diabetes dataset.

Table 5 Algorithm's Performance Evaluation Metrics

Performance Algorithm	Algorithm								
	Random Tree			**Simple Logistic**			**Naïve Bayes**		
	A	B	C	A	B	C	A	B	C
TP Rate	0.885	0.890	0.888	0.679	0.930	0.852	0.872	0.860	0.864
FP Rate	0.110	0.115	0.114	0.070	0.321	0.242	0.140	0.128	0.132
Precision	0.784	0.944	0.894	0.815	0.865	0.849	0.739	0.937	0.875
Recall	0.885	0.890	0.888	0.679	0.930	0.852	0.872	0.860	0.864
F-Measure	0.831	0.916	0.890	0.741	0.896	0.848	0.800	0.897	0.867
MCC	0.751	0.751	0.751	0.644	0.644	0.644	0.703	0.703	0.703
ROC Area	0.957	0.957	0.957	0.939	0.939	0.939	0.952	0.952	0.952
PRC Area	0.899	0.977	0.953	0.872	0.973	0.942	0.901	0.979	0.955

Performance Algorithm	J48		
	A	**B**	**C**
TP Rate	0.654	0.953	0.860
FP Rate	0.047	0.346	0.253
Precision	0.864	0.859	0.86
Recall	0.654	0.953	0.86
F-Measure	0.745	0.904	0.854
MCC	0.663	0.663	0.663
ROC Area	0.875	0.875	0.875
PRC Area	0.79	0.918	0.878

And it has the lowest percentage of incorrectly classified instances, which is 11.2%, indicates that the model is performing well in terms of accuracy and precision. In terms of Kappa Statistics it shows that the Random Tree has the highest kappa statistic among the other algorithms used. Its value is 0.7479 and with Cohen's kappa it is interpreted as substantial since it is between 0.61–0.80. The mean absolute error ranges from 0.1214 – 0.1854 and the root mean squared error ranges from 0.2657 – 0.3259, with Random Tree having the lowest value. The relative absolute error ranges from 28.2416% - 43.1083% with random tree that has the lowest values and in terms of the root relative squared error, the algorithm has the smallest percentage with 57.3343%. The total number of instances of all algorithms is 250.

The table 4 provides the performance evaluation of the four different machine learning algorithms which are the Random Tree, Simple Logistic, Naïve Bayes and J48. It shows

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

the performance evaluation of each algorithm in terms of the several metrics which includes TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area, and PRC Area.

The Random Tree has the highest rating in terms of TP Rate with the value of 0. 888, it shows that the algorithm has correctly identified the positive instances of interest at a higher rate compared to other. The lowest FP Rate was attained by Random Tree with 0.114, which indicates that the model is effectively minimizing the number of false positive predictions or instances that are incorrectly classified as positive. The algorithm that has the highest precision is Random Tree with the value of 0.894, it means that it has a low rate of false positives. The Random Tree has the highest value in recall which is 0.888, shows that it has a low rate of false negatives. The highest rating for the F-Measure was achieved by Random Tree with a value of 0.890, indicates that the model achieves both high precision and high recall simultaneously, striking a balance between correctly identifying positive instances and minimizing false positives and false negatives. Random Tree has the highest value in MCC which is 0.751, which depicts that it has strong overall performance considering both true positives, true negatives, false positives, and false negatives.

The Random Tree has the highest ROC Area which is 0.957, it indicates that it can effectively control the false positive rate while maintaining a high true positive rate, suggesting its capability to make accurate predictions across different classification thresholds. The highest PRC Area was bagged by Naïve Bayes with a value of 0.953, which defines a strong performance of a binary classification model in terms of precision and recall trade-off.

In summary, the table shows that the Random Tree was the best performing algorithm among the other algorithms used to the test the dataset. The Random Tree has attained the highest value in terms of TP Rate, Recall, F-Measure, ROC Area, precision and MCC. While the Naïve Bayes achieved the highest PRC Area.

Table 6 Confusion Matrix

Random Trees		**Simple Logistic**		**Naïve Bayes**		**J48**	
a	**b**	**a**	**b**	**a**	**b**	**a**	**b**
63	15	53	25	68	10	51	27
18	154	12	160	24	148	8	164

Legend
a = Tested_Negative
b = Tested_Positive

The table 5 above shows the confusion matrix that compares the results of the four different algorithms.The confusion matrix serves as a performance evaluation tool that is commonly used to evaluate the accuracy of a classification model. The matrix provides a tabular representation of the number of correct and incorrect predictions made by the model, where the rows represent the actual class labels (positive or negative), and the columns represent the predicted class labels (positive or negative). To visualize the performance of the model through the confusion matrix, it is represented by a legend of a for tested_negative and b for tested_positive.

The Random Tree correctly predicted 63 cases of tested_negative and 154 cases of tested_positive. Simple Logistic accurately predicted 53 cases of tested_negative and 160 cases of tested_positive. Naïve Bayes precisely predicted 68 cases of tested_negative and 148 cases for tested_positive. While, J48 correctly predicted 51 cases of tested_negative and 164 cases of tested_positive.

- *Experimenting the Dataset with Random Tree*

This experiment includes changing the parameters of the given technique to avoid over fitting. The experiment was done by using the preprocessed diabetes dataset with 250 samples and 7 attributes. The changing of parameters was done in WEKA by changing the value of maxDepth from 0 to 3 and minNum from 1.0 into 2.0. Likely, the parameter of the Random Tree has been changed since it is the most performing algorithm among the other algorithms, and to assess its impact on the dataset performance.

Table 7 Experimenting the Dataset with Random Tree

Summary	**Result of the Experiment**
Correctly Classified Instances	86.8%
Incorrectly Classified Instances	13.2%
Kappa Statistic	0.6893
Mean Absolute Error	0.1915
Root Mean Absolute Error	0.3177
Relative Absolute Error	44.5398%
Root Relative Squared Error	68.5545%
Total Number of Instances	250

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

The table 6 above shows that after changing the parameters of the Random Tree, there was a change in the result. In terms of correctly classified instances it has 86.8% and based on the previous result of the Random Tree which is 88.2 %, it has a difference of 1.4%. With the incorrectly classified instances it has 13.2% which is much higher than the previous result that is not being experimented yet. In terms of Kappa Statistic it is 0.6893 which is good since based on Cohen's kappa it is interpreted as substantial since it is between 0.61–0.80. For the mean absolute error it has 0.1915 and its relative mean absolute error is 0.3177. While, its relative absolute is 44.5398% and the root relative squared error is 68.5545%, and the total number of instances is 250.

Table 8 Random Tree Performance after Experiment

Performance Algorithm	Random Tree		
	A	B	C
TP Rate	0.769	0.913	0.868
FP Rate	0.913	0.213	0.186
Precision	0.800	0.897	0.867
Recall	0.769	0.913	0.868
F-Measure	0.784	0.905	0.867
MCC	0.690	0.690	0.690
ROC Area	0.922	0.922	0.922
PRC Area	0.844	0.961	0.924

The table 7 above shows the Random Tree performance after the experiment. It shows that in terms of TP Rate it has a weighted average of 0.868, which shows that the algorithm has correctly identified the positive instances. For the FP Rate it has 0.186, and in terms of Precision and Recall it has obtained a good result, indicates that the model has performed well in terms of correctly identifying and classifying instances of a specific class. In ROC Area, it has 0.922 which means that the model has a strong discriminatory power and is effective in distinguishing between positive and negative instances.

V. CONCLUSION AND RECOMMENDATION

The effectiveness of the random tree classifier can be compared with the performance of other classifiers, leading to knowledge discovery about the advantages and disadvantages of each classifier and how it works for the dataset. The Random Tree Algorithm's performance as measured by appropriate metrics, including TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area, and PRC Area, demonstrates that it is capable of accurately predicting the presence or absence of diabetes based on the attributes of the dataset. The high-performance evaluation of the random tree model suggests that it is a good fit for the data and can be used to make accurate predictions. Also, the random tree classifier can offer a measure of feature relevance, making it possible to identify the features that are most essential in diagnosing diabetes. Using this information, more precise and efficient diabetes screening techniques can be created. This can be utilized to increase the precision of diabetes diagnosis and care, improving patient outcomes. In addition to recognizing the underlying biological pathways that contribute to the disease, this knowledge can be used to create models for diabetes diagnosis and treatment that are more precise and efficient. Therefore, the random tree algorithm is helpful in a diabetes dataset due to its capacity to handle complex, non-linear relationships between the independent and dependent variables, interpretability of decision rules, identification of significant features, generalizability to new data, and scalability to handle large datasets.

For future work, the same method could be considered and many other machine learning classifier algorithms could be considered to compare the most accurate one. This method can also be implemented on other disease-related and medical datasets. In this study, only a small sample dataset of 250 instances was taken into account. However, the same method could be applied to much larger datasets, which would greatly expand the scope of disease prediction. It may also provide much-needed early detection, diagnosis, and timely help to keep health issues under control and possibly find a way to completely eradicate them in the future.

REFERENCES

[1]. A. Mujumandar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Computer Science, vol. pp. 292-299, Feb.27,2020. https://www.sciencedirect.com/science/article/pii/S1877050920300557 (accessed Mar. 24, 2023).

[2]. B. Shamreen Ahamed, M. Arya, S. K. B. Sangeetha, N. Auxilia Osvin, "Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers", Applied Computational Intelligence and Soft Computing, vol. 2022, Article ID 7899364, 11 pages, 2022. https://www.hindawi.com/journals/acisc/2022/7899364/ (accessed Mar. 24, 2023).

[3]. CDC, "What is Diabetes," Center for Disease Control and Prevention, 2022. https://www.cdc.gov/diabetes/basics/diabetes.html (accessed Mar. 24, 2023).

[4]. H. Rashid Abdulqadir, A. Mohsin Abdulazeez, and D. Assad Zebari, "Data Mining Classification Techniques for Diabetes Prediction", QAJ, vol. 1, no. 2, pp. 125–133, May 2021.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25oct294

https://journal.qubahan.com/index.php/qaj/article/view/55 (accessed Mar. 24, 2023).

[5]. IBM SPSS, "What is random forest?," *IBM*. https://www.ibm.com/topics/random-forest (accessed Mar. 29, 2023).

[6]. IBM SPSS, "What is logistic regression?," *IBM*. https://www.ibm.com/topics/logistic-regression (accessed Mar. 29, 2023).

[7]. J. Brownlee, "Naive Bayes for Machine Learning," *Machine Learning Mastery*, Aug. 15, 2020. https://machinelearningmastery.com/naive-bayes-for-machine-learning/ (accessed Mar. 29, 2023).

[8]. J. Khanam, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no.4, pp.432-439, Feb. 20, 2021. https://www.sciencedirect.com/science/article/pii/S2405959521000205 (accessed Mar. 24, 2023).

[9]. M. Javaid, A. Haleem, R. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," International Journal of Intelligent Networks, vol. 3, pp.58–73,2022. https://www.sciencedirect.com/science/article/pii/S2666603022000069 (accessed Mar. 24, 2023).

[10]. M. Chandrasekaran, "Logistic Regression for Machine Learning," Capital One. https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/ (accessed Mar. 29, 2023).

[11]. MicroFocus, "What is Machine Learning?," Open Text Corporation, 2023. https://www.microfocus.com/en-us/what-is/machine-learning (accessed Mar. 24, 2023).

[12]. N. Chauhan, "Naïve Bayes Algorithm: Everything You Need to Know," *KD Nuggets*, Apr. 08, 2022. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html (accessed Mar. 29, 2023).

[13]. N. Khanna, "J48 Classification (C4.5 Algorithm) in a Nutshell," *Medium*, Aug. 18, 2021. https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e (accessed Mar. 30, 2023).

[14]. O. Adigun, F. Okikiola, N. Yekini, and R. Babatunde, "Classification of Diabetes Types using Machine Learning," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 13, no. 9, 2022. https://thesai.org/Downloads/Volume13No9/Paper_18- Classification_of_Diabetes_Types_using_Machine_Learning.pdf (accessed Mar. 24, 2023).

[15]. S. Saru and S. Subashree, "Analysis and Prediction of Diabetes Using Machine Learning," International Journal of Emerging Technology and Innovative Engineering, vol. 5, no.4, Apr. 23, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368308 (accessed Mar. 24, 2023).

[16]. SAS, "Machine Learning: What it is and why it matters," SAS Insights,2023. https://www.sas.com/en_us/insights/analytics/machine-learning.html (accessed Mar. 24, 2023).

[17]. WHO, "Diabetes," *Health Topics*, 2023. https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed Mar. 24, 2023).

[18]. X. Li and et. al, "Artificial intelligence-assisted reduction in patients' waiting time for outpatient process: a retrospective cohort study," BMC Health Services Research, Mar. 17, 2021. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-021-06248-z (accessed Mar. 24, 2023).