Survey paper on machine learning algorithms for diabetes prediction

Radhika Sreedharan[1]

Assistant Professor, Department of Computer Science and Engineering, Presidency University,
Bangalore, India

[1]radhika.sreedharan@presidencyuniversity.in

Parkavi B[2]

Assistant Professor, Department of Computer Science and Engineering, CMR University,
Bangalore, India

[2]kavitha.oec@gmail.com

***Abstract*— A long-standing ill health which straight away attacks the pancreas, and the body doesn't have the ability to produce insulin is diabetes. Insulin is principally important to perpetuate the level of blood glucose. Diabetes can be caused to patients due to a lot of characteristics like uncontrolled weight of body, lack of physical exercise, high blood pressure, and unusual cholesterol level. A lot of complicatedness can be caused, however one of the most common ones is increase in urination. Diabetes can harm the skin, nerves, and eyes. If left untreated, it can also result in kidney failure and the eye condition diabetic retinopathy.**

537 million people worldwide had diabetes in 2021, according to data from the International Diabetes Federation. Based on 2019 statistics, 7.10 million persons in Bangladesh had contracted this disease.

The majority of the population is affected by diabetes, which is the main disease increasing the ratios and leading to renal failure, amputations, blindness, lower limb amputations, stroke, and heart issues. Numerous factors, including a lifestyle devoid of regular exercise, eating unhealthy foods, being overweight, heredity, and more, could be the cause. Foods are converted into glucose by the human body. A higher blood glucose level is used to categorize a group of diseases known as diabetes. Individuals with diabetes, the pancreas is unable to produce insulin. In this case, machine learning techniques are crucial for forecasting this illness. As a supervised machine learning technique, classification is the best-preferred methodology for the labelled data categorization. The medical field uses a variety of machine learning approaches to anticipate and identify illnesses. The main aim of this paper is to compare various algorithms of machine learning for diabetes prediction.

***Keywords*— Machine learning, Diabetes, prediction, SVM**

I. Introduction

Diabetes, also known as diabetes mellitus (DM) or just diabetes, is a group of metabolic disorders mostly brought on by abnormal insulin production. When cells and/or the pancreas are unable to create enough insulin, blood sugar levels rise and damage several organs, including the kidneys, nerves, and eyes. Diabetes is sometimes referred to as the "silent killer" for this reason. Type I, type II, and gestational diabetes are the three subtypes of diabetes. In type I diabetes, the pancreas secretes either very little or none at all. The pancreatic cells are attacked by type 1 diabetes, which results in their shutdown. Between 5% and 10% of people have type I diabetes, which can develop at any age, including infancy and adolescence. More than 90% of cases of diabetes are type II diabetes. More than 90% of diabetes cases worldwide are type II diabetes. It appears when the body doesn't produce enough insulin. Type II diabetes is a disease that can affect both adults and kids. A third form of diabetes that is comparable to type II diabetes in that it is brought on by an imbalance between insulin secretion and responsiveness is called gestational diabetes mellitus (GDM). The long-term development of this medical condition is caused by hypertension and elevated blood pressure. About 2–10% of pregnant women have gestational diabetes, which can worsen or go away after birth. Diabetes may indicate the start of further illnesses. Researchers from all over the world are putting in endless effort to combat the condition by developing efficient techniques for prediction and detection as well as workable treatments. In this case, machine learning techniques are crucial for forecasting this illness. As a supervised machine learning technique, classification is the best-preferred methodology for the labelled data categorization. The medical field uses a variety of machine learning approaches to anticipate and identify illnesses.

Diabetes is one such condition for which machine learning methods are used to determine the best course of action. Because machine learning approaches may yield consistent, reliable, and accurate results, they are used to address practical difficulties in almost every aspect of life. Given its deleterious impact on essential organs, diabetes is frequently considered a systemic disease with wide-ranging implications. People who are diagnosed with diabetes are more vulnerable to a number of issues, such as miscarriage, visual impairment, myocardial infarction, renal failure, and other long-term, potentially fatal illnesses. Thus, in order to stop or postpone the beginning of these consequences, it is imperative that diabetes mellitus (DM) be diagnosed as soon as possible. By analysing various health indicators like plasma glucose concentration, serum insulin resistance, and blood pressure, machine learning algorithms (like support vector machine (SVM), k-nearest neighbors (k-NN), random forest, and artificial neural network (ANN)) can aid in the early diagnosis and accurate prediction of diabetes mellitus. Accurate diagnosis and timely prognosis of diabetes mellitus are critical for enabling successful treatments and the best possible care for the condition. Making use of machine learning algorithms' capacity to handle enormous amounts of data allows for the identification of complex patterns that may be invisible to human specialists. But more investigation is required, as well as better techniques, to improve the precision of diagnosis and individualized treatment plans

II. METHODOLOGY

By serving as a resource, artificial intelligence and machine learning approaches can assist them in learning more about this illness and lessen their workload. Numerous studies have been conducted to automatically predict diabetes through the use of ensemble and machine learning approaches. The open-source Pima Indian dataset was used in the majority of these projects. In the lines that follow, a few of these articles on automated diabetes prediction using the Pima Indian dataset are briefly reviewed. For example, Kumar et al. created a system that can rapidly and correctly diagnose diabetes using the random forest method. The UCI learning repository provided the dataset that was used in this study. Initially, the authors employed standard preprocessing methods for data, such as reduction, integration, and cleaning. Using the random forest approach, the accuracy level was 90%, which is significantly greater than previous algorithms. Using the Pima Indian Diabetes Dataset, Mohan and Jain employed the SVM algorithm in a recent study to assess and forecast diabetes. Four different kernel types—linear, polynomial, RBF, and sigmoid—were employed in this study to predict diabetes using a machine learning platform. The writers' accuracies, which ranged from 0.69 to 0.82, varied depending on the kernel. With a radial basis kernel function, the SVM method produced the maximum accuracy, 0.82. To identify diabetes, Goyal and his colleagues developed a smart home health monitoring system. For their investigation, the writers also used the Pima Indian dataset. They utilized conditional decision making for predicting status of blood pressure and KNN, SVM, and decision trees for predicting diabetes. Performance of SVM was better than the other models, with an accuracy rate of 75%, compared to other classification methods. Using the Pima Indian dataset and various ensemble method-based machine learning methods, Hassan et al. tried to predict diabetes. The authors' accuracy metric was the area under the ROC curve, or AUC. In the end, the suggested ensemble classifier achieved an AUC of 0.95. Jackins et al. suggested a multi-disease prediction method, incorporating diabetes utilizing machine learning techniques and the Pima Indian dataset. The authors claim that at accuracy increments of 0.43%, the Naive Bayes method outperformed the random forest technique. Mounika et al. used machine learning approaches to predict the probability of diabetes. Several machine learning frameworks and the publicly available Pima Indian dataset were used in this study. An ensemble strategy based on soft voting classifiers was attempted to be applied for diabetes prediction by Kumari et al. With an F1 score of 0.716 and overall greatest accuracy of 0.791, the suggested soft voting classifier was victorious. Using the deep belief network model, Prabhu and Selvabharathi were able to predict diabetes using the Pima Indian diabetes dataset, which is available for public use. The model was built by the authors in three stages: first, they pre-processed the data making use of min–max normalization, then they built the network model, and finally, they used NN-FF classification to fine-tune the test dataset in order to eliminate any bias. Lastly, MATLAB was used by the authors to complete all of the model's implementation and simulation. The authors found that, when compared to the other classification techniques, the F1 score of 0.808 represented the best performance statistic.

A diabetes prediction system was developed by Orabiet al. Many algorithms related to genetic programming are presented by Pradhan and Bamnote, and numerous tests are run on this dataset. In order to anticipate diseases, algorithms of machine learning have been employed in numerous study studies. In order to do further research on the diagnosis of diabetes mellitus, experts have looked at a variety of datasets, algorithms, and methods. The main literary genres that are covered below are listed below. To obtain an accuracy of 74.80% in the paper, SVM with Radial Basis Function (RBF) kernel was utilized, and entries with mean were substituted for zero values. The authors recovered three of the eight traits. After that, the author classified the data using linear discriminant analysis (LDA), which had a 74.30% accuracy rate. An alternative approach involved eliminating the items that had zero values, resulting in 460 entries remaining out of a total of 768. After 200 submissions were used for training and the remaining 460 for testing, the accuracy of these 460 entries was 75.5%. The diabetic dataset was classified by the authors using SVM in a Feed Forward Neural Network with an accuracy of 75.65%. Two features were chosen out of eight using Linear Discriminant Analysis (LDA). When identifying the diabetic dataset using Automatic Identification System approaches, the ANN strategy yielded a 76% accuracy rate. The authors employed the correlation-based feature selection strategy to choose two out of the eight available features, and then utilized the naïve Bayes and decision tree method to categorize diabetes. They reported an accuracy of 74.79% when they utilized the average to fill in the missing numbers. The output accuracy was 81.19% when Multi-Layer Perceptron (MLP) and Bayes Net classifiers were employed. Using the J48 classification technique, an accuracy of 76.58% was obtained. Utilized the Random Forest, J48, NB, and LR classification algorithms, producing results with an accuracy of 80.43%. It was demonstrated that Gaussian Process Classifier (GPC), which uses the radial basis kernel to achieve the greatest accuracy of 82%, outperformed Naive Bayes in the classification of diabetes patients. first presented in a Hierarchical Multi-Level Classifier with Multi-Objective

Voting Technique (HM-Bag) for Classification and their approach was contrasted with classifiers that use, support vector machines, artificial neural networks, logistic regression, quadratic discriminate analysis, Naïve Bayes, and logistic regression. They found that, at 77.21%, HM-Bag offered the best degree of accuracy. A few examples are SVM, Naive Bayes, Random Forest, AdaBoost, etc. Example Using logic, noteworthy approaches that increase the precision of diabetes prediction are identified. The disadvantages of nearly every solution—including Decision Trees, Spiral Premise Works, Genuine Valued Neural Networks, and Complicated Valued Neural Networks—are also listed. Using the PIMA, NB, LR, and GB—as well as other machine learning approaches and algorithms—proposed a model for future diabetes risk levels. To ascertain how closely the attributes are related, the Boruta technique has been applied. With an accuracy of 86%, GB was the most accurate classifier of all. It was decided to use a Probabilistic Neural Network (PNN) to forecast diabetes diseases. The algorithm was applied to the "PIMA Indian dataset." The author did not employ pre-processing. The sentence goes on, "The dataset is separated into 90% for the training set and 10% for the standard setting." The proposed technique yielded accuracy rates of 81.49% and 89.56%, respectively, for testing and taxing data. Estimation of ambiguity throughout the prediction process was streamlined by utilizing a neural network to anticipate outcomes based on blood glucose levels. The dataset in use is the Type 1 diabetes dataset, which accounts for blood sugar levels. The surveillance-error-grid (SEG) approach and the root-mean square error (RMSE) measure are evaluated using this methodology. Table 1 describes how the related work, which is shown below, used eleven MLT to build a model to predict the onset of diabetes disease. Using the UCI dataset, the support vector machine yielded the best results of all the classifiers considered, with an accuracy rate of 83.49%. The approach reduced prediction error rate and increased accuracy for other indicators. Several metrics are looked at in order to validate the proposed framework. The authors offered a system for predicting diabetes disease using machine learning-based techniques and suggested employing a hybrid technique to build on earlier studies for a more accurate prediction. The approach reduced prediction error rate and increased accuracy for other indicators. Several metrics are looked at in order to validate the proposed framework. To build on earlier studies and achieve more precise prediction, the authors suggested use a hybrid approach. A system that uses machine learning techniques to anticipate the onset of diabetes was proposed. The author compared different machine learning methods to create a model for diabetic illness prediction. Seven machine learning classifiers were subjected to performance and comparative analyses. With an accuracy rate of 88.4%, the hybrid random forest with linear model produced the best results of all the classifiers. Without using any techniques for data preparation, the output of the proposed model was enhanced. For future possibilities in this research, the authors suggested utilizing large datasets and a range of machine learning algorithms. Prior to creating machine learning algorithms, most earlier studies did not fully utilize data pre-processing. The end outcome was subpar outputs. examined the application of a range of machine learning models, including ensemble models that were based on SVM rule extraction and traditional models like support vector machines, random forests, and decision trees, for the purpose of predicting diabetes diagnoses. When this technique was coupled with a random forest classifier, the precision scores were found to be greater than those of the base random forest and support vector models (89.5 % versus 81.2 % and 88.4%, respectively). Conversely, they found that although the recall scores of the ensemble classifier (44.3%) were greater than those of the base support vector models (40.0%), they were lower than those of the base random forest model (49.0%). The ensemble's usage of random forest rules and support vector models—which were simpler than those in the original random forest model—was the cause of this. In the end, they concluded that both base models were still inferior to the support vector + random forest ensemble. A framework for disease prediction in healthcare has been built using machine learning techniques.

TABLE I
A REVIEW OF EXISTING RESEARCH ON DIABETES DETECTION

Algorithms	Data Set	Reported Accuracy
DT, LR, SVM	PIMA data set	79.86%in SVM
DT, Gradient Boost	Clinical dataset	82.6 % in Gradient Boost
DT, LR, SVM	PIDD set	89.7 % in DT
KNN, SV, DT, NB	PIMA data set	81.7 % in NB
RF, DT, NB	PIMA data set	83.5 % in RF
Gradient Boost, NB, SVM	PIMA data set	84.9 % in SVM
LR, SVM, KNN, DT	Clinical dataset	89.9 % in LR

Kumar et al. predicted type-2 diabetes using a variety of classification algorithms, such as SVM, ANN, and classification trees, and their accuracy ranged from 73.00% to 80.00%. An analysis of the primary risk variables for type 2 diabetes was conducted by Miah et al. By employing correlation analysis, more significant characteristics regarding type-2 diabetes and its impact on quality of life were found. The effectiveness of well-known machine learning techniques (ANN, K-NN, and decision trees) for diabetes mellitus prediction was assessed by certain authors. Two databases were used for the experiments: one was taken from a hospital in Frankfurt, and the other was an open-source PIMA Indian dataset. The best total accuracy, according to the data, was

90.00. Furthermore, Tafa et al.'s model predicts diabetes by combining SVM and Naive Bayes. The model was tested using a collection of data collected from three distinct locations inside Kosovo. Of the 402 participants in the study, 80 had type 2 diabetes, and the dataset included 8 important characteristics. They divided the dataset in half (50%) for the training set and the other half (50%) for the testing set in order to conduct the validation test. The SVM's accuracy was 95.50%, according to the authors, whereas the Naive Bayes classifier's accuracy was 90.00%. Practitioners and officials in the healthcare industry may find great use in an ANN model as described by certain authors. The disease's very fatal consequence served as the author's inspiration. To lower the training error function, they used an ANN model. As a result, the accuracy attained by ANN was 87.30%, and the average error function that was determined was 0.01%. A probabilistic neural network (PNN)-based diabetic prediction system was presented by Soltani et al. 90% of the data for training and 10% for testing were taken from the Pima Indians Medical Diabetes (PIMA), which was used in the experiment. An overall training accuracy of 89.50% and a testing accuracy of 82.00% were attained by the suggested network. Feature selection was used to diabetes (type 1) patients by Rodriguez et al. using variables like sleep, routine, diet, exercise, insulin, and heart rate. The Sequential Input Selection Algorithm (SISAL) and time-series data were utilized by the authors to rate each feature according to its prediction value for blood glucose levels. Aiswarya Iyer (2015) examined latent patterns in a diabetes dataset by using a classification algorithm. This model made use of Decision Trees and Naïve Bayes. The two algorithms' performances were compared, and the outcome showed how effective each method was. V. Sangeetha and K. Rajesh (2012) employed a categorization technique. For effective classification, they employed the C4.5 decision tree technique to extract hidden patterns from the dataset. Fuzzy logic and artificial neural networks (ANNs) were employed by Humar Kahramanli and Novruz Allahverdi (2008) to forecast diabetes. B.M. Patil, R.C. Joshi, and Durga Toshniwal (2010) presented the Hybrid Prediction Model, which consists of applying a classification algorithm to the output of the clustering technique after applying the Simple K-means clustering algorithm. The C4.5 decision tree algorithm is used to create classifiers. A Random Forest Classifier model was suggested by Mani Butwall and Shraddha Kumar (2015) to predict diabetes behavior. Neural networks, K-means clustering algorithms, C4.5 decision tree algorithms, and visualization were employed by Nawaz Mohamudally1 and Dost Muhammad (2011) to predict diabetes. Sadeghi et al. used Tehran Lipidand Glucose Study (TLGS) cohort data to predict diabetes using deep neural network (DNN), extreme gradient boosting (XGBoost), and random forest (RF). DNN performed best with the maximum accuracy. Applied machine learning approaches were evaluated on PIDD; the best accuracy of 90.91% was demonstrated by generalized boosted regression modeling. To predict short-term blood glucose, Zecchin et al. and Reddy et al. used a polynomial model and a neural network (NN). This plan necessitates ongoing sample distribution and monitoring, both of which take time. A public dataset on diabetes in the Pima indigenous population was utilized, along with independent algorithms like Decision Tree, Polynomial SVM (Poly-SVM), Linear SVM (L-SVM), Radial basis function SVM (RBF-SVM), and Polynomial SVM, as well as the KNN algorithm as a meta-classifier to combine the data. Singh y Singh established a system on the basis of stacking called "NSGA–II–Stacking" for forecasting the onset of type 2 diabetes mellitus over a period of five years. In order to do this, a public dataset on diabetes in the Pima indigenous population was used. The KNN algorithm was employed as a meta-classifier for combining the forecasting of the base models, and the independent algorithms Linear SVM (L-SVM), Radial basis function SVM (RBF-SVM), Polynomial SVM (Poly-SVM), and Decision Tree were used. With respect to performance measures, the suggested system yielded 83.8 percent accuracy, 96.1 percent sensitivity, 79.9 percent specificity, 88.5 percent F1-Score, and 84.9 percent ROC curve. Similarly, Kumari et al. examined the effects of employing an ensemble model to predict diabetes. The proposal involved combining three binary classifiers, namely Naïve Bayes, Random Forest, and Logistic Regression. The Ensemble model produced the best results, with accuracy values of 79.04 percent, precision of 73.48 percent, recall of 71.45 percent, and F1 Score of 80.6%. Rajendra and Latifi combined the Vanderbilt and PIMA diabetes datasets in a subsequent study to create predictive models for diabetes. Logistic regression and two ensemble approaches, max voting and stacking, were used to build the primary models. Logistic regression and two ensemble approaches, max voting and stacking, were used to build the primary models. SVM, Decision Tree, KNN, and Naive Bayes were then added to the ensemble models. The ensemble model performed better in both datasets than the logistic regression model, with accuracy values of 77.83% and 93.41%, respectively. Using a set of data from patients in Nanjing, including 3,845 confirmed cases of DMTS and 8,000 non-diabetic patients, Xiong et al. applied an ensemble-based technique to assess the risk of type 2 diabetes mellitus in the Chinese urban population. They used five machine learning algorithms: Multilayer Perceptron (MLP), Adaboost, Random Forest, SVM, and Gradient Tree Boosting (GTB). The Logistic regression and two ensemble approaches, max voting and stacking, were used to build the primary models. SVM, Decision Tree, KNN, and Naive Bayes were then added to the ensemble models. The ensemble model performed better in both datasets than the logistic regression model, with accuracy values of 77.83% and 93.41%, respectively. In order to determine the risk of type 2 diabetes mellitus in the Chinese urban population, Xiong et al. used an ensemble-based approach. They used a set of data from patients in Nanjing, which included 8,000 non-diabetic patients and 3,845 confirmed cases of DMTS. Using a set of data from patients in Nanjing, incorporating 3,845 confirmed cases of DMTS and 8,000 non-diabetic patients, Xiong et al. applied an ensemble-based technique to assess the risk of type 2 diabetes mellitus in the Chinese urban population. They utilized five machine learning algorithms: Adaboost, Multilayer Perceptron (MLP), Adaboost, SVM, Gradient Tree Boosting (GTB) and

Random Forest. The empirical findings demonstrated that the ensemble-based classifier combination produced better outcomes, with 91% accuracy, 95% specificity, 83% sensitivity, 97% AUC, and 88% precision. Ahmad et al.'s study used machine learning techniques to investigate the influence of health-related characteristics on the prediction of type 2 diabetes mellitus. The dataset included 16 attributes and consisted of 3000 patient records from different Saudi hospitals. The modeling techniques Random Forest, Decision Tree, Ensemble Majority, Logistic Regression, and SVM were then used. These models were assessed twice by Cross Validation with ten repetitions, one with nine attributes and the other with eight. SVM performed better than the others in the initial set of data, obtaining 82.1% accuracy with both nine and eight attributes. However, the second dataset's results demonstrated that Random Forest had the best accuracy, scoring 87.65 percent with eight attributes and 88.27 percent with nine. utilizing the Pima diabetes dataset, Li et al. (2020) created a model to predict diabetes utilizing ensemble learning techniques to improve disease prediction. With an accuracy rate of 80.20%, they obtained the best results using extreme gradient boosting (XGBoost). The authors' enhanced feature combination classifier, which makes use of the XGBoost model, can be investigated to enhance disease prediction in the medical field. Mahabub (2019) examined various ensemble learning approaches, including AdaBoost, gradient boost, XGBoost, random forest, etc., to forecast diabetes while taking into account a number of clinical factors, including blood pressure, body mass index (BMI), pregnancy, skin thickness, glucose, insulin, and diabetes pedigree function. They utilized the multilayer perceptron technique to attain the greatest accuracy rate of 84.42%. Using the Pima diabetes dataset, Mushtaq et al. (2022) developed an optimized model that uses a vote classification based on the ensemble approach to predict diabetes. A two-phase model selection procedure was employed in this research project to create the model. Out of all the classifiers, the voting classifier achieved the highest accuracy rate of 81.50%. In addition, data balancing methods such as Tomek and synthetic minority oversampling technique (SMOTE) were employed to eliminate biases from the dataset. The researchers recommended carrying out more research to determine the probability that those without diabetes will eventually develop this illness. Several boosting techniques were used by Beschi Raja et al. (2019) to construct the diabetes prediction model. Out of all the classifiers, the gradient boosting technique achieved the greatest accuracy rate of 89.70%. In order to verify the suggested model, additional statistical measures have also been examined.

Khan et al. (2021) used the boosting method to construct a diabetes prediction model. For predictive analytics, the authors investigated a variety of classifiers, including ANN, j48, deep learning, naive Bayes, gradient boosting, and hybrid k-nearest neighbor (kNN). The gradient boosting technique produced the best results out of all the classifiers. Furthermore, the k-fold cross-validation approach was employed to validate the outcomes. The model's creators proposed using it as a prognostic tool for early illness prediction in the healthcare sector.

A comprehensive framework for the predictive analysis of diabetes was created by Lai et al. in 2019. In particular, for class balancing, the gradient boosting machine approaches were employed with hyperparameter adjustment, which minimized the loss of classification prediction probabilities.

Singh et al. (2021) presented eDiaPredict, an ensemble approach-based framework for predicting a patient's diabetes condition. XGBoost, random forest, support vector machine (SVM), neural network, and decision tree are all included in the suggested methodology. Applying XGBoost and random forest together yielded accuracy, precision, and sensitivity of 95%, 88%, and 90.32%, respectively, on the PIMA Indian diabetes dataset, demonstrating the effectiveness of eDiaPredict.

A framework for diabetes prediction employing kNN, decision trees, random forests, AdaBoost, Naive Bayes, XGBoost, and multilayer perceptron was presented by Hasan et al. (2020). To increase the accuracy of the predictions, they tested with the PIMA Indian diabetes dataset and a weighted ensemble of machine learning models. The AUC and specificity of the suggested ensemble model were considerably higher, at 0.950 and 0.934, respectively. But at 88.84%, 84.32%, and 78% accuracy, precision, and sensitivity, respectively, it performed worse.

III. Conclusions

Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. According to the experimental findings, gradient boosting had the greatest accuracy rate of 96%. Additionally, it did well in terms of other evaluation criteria like ROC curve, precision, recall, and f1-score

References

[1] Isfafuzzaman Tasin,Tansin Ullah Nabil,Sanjida Islam,and Riasat Khan, *Diabetes prediction using machine learning and explainable AI techniques*, Healthcare Technology Letters, Volume 10, Issue 1-2 p. 1-10 Wiley, 2022

[2] Monalisa Panda, Debani Prashad Mishra, Sopa Mousumi Patro, Surender Reddy Salkuti2., *Prediction of diabetes disease using machine learning algorithms*, IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 11, No. 1, March 2022, pp. 284~290, ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i1.pp284-290 284

[3] Salliah Shafi Bhat, Madhina Banu, Gufran Ahmad Ansari, Venkatesan Selvam, "*A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms*," Healthcare Analytics Volume 4 December 2023, 100273.

[4] Kiran Kumar Patro1, Jaya Prakash Allam, Umamaheswararao Sanapala1, Chaitanya Kumar Marpul, Nagwan Abdel Samee3, Maali Alabdulhafth and Pawel Plawiak, "*An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques*" in BMC Bioinformatics 24, Article number: 372 (2023)

[5] Aishwarya Mujumdar, V Vaidehi Dr, "Diabetes Prediction using Machine Learning Algorithms Procedia Computer Science Volume 165, 2019, Pages 292-299

[6] Mohammad Atif, Faisal Anwer, Faisal Talib, Rizwan Alam, Faraz Masood, "Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage*",* IAES International Journal of Artificial Intelligence (IJ-AI) *,Vol. 12, No. 3, September 2023, pp. 1302~1311, ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i3.pp1302-1311 1302*

[7] Alfredo Daza, Carlos Fidel Ponce Sánchez, Gonzalo Apaza-Perez, Juan Pinto, “*Stacking ensemble approach to diagnosing the disease of diabetes”*, Informatics in Medicine Unlocked, Volume 44, 2024, 101427

[8] M. Sivaraman; J. Sumitha,, “Prediction of diabetes using optimized RBFNN algorithm” , AIP Conf. Proc. 2901, 060001 (2023)

[9] Aman, Rajender Singh Chhilla, “Optimized stacking ensemble for early-stage diabetes mellitus prediction”, *International Journal of Electrical and Computer Engineering (IJECE)* ,Vol. 13, No. 6, December 2023, pp. 7048~7055 ,ISSN: 2088-8708