Machine Learning for Diabetes Detection in Females

AONDOFA ODOT ADDAI
Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria

Abstract*— Diabetes millitus commonly referred to as diabetes is a chronic medical deficiency characterized by abnormalities in the production and secretion of insulin in the human body system. Diabetes is a critical illness that can cause misfunctioning of several vital organs in the body system including the eyes, nerves, kidney, and the heart. The severity of symptoms can vary subject to the duration and type of diabetes. Individuals with high blood sugar levels particularly those with a complete lack of insulin such as children may experience symptoms such as increased appetite, polydipsia, weight loss, increased appetite, and vision problems. Diabetes affects approximately 9% of the entire adult population globally. Treatment of diabetes requires accurate and timely diagnosis. The use of Machine Learning techniques to support diabetes diagnosis has been widely adopted as an effective and efficient Artificial Intelligence approach in line with modern healthcare standards. Thus, this study presents the diabetes epidemic, prevalence, and diagnosis with emphasis on the use of Artificial Intelligence driven diagnosis. Different machine learning techniques are experimented on diabetes diagnosis for female patients particularly due to the greater risk of female patients to experience severe complications such as blindness from diabetic retinopathy and death from cardiovascular disease. A comparative analysis is performed for all experimented techniques to ascertain the optimal technique for diabetes diagnosis.

Keywords*—Artificial Intelligence, Diabetes, Machine Learning

I. INTRODUCTION

The term diabetes refers to a metabolic disorder of multiple aetiology characterized by chronic hyperglycaemia with inconsistencies of fat, carbohydrate and protein metabolism resulting from defects in the secretion of insulin [1]. Diabetes in long term causes dysfunction and failure of various organs including the heart, kidney, eyes, and liver. Common symptoms of diabetes include blurry vision, thirst, polyuria, and weight loss. Severe diabetes medical condition can lead to coma and death in the absence of treatment. Long term effects of diabetes lead to progressive development of associated complications retinopathy with potential blindness, nephropathy, and neuropathy with risk of foot ulcers, amputation, and features of autonomic dysfunction [2]. Patients with diabetes are at increased risk of cardio vascular, peripheral vascular and cerebrovascular disease. There are several pathogenetic processes involved in the development of diabetes. These include processes which destroy the beta cells of the pancreas with consequent insulin deficiency [3]. There are two major types of diabetes; type 1 and type 2 diabetes. In type 1 diabetes often referred to as insulin-dependent, the body does not produce insulin. Type 1 diabetes is usually developed before the age 40. Approximately 10 % of all diabetes cases are type 1. Type 2 diabetes, the body does not produce enough insulin for proper function or the cells in the body become insulin resistant. Approximately 90 % of all cases of diabetes worldwide are type 2 [2]. Diabetes is one of the fastest growing global epidemics of the 21st century. According to the International Federation Diabetes (IDF) atlas 2021, it is estimated that a total of 537 million people have diabetes presently and expected to reach 643 million cases by 2030, and 783 million people by 2045. It is also that 6.7 million people of the age range 20 – 79 years died from diabetes related causes in 2021 [4]. Many studies report women to have a higher risk of complications from diabetes than in men especially in complications such as coronary heart disease, and in gestational diabetes in pregnant women [5]. This study therefore focuses on diagnosis and treatment of diabetes for women. Diabetes can be effectively managed and treated in some cases. An important step to effective treatment is timely and accurate diagnosis. With the advent of modern technology and widespread adoption of Artificial Intelligence and Machine Learning techniques for clinical diagnosis as a standard for modern healthcare, machine learning has become an indispensable tool to facilitate the timely and accurate diagnosis of diabetes (Fadziso & Adusumalli, 2020; Goel et al., 2022; Pewekar et al., 2024). Some key indicators used in the diagnosis of diabetes for women include patient glucose level, no of pregnancies, blood pressure, skin thickness, insulin level, Body Mass Index (BMI), Diabetes Pedigree Function (DPF), and age.

Machine Learning is a branch of Artificial Intelligence that constructs computer programs capable of learning from data and making inference from learned knowledge. It is a multi-disciplinary field of computer science, statistics, cognitive science, and information theory. A machine learning algorithm stores learned information in some knowledge representation structure referred to as inductive hypothesis typically referred to as a model. A machine learning model uses data referred to as training data to obtain valid generalization about a concept and extends the generalization to new data instances referred to as test data [9]. Fig. 1 depicts the general concept of machine learning.

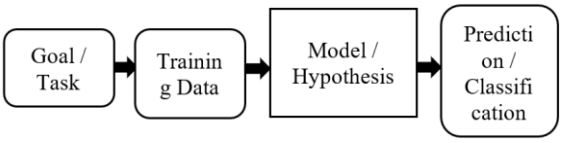

Fig. 1. Concept of Machine Learning

Over the years, machine learning has been applied to many real-world problems fraud detection, language recognition and processing, customer relationship management, weather forecasting, image recognition, and medical diagnosis [10]. Machine learning can be typically divided into two main purposes: classification and regression. Classification seeks to separate and label the result of a problem as one of two or more cases whereas prediction predicts the outcome of an uncertain relationship. Machine learning approaches are typically categorized into three main sub categories including supervised, unsupervised and semi-supervised learning [11].

II. MACHINE LEARNING ALGORITHMS

There are a wide range of machine learning algorithms. Some of the well-known machine learning algorithms are Support Vector Machines (SVMs), Random Forests, Decision Trees, Logistic Regression, Linear Regression, Naïve Bayes, and k Nearest Neighbours algorithms.

A. Support Vector Machines

The support vector machine algorithm was introduced in 1992 by Guyon and Vapnik during the Fifth Annual Association for Computing Machinery Workshop on Computational Learning Theory. SVM relies on the complexity for the hypothesis space and empirical error [12]. Support vector machines can be used for both binary classification and regression analysis. In binary classification, a linear hyper-plane or decision boundary is created to distinguish instances of one class from the other class. Separation between the classes is optimized by obtaining the separating hyperplane defined as the plane having the largest distance or margin to the nearest training datapoint of any class [13]. Support vector machines can be modelled as follows:

Given l training examples $\{X_i, Y_i\}$, $i = 1, \ldots, l$, where each example has d inputs ($X_i \in R^d$), and a class label with one of two values ($y_i \in \{-1,1\}$), Now, all hyperplanes in R^d are parameterized by a vector (w), and a constant (b), expressed in (1)

$$\boldsymbol{w} . \boldsymbol{x} + b = 0 \qquad (1)$$

Where w is the vector orthogonal to the hyperplane. Given such a hyperplane (w, b) that separates the data, this gives the function:

$$f(x) = sign\,(\boldsymbol{w} . \boldsymbol{x} + b) \qquad (2)$$

The canonical hyperplane is then defined to be that which separates the data from the hyperplane by a distance of at least 1. That is, we consider those that satisfy:

$$x_i . \boldsymbol{w} + b \geq +1 \; when \; y_i = +1 \qquad (3)$$

$$x_i . \boldsymbol{w} + b \leq -1 \; when \; y_i = -1 \qquad (4)$$

B. Logistic Regression

Logistic Regression sometimes referred to as the logit model analyses the relationship between multiple independent variables and a categorical dependent variable [14]. Logistic regression estimates the probability of occurrence of an event by fitting data to a logistic curve [15]. Logistic model is popular because the logistic function on which the logistic model is based provides estimates in the range 0 to 1 and an appealing S-shaped description of the combined effect on several risk factors on the risk for an event [16]. Since logistics regression calculates the probability of an event occurring over the probability of an event not occurring, the impact of explained variables is typically interpreted in terms of odds. In logistic regression, the mean of the dependent variable p in terms of an independent variable x is modelled correlating p and x in (5):

$$p = \alpha + \beta x \qquad (5)$$

To ensure (5) produces values between the range 0 and 1, Logistics Regression transforms the odds using the natural logarithm. The natural log odds as a linear function of the independent variable are modelled as follows:

$$Logit(y) = \ln(odds) = \ln(\frac{p}{1-p}) = \alpha + \beta x \quad (6)$$

where p is the probability of outcome of interest, and x is the independent variable. The parameters of the Logistic Regression are α and β.

Taking the antilog of equation 1.6 on both sides, an equation for the prediction of the probability of the occurrence of interested outcome as:

$$p = P\left(Y = \frac{interested\ outcome}{X}\right) = \quad (7)$$

$$= \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}} \quad (8)$$

C. k Nearest Neighbours

k Nearest Neighbours (kNN) is a machine learning algorithm used for data classification by the utilization of training records from the nearest neighbours. Where k specifies the number of closest neighbours used in the classification process [17]. Compared to other machine learning algorithms, k Nearest Neighbours is classified by simplicity. In actual principle, kNN calculates the distance between the training data and test data and uses the smallest possible distance value to classify the test data [18]. KNN classifies a class an object with the majority class of the neighbouring class. If K = 1, then the object is assigned the class of that 1 single neighbour. kNN can be used for both classification and regression and is a supervised learning machine learning approach. Generally, kNN uses the Euclidean distance to between two data points calculated using the Pythagorean formula. Typically, it is the length of a straight line that connects two points in a space. Given a data point that lies between the x and y cartesian planes; where x represents the training data and y represents the test data, the Euclidean distance between the two points can be calculated as follows using (9).

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (9)$$

where d specifies the calculated distance be the points x and y. $x_1, x_2, x_3, \ldots, x_n$ represent the x_n dimension and $y_1, y_2, y_3, \ldots, y_n$ represent the y_n dimension on the data plane.

D. Decision Trees

A decision tree is a machine learning classifier expressed as a recursive partition of the instance space. The concept of decision tree dates back to the mid-20th century. Introduce by Charles J. Clopper and Egon S. Pearson in 1934 who introduced the concept of binary decision processes [19]. The learning process of the decision tree begins with data split into homogenous subsets. The decision tree consists of nodes from a root tree and proceeds to all other nodes with exactly one directed edge [20]. All other nodes in the decision tree except the root node are referred to as leaves of the decision tree. In a decision tree, each internal node splits the instance space represented by the root node into two or more sub-spaces subject to a certain discrete function or criteria of the input attribute variables. Each leaf in a decision tree is assigned to one class representing the most appropriate target value [21]. The success of decision tree techniques mainly depends on several factors contributing to their performance, interpretability, and applicability to a wide range of problems. These factors include data quality, tree depth, splitting criteria, and tree pruning method. In the decision tree algorithm, the choice of splitting criteria is an import success factor. Different decision tree architecture uses different choices for this purpose. Some of the well-known choices for splitting in decision trees include Gini Index, Information gain, Information gain ratio, and Chi-square. There are different tree pruning methods employed by the decision tree such as pre-pruning, and post-pruning [22].

III. METHODOLOGY

The diabetes dataset comprises 768 samples obtained from patients. The key indicators (independent variables) present in the dataset are pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index, diabetes pedigree function, and age. The interest variable (outcome, dependent or response variable consist of two major classes 0 and 1. Where 0 represents a negative diagnosis and 1 represents a positive diagnosis. The dataset was obtained from the Kaggle machine learning and data science repository. The data preprocessing step involves checking for missing values and handling of data class imbalacing. The Scikit-learn library was used to implement the SVM, Logistic Regression, Decision Trees, Random Forest, k Nearest Neighbours, and Naïve Bayes algorithms. Python 3.13.5 version was used for implementation. Anaconda package and jupyter notebook were used for implementation. Fig. 3 depicts the class distribution of the data. Fig. 4 depicts the age distribution of the dataset, and Fig. 5 presents the scientific procedure of experiments in this study.

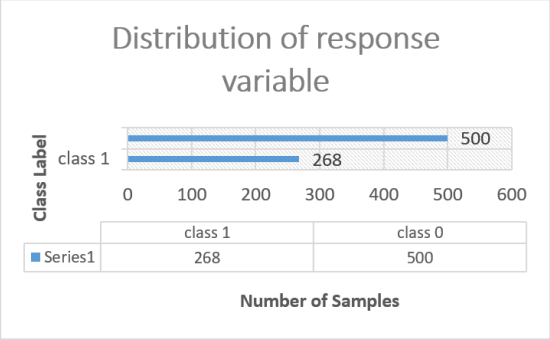

Fig. 3. Class distribution of the diabetes dataset

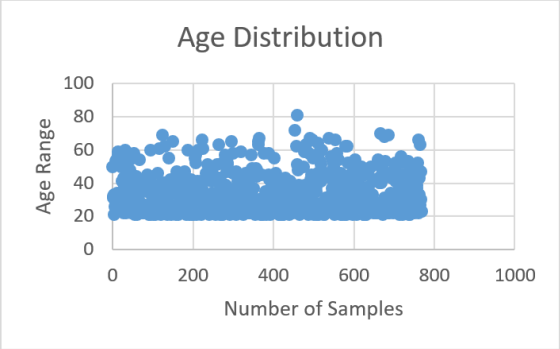

Fig. 4. Age distribution of the diabetes dataset.

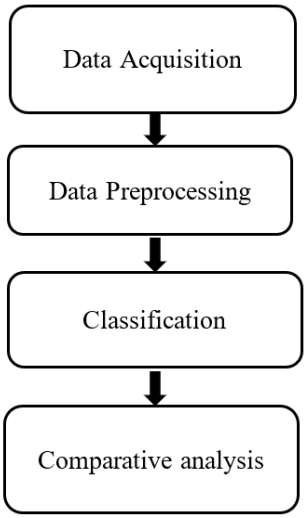

Fig. 5. Experimental Procedure

IV. RESULTS

This section presents the results of the experiments performed in this study. The following are the accuracies obtained by the respective algorithms: k Nearest Neighbours (77 %), Support Vector Machine (76 %), Logistic Regression (77 %), Naïve Bayes (78 %), Decision Trees (81 %), Random Forest (80 %). The Decision Tree classifier performs comparatively better than all other classifiers. Table I. represents a summary of performance report.

Table I. Tabulated summary of experimental results

Algorithm	SVM	Logistic Reg.	kNN	Decision Tree	Random Forest	Naïve Bayes
Accuracy (%)	76	77	77	81	80	78

Fig. 5 presents a visual comparative analysis of experimented results.

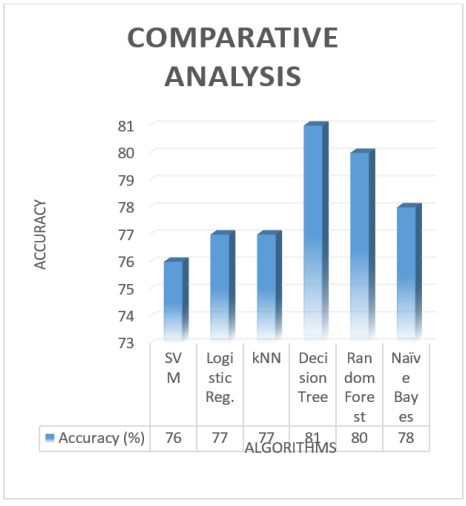

Fig. 5. Comparative analysis of experiments

V. CONCLUSION

This study presents the epidemic of diabetes in female patients, as a global threat to human and social well-being. The concept of diabetes diagnosis with machine learning was introduced to illustrate the potential of machine learning to solve the problem of accurate and timely diabetes diagnosis. Experiments reveal the machine learning approach as a viable solution to accurate diabetes diagnosis. A comparative analysis was performed on the various experimented machine learning algorithms. The results of experiments reveal a classification accuracy of 76 % (SVM), 77 % (Logistic Regression), 77 % (kNN), 81 % (Decision Trees), 80 % (Random Forest), and 78 % (Naïve Bayes). The decision tree algorithm emerged superior to relative techniques with an accuracy of 81 %.

VI. ACKNOWLEDGMENT

I hereby wish to acknowledge the entire staff of computer science for their indelible contribution and support towards my academic endeavours and carrier. Profound acknowledgement to Ahmadu Bello University for a conducive and accommodative environment with the right resources to carry out this research.

REFERENCES

[1] A. O. Sanyaolu, A. Marinkovic, S. Prakash, and M. Williams, “Diabetes mellitus: An overview of the types, prevalence, comorbidity, complication, genetics, economic implication, and treatment,” no. June, 2023, doi: 10.13105/wjma.v11.i5.134.

[2] S. A. Antar *et al.*, “Biomedicine & Pharmacotherapy Diabetes mellitus : Classification , mediators , and complications ; A gate to identify potential targets for the development of new effective treatments,” *Biomed. Pharmacother.*, vol. 168, p. 115734, 2023, doi: 10.1016/j.biopha.2023.115734.

[3] L. C. Martinez and T. Zahra, “Chronic Complications of Diabetes,” no. April, 2022, doi: 10.33590/emjdiabet/21-00180.

[4] N. H. Cho *et al.*, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, 2018, doi: 10.1016/j.diabres.2018.02.023.

[5] L. B. Ribeiro, C. M. Pieper, G. A. Frederico, M. A. Gamba, and A. da S. Rosa, “The relationship between women with diabetes and their body: the risk of diabulimia.,” *Anna Nery Sch. J. Nurs. / Esc. Anna Nery Rev. Enferm.*, vol. 25, no. 4, pp. 1–8, 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=ccm&AN=149825312&lang=pt-pt&site=ehost-live

[6] T. Fadziso and H. P. Adusumalli, “Automatic Diagnosis of Diabetes Using Machine Learning: A Review,” vol. 7, no. 2, pp. 2013–2018, 2020.

[7] D. Goel, M. Vats, Ayush, P. Baliyan, and P. Mittal, “Prediction and Detection of COVID-19 Using Machine Learning,” *Lect. Notes Networks Syst.*, vol. 341, pp. 91–98, 2022, doi: 10.1007/978-981-16-7118-0_8.

[8] S. Pewekar, M. Tirkey, A. Mallik, R. Shaikh, and S. A. Wagle, “Diabetes Prediction Using Machine Learning,” *Lect. Notes Electr. Eng.*, vol. 1196 LNEE, pp. 67–76, 2024, doi: 10.1007/978-981-97-7862-1_5.

[9] A. Ławrynowicz and V. Tresp, “Introducing machine learning,” *Perspect. Ontol. Learn.*, vol. 18, no. November, pp. 35–50, 2014, doi: 10.1007/978-3-030-67626-1_8.

[10] G. Tzanis, I. Katakis, I. Partalas, and I. Vlahavas, “Modern Applications of Machine Learning Extreme Classification View project M Competitions View project Modern Applications of Machine Learning,” no. May 2014, pp. 1–10, 2006.

[11] M. Mohammed, M. B. Khan, and E. B. M. Bashie, *Machine learning: Algorithms and applications*, no. July. 2016. doi: 10.1201/9781315371658.

[12] M. Awad and R. Khanna, “Efficient learning machines: Theories, concepts, and applications for engineers and system designers,” *Effic. Learn. Mach. Theor. Concepts, Appl. Eng. Syst. Des.*, no. April 2015, pp. 1–248, 2015, doi: 10.1007/978-1-4302-5990-9.

[13] S. R. Bansal, S. Wadhawan, and R. Goel, “mRMR-PSO: A Hybrid Feature Selection Technique with a Multiobjective Approach for Sign Language Recognition,” *Arab. J. Sci. Eng.*, vol. 47, no. 8, pp. 10365–10380, 2022, doi: 10.1007/s13369-021-06456-z.

[14] L. Khairunnahar, M. A. Hasib, R. H. Bin Rezanur, M. R. Islam, and M. K. Hosain, “Classification of malignant and benign tissue with logistic regression,” *Informatics Med. Unlocked*, vol. 16, no. August 2018, p. 100189, 2019, doi: 10.1016/j.imu.2019.100189.

[15] S. Sperandei, “Lessons in biostatistics Understanding logistic regression analysis,” no. February, 2014, doi: 10.11613/BM.2014.003.

[16] C. Starbuck, “Logistic Regression,” 2023.

[17] M. Bistroń and Z. Piotrowski, “Comparison of Machine Learning Algorithms Used for Skin Cancer Diagnosis,” *Appl. Sci.*, vol. 12, no. 19, 2022, doi: 10.3390/app12199960.

[18] R. K. Halder, M. N. Uddin, A. Uddin, and S. Aryal, “Enhancing K - nearest neighbor algorithm : a comprehensive review and performance analysis of modifications,” *J. Big Data*, 2024, doi: 10.1186/s40537-024-00973-y.

[19] I. D. Mienye and N. Jere, “A Survey of Decision Trees : Concepts , Algorithms , and Applications,” *IEEE Access*, vol. 12, no. June, pp. 86716–86727, 2024, doi:

10.1109/ACCESS.2024.3416838.

[20] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, “Machine learning with applications in breast cancer diagnosis and prognosis,” *Designs*, vol. 2, no. 2, pp. 1–17, 2018, doi: 10.3390/designs2020013.

[21] C. R. Dhivyaa, K. Sangeetha, M. Balamurugan, S. Amaran, T. Vetriselvi, and P. Johnpaul, “Skin lesion classification using decision trees and random forest algorithms,” *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02675-8.

[22] Y. Song and Y. Lu, “Decision tree methods : applications for classification and prediction,” vol. 27, no. 2, pp. 130–136, 2015.