FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 3, June, (Special Issue) 2024, pp 331 - 343
DOI: https://doi.org/10.33003/fjs-2024-0803-2510

UTILIZING A FUSION OF MACHINE LEARNING TECHNIQUES FOR DIABETES MELLITUS SUBTYPES CLASSIFICATION AND IDENTIFICATION

***[1]Malik Adeiza Rufai, [2]Muhammad Bashir Abdullahi, [2]Opayemi Aderike Abisoye and [2]Oluwaseun Adeniyi Ojerinde**

[1]Department of Computer Science, Faculty of Science, Federal University Lokoja.
[2]Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, Minna.

*Corresponding authors' email: rufai.malik@fulokoja.edu.ng

ABSTRACT

Diabetes Mellitus (DM) is one of the most common health challenges in the world we live in today. It is a deadly disease which prevents the body from making enough insulin. Diabetes Type1 and Type2 are the two major types, which have some similarity in symptoms. Identifying Diabetic Patients with respect to type plays a very significant role in the management process. Misdiagnosis of these types leads to serious impediments. Research shows that the overlapping nature of features contributed to the difficulty in identifying the types and the classification into sub-types. This is still an area of concern (Hassan, et *al*, 2020; Albahli, 2020). In this research, we proposed a method of Support Vector Machine (SVM) and Random Forest Tree (RFT) for the classification of Diabetes sub-types. To reduce the dimensions of the feature set, the Principal Component Analysis (PCA) and Logistic Regression (LR) were used. For effective research, data is sourced from the Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital and Medical City Hospital's public laboratory Dataset to ensure wide coverage. The dataset consists of 834 patient records with eight features and an output column labelled "Type I" or "Type II." This study conducted the experiment using Python, and the results show that the hybrid model outperformed the other prediction methods.

Keywords: Classification, Prediction, Diabetes subtypes, Support Vector Machine, Random Forest, Misdiagnosis

INTRODUCTION

Diabetes mellitus is one of the major global health challenges and this chronic disease has been on the rise in both developing and developed countries (Choubey and Paul, 2016; Pavate *et al.*, 2019; Albahli, 2020; Ganie et *al*, 2022).

People of all age groups are affected by Diabetes Mellitus (DM), a chronic disease that has been affecting people for centuries. The exact cause of the disease is still not known. Age, family history, other relative illnesses, pregnancy, changing glucose levels, blood pressure, etc. are some of the factors or causes (Dash et al., 2019; Annamalai and Nedunchelian,2021; Iparraguirre-Villanueva et *al*, 2023). Diabetes is a disease that medication can managed. A complete cure through medication is not possible (Ganie et *al*, 2022). Type-1, type-2, gestational diabetes, and prediabetes are the four main forms of diabetes (Nibareke and Laassiri, 2020).

Chronic diseases for instance "Diabetes Mellitus" is a global health problem which can lead to several health complications or impediments such as Cardio Vascular diseases, renal failure, Visual impairment (Yuvaraj and Sripreetha, 2019; Jiby, 2021; Saxena et *al*, 2021; Kibria et *al*, 2022). "Insulin is a natural hormone which is secreted by pancreas in the human body". The situation in which this natural hormone cannot efficiently works lead to the accumulation of sugar in the blood stream (Chowdary and Kumar, 2021; Agliata et *al*, 2023). Because of this situation, blood glucose level starts increasing and the person develops Diabetes Mellitus (Hussain and Naaz, 2020; Jiby, 2021; Kibria et *al*, 2022).

Diabetes is characterized by increased blood glucose (sugar) levels (Zou *et al.* 2018; Jiby, 2021). Either an insufficient amount of the hormone insulin, which regulates blood glucose levels, or an improper response of the body's tissues to insulin can cause this (Shuja et *al*. 2019; Hussain and Naaz, 2020).

Diabetes is a serious health disease that affects people of all ages and causes a variety of difficulties (Jiby, 2021; Iparraguirre-Villanueva et *al*, 2023).

Over 425 million individuals are determined to have diabetes in the globe and is anticipated to rise exponentially and estimated to double by 2035 (Hassan, et *al*, 2020; Chang et al. 2023; Rajamani and Sasikala, 2023).

The International Diabetes Federation reported that there are over 400 million People living with Diabetes worldwide and this is expected to rise to about 40% within the next 20 years (Albahli, 2020; Muhammad, 2020; Chang et al. 2023).

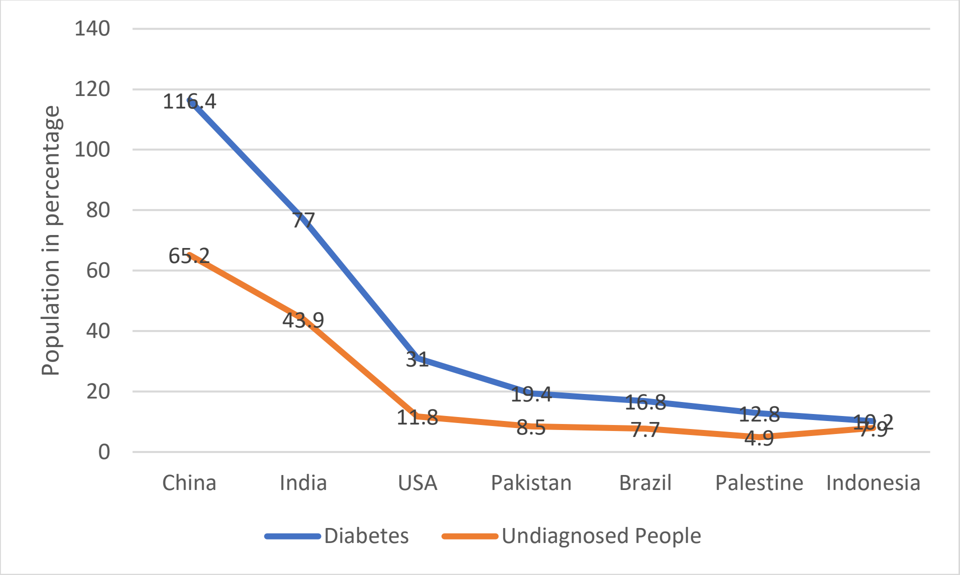


Figure 1: Diabetes Statistics in some top leading Countries around the world

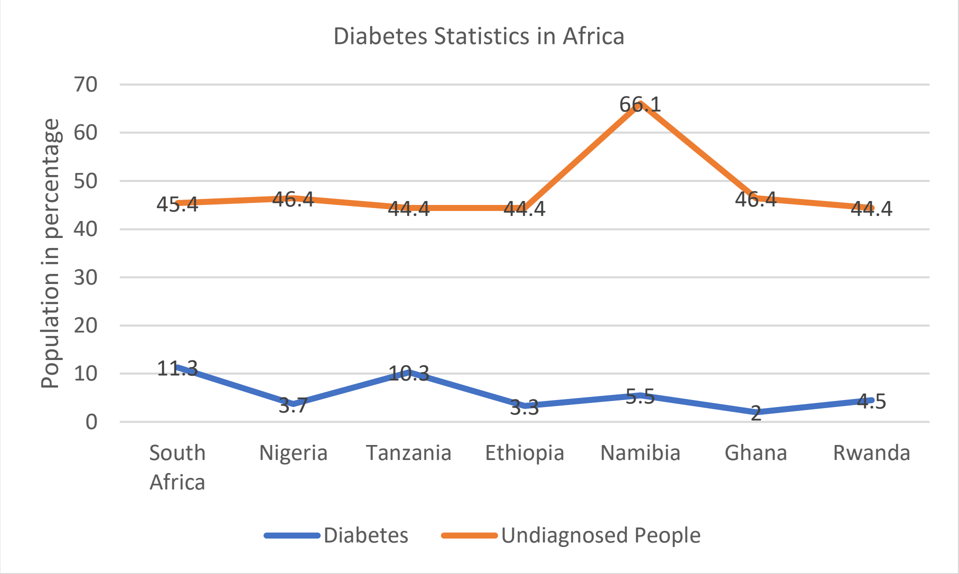


Figure 2: Diabetes Statistics in some African Countries

Many studies are being conducted at various levels to predict or diagnose diabetes or diabetes sub-types earlier in order to reduce the rate of diabetes increase in the coming years (Jiby, 2021; Kibria et *al*, 2022).

According to the Centers for Disease Control and Prevention, Type 2 diabetes, the most prevalent type, accounts for about 80% of cases and is largely caused by excessive weight and sedentary lifestyles. Insulin deficiency is the primary cause of type 1 diabetes, the most common form of the disease that affects children. Type 1 diabetes is rapidly fatal without insulin (Muhammad, 2020; Krishnamoorthi et *al*., 2022).

Insufficient insulin production in the body is the primary cause of Type 1 diabetes. Immune system attacks and pancreatic dysfunction frequently cause low insulin production in the diabetic population. Researchers observed this kind of diabetes in both adults and children.

(Krishnamoorthi et *al*., 2022). The most significant risk factors for type 1 diabetes are pancreatic disease, infection, and family history.

Type II is the next stage of diabetes. It develops when the body's insulin is not properly utilized. This kind of diabetes is most frequently observed in middle-aged adults, although it can affect people of any age (Krishnamoorthi et *al*., 2022). Type 2 diabetes is linked to fatigue, insulin resistance, low glucose tolerance, and gestational diabetes. Other type 2 risk factors include age, sedentary lifestyle, and polycystic ovary syndrome (PCOS).

There are several computerized methods for diagnosing Diabetes Mellitus, but the main disadvantages of these approaches are: patients must undergo several medical tests as a result of the input values required, which is very expensive, with accuracy issues, inconsistency results in large amounts of data, and time-consuming. With the rapid growth in the field of Artificial Intelligence and Machine Learning, there are many approaches and algorithms that can be effectively used for the prediction and diagnosis of Diabetes to achieve optimal accuracy (Zou *et al.* 2018; Contreras et *al*. 2020; Laila et *al*, 2022).

Recently, data mining tools and Machine techniques are widely used in almost every field like healthcare system (Albahli, 2020), weather forecasting, E- business, etc. The Healthcare System is one of the new trending research areas where these techniques and algorithms of Machine Learning can be effectively applied (Albahli, 2020 Saxena et al., 2021).

Diabetes sub-types Prediction, classification and diagnosis at an early stage are critical because chronic hyperglycemia destroys the eye, kidney, nerves, heart, and blood vessels, resulting in lifelong damage. Because of this possibility, the diagnosis/classification is critical (Pekel and Zcan 2018; Saxena et al., 2021; Laila et *al*, 2022).

Much research has been published in the literature to address this issue (Pekel and Zcan 2018; Abaker and Saeed, et *al*, 2021; Kangra and Singh, 2023). Hence, the need to improve on the existing study techniques from the literature for optimal results.

The prime goal is to develop a machine learning model for the prediction of Diabetes Mellitus subtypes so that higher accuracy results can be achieved. This research focuses on how machine learning approaches can be used to classify and diagnose Diabetes Mellitus of sub-types I or II.

The categorisation by types of diabetes mellitus is difficult to distinguish from each other, especially during the initial phase of attack (Albahli, 2020; Nagaraj and Deepalakshmi, 2021; Laila et *al*, 2022). This, in turn, can result in incorrect diagnosis and treatment/management (Ishaq et al., 2018; Nagaraj and Deepalakshmi, 2021). As a result of this, diabetes has become a major contributor to the high mortality rate (Ishaq *et al*., 2018; Pethunachiyar, 2020; Laila et *al*, 2022; Kangra and Singh, 2023). This situation cuts across every society. However, the challenge is more prominent in developing nations (Laila et *al*, 2022; Kangra and Singh, 2023).

Although there have been recent studies that used different learning approachings for the classification tasks and prediction of diabetes mellitus, classification of diabetes into major sub-types is still an area of concern (Iparraguirre-Villanueva et *al*, 2023). Misclassification can have adverse effect on patients' record, research, treatment and outcome of their health (Qureshi and Qureshi, 2017, Emmanuel, et *al*. 2021, Ahamed, et *al*. 2022). Each subtype presents its own set of challenges; thus, research evolves. (Maniruzzaman et *al*., 2020; Emmanuel, et *al*. 2021; Krishnamoorthi et *al*., 2022,).

The complexity of Diabetes subtypes resulting from various factors such as environmental, genetics, lifestyles, insulin levels and clinical characteristics make the classification subtypes difficult due to significant overlapping nature (Emmanuel, et *al*. 2021; Krishnamoorthi et *al*., 2022).

Failure to know a patient's status can lead to complications, such as renal neuropathy and retinopathy, which can eventually lead to death (Emmanuel, et *al*. 2021; Krishnamoorthi et *al*., 2022). There should be some definite measures to reduce the chances of failure and unwanted outcomes.

Recent studies reported that (Sisodia and Sisodia, 2018) used some machine learning algorithms which utilized the PIDD – Pima Indians Diabetes Dataset for its classification, prediction and results show Naive Bayes performed better, achieving the highest accuracy of approximately 77% compared with other techniques used in that research. Other performance metrics such as precision and sensitivity were not into consideration.

Although technology advancements have demonstrated that most diseases can be cured in the current medical era, certain diseases, such as diabetes, can only be prevented and managed rather than cured (Raj et al., 2019; Emmanuel, et *al*. 2021). Predicting diabetes subtypes with classification algorithms at an early stage is critical to this research (Raj et al., 2019; Contreras et al., 2020; Sexana et al., 2021; Laila et *al*, 2022).

Thus, there is a need to improve the accuracy. To achieve it, the research seeks to focus on developing a machine learning prediction model for the classification of Diabetes Mellitus into Type I or Type II, capable of achieving higher accuracy.

Therefore, it is aimed to develop an integration of Support Vector Machine (SVM) and Random Forest (RF) classification model for the Diabetes Mellitus sub-types.

The specific objectives are:

i. Identify the most significant features associated with the classification of Diabetes Mellitus sub-types using Principal Component Analysis (PCA) and Logistic Regression (LR).
ii. To develop a hybrid prediction model for the rapid classification of diabetes mellitus sub-types using modified Support Vector Machine and Random Forest Techniques.
iii. To evaluate the models using some performance metrics.

The result of the research work will be useful for the doctors and other healthcare providers, patients and the general public.

This research will contribute to the body of literature on developing a model for the classification of diabetes into subtypes for better prevention and management plans, thereby constituting the empirical literature for future research in the subject area. This research will also reduce the mortality rate associated with diabetes and other related complications.

Padma et al.; (2018) did a review on classification and prediction techniques in data mining for diabetes mellitus. They talked about how different methods, such as decision trees, Naïve Bayes, Support Vector Machines (SVM), clustering, K-Nearest Neighbours, K-Means, K-Medoids, Neural Networks, Association rule mining, and Multilayer Preceptrons, would be used to create diabetic models. They conducted a thorough analysis of these techniques, finding that the Naïve Bayes and C4.5 algorithm systems performed better and produced satisfactory results, with the C4.5 algorithm's accuracy being 78% and the Naïve Bayes system's 86.37%. Their review provides an in-depth analysis of data mining techniques and suggests that analysts and specialists collaborate to generate simple clinical datasets for the data mining models.

The model's data was retrieved from the Pima Indian diabetes database (PIDD), which is sourced from the UCI machine learning database and contains 768 records. For the condensed dataset with the nine attributes discovered through the comparison of the results of multiple models, the SVM algorithm can be made best with an accuracy of 76 percent (Emmanuel, et *al*. 2021).

Predicting Diabetes Mellitus with Machine Learning Techniques presented by (Qunan, et *al*., 2018) used Decision tree, Random forest, and Neural network to predict diabetis mellitus. They obtained their dataset from hospital physical examination in China. Five-fold cross validation was used to examine the models. They divided their dataset into 2 parts: The healthy people and the diabetics. The healthy people dataset was used to train the model while the diabetics was used as the independent test set. After randomly extracting five times data, the result was the average of the five (5) experiments. They attained an accuracy level of 0.8084 when all the attributes were used. The drawback of this system is that there are 3 types of diabetes but their work only predicted two types which are the type 1 and type 2 diabetes. It cannot be used on the other type of diabetes which is the gestational diabetes that happens to pregnant women.

Iyer, *et al*., (2015) discovered that model(s) for diabetes has been an active research area for many years ago. Most of the models found from the literature were based on classification algorithms and clustering algorithms as well.

In this research work, SVM and RFT will be used to analyze the diabetes parameters and to establish a relation between the two approaches.

Maniruzzaman *et al*. (2020) applied LDA, Quadratic Discriminant Analysus (QDA), Gaussian Process Classifier (GPC), and Naïve Bayes classification algorithms for diabetic patient classification and found that GPC gave the highest accuracy of approximately 82% using the radial basis kernel.

Kumari et al. (2021) proposed an ensemble approach for the classification and prediction of diabetes mellitus using soft voting classifiers. The Pima Indians diabetes dataset has been considered for experimentation, which gathers details of patients with and without diabetes, and the second dataset is the breast cancer dataset, which classifies the dataset into benign and malignant. The proposed ensemble soft voting classifier gives binary classification and uses the ensemble of three machine learning algorithms, viz., Random Forest, Logistic Regression, and Naive Bayes, for the classification. Empirical evaluation of the proposed methodology has been conducted with state-of-the-art methodologies and base classifiers such as AdaBoost, Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, Bagging, Gradient Boost, and XGBoost. By taking accuracy, precision, recall, and the and the F1-score as the evaluation criteria, The proposed ensemble approach gives the highest accuracy, precision, recall, and F1_score value with 79.04%, 73.48%, 71.45%, and 80.6%, respectively, on the PIMA diabetes dataset.

Maniruzzaman *et al*. (2020) firstly replaced the zero entries with the median values and the outliers were detected using the Inter Quartile Range (IQR) method. If the outliers were detected they were then replaced with the median values. Six feature selection techniques, consisting of the Principal Component Analysis (PCA), Logistic Regression, mutual information, analysis of variance and the Fisher Discriminant Ratio (FDR) were applied in combination with ten classification algorithms (Random Forest, Linear Discriminant Analysis, Gaussian Process, Naïve Bayes, Quadratic Discriminant Analysis Classifier, Artificial Neural Network, Support Vector Machine, Decision Tree, Logistic Regression and AdaBoost). They found that the Random Forest classification with Random Forest feature selection gave the highest accuracy of 82.26%.

A survey on the classification techniques for the diagnosis of diabetes was done by (Choubey & Paul 2016). The authors addressed series of problems on research works as reviews were based on several existing papers, mostly on related areas. The implementation for the classification for the diagnosis of diabetes using SVM yields a very high accuracy on popular diabetes dataset – Pima Indian Diabetes Datasets. The paper summarizes and make a lot of comparisons and the techniques are analyzed and compared on the basis of their benefits, challenges and classification accuracy. Guarantee is not assured on the efficient and high accuracy results yielded from popular Pima Indian Diabetes Dataset to be the same as those ones conducted on other datasets. A preliminary benchmark experiment was conducted which indicated a lack of consensus on the best methods for the Diabetes diagnosis Predictions and Identifications. However, the survey was also limited to the publications that were based on classification techniques of diabetes patients.

Kavakiotis et *al*, (2017) used 10 fold cross validation as an evaluation method in three different algorithms, including Logistic regression, Naive Bayes, and SVM, where SVM provides better performance and accuracy of 84 % than other algorithm.

Machine learning plays a vital part in diabetes research in recognizing disorders at an early stage. More machine learning methods were applied in the study. The most successful and commonly used algorithm is Support Vector Machines (SVM) (Pethunachiyar, 2020). SVM with several kernel functions is used in this paper. For diabetes categorization, SVM with linear kernel had the highest accuracy value.

Table 1: Summary of the related existing techniques on diabetes classifications

S/N	Author(s)	Paper title	Method(s)	Results	Contributions	Observed Limitation(s)
1	Kumar et al (2019)	An optimized Random Forest classification for diabetes mellitus	Random Forest in conjunction with Genetic algorithm	Optimized Random forest achieved higher accuracy result of 92.3%	Hybrid optimized Random forest with Genetic algorithm	Evaluation was based on UCI dataset only
2	Kumari and chitra (2013)	Classification of diabetes disease using support vector machine (SVM).	Support vector machine (SVM).	SVM can be successfully used to achieve a higher accuracy. Accuracy, sensitivity are found to be higher using SVM.	The method used focuses on classifying diabetes from high dimensional medical dataset	It can be improved by future subset selection process.
3	Zou,et *al(2018)*	Predicting diabetes mellitus with machine learning techniques.	Decision trees, random forest and neutral network.	Random forest gives highest accuracy when all the attributes were used.	By using a Principal Component Analysis (PCA) and minimum Redundancy Maximum Relevance (mRMR) to reduce the dimensionality	Diabetes sub-types prediction not possible due to nature of database. Large data is expected to be used to optimize higher accuracy.
4	Kumar and Gunavathi (2016)	A Survey on data mining approaches to Diabetes diagnosis and prognosis.	Random forest test, SVM, ANN, Bayesian and Decision tree.	SVM lead the accuracy result 94% accuracy	Multiple techniques were reviewed on different datasets.	Class imbalance and Dimensionality issues
5	Khurana and Kumar(2019)	Improving accuracy for Diabetes Mellitus prediction using data pre- processing and various new learning models.	Algorithms like Decision tree, KNN ,Naïve Bayes, Random forest, logistic regression etc.	Based on comparative study, logistic regression was found to be better than others.	Developing a model tested via dataset with noise (Preprocessing) and dataset without the noise (after preprocessing)	Large data set should be used especially hospital real and very recent data expected to be used, instead of UCI.
6	Choudhury and Gupta (2019)	A survey on medical diagnosis of diabetes using machine learning techniques.	Machine learning algorithms like Decision trees, Random forest, Naïve Bayes, KNN, SVM and logistic regression	Logistic regression gives most accurate results to classify the diabetic and non- diabetic samples.	Comparative study on various machine learning approaches	Focus should be based on classifying type I and type II using a single classifier.
7	Alcala- RMZ, et *a*l (2019)	Identification of diabetic Patients through clinical and para-clinical features in Mexico: An approach using Deep Neural Networks.	Neural Network (ANN)	It gives higher accuracy result of 94%.	Building a model on two separate datasets	More attributes are expected to be used. Increase the dataset only on clinical data.

8	Saxena et. *al*(2014)	Diagnosis of diabetes mellitus using K-Nearest Neighbour Algorithm	K-Nearest Neighbour	Result shows that as K increases, accuracy and errors increases as well. Efficient and higher accuracy obtained.	Building a model for the diagnosis of diabetes using KNN approach.	Focus should be on hybrid classification models using KNN and other techniques and simulation can be better of WEKA, R and Python for more accurate results
9	Huang and Lu (2018)	Intelligent diagnosis of diabetes based on information gain and Deep Neutral Network.	Information gain and Deep Neutral Network	The results shows that the methods used has a classification accuracy of 90.20%.	Development of a deep learning model	Large dataset should be used especially real diabetes patients data from hospitals.
10	Maniruzza man, et. *a*l (2020)	Classification and prediction of diabetes disease using machine learning paradigm	Naïve Bayes, Decision tree, Adaboost and Random Forest	Overall ML results gave accuracy of 90.62% and combination of LR- based and RF- classifiers yields 94.25%	The hybridization of LR-based selection and RF-based classifier perform better accuracy of and 95% AUC for K 10 protocol	Dimensionality issues. Large dataset is required to enhance more accuracy
11	Han, *et al.*, (2018)	Type 2 Diabetes Mellitus prediction Model Based on Data Mining	K- means algorithm and Logistic Regression for feature extraction	85.42% accuracy result achieved	Using logistic regression for feature extraction in other to reduce dimensionality	Focused should be fully on Diabetes dataset only
12	Peter, (2014)	An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus	Clustering approach, Neutral network approach, Support Vector Machine approach, Hybrid approach (Cascading K- means clustering and K- Nearest Neighbour classifier).	Hybrid approach yield an efficient and reliable result of 88.68%	Evaluation were based on certain parameters: Convergence Behaviour, Processing Time, Classification Accuracy	Processing time/ computational speed needs to be improved.
13	Aiswarya, *et al*. (2015)	Diagnosis of Diabetes Using Classification Mining Techniques	Naïve Bayes and Decision Tree (J48)	The Naïve Bayes technique gave an accuracy of 79.56%. while the percentage split for J48 gave an accuracy of 76.95%	Cross-validation techniques and percentage split technique (70:30) are the contributions to improve on the result	It is limited for being used for pregnant women only. The dataset is not dynamic.

14	Quanan, *et al.* (2018)	Predicting Diabetes Mellitus with Machine Learning Techniques	Decision tree, Random forest, and Neutral network to predict Diabetes Mellitus	Attained an accuracy level of 80.84% when all the attributes were used.	Dataset from hospital physical examination in China and classified into two separate data Diabetic and Non-diabetic	The drawback of this research is that there are 3 types of diabetes but their work only predicted two types which are the type1 and type2 Diabetes.
15	Choubey and Paul (2016).	Classification techniques for diagnosis of diabetes: a review	Various machine learning techniques	SVM yields a very high accuracy on popular diabetes dataset- Pima Indian Diabetes Datasets.	A Preliminary benchmark experiment was conducted which indicated a lack of consensus on the best methods for the Diabetes diagnosis Predictions and Identifications.	The survey was also limited to the publications that were based on classification techniques of Diabetes patients.
16	Alhassan, *et al.,* (2015)	Performance Analysis of Artificial Neutral Network (ANN) with Decision Tree Algorithm (DTA) in Prediction of Diabetes Mellitus	ANN and DTA	DTA yields higher accuracy than ANN based on some parameters	The Dataset evaluation was based on some certain parameters	Large dataset not used in this research

The literature evaluation revealed research gaps, indicating that an algorithm for improving SVM using a feature selection approach can be developed. It is clear that a modified SVM (SVM) with a customized instance (C), loss function (ε), penalty parameter (bit), and Random Forest Tree does not exist (RFT). As a result, we would like to research a hybrid algorithm based on this better combination of lowered penalty parameters and loss function in the hopes of acquiring important insight into the field of DM prediction and classification in order to provide precise classification

Table 2: Research gaps

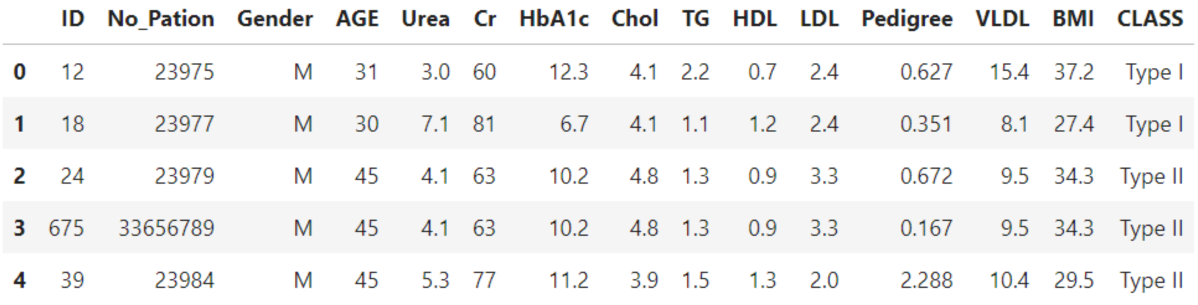
S/N	Authors'	Methodology	Research Gaps identified
1	Sisodia and Sisodia (2018)	Naïve Bayes classification algorithm	Unavailability of feature selection algorithm
2	Nagaraj and Deepalakshmi, 2021	SVM -NN	ESVM with feature selection process not applied. Non-applicability of optimization algorithm and PCA.
3	Haritha *et al*., (2018)	Cuckoo-Fuzzy KNN	Improvement in the learning rate of optimization algorithm
4	Zhu *et al*., (2019)	PCA and K-means techniques	Lack of mechanism for selection of number of Principal components
5	Perveen *et al*., (2019)	SVM	Performance improvement for all sampling cases
6	Sivakumar *et al*., (2020)	Naïve Bayes and Random Forest	Increased misclassification rate
7	Edla et al., (2017)	RBFNN	Tradeoff between number of hidden layer and accuracy
8	Srivastava et al., (2020)	PCA with SVM	Need for specifying number of accuracies

MATERIALS AND METHODS

Dataset

The Specialized Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital and Medical City Hospital's public laboratory served as the training and testing grounds for the Machine Learning models (Rahid, 2020). This dataset includes 934 type I and type II diabetic patients with twelve different features and an outcome feature. Table 1 displays the attribute descriptions and a brief statistical overview.

Table 3: Dataset description



	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	Pedigree	VLDL	BMI	CLASS
0	12	23975	M	31	3.0	60	12.3	4.1	2.2	0.7	2.4	0.627	15.4	37.2	Type I
1	18	23977	M	30	7.1	81	6.7	4.1	1.1	1.2	2.4	0.351	8.1	27.4	Type I
2	24	23979	M	45	4.1	63	10.2	4.8	1.3	0.9	3.3	0.672	9.5	34.3	Type II
3	675	33656789	M	45	4.1	63	10.2	4.8	1.3	0.9	3.3	0.167	9.5	34.3	Type II
4	39	23984	M	45	5.3	77	11.2	3.9	1.5	1.3	2.0	2.288	10.4	29.5	Type II

Research Framework

This research methodology covers the methods applied in the accomplishment of a machine learning model for the prompt classification of diabetes diseases. It further explains the method of application of a machine learning algorithm to correctly predict a patient's ailment status with the provision of very large datasets.

Preprocessing

Outlier rejection (OR), filling in missing values (MV), and feature selection of the attribute are included in the preprocessing step of the suggested framework. These steps are briefly described as follows: An outlier is an observation that deviates significantly from other observations (Bansal et al., 2016; Hasan et *al*., 2020). Given that classifiers are particularly sensitive to the data range and attribute distribution, it must be excluded from the data distribution.

The attributes with missing or null values were processed to fill the null values as this could cause any classifier to make an incorrect prediction. In the framework, the missing values were calculated by the mean values of the attributes rather than dropping, which can be formulated as in the equation below.

$$MV(x) = \begin{cases} mean(x), if\ x = null/missed \\ x, \qquad otherwise \end{cases} \qquad (1)$$

The Feature Selection Technique (FST) consistently reduces computing overhead while increasing classification accuracy. Additionally, FST removes the less significant features and lowers the time complexity of the machine learning techniques.

With each increase in attribute dimension, the classifiers' accuracy rises. When the attribute's dimension rises without the sample size, the performance of the classifiers will, however, tend to decline. In this literature, Principle Component Analysis (PCA), the most widely used method for feature selection, was used to compare their performance for the dataset used. The detailed algorithm of the PCA-based technique was used to compare their performance.

Model Validation

In this research, 10-fold cross validation was used to evaluate the capability of the model.

The data is divided into 10 equal portions using cross validation. The remaining nine subsets are combined to create a training set, while one portion of the 10 subsets is used for testing. Now, the components into which we divided the dataset continue to interact to create various pairings of training and testing data. In the comparison table below, several accuracy scores for each combination are displayed. The benefit of this technique is that it decreases errors induced by bias linked with the random sampling technique.

Machine Learning Models and Ensembling

Two different ML models, such as Support Vector Machine (SVM) and Random Forest Tree (RFT), have been trained and tested in the proposed framework independently. The SVM was modified and then hybridized with the Random Forest for optimal results. The essence of combining the models is to boost the performance of the result.

For over two decades, researchers have evaluated diabetes using a variety of machine learning techniques. Classification algorithms such as the random forest, support vector machines (SVMs), are mostly used in the classification and prediction tasks (Alghamdi et al., 2017; Chowdary and Kumar, 2021), thus the choice of the two techniques.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method where it categorizes new example with an optimal hyperplane. Support vector machines are frequently preferred due to their great accuracy and low processing resources/computational power (Hassan, et *al*., 2020).

More machine learning methods were applied in the diabetes classifications. The most successful and commonly used algorithm is Support Vector Machines (SVM) (Pethunachiyar, 2020).

The mathematical concept on how SVM classifies new data points into type I and type II diabetes based on some features is stated as follows:

Let's denote; x as the feature vector representing a data point X = [x_1, x_2, x_3,.......xn] where :

x_1 represents age, x_2 represents HbA1c, x_3 represents VLDL let's denote:

w as the weight vector // to the hyperplane, b as the bias term

The hyperplane can be represented as;

$$w.x + b = 0 \tag{2}$$

The classification decision for a new data point x^1 is :

$$f(x^1) = \pm(w.x^1 + b) \tag{3}$$

If $f(x^1)\ is + ve : Type\ I\ and\ if\ f(x^1)\ is - ve : Type\ II$

The optimization objective of SVM mathematically represented as

$minimise(\frac{1}{2}\|w\|\text{^}2)$

Subject to :

$y_i \quad (w.x_i + b) \quad \geq 1\ where\ yi\ is\ the\ label\ of\ the\ i-th\ data\ point\ \ (+1\ for\ type\ 1\ and - 1\ for\ type\ II), and\ x_i\ is\ the\ feature\ vector\ of\ the\ i-th\ data\ point.$

Random Forest

The Random Forest uses numerous decision trees for classification. RF is a multifunctional machine learning method. It is capable of carrying out regression and prediction tasks. Additionally, bagging-based RF is a key component of ensemble machine learning. RF has been employed in several biomedical research projects. In contrast to other decision tree algorithms, RF generates a large number of decision trees. In the regression problem, the RF output is the average value of the output of all decision trees (Zou, et *al*., 2018; Hasan, et *al*., 2020).

The mathematics representation of Random Forest based on some selected features: Age, HbAIc, and VLDL are:

Let X represent the input feature vector for a patient, where $X =$ (x_1, x_2, x_3……x_n)

x_1 =Age, x_2= HbA1c and x_3 =VLDL and Y represent the output variable, where Y can take on two values representing the classes.

Y = {Type I, Type II}.

Let D represent the training dataset, where each entry consists of an input feature vector X and its corresponding class label Y.

D = {(X_1, Y_1), (X_2, Y_2), (X_n, Y_n)}

Let T represent Random Forest model which is an ensemble of decision trees. Each decision tree t_i in the Random Forest T is represented as a function t_i (X)

The Random Forest model T combines the predictions of all decision trees in the ensemble using a majority voting mechanism to make the final prediction.

Mathematically,

T(X) = MajorityVote (t_1(X), t_2(X), …..t_k (X))

Where k is the number of decision trees in the Random Forest and t_i(X) represents the prediction of the ith decision tree for input feature vector X.

RESULTS AND DISCUSSION

In experimental studies, the dataset has been partitioned between 70–30 % (583–351) for training and testing purposes. Tab. 7 shows that the proposed model performed well with an accuracy of 99.99%. The combined model has higher accuracy, sensitivity and specitivity respectively and has the lowest RMSE value of 26.52%. The more the area covered, the better the classifier. These measurements are taken by using the Python on the Diabetes Dataset taken from the KDnugget repository. The results are shown in Tab. 7. The results may be improved by applying large-scale updated datasets. However, we need to apply other machine learning algorithms using real data sets before generalizing the results.

Table 4: Performance metrics on the used dataset for the research

Model	Accuracy	F1-Score	Sensitivity	Specificity	Precision	Recall	AUC
SVM	86.79	97.93	100.0	86.66	95.94	100.0	93.33
RF	89.46	99.64	100.0	97.77	99.30	100.0	98.88

Table 5: Performance metrics on augmented dataset

Model	Accuracy	F1-Score	Sensitivity	Specificity	Precision	Recall	AUC
SVM	98.0	98.0	90.14	64.44	88.88	90.14	77.29
RF	99.19	99.0	100.0	100.0	99.98	100.0	100.0

Table 6: Accuracy comparison on Diabetes Type I and Type II on augmented Dataset

Diabetes types	Model	Accuracy
Type I	SVM	94.30%
Type II		89.56%
Type I	RF	96.77%
Type II		95.87%

Table 7: Summary of performance evaluation of the ML Algorithms

Method	Precision %	Recall %	F1- Score%	Accuracy %	Sensitivity%	Specificity%
Random Forest	99.98	100	99.0	99.19	90.14	64.44
SVM	97.01	100	98	98.00	99.10	98.91
SVM + RFT	99.09	99.99	98.65	99.99	100.00	99.99

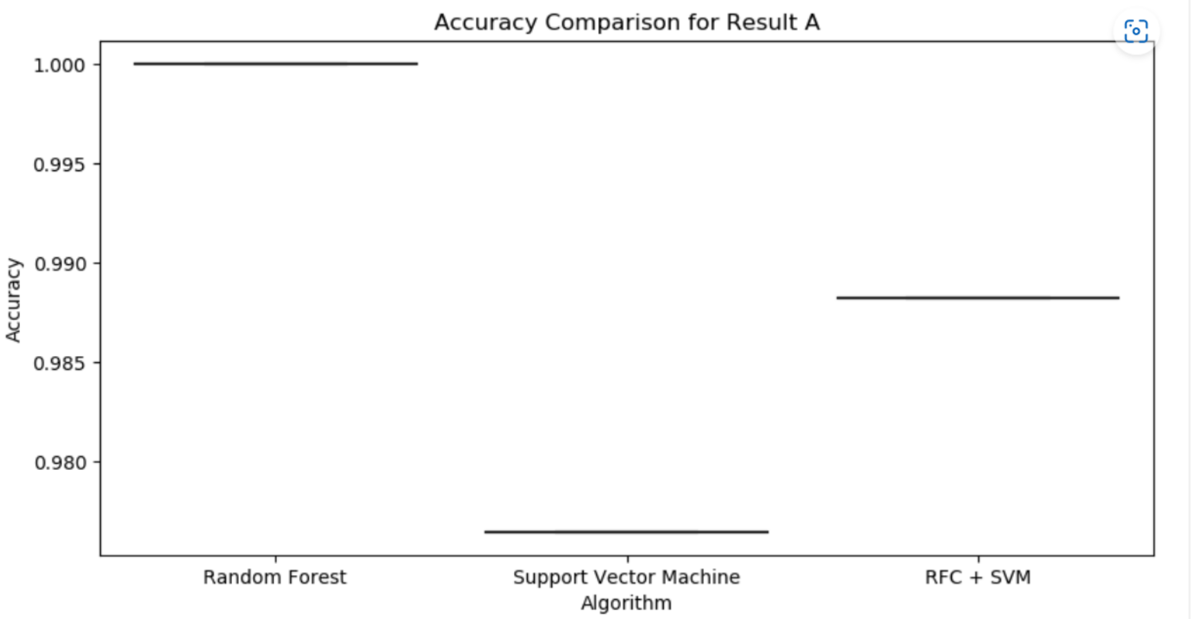


Figure 2: Performance evaluation of the proposed model based on accuracy

CONCLUSION

This study introduced a novelty by integrating modified Support Vector Machines (SVM) with Random Forest (RF) for the classification of type I and type II diabetes. Our research addresses the challenges of accurate diabetes subtypes classification by leveraging on the strengths of both machine learning approaches.

Our results demonstrated that the integrated modified SVM-RF model achieved optimal performance of 99.99%, 100.00%, and 99.99% respectively, in terms of accuracy, sensitivity, and specificity in classifying diabetes types. This model performs better compared to other approaches so far in the literature related to this study. The modified SVM through kernels in the SVM component proved particularly effective, and these equally handled the complex features and class imbalance.

The improved classification accuracy of our model has significant implications for diabetes diagnosis and management. It offers the potential for earlier and more accurate identification of diabetes types, which is crucial for appropriate treatment planning and patient care. Furthermore, the model's ability to resolve the overlapping nature of the features and handle diverse input data shows its potential applicability across various issues relating to health.

Despite its promising results, our study had some limitations. Computational complexity might be a challenge, and this can serve as a direction for future research. The study recommends that future research focus on the interpretability of the model's performance on larger and more diverse datasets and the integration of additional relevant predominant features. Additionally, clinical validation studies will be crucial to assessing the model's real-world applicability and impact on patient outcomes.

In conclusion, the hybridized SVM-RF model represents a significant step forward in the application of learning-based techniques to diabetes classification. By leveraging the strengths of the random forest and SVMs, we have developed a promising tool that could contribute to more accurate and timely diabetes classifications. It is sure that the study will play an increasingly important role in improving diabetes classifications and prompting subtypes' identification and management.

REFERENCES

Abaker, A. A., & Saeed, F. A. (2021). A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. *Informatica*, *45*(1).

Ade-Ojo, Toluwani (2018). Development of an intelligent decision support system for prompt diagnosis of Ebola and Lassa fever disease (Doctoral dissertation, Federal University Oye- Ekiti).

Agliata, A., Giordano, D., Bardozzo, F., Bottiglieri, S., Facchiano, A., & Tagliaferri, R. (2023). Machine learning as a support for the diagnosis of type 2 diabetes. *International Journal of Molecular Sciences*, *24*(7), 6775.

Ahamed, B. S., Arya, M. S., Sangeetha, S. K. B., & Auxilia Osvin, N. V. (2022). Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers. *Applied Computational Intelligence and Soft Computing*.

Ahuja, R., Sharma, S. C., & Ali, M. (2019). A Diabetic Disease Prediction Model Based on Classification Algorithms. *Annals of Emerging Technologies in Computing (AETiC)*, *3*(3).

Albahli, S. (2020). Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection. *Journal of Medical Imaging and Health Informatics*, *10*(5), 1069-1075.

Amoo, A. O., Oyegoke, T. O., Balogun, J. A., & Bamidele, S. A. (2020). Survival Model for Diabetes Mellitus Patient Receiving Treatm. *International Journal of Computers*, *5*.

Annamalai, R., & Nedunchelian, R. (2021). Diabetes mellitus prediction and severity level estimation using OWDANN algorithm. *Computational Intelligence and Neuroscience*, *2021*.

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*(22), 16157-16173.

Chowdary, P. B. K., & Kumar, R. U. (2021). An Enhanced NAÏVE BAYES Classification Algorithm to Predict Type II Diabetes. *Journal of Engineering Science and Technology*, *16*(4), 2927-2937.

Contreras, I., Bertachi, A., Biagi, L., Oviedo, S., Ramkissoon, C., & Vehi, J. (2020). Artificial intelligence-based decision support systems for diabetes. In *Artificial Intelligence in Precision Health* (pp. 329-357). Academic Press.

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. J. Big Data 6, 54. doi: 10.1186/s40537-019-0217-0

Edla, D. R., & Cheruku, R. (2017). Diabetes-finder: A bat optimized classification system for type-2 diabetes. Procedia Computer Science, 115, 235–242. doi:10.1016/j.procs.2017.09.130

Emmanuel, G., Hungilo, G. G., & Emanuel, A. W. R. (2021, March). Performance evaluation of machine learning classification techniques for Diabetes disease. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1098, No. 5, p. 052082). IOP Publishing.

Ganie, S. M., Malik, M. B., & Arif, T. (2022). Machine Learning Techniques for Diagnosis of Type 2 Diabetes Using Lifestyle Data. In *International Conference on Innovative Computing and Communications* (pp. 487-497). Springer, Singapore.

Han, W., Shengqi, Y., Zhangqin, H., Jian, H., & Xiaovi, W. (2018). Type 2 Diabetes Mellitus Prediction Model Based on Data MIning . *Informatics in Medicine Unlocked 10*, 100-107.

Haritha, R., Babu, D. S., & Sammulal, P. (2018). A Hybrid Approach for Prediction of Type-1 and Type-2 Diabetes using Firefly and Cuckoo Search Algorithms. International Journal of Applied Engineering Research: IJAER, 13(2), 896–907.

Hassan, A. S., Malaserene, I., & Leema, A. A. (2020). Diabetes Mellitus Prediction using Classification Techniques. *Int. J. Innov. Technol. Explor. Eng*, *9*(5), 2080-2084.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531.

Hussain, A., & Naaz, S. (2020). Prediction of Diabetes Mellitus: Comparative Study of Various Machine Learning Models. In *International Conference on Innovative Computing and Communications* (pp. 103-115). Springer, Singapore.

Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. *Diagnostics*, *13*(14), 2383.

Ishaq, F.S., Muhammad, L.J., Yahaya, B.Z and Atomsa, Y (2018) Data Mining Driven Models for Diagnosis of diabetes Mellitus: A Survey. Indian Journal of Science and Technology. Vol.11 (42). pp 1-9.

Jiby, T. C. (2021) A Study on Various Machine Learning Classification Algorithms for Diabetes Prediction. *International Journal of Engineering Research & Technology (IJERT)*, *10*(8). 425-427.

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, *12*(3), 1728-1737.

Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors*, *22*(19), 7268.

Korzun, D.G. (2017) Internet of things meets mobile health systems in smart spaces: An overview. In Internet ofThings and Big Data Technologies for Next Generation Healthcare; Springer: Cham, Switzerland,pp. 111–129.

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel

Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering*, *2022*.

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.

Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. (2022). An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. *Sensors*, *22*(14), 5247.

Lican, H., & Chuncheng, L. (2018). Intelligent Diagnosis of Diabetes Based on Information Gain and Deep Neural Network. *Proceeding of CCIS2018*.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer methods and programs in biomedicine*, *152*, 23-34.

Manu, G., Neil, D. R., Adrian, K. D., Jennifer, S., & Moi, H. Y. (2017). DFUNet: Covolutional Neural Networks for Diabetic Foot Ulcer Classification. *arXiv:1711.10448v2 [cs.CV]*.

Morris, A.H (2019). “Decision Support and Safety of Clinical Environments”. British Medical Journal. Vol. 11(1).pp:69-75. https://dx.doi.org/10.1136/qhc-11.1.69. Epidemiology and Community Health.

Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. *SN Computer Science*, *1*(5), 1-10.

Nagaraj, P., & Deepalakshmi, P. (2021). Diabetes Prediction Using Enhanced SVM and Deep Neural Network Learning Techniques: An Algorithmic Approach for Early Screening of Diabetes. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, *16*(4), 1-20.

Nibareke, T., and Laassiri, J. (2020). Using big data-machine learning models for diabetes prediction and flight delays analytics. J. Big Data 7, 78. doi: 10.1186/s40537-020-00355-0

Olokoba, A. B., Obateru, O. A., & Olokoba, L. B. (2012). Type 2 diabetes mellitus: a review of current trends. *Oman medical journal*, *27*(4), 269.

Padma,T.,Uma,N.M., R, J.G. (2018). A Survey on Classification and Prediction Techniques in Data Mining for Diabetes Mellitus. *International Journal of Trend In Scientific Research and Development (IJTSRD)*, 496-504.

Patil, R., & Tamane, S. (2018). A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. *International Journal of Electrical and Computer Engineering*, *8*(5), 3966.

Pavate, A., Nerurkar, P., Ansari, N., & Bansode, R. (2019). Early prediction of five major complications ascends in diabetes mellitus using fuzzy logic. In *Soft Computing in Data Analytics* (pp. 759-768). Springer, Singapore.

Pekel, E., & ÖZCAN, T. (2018). Diagnosis of Diabetes Mellitus using Statistical Methods and Machine Learning Algorithms. *Sigma: Journal of Engineering & Natural Sciences/Mühendislik ve Fen Bilimleri Dergisi*, *36*(4).

Perveen, S., Shahbaz, M., Keshavjee, K., & Guergachi, A. (2019). Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. IEEE Access. IEEE, 7, 1365–1375. doi:10.1109/ACCESS.2018.2884249

Peter, S. (2014). An Analytical Study On Early Diagnosis and Classification of Diabetes Mellitus. *Bonfring International Journal of Data Mining*.

Pethunachiyar, G. A. (2020, January). Classification Of Diabetes Patients Using Kernel Based Support Vector Machines. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-4). IEEE.

Qunan, Z., Qu, K., Yamei, L., Dehui, Y., Ying, J., & Hua, T. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in Genetics*.

Rajamani, S., & Sasikala, S. (2023). Artificial Intelligence Approach for Diabetic Retinopathy Severity Detection. *Informatica*, *46*(8).

Raj, R. S., Sanjay, D. S., Kusuma, M., & Sampath, S. (2019). Comparison of support vector machine and Naive Bayes classifiers for predicting diabetes. In *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)* (pp. 41-45). IEEE.

Resul T.I, Abdulkadir (2008), “Effective diagnosis of heart disease through neural networks ensemble”. An International Journal ELSEVIER Vol.(36). pp: 7675-7680.

Rodrigues, J. and Compte, S. (2016) Health Level In e-Health Systems. Theory and Technical Applications; Elsevier: Amsterdam, the Netherlands. 21–31.

Sarda-Espinosa, A., Subbiah, S., & Bartz-Beielstein, T. (2017). Conditional inference trees for knowledge extraction from motor health condition data. *Engineering Applications of Artificial Intelligence*, *62*, 26-37.

Saxena, R., Sharma, S. K., & Gupta, M. (2021). Analysis of machine learning algorithms in diabetes mellitus prediction. In *Journal of Physics: Conference Series* (Vol. 1921, No. 1, p. 012073). IOP Publishing.

Shuja, M., Mittal, S. and Zaman, M (2019) Diabetes Mellitus and Data Mining Techniques: A survey. International Journal of Computer Sciences and Engineering. Vol. 7(1). pp: 2347-2693

Sisodia, D.and Sisodia, D. S. (2018) Prediction of Diabetes using Classification Algorithms.
International Conference on Computational Intelligence and Data Science (ICCIDS 2018); Elsevier: Raipur, India. 132 pp.1578-1585

Sivakumar, S., Venkataraman, S., & Bwatiramba, A. (2020). Classification Algorithm in Predicting the Diabetes in Early

Stages. Journal of Computational Science, 16(10), 1417–1422. doi:10.3844/jcssp.2020.1417.1422

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1), 13.

Srivastava, A. K., Kumar, Y., & Singh, P. K. (2020). A Rule-Based Monitoring System for Accurate Prediction of Diabetes: Monitoring System for Diabetes. International Journal of E-Health and Medical Communications, 11(3), 32–53.

WorldHealth Statistics 2017: MonitoringHealth for the SDGs;WorldHealthOrganization: Geneva, Switzerland, 2017.

Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Cluster Computing*, *22*(1), 1-9.

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179.

Zou, Q., Qu, K., Luo, Y.,Yin, D., Ju, Y. and Tang, H (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Fronties in Genetics. Vol. 9 (515).