ISSN 2278-3091
Volume 10, No.1, January - February 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse641012021.pdf
https://doi.org/10.30534/ijatcse/2021/641012021

# Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review

**Oluwafemi Samuel Abe[1], Olumide O. Obe[2], Olutayo K. Boyinbode[3],Olagbuji N. Biodun[4]**
[1]Department of Computer Science. Federal University of Technology, Akure, Nigeria. E-mail: osabe@futa.edu.ng
[2] Lecturer, Department of Computer Science. Federal University of Technology, Akure, Nigeria. E-mail: ooobe@futa.edu.ng
[3] Lecturer, Department of Computer Science. Federal University of Technology, Akure. Nigeria. E-mail: okboyinbode@futa.edu.ng
[4] Department of Obstetrics and Gynecology, Ekiti State University, Ado-Ekiti, Nigeria.E-mail: ebiodun_olagbuji@yahoo.com

## ABSTRACT

Timely disease prediction means a lot to the improvement of the health care services and this will go a long way to assist populaces to avoid unsafe health circumstances before resulting in complex medical situations. Diabetes Mellitus as one of the deadly diseases that is described by hyperglycemia taking place due to defects in insulin secretion which allow an irregular increase in glucose level. Diabetes Mellitus can lead to loss of sight, non-traumatic lower extremity amputation, chronic kidney disease, coronary heart disease, stroke, etc. Hence, prompt diagnosis of the Diabetes Mellitus disease has to pay more attention to in the recent area of research. Presently, there is a great and wide work carried out on Machine Learning with a focus on medical and its application. This paper reviewed recent journals that made use of Artificial Intelligence techniques, different classifiers and ensemble methods to assist in the management of diabetes. Classifiers algorithms such as Naive Bayes, Decision Tree, Artificial Neural Network, Support Vector Machine, K-Nearest Neighbour and Multi-Layer Perceptron. The results from over 1,137 most related reviewed journals reveled that Ensemble Models have the highest average accuracy of 87.09 % in respect to prediction of diabetes mellitus.

**Key words:-** Diabetes Mellitus, Classifiers, Ensembles, Machine Learning, Prediction

## 1. INTRODUCTION

Considering the rapid increase in the diseases that is common to the elderly one in the developing and developed countries, it was discovered from statistics that Diabetes Mellitus is more common. This disease is a "metabolic illness described by chronic hyperglycemia that results from disturbances of carbohydrates, fat and protein metabolism as a result of imperfections in insulin ooze, insulin action or both" [1].Due to diabetes, a high rate of considerable morbidity, healthcare deployment and mortality have been recorded. Based on the record of [2] and work of [3], it was discovered that in "Nigeria, over 29 percent of the death was estimated to have caused by Diabetes Mellitus". "And globally, it was estimated in 2017 that 425 million people had diabetes; it has been predicted that by the end of 2045 this will increase to 629 million"[4].

Generally, there are four different types that Diabetes.Type-1 Diabetes (T1D), "this occurred as a result of β-cells autoimmune damage which causes suppression or cessation of the insulin secretion" [5]. Therefore, the pancreas stops the production of insulin due to the weakness of the organ. This type is common among the young and growing up people. This is also referred to as juvenile-onset DM. The T1D causes narrow of blood vessels in the kidney (diabetic nephropathy), heat failure and stroke at the complicated stages.Type-2 Diabetes (T2D) is also referred to as " 'Non-Insulin-Dependent Diabetes Mellitus (NIDDM) or Adult-onset Diabetes (AOD)' this is as a result of the absence of insulin or Insulin Resistance (IR)"[6] or in simple term or low level of insulin. T2D is associated with least or sometimes no symptoms [7]. Those that have obesity risk the tendency of having T2D due to their overweight. Another type is Gestational Diabetes Mellitus (GDM):This occurs majorly at pregnancy stage. When compared to other types of diabetes, it does occur as a result of having too little insulin. But "GDM is a result of hormones from the placenta that hinder the body making use of the insulin" [7],[8]. Lastly, Pre-diabetes: "This category occur before TD2. In pre-diabetes, the individual glucose level is higher normal and yet not the level of TD2. Patients with pre diabetes have a higher tendency of having TD2" [8].

In any of the types of diabetes, early diagnosis, good self-management by the patient and nonstop medical attention are necessary to avoid acute complications (e.g. ketoacidosis). This also reduces the risk of prolonging complications such as stroke, cardiovascular disease, diabetic foot, diabetic nephropathy, or diabetic retinopathy. "The scientists believe that cognitive ability and environmental lifestyle contributed a lot of treatment and management of diabetes patients"[9]. "Computer Science and Machine Learning has made giant strides and is now being used in medical settings"[10]. For the better diagnosis and treatment of patients. Therefore, the therapeutic decision has to be taken into account due to the complexity of diabetes therapy and to make the patient fit for daily responsibilities, lifestyle-related activities must be optimized. In line with this challenge, more researchers are

working hard to address and create a system and tools that will predict diabetes more accurately.
Prediction through Machine Learning (ML) which is a branch of Artificial Intelligence, "is a system that allows the computers to learn and gain intelligence based on experience with the development of algorithms"[11],[14].
AI has played of a lot of roles in various fields and now a "key point of focus applied in various medical specialized areas" [12]. AI is involved cleverly calculations and methods. For example, Machine Learning, Fuzzy Logic, Natural Language Processing, Robotics, Knowledge Base, Expert Systems and the combination of two or more strategies called multi-methods [13].This paper focused on ML, the major goal of ML is to develop a computer system that respond from their prior observation that it learnt. There are three categories of ML, Supervised learning (SL), Unsupervised Learning (UL) and Reinforcement Learning (RL).

## 2. METHOD AND ARTICLE SELECTION PROCESS

The searching process was focused on Artificial Intelligence Methods and the application of DM. IEEE, Pubmed, Elsevieretc. database were searched for articles. These were selected as a result of large collection of high impact academic research papers. To be precise on get more relevant papers, more attention was focused on Diabetes and Artificial Intelligence (DAI); Diabetes Machine Learning and Ensemble Method. This was followed by manual review. The year range was limited to the work of 2016 – 2020. The process reduced the papers from 1,135 to 137 and then to 51which is most relevant. The details were spelled out on table 1.Figure 1 shows the workflow of the entire collection of the reviewed papers.

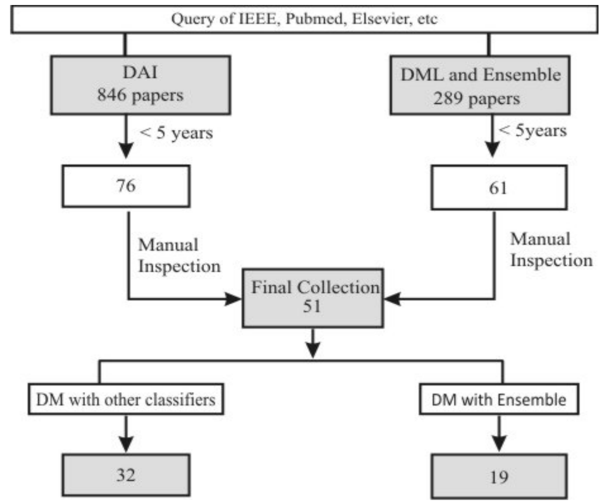


Figure 1. The workflow of the entire collection of the reviewed papers.

## 3.RELATED WORKS

In survey of earlier research on diabetes mellitus, many scholars have done a lot of study on the medical data of the disease. Also in the area of making use of various machine learning algorithms and data mining technology to construct various prediction and analysis models. Table 1 shows the 2016-2020 most related researcher with the dataset and tool used.

Table 1: Researchers Dataset and Tools used

| Ref. No. | Author(s) | Year | Title | Method | Data Set | Method with best performance | Best performance Result |
|---|---|---|---|---|---|---|---|
| [15] | Casanova *et al.* | 2016 | Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning. | RF CART | Un disclosed | RF CART | Accuracy = 75% |
| [16] | Anderson *et al.* | 2016 | Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. | LR and RF | Un disclosed | LR | Accuracy = 75% |
| [17] | Lee and Kim | 2016 | Identification of type 2 diabetes risk fac- tors using phenotypes consisting of anthropometry and triglycerides based on machine learning. | NB | Un disclosed | LR | Accuracy = 0.698 |
| [18] | Jelinek *et al.* | 2016 | Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis. | DT | Un disclosed | DT | Accuracy = 85.63% |
| [19] | Yu *et al.* | 2016 | Artificial neural networks for estimating glomerular filtration rate by urinary dipstick for type 2 diabetic patients. | ANN | Un disclosed | ANN | Accuracy = 87% |
| [20] | Luo | 2016 | Automatically explaining machine learning pre-diction results: a demonstration on type 2 diabetes risk prediction. | SVM | Un disclosed | SVM | Sensitivity = 84.4% |
| [21] | Cai *et al.* | 2017 | Predicting DPP-IV inhibitors with machine learning approaches. | NB | Un disclosed | NB | Accuracy = 87.2% |
| [22] | Chen*et al.* | 2017 | A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree | K-means and DT | UCI Pima | DT | Accuracy = 90.4% |
| [23] | Kagawa *et al.* | 2017 | Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach. | SVM Rule base | Un disclosed | SVM | |
| [24] | Zheng *et al.* | 2017 | A machine learning-based framework to identify type 2 diabetes through electronic health records. | KNN, NB, DT RF, SVM and LR | Un disclosed | LR | Accuracy = 99% |
| [25] | Sayadi *et al.* | 2017 | Simple prediction of type 2 diabetes mellitus via decision tree modeling. | DT and LR | Un disclosed | DT | Accuracy = 89% |
| [26] | Uswa and Naeem | 2017 | Predicting Diabetes in Medical Datasets Using Machine Learning Techniques | NB, DT and KNN | Pima Indians Diabetes Database | DT | Accuracy = 94.44% |
| [27] | Priya *el al.* | 2017 | Analyze Data Mining Algorithms For Prediction Of Diabetes | Gaussian NB, KNN, SVM and DT | Pima Indians Diabetes Database | KNN | Accuracy =70.87% |
| [28] | Deepti and Dilip | 2018 | Prediction of Diabetes using Classification Algorithms | DT, SVM and NB | Un disclosed | NB | Accuracy = 76.30% |
| [29] | Das *et al.* | 2018 | Automatic Diabetes Prediction Using Tree Based Ensemble Learners. | RF and Gradient Boosting | Un disclosed | Ensemble | Accuracy = 90% |
| [30] | Patil R and Sharvari C. T. | 2018 | A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes | NB, LR, RF, KNN, Gradient Boost, DT, Linear SVM, Neural Net | UCI Pima | LR | Accuracy = 79% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **[31]** | Kemal and Baha | 2018 | Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms | AdaBoost, Gradient Boosted Trees and RF ensemble | UCI Pima | AdaBoost | Accuracy = 73.88% |
| **[32]** | A. Mir and S. N. Dhage | 2018 | Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare | NB, SVM, RF and Simple CART algorithm | UCI Pima | SVM | Accuracy = 78.4% |
| **[33]** | Singh and Singh | 2019 | Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus | Stacking | Un disclosed | Ensemble | Accuracy = 83.8%, |
| **[34]** | Ramya*et al.* | 2019 | Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type2 Diabetes | SVM, RF, and Gradient Boosting | UCI Pima | Ensemble | Accuracy = 89% |
| **[35]** | Prema*et al.* | 2019 | Prediction of Diabetes using Ensemble Techniques | KNN, LR, DT, NB, Linear SVM, RBF SVM, Gaussian Process, Adabost RF, Voting - Ensemble | UCI Pima | Ensemble | Accuracy = 80% |
| **[36]** | El-Sappagh *et al.* | 2019 | A Comprehensive Medical Decision–Support Framework Based on a Heterogeneous Ensemble Classifier for Diabetes Prediction. | K-NN, NB, DT, SVM, fuzzy DT, ANN, and LR | Un disclosed | Ensemble | Accuracy = 90%, |
| **[37]** | Xu. and Wang | 2019 | A Risk Prediction Model for Type2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. | C4.5, NB, AdaBoost, RF | UCI Pima | Ensemble (AdaBoost) | Accuracy = 93.75% |
| **[38]** | Rubul and Anindiya | 2019 | R-Ensembler: A Greedy Rough set based Ensemble Attribute Selection Algorithm with KNN Imputation for Classification of Medical Data | NB, DT and RF | Un disclosed | Ensemble | Accuracy = 90.36 |
| **[39]** | Fitriyani*et al.* | 2019 | Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension | Isolation forest (iForest) | Un disclosed | Ensemble | Accuracy = 96.74%, |
| **[40]** | Sneha and Gangil | 2019 | Analysis of diabetes mellitus for early prediction using optimal features selection | DT, RF, NB, KNN and SVM | UCI Pima | DT and RF | Accuracy = 98.2% |
| **[41]** | Karun *et al.* | 2019 | Comparative Analysis of Prediction Algorithms for Diabetes. | LR, NB SVM, DT, KNN, NN AND RDF | UCI Pima | KNN | Accuracy = 75% |
| **[42]** | Alehegn, Joshi and Mulay | 2019 | Diabetes Analysis and Prediction using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach | RF, KNN, NB, J48 (DT) With Bagging | UCI Pima and 130_US hospital diabetes data sets | RF | Accuracy = 93.62% |
| **[43]** | Neha and Tigga | 2019 | Prediction of Type 2 Diabetes using Machine Learning Classification Methods | LR, KNN, SVM, NB, DT | Pima Indian Diabetes database and (online & offline) questionnaire | RF | Accuracy = 94.10% |
| **[44]** | Prema and Pushpalatha | 2019 | An Ensemble Model for the Prediction of Gestational Diabetes Mellitus (GDM). | KNNRandom-forestLogistic Regression | Pima Indian Diabetes database | RF | Accuracy = 93.8 |
| **[45]** | Vandana Rawat and Suryakant | 2019 | A Classification System for Diabetic Patients with Machine Learning Techniques | AdaBoost, LogicBoost, RobustBoost, NB and Bagging | Pima Indian Diabetes database | Ensemble | Accuracy = 81.77% |
| **[46]** | Sujit *et al.* | 2019 | Automatic Diabetes Prediction Using Tree Based Ensemble Learners | RF and Gradient Boosting classifiers | Pima Indian Diabetes database | Ensemble | Accuracy = 90% |
| **[47]** | Hasan *et al.* | 2020 | Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers | KNN, DT., RF, AdaBoost (AB), NB, and XGBoost (XB) and MLP | Pima Indian Diabetes | Ensemble | Accuracy = 99% |
| **[48]** | Ankit | 2020 | Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques | RF, SVM., DT, MP, and NB, Voting and Stacking Ensemble | UCI Pima | Ensemble | Accuracy =79.87% |
| **[49]** | Yang T,*et al* | 2020 | Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation | Linear discriminant analysis, SVM, RF and Ensemble | National Health and Nutrition Examination Survey from 2011-2016ds | Ensemble | Accuracy =73.0% |
| **[50]** | Naveen*et al.* | 2020 | Prediction Of Diabetes Using Machine Learning Classification Algorithms | SVM, DT, KNN, LR, RF | Pima Indians Diabetes Database | RF | Accuracy = 74.4% |
| **[51]** | Olivia and Tigeborn | 2020 | Detecting diabetes with Machine learning: A study of Naive Bayes and Decision Tree | NB and DT | Pima Indians Diabetes Dataset | NB | Accuracy = 80.0% |
| **[52]** | Jingyu Xue*et al.* | 2020 | Research on Diabetes Prediction Method Based on Machine Learning | SVM, NB and LightGBM | UCI | SVM | Accuracy = 96.54% |
| **[53]** | Parameswari and Rajathi | 2020 | Comparative Study of Machine Learning Approaches in Diabetes Prediction | RF, MP and J48 | UCI | RF | Accuracy = 97.5% |
| **[54]** | Badiuzzaman *et al.* | 2020 | Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh | KNN, DT , RF and NB | PIMA Indian | RF | Accuracy = 77.9% |
| **[55]** | Ogundele | 2020 | An Intelligent Diabetes Diagnostic Prediction System Using Ensemble Classifier | DT, NB, KNN and ensemble | Pima Indian Diabetes Dataset | Ensemble | Accuracy = 94.48% |
| **[56]** | Jyoti | 2020 | Diabetes Mellitus Prediction using Ensemble Machine Learning Techniques | RF, SVM, DT, MP, and NB | Pima Indian Diabetes Dataset | Ensemble | Accuracy = 98.5% |
| **[57]** | Mani*et al.* | 2020 | Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network | DT, LR and Ensemble model | Pima Indian Diabetes | Ensemble | Accuracy = 83.08% |
| **[58]** | Samah and Kamal | 2020 | Assessing Advanced Machine Learning Techniques for Predicting Hospital Readmission | KNN and ensemble-based learning | Undisclosed | Ensemble | Accuracy = 93.27% |
| **[59]** | Shawni and Bandyopadhyay | 2020 | Diabetes Prediction Using Ensemble Classifier | Multinomial NB , Perceptron , KNN, DT and Ensemble Method | Pima Indian women dataset | Ensemble | Accuracy = 79.87% |
| **[60]** | Mitushi and Sunita | 2020 | Diabetes Prediction using Machine Learning Techniques | SVM, KNN, RF, DT, LR and Gradient Boosting | Pima Indian Diabetes | RF | Accuracy = 77% |
| **[61]** | Faisal *et al.* | 2020 | Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation | SVM, NB, KNN and DT | Ulster Community and Hospitals Trust (UCHT) | DT | Accuracy = 73.5% |
| **[62]** | Fareeha *et al.* | 2020 | A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey | DT, RF, DNN and SVM | Pima Indian Diabetes | DNN | Accuracy = 98.35% |
| **[63]** | Preety *et al* | 2020 | Diabetes Prediction Method using the Ensemble Classification | SVM, NB and Ensemble | UCI | Ensemble | Accuracy = 92.9% |
| **[64]** | Sandhiya and LookmanSithic | 2020 | Design And Development Of Supervised Learning Algorithm For Diabetes Diagnosis | SVM, LR, RFand DT | PIMA database | SVM | Accuracy = 78.00% |
| **[65]** | Bhavya | 2020 | Diabetes Prediction using Machine Learning | KNN | Pima Indian Diabetes | KNN | Accuracy = 98% |

It is was observed that most of the paper reviewed made use of the Pima Indian diabetes database was acquired from UCI repository, US Hospital diabetes dataset and few undisclosed sources dataset.

### 3.1 Feature Selection

Having immaterial features in data can cause a decrease in the precision of the models and cause the model to learn based on unessential features. Data Normalization, Chi-square and information gain were examples of feature selection methods mostly used by the reviewed papers.

**a. Data Normalization:** This method is applied to data for pre-processing“ to execute the machine learning more proficiently, it is referred to as ‘minimum-maximum normalization’ which guards the connections in the original value” [67].

**b. Chi-Square ($\chi^2$):** To use chi-square ($\chi^2$) for feature selection, $\chi^2$ as shown in equation 1 must be calculated using each feature and the target. The preferred number of features with the best $\chi^2$ scores will be selected. The intuition is that if a feature is independent to the target, it is uninformative for classifying observations. *x* is the feature attribute values and the output *y* the class labels [68].

$$\chi^2 = \sum_{i=1}^{x}\sum_{j=1}^{y}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (1)$$

where: $O_i$ is the amount of observation of class *i*, $E_i$ is the number of expected observations in class *i* if there was no relationship between the feature and the target.

**c. Information Gain:** This is the difference between the prior uncertainty and expected posterior uncertainty. Information gain is maximal for equal probable classes, and uncertainty is minimal. Shannon entropy is broadly used for uncertainty measures [74].

**d. Training and Testing of Dataset:** At this stage, the dataset is rearranged “into 60-40%, 70-30% and 80-20% split of training and test” [74] sets separately. They are utilized for learning.

### 3.2 Model Construction

To setup a model, training set of dataset is use to build the model while testing set is use to validate the model. Classifications with higher accuracy e.g. Artificial Neural Network, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Naive Bayes, and Logistic Regression were considered in most paper reviewed.

**a. Support Vector Machine**

“Support Vector Machines (SVMs) are supervised learning approaches that examine data and identify patterns”[67].

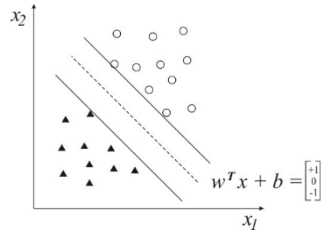


Figure 2. Maximum-Margin Hyperplane[69].

SVM algorithm forecasts the event of diabetes by plotting the disease predicting attributes in “multidimensional hyperplane and characterizes the classes ideally by making the margin between two data groups” [69].

The SVM targets fulfilling two prerequisites; the SVM will maximize the separation between the two choice limits as shown in figure 2. Mathematically, as expressing in equation 2 and 3, this means to expand the separation between the hyperplane at that point

$w^T x + b = -1$;(2)

the hyperplane expressed by

$$w^T x + b = 1 \quad (3)$$

Where$w^T = weight\ vector$ (distancebetween hyperplane and attribute). $b$ = hyperplane bias, x = data feature vector

**b. Decision Tree (DT)**

This is a predictive model also called classification or reduction tree. “The process is a mapping process from observation of an item extended to the target rate”[70].ID3 (Iterative Dichotomiser 3) is a good example.There are two mathematical tools needed to complete ID3 algorithms. Entropy (equation 4)and Information Gain expressed in equation 5.It is the difference between original information required and the new requirement

$H(a) = -\ p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$(4)

where, $a$ is training set, -$p_{(+)}$ / $p_{(-)}$ ... % of positive/negative example in “$a$”

$Gain\ (a, b) = H(a) - \sum_{i \in value\ (b)} \frac{|\Delta b_i|}{|\Delta b|} H(b_i)$(5)

$\Delta bi -\ Possible\ Value\ of\ b, \Delta b -\ Set\ example\ of\ (x)$, b*i* – Subset where Xb = I,

$i \in value -$ partition of the training set data

**c. K-Nearest Neighbour (KNN)**

K-Nearest Neighbour classifies an unlabeled instance (tuple) in the dataset by allocating them to the class of the most related labeled instances in the training dataset[71].KNN is “built on Euclidean separation between the training set and the testing set.

**d. Naïve Bayes (NB)**

"The Naïve Bayes is a classifier that is based on Bayes theorem coupled with probability-based classifier"[72]. The NB algorithm is fitting for characterizing high dimensional datasets.

**e.A Multi-Layer Perceptron (MLP):**This a deep Artificial Neural Network that creates a lot of outputs from a lot of sources (inputs).Every neuron *j* in the hidden layer supplements its input signals $x_i$ once it weights them with the qualities of the individual connections $w_{ji}$ from the input layer and decides its output $y_i$ as a function or capacity *f* of the sum, given in equation6 as:

$$y_i = f(\sum_{i=1}^{n} w_{ji} x_i) \qquad (6)$$

At this instant, *f* is a threshold function, for example, a sigmoid, or a hyperbolic tangent function. The output of neurons in the output layer is resolved in an indistinguishable style. Input data is provided to the input layer for processing, which creates an anticipated (predicted) output. To get the error value, the anticipated output will be deducted from the genuine output[73].

**f. Logistic Regression**

Logistic regression is a meta-level learning classifier that predicts the outcome of a categorical dependent variable from a set of predictor or independent variables [36]. The given data can be used to calculate the probability of a discrete outcome. [74]. The predictive attribute also called the independent variables, while the target attribute is also called response variable.

**g.Linear Regression(LR)**

In LR, a target is set to observe the data use for analysis. This is achieved by modifying the input and output variables. In view to qualify the relationship between the input variables*(x)* and output variable *(y)*, *x* will determine the *y*. This is the goal of ML. This is express as $y = a + bx$. Thus, the value of coefficients of *a* and *b* is to be calculated. Where *a* is the intercept and b is the slope of the line [74].

**3.3. Ensemble Method**

Ensemble Method is the process use to improve the accuracy of predictive analytics and data mining application. This helps to improve machine learning [75].It is also referred to as committee based learning or multiple classifier systems or classifier combination. Boosting, Bagging, Stacking and Voting the mostly used ensemble learning algorithm.

**a. Stacking**

Stacking algorithm is a type of ensemble learning system that merges extra classifier in a ranked or hierarchical architecture. The predictions of *level-0 classifiers* represent the attributes in a new training set (*level-1 data*), which keeps the original class labels [74]. The selection of the classifier used was based on the strength and weaknesses of the individual classifiers as revealed in the related works reviewed. After the training of the separate classifiers, cross-validation technique can be used for the training of the entire stacked architecture[33].

**b. Voting Ensemble**

In the majority voting ensemble, the last class label will be predicted as the class label that has been predicted most often by different classification models. Consider a dataset $D$ as shown in equation 8 instances $N$ number of instances and $C$ as the class label as stated [76].

$$D = \{(y_m, x_n),\ m = 1,\ 2\ and\ n = 1, \ldots, N\} \qquad (8)$$

where$y_m$ is the aim class; and $x_n$signifies feature vectors of the nth instance. Also, define a set of classifiers = $\{m_1,\ m_2,\ m_3\}$.Each instance $x \in$ D is allotted to have one of the C classes and ∈an element of the dataset. All classifier has its prediction for each instance. The concluding class assigned to each instance is the class predicted by the majority of classifiers (gaining the majority votes) [76]for that instance.

**c. Boosting**

Boosting, as an Ensemble algorithm. At the principal occasion, the main dataset is separated into different subsets. At that point, the ensembling procedure boosts their performance by combining the weak models using a cost function.

**d. Bagging -** Bootstrap Aggregation.

Bagging is established on bootstrapping and aggregation approaches. Both bootstrapping and aggregation methods have gainful properties[33]. Bootstrapping comprises of acquiring irregular examples with substitution of a similar size as the first set.

**3.4PERFORMANCE METRICS**

The efficiency of the ensemble model can be evaluated and validated using confusion matrix and other statistical methods such as MAE, RMSE etc.

**a. Confusion Matrix**

Confusing Matrix holds genuine information on the predicted classification system. As shown in Table 2, the abbreviations in the confusion matrix table can be defined as:"TP - the quantity of right forecasts that an example is certain or positive; FN - the quantity of wrong expectations that an occurrence is negative FP - the quantity of wrong expectations that a case is certain and TN - the quantity of right forecasts or predictions that an occasion is negative" [39].

Table 2. Evaluation Metrics

| | | Predicted | |
|---|---|---|---|
| | | P | N |
| Actual | P | TP | FN |
| | N | FP | TN |

While P-Positive, N-Negative, TP-True Positive, FN-False Negative, FP-False Positive and TN-True Negative.

**i. Accuracy:**It is the measurement of how well a system identifies target conditions of interest [39]. As shown in equation 8.

$$\text{Accuracy} = \frac{TP+FN}{TP+FN+FP+TN} \quad (8)$$

**ii. Precision:** This is the ratio of correct positive observations. This is also referred to as *Positive predicted value (PPV)* expressed in equation [30].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

**iii. Negative Predictive Value (NPV):**"This occurs when the test is negative and is considered as the probability of a classifier that the disease is absent"[77]. It can be computed in equation (10).

Negative Predictive Value (NPV)

$$= \frac{TN}{TN+FN} \quad (10)$$

**iv. Sensitivity:** This also referred to as Recall or True Positive rate. This can be defined as the ability of the test to adequately identify the patients with the illness. In [78]and as shown in equation 11, this was expressed as the ratio of correctly predicted positive events.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (11)$$

**v. Specificity:** This is the ability of the test to correctly identify the patients without the disease and the equation 12 expresses the specificity.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (12)$$

**vi. F1 Score:** It is interpreted as the weighted average (or harmonic mean) of the precision and recall. "An F1 score of 1 (one) is considered as best while 0 (zero) is worst; F-measures do not take the TNs into account"[77]. Therefore, F1 can be considered as expressed in equation (13).

$$F1\,Score = 2\left(\frac{Precision\ X\ Recall}{Precision+Recall}\right) \quad (13)$$

**vii. AUC:**Area Under the Receiver Operating Characteristic (ROC) curves link sensitivity versus specificity across a range of values of the ability to predict a unique different outcome. It is also the technique of conceptualizing, organizing and selecting classifiers of the basis of their performance [79]. AU-ROC is an astounding measure for performance or execution assessment since it analyzes the exhibition over whole scope of class distributions and error value. This is expressed with equation 14.

$$AU-ROC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (14)$$

**b. Statistical Methods**

**i. Mean Absolute Error (MAE):**This method as expressed in equation 15 did not consider the "direction while measuring the average magnitude of the errors in a set of predictions. $(y_i - x_i)$is the arithmetic average of the absolute errors, where $y_i$ is prediction and $x_i$,is the true value" [79].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) \quad (15)$$

**ii. Root mean square error (RMSE):**This method measures the average magnitude of the error. Equation 16expresses its definition as the square root of the average of squared differences between prediction$y_i$ and actual observation$x_i$. [79].

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (16)$$

## 4. EXPERIMENTAL RESULT

In this review, table 1 spelled out the number of selected articles in respect to Diabetes using different classifiers and ensemble methods. Figure 3 shows the average accuracy of Ensemble methods and average accuracy of other individual classifiers. The result shows that ensemble method have 87.09% which higher when compare to the average accuracy of other classifiers.

Table 3: The Average Accuracy performance of classifier methods mostly used from 2016 to 2020.

| Method | RF | LR | DT | SVM | NB | KNN | En. |
|---|---|---|---|---|---|---|---|
| Averages (Accuracy) | 86.62 | 80.70 | 85.64 | 84.54 | 81.17 | 81.29 | 87.09 |

- *En – Ensemble*

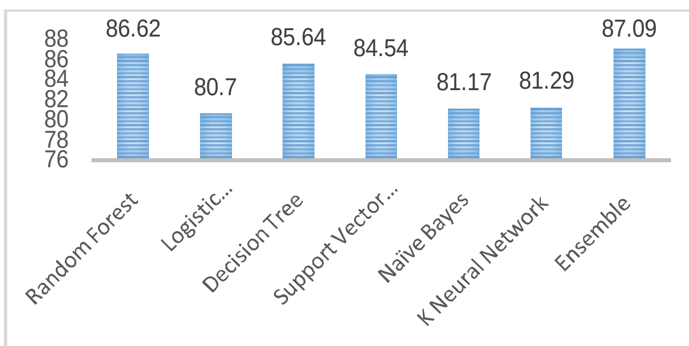


Figure 3. The graphical representation of Average Accuracy performance

## 5. FUTURE SCOPES AND CHALLENGES

It was observed that Pima Indians Diabetes Database was used by majority of the related 1,137 journals reviewed. In a way to meet up with future expectation and further investigation, more data should be made available by medical agencies to improve on prediction of DM. Training with insufficient data of can affect the result of some automated optimization technique like deep learning algorithms i.e. Deep Neutral Network, Convolutional Neutral Network and Recurrent Neutral Network.

## 6. CONCLUSION

Diabetes mellitus diseases cause severe harm to human heart, blood, eyes, kidneys and nerves. This resulted in death among the people. Evidence of several research activities targeted at developing artificial intelligence-powered tools for prediction and prevention of complications associated with diabetes was observed. Looking through the various contributions and accomplishments of current studies in examining digital clinical data, evaluating clinical data is still an essential and challenging task. The precise and productive chance expectation has continuously been a basic matter luring numerous researchers' interest.Different methods of Machine Learning that mostly used in most paper reviewed were considered and briefly explained. Using classifiers such as Naïve Byes, Decision Tree, SVM, KNN, ANN and Deep Forest LR and Linear Regression. The ensemble algorithm include Boosting, Bagging, Stacking and Voting. They were applied to based classifier or classifiers. Based on the result shown in figure 7, an ensemble algorithm provides more accuracy than a single algorithm. Most of the researchers' work tends towards using a combination of individual techniques to aid performance and accuracy. The performance evaluation criteria for a model can be carried out statistically using MAE, RMSE and Confusing Matrix i.e. the sensitivity, specificity, accuracy and precision. With higher accuracy recorded in the recent paper reviewed, machine learning algorithm can offer improved evidence to medical personnel at the point of patient care. In the future work, more recent articles can be consider to improve performance metrics.

## REFERENCES

[1] World Health Organisation ***Global Health Observatory***. Geneva, WHO Press. World Health Organisation Diabetes country profile (2016). Available on: www.int/ diabetes/country-profiles/en

[2] World Health Organisation. 2018. **Non-Communicable Diseases (NCDs) country profiles**, Available on https://www.who.int/nmh/countries/nga_en.pdf (Accessed December 2018).

[3] Bosun-Arije, S., Ling J., Graham Y., Hayes, C. **A systematic review of factors influencing Type 2 Diabetes Mellitus management in Nigerian public hospitals**. International Journal of Africa Nursing Sciences. 11. 100151. 10.1016/j.ijans.2019.100151.

[4] Binh, P., Nguyena, I., Hung, N.P., Hop, T., Nhung, N., Quang, H., Nguyenb, Trang, T.T., Do d., Cao, T.T., Colin, R.S.,**Predicting The Onset Of Type 2 Diabetes Using Wide And Deep Learning With Electronic Health Records.**Computer Methods and Programs in Biomedicine Duke University of Medical Centre Publication.

[5] Ribeiro, C, de-AlencarMota, C.S., Voltarelli, F.A., de-Araújo, M.B., Botezelli, J.D.**Effects of Moderate Intensity Physical Training in Neonatal Alloxan-Administered Rats**. J Diabetes Metab. 2010; 1:107.

[6] Dattatreya, Adapa, Sarangi, T.K. **A Review on Diabetes Mellitus: Complications, Management and Treatment Modalities.** Research and Reviews: Journal of Medical and Health Sciences. RRJMHS| Volume 4 | Issue 3 | May-June, 2015. p-ISSN: 2322-0104.

[7] Nair, U.l., Syed, H.**Prediction and Management of Diabetes using Machine Learning: A Review**. International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177 Volume 7 Issue V, May 2019.

[8] Lemos Costa, T.M.R, Detsch, J.M., Pimazoni-Netto, A., de-Almeida, A.C.R., Sztal-Mazer, S. **Glycemic Variability and Mean Weekly Glucose in the Evaluation and Treatment of Blood Glucose in Gestational Diabetes Mellitus; Evidence for Lower Neonatal Complications**. J Diabetes Metab. 2011; 2:137.

[9] Sneha, N., Gangil, T.**Analysis of Diabetes Mellitus for Early Prediction using Optimal Features Selection. Journal of big data**, 6(1).doi:10.1186/s40537-019-0175-6.

[10] Firdaus, H., Hassan, S.I., Kaur, H.**A Comparative Survey of Machine Learning and Meta-Heuristic Optimization Algorithms For Sustainable And Smart Healthcare.** African Journal of I. Computing &ict, p.1.

[11] Park, S., Choi, D., Kim, M., Cha, W., Kim, C., Moon I. C., **Identifying Prescription Patterns with a Topic Model of Diseases And Medications.** Journal of Biomed. Informat. 75 (2017) 35-47.

[12] Contreras, I., Vehi, J. **Artificial Intelligence for Diabetes Management and Decision Support: Literature Review.** J Med Internet Res 2018;20(5):e10775.

[13] Kaur, H., Lechman, E., Marszk, A.**Catalyzing Development through ICT Adoption: The Developing World Experience.** Springer Publishers, Switzerland. 2017

[14] Chaki, J., Thillai Ganesh, S., Cidham, S.K, Ananda Theertan, S., 2020. **Machine Learning and Artificial Intelligence based Diabetes Mellitus Detection and Self-Management: A Systematic Review.** Journal of King Saud University - Computer and Information Science. doi: https://doi.org/10.1016/j.jksuci. 2020.06.013

[15] Casanova, R., Saldana, S., Simpson, S.L., Lacy, M.E., Subauste, A.R., Blackshear, C., Bertoni, A.G., **Prediction of Incident Diabetes in the Jackson Heart Study Using High-Dimensional Machine Learning.** PLOS ONE, 11(10). e0163942. doi:10.1371/ journal.pone.0163942, 2016

[16] Anderson, A.E., Kerr, W.T., Thames, A., Li, T., Xiao, J., Cohen, M. S.**Electronic Health Record Phenotyping**

**Improves Detection and Screening of Type 2 Diabetes in the General United States Population: A Cross-Sectional, Unselected, Retrospective Study**. J Biomed Inform 2016;60:162-8.

[17] Lee, B.J., Kim, J.Y.**Identification of Type2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides Based on Machine Learning.** IEEE J Biomed Health Inform 2016;20(1):39-46.

[18] Jelinek, H.F., Stranieri, A., Yatsko, A., Venkatraman, S., **Data Analytics Identify Glycated Haemoglobin Co-Markers for Type 2 Diabetes Mellitus Diagnosis**. ComputBiol Med 2016;75:90-7.

[19] Yu, C.S., Liu, C.S., Chen, R.S., and Lin, C.W.**Artificial Neural Networks For Estimating Glomerular Filtration RateBy Urinary Dipstick for Type 2 Diabetic Patients**. Biomed Eng (Singapore)

[20] Luo, G.**Automatically Explaining Machine Learning Prediction Results: A Demonstration on Type 2 Diabetes Risk Prediction.** Health InfSciSyst 2016;4:2.

[21] Cai, J., Li, C., Liu, Z., Du, J., Ye, J., Gu, Q. **Predicting Dpp-Iv Inhibitors with Machine Learning Approaches.** J Comput Aided Mol Des 2017;31(4):393-402.

[22] Chen, W., Chen, S., Zhang, H., Wu, T.**A Hybrid Prediction Model for Type 2 Diabetes Using K-Means and Decision Tree.** 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS).

[23] Kagawa, R., Kawazoe, Y., Ida, Y., Shinohara, E., Tanaka, K., Imai, T.**Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge And Machine Learning Approach.** J Diabetes SciTechnol 2017;11(4):791-9.

[24] Zheng T., Xie W., Xu L., Xiaoying H., Ya Zhang, Mingrong Y., Gong Y., You C. **A Machine Learning-Based Framework to Identify Type 2 Diabetes Through Electronic Health Records.**International Journal of Medical Informatics. 2017 Jan;97:120-127. doi: 10.1016/j.ijmedinf.2016.09.014. Epub 2016 Oct 1.

[25] Sayadi, M., Zibaeenezhad, M., TaghiAyatollahi, S.M., **Sim-ple prediction of type 2 diabetes mellitus via decision tree modeling.** IntCardiovasc Res J 2017;11(2):71-6

26] Uswa, A.Z., Naeem, K.**Predicting Diabetes in Medical Datasets Using Machine Learning Techniques.** International Journal of Scientific & Engineering Research. Volume 8, Issue 5, May-2017. ISSN 2229-5518.

[27] Priya, B.P., Parth, P.S., Himanshu, P.**Analyze Data Mining Algorithms For Prediction Of Diabetes.** International Journal of Engineering Development and Research IJEDR | Volume5, Issue 3| ISSN: 2321-9939.

[28] Deepti, S., Dilip, Singh. S., **Prediction Of Diabetes Using Classification Algorithms**. International Conference on Computational Intelligence And Data Science (ICCIDS 2018).

[29] Das, S., Kumar, A., Roy, P.**Automatic Diabetes Prediction Using Tree Based Ensemble Learners.** International Conference on Computational Intelligence &IoT(ICCIIoT) 2018.

[30] Patil, R., Tamane, Sharvari. **A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes**. International Journal of Electrical and Computer Engineering. Vol. 8, No. 5, October 2018, pp. 3966~3975 ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp3966-3975.

[31] Kemal, A., Baha. Şen.**Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms.**International Journal of Modern Education and Computer Science (IJMECS), Vol.10, No.6, pp. 10-16, 2018.DOI: 10.5815/ijmecs.2018.06.02.

[32] Mir, A., Dhage, S. N. **Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare.** 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697439.

[33] Singh, N., Singh, P. **Stacking-Based Multi-Objective Evolutionary Ensemble Framework For Prediction Of Diabetes Mellitus**. Biocybern Biomed Eng (2019).

[34] Ramya, A., Ivan G., Akula, R. **Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes.**Institute of Electrical and Electronics Engineers 978-1-7281-0137-8/19 ©2019 IEEE.

[35] Prema, N.S.,Varshith V., Yogeswar J. **Prediction of Diabetes using Ensemble Techniques.**International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S4, April 2019.

[36] El-Sappagh, S., Elmogy, M., Ali, F., Abuhmed, T., Islam, S.M.R., Kwak, K. **A Comprehensive Medical Decision-Support Framework Based on a Heterogeneous Ensemble Classifier for Diabetes Prediction. Electronics.** 8. 635. 10.3390/ electronics8060635.

[37] Xu, Z., Wang, Z. **A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier.**pg 278-283. 10.1109/ ICACI. 2019.8778622.

[38] Rubul, K.B., Anindya, H. **Rough set based Ensemble Attribute Selection Algorithm with kNN Imputation for Classification of Medical Data.** Computer Methods and Programs in Biomedicine. doi.org/10.1016/j.cmpb. 2019. 105122.

[39] Fitriyani, N.L., Syafrudin, M., Ganjar, Rhee J. **Development of DPM Based on Ensemble Learning Approach for Diabetes and Hypertension.** IEEE

Access, Multidisciplinary Rapid Review Open Access Journal, Digital Object Identifier 10.1109/ACCESS.2019. 2945129.

[40] Sneha, N., Gangil, T. **Analysis of Diabetes Mellitus For Early Prediction Using Ptimal Features Selection.** Journal of Big Data, 6(1).doi:10.1186/s40537-019-0175-6.

[41] Shweta, K., Aishwarya Raj, GirijaAttigeri, **Comparative Analysis of Prediction Algorithms for Diabetes.**Advances in Computer Communication and Computational Sciences, 2019, Volume 759. ISBN : 978-981-13-0340-1.

[42] Minyechil, A., Rahul Raghvendra Joshi, PreetiMulay, **Diabetes Analysis and Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach.** International Journal of Scientific & Technology Research Volume 8, Issue 09, September 2019 ISSN 2277-8616.

[43] Neha, P.T., Shruti, G., **Prediction of Type 2 Diabetes using Machine Learning Classification Methods.** International Conference on Computational Intelligence and Data Science. (ICCIDS 2019) Procedia Computer Science 167 (2020) 706–716.

[44] Prema, N.S., Pushpalatha, M.P. **An Ensemble Model for the Prediction of Gestational Diabetes Mellitus (GDM).** Indian Journal of Public Health Research & Development, September 2019, Vol. 10, No. 9. DOI Number: 10.5958/0976-5506.2019.02399.4

[45] Vandana, R., Suryakant.**A Classification System for Diabetic Patients with Machine Learning Techniques.** International Journal of Mathematical, Engineering and Management Sciences. Vol. 4, No. 3, 729–744, 2019.

[46] Kumar, D., Sujit, K. M., Arnab and Roy P.**Automatic Diabetes Prediction Using Tree Based Ensemble Learners.** International Journal of Computational Intelligence &IoT, Vol. 2, No. 2, 2019.

[47] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., Hasan M., 2020. **Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers**. IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[48] Ankit N. S.**Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques.**SSRN Electronic Journal. DOI: 10.2139/ssrn.3642877.

[49] Yang, T, Zhang, L., Yi, L., Feng, H., Li, S., Chen, H., Zhu, J., Zhao, J., Zeng, Y., Liu, H.**Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation.** JMIR Med Inform 2020;8(6):e15431.doi: 10.2196/15431PMID: 32554386P MCID: 7333074.

[50] Naveen, K. G., Rajesh, V., VamsiAkki Reddy, A., Sumedh, K., Rajesh, S. T.**Prediction of Diabetes Using Machine Learning Classification Algorithms.** International Journal of Scientific and Technology Research Volume 9, Issue 01, January 2020. ISSN 2277-8616.

[51] Olivia, S., Frida W., **Detecting diabetes with Machine learning: A study of Naive Bayes and Decision Tree.** Independent thesis Basic level (degree of Bachelor). 2020

[52] Jingyu X., Fanchao M., Fengying M. **Research on Diabetes Prediction Method Based on Machine Learning.** Journal of Physics: Conference Series, Volume 1684, The 2020 International Seminar on Artificial Intelligence, Networking and Information Technology 18-20 September 2020, Shanghai, China.

[53] Parameswari P., Rajathi N. **Comparative Study of Machine Learning Approaches in Diabetes Prediction.**Biosc. Biotech. Res. Comm. Special Issue Vol 13 No 11 (2020) Pp-42-46.

[54] Pranto, B., Mehnaz, S.M., Mahid, E.B., Sadman, I.M., Rahman, A., Momen,**Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh**. Information **2020**, 11, 374. https://doi.org/ 10.3390/info11080374.

[55] Ogundele, I.O., Sokunbi, M.A., Akinsola, A.F,,Akinade, A.O., Adebayo A.A., **An Intelligent Diabetes Diagnostic Prediction System Using Ensemble Classifier.** International Journal of Scientific & Engineering Research. Volume 11, Issue 6, June-2020 318 ISSN 2229-5518.

[56] Jyoti, P.A. **Diabetes Mellitus Prediction using Ensemble Machine Learning Techniques.** International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-2, July 2020.

[57] Mani, A., Anita B., Touraj B.**Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network.** International Journal of Advanced Research in Computer and Communication Engineering. Vol. 9, Issue 7, July 2020. ISSN (Online) 2278-1021 ISSN (Print) 2319-5940.

[58] Samah, A., Kamal J.**Assessing Advanced Machine Learning Techniques for Predicting Hospital Readmission.** (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 2, 2020.

[59] Shawni. D., Bandyopadhyay, Kumar, S. **Diabetes Prediction Using Ensemble Classifier.** International Journal of Medical and Health Sciences. ISSN:2277-4505, April 2020,Vol-9;Issue-2.

[60] Mitushi, S., Sunita V. **Diabetes Prediction using Machine Learning Techniques.** International Journal of Engineering Research & Technology (IJERT). ISSN: 2278-0181 IJERTV9IS090496. Vol. 9 Issue 09, September-2020.

[61] Faisal, F., Asaduzzaman1, S., Minhaz, H., Hasan, F., Iqbal, H., Sarker.**Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation.** EAI Endorsed Transactions on Pervasive Health and Technology. 08 2019 - 05 2020 | Volume 5 | Issue 20 |e7.

[62] Fareeha, A.A, Qurat-Ul-Ain a, Muhammad, Y.E., Amir, M.**A Comparative Analysis on Diagnosis of Diabetes Mellitus Using Different Approaches – A Survey.** Informatics in Medicine Unlocked. Elsevier, ISSN 2352-9148.

[63]Preety, E., Parminder, S., Naveen D. **Diabetes Prediction Method using the Ensemble Classification.**International Journal of Review in Electronics & Communication Engineering (IJRECE). Vol. 8 issue 1, ISSN: 2393-9028 (print) | ISSN: 2348-2281 (online).

[64] Sandhiya, K., LookmanSithic, H.**Design and Development of Supervised Learning Algorithm for Diabetes Diagnosis.** Journal of Information and Computational Science. Volume 10 Issue 1 - 2020 ISSN: 1548-7741.

[65] Bhavya, M.R., Sanjay, H.C., Suraj, S.K., Savant, A. S. R., Sanjay, M. **Diabetes Prediction using Machine Learning.** International Journal of Advanced Research in Computer and Communication Engineering. Vol. 9, Issue 7, July 2020. ISSN (Online) 2278-1021ISSN (Print) 2319-5940.

[66] Han, J., Kamber, M., Pei, J. ***Data Mining: Concepts and Techniques***. 3rd ed., Waltham, MA, USA.2012.

[67] Awad M., Khanna, R. **Efficient Learning Machines Theories, Concepts, and Applications for Engineers and System Designers.** Published by Apress Open. 2015.

[68] Liu H, Li J, Wong L.,**A Comparative Study on Feature Selection And Classification Methods Using Gene Expression Profiles and Proteomic Patterns.** Genome inform. 13:51-60. PMID: 14571374.

[69] Santhanam, T., Padmavathi, M.S.**Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis.** Procedia Computer Science, vol. 47, pp. 76-83, 2015.

[70] Chandan, K., Nanhay, S., Jaspreet, Singh.**Prediction of Diabetes using Data Mining Algorithm.** International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 7 Issue II, Feb 2019- Available at www.ijraset.com ©IJRAS 773 ET.

[71] Ramesh, D., Singh, K.Y. **Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach.** Health and Technology, IUPES Mand Springer-Verlag GmbH Germany, part of Springer Nature 2019 https://doi.org/10.1007/s12553-019-00299-3

[72] Ramzan, M. **Comparing and Evaluating the Performance of Weka Classifiers on Critical Diseases.** In 2016 1st India International Conference on Information Processing (IICIP) (pp. 1- 4). IEEE.

[73] Atkov, O.Y.D., Pittman, S., Zhou, S. **Coronary Heart Disease Diagnosis by Artificial Neural Networks Including Genetic Polymorphisms and Clinical Parameters.** Japanese College of Cardiology Journal of Cardiology, 59(2), p. 190–194.

[74] Ayeni, B.O.**Development of Ensemble Predictive Models for Coronary Heart Disease (CHD),** A Dissertation in the Department of Computer Science, School of Computing, Federal University of Technology, Akure, Nigeria. 2019.

[75] Bhavana, N., Meghana, S., C.A Pradeep, K.R. **A Review of Ensemble Machine Learning Approach in Prediction of Diabetes Diseases.** International Journal on Future Revolution in Computer Science and Communication Engineering ISSN: 2454-4248. 2018.

[76] Ali, S., Majid, A., Can–Evo–Ens.**Classifier Stacking Based Evolutionary Ensemble System for Prediction of Human Breast Cancer using Amino Acid Sequences.** J Biomed Inform 2015;54:256–69.

[77] Khalid, R. **Improving The Prediction Accuracy of Heart Disease With Ensemble Learning And Majority Voting Rule.** U-Healthcare Monitoring Systems. Elsevier Inc. 2019.

[78] Kinge, D., Gaikward, S.K. **Survey on Data Mining Techniques for Disease Prediction.** International-Research Journal of Engineering and Technology (IRJET), Vol. 05 issue 01, Jan. 2018 p-ISSN: 2395-0056.

[79] Rahman, R.M., Afoz, F.**Comparison of Various Classification Techniques with different Data Mining Tools for Diabetes Diagnosis.** Journal of Software Engineering and Application. Vol 6. Pp. 85-97, 2013.