TIERS Information Technology Journal
Vol.4, No.2, December 2023, pp. 150~164
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v4i2.3617

A Review of Diverse Diabetic Prediction Models: A Literature Study

Amina Zafar [1], Areeg Tahir [2], Umer Asgher [3]
azafar.cse21ceme@student.nust.edu.pk, atahir.cse21ceme@student.nust.edu.pk,
umer.asgher@ceme.nust.edu.pk
Software Engineering, NUST College of Electrical and Mechanical Engineering, Islamabad Capital Territory, Pakistan

ABSTRACT

Diabetes is a disease described by extreme glucose measurement in the blood and can trigger an excessive number of problems likewise in the body, like the failure of internal organs, retinopathy, and neuropathy. As per the forecasts made by World Health Organization, the figure might reach roughly 642 million by 2040, and that implies one in ten might experience diabetic diseases due to various reasons such as low activity levels, unhealthy routines, and schedules, rising tension levels and so on. Many researchers in the past have explored widely on diabetes disease through AI calculations and ML algorithms. The possibility that had persuaded us to introduce a survey of different prediction models of diabetic disease is to address the diabetes issue by recognizing and coordinating the discoveries of all-important, individual examinations. In this research, we have analyzed the different prediction algorithms and techniques by different researchers that how they predict diabetic disease. Also, we have analyzed the PIMA and symptom and other datasets and how they reach their resultant accuracy by applying different classifiers. Because of non-linear, correlated, and complex structured data in the medical field, diabetic data analysis is very difficult. That's why Ml-based algorithms have been utilized for the prediction of diabetic disease and handle a large amount of data and it needs a different approach from others at the initial stage. We emphatically suggest our review since it involves articles from different sources that will assist different specialists with different models of prediction for diabetes.

Keywords: Machine learning; Data Mining; Diabetes; Random Forest; Voting Classifier; Naïve Bayes; SVM; AdaBoost Decision Tree

Article Info

Accepted : 10-12-2023
Revised : 07-07-2023
Published Online : 25-12-2023

Correspondence Author:

Amina Zafar
Software Engineering,
NUST College of Electrical and Mechanical Engineering,
Main Peshawar Rd, adjacent Daewoo Terminal, Rawalpindi, Islamabad, Islamabad Capital Territory, Pakistan.
Email: azafar.cse21ceme@student.nust.edu.pk

1. INTRODUCTION

Good Diabetes is a persistent (durable) clinical issue that influences how the functions of the human body transform food into energy. The meals that the human body eats are converted into sugar (called glucose) and transmitted into your circulation system. When glucose rises to a high level, it signals your pancreas to convey insulin. Insulin acts as the main role in diabetes by passing the glucose into cells of the human body. So, we can say that the massive quantity of sugar existing in the blood becomes the cause of diabetes. Somehow, the pancreas can't change over the food into insulin; accordingly, sugar remains unabsorbed, which becomes the cause of diabetic disease. The diabetic disease can influence blood vessels, kidneys, eyes, sensory system, etc. Three types of diabetes, the first one is juvenile (type 1 Mellitus) diabetes which mostly occurs in children due to different reasons and annihilates the cells which generate insulin in the pancreas. It is the most complicated

disease from which children are suffering in their childhood. Focusing on kids with Juvenile diabetes is ordinarily the job of the mother. Mothers should focus on the children of their good dietary patterns, invigorate active work, screen glucose levels, make successive visits to doctors, and show emotional support to them [1].

The second type of diabetes is Type 2 and people of different ages suffer from the type 2 diabetes [2] disease and it may happen due to different reasons like their unhealthy lifestyle, not doing any physical activity, eat unhealthy food without any measures. Diabetes is a chronological disease and can be harmful in case of carelessness but can cure and control this disease by taking proper medications, doing exercise, daily walking, and a healthy diet. Gestation is the third type of diabetic disease [3], which happens during the period of childbirth because of a change of hormones in the body, and after childbirth, it disappears. In this research, we are going to review and combine multiple studies that have been done until now regarding ML algorithms used for the prediction of diseases such as diabetic disease. ML is the mathematical measurement of logics and Statistics models that PC frameworks use to play out an explicit assignment without being expressly modified. ML algorithms are used as a daily schedule in many applications [4].

ML algorithms are utilized for different purposes like Digital Image processing, Data Mining, prediction, and so on. It can automatically learn what should do with data and after training, it can take care of its responsibilities consequently. ML is a branch of artificial intelligence and it has two types supervised and unsupervised learning. In the past few years, many researchers have introduced their effort on diabetic disease prediction by utilizing the algos of Machine Learning. In this article, we have concentrated on different diabetic disease prediction techniques utilizing the AI concept and introduced a detailed comparative study of a couple of strategies in our article. The actual aim of this article is:

1. To familiar with different diabetic prediction models.
2. To analyze existing implemented models based on their resultant accuracies.
3. To distinguish the gap in the current study.
4. To present a similar investigation of different models of prediction.
5. To gather maximum information for better analysis of multiple prediction models.

The possibility that had persuaded us to survey the different diabetic prediction models is to tackle the diabetics' issues by recognizing, fundamentally assessing, and coordinating the discoveries of all applicable, excellent individual examinations. To accomplish our inspiration for this survey cycle, we have concentrated on different research paper on diabetic prediction models, and we have chosen those research papers that has fulfilled the accompanying rules: Article high priority talked about different prescient strategies and AI calculations for the grouping of diabetic information. Article high priority talked about different preprocessing methods for filtering the noise data. Researchers have approved their model against a couple of execution boundaries.

Literature Review

Xu et. al. [5] proposed the diabetic prediction system by the analysis of correlated symptoms using exhaustive grid search techniques applied on the Random-Forest and Decision Tree, both achieve the accuracy of 98%, but correlation removal techniques have not been used to balance the dataset in this proposed study. Xue et. al. [6] proposed the diabetes prediction system to classify early diabetes by the usage of a Support vector machine, achieved an accuracy of 96.54%, and limited recommendation techniques for diabetes prediction have been used in this paper. Emon et. al. [7] proposed the diabetic prediction system to lower the health risks. In this study, 25% of the dataset has been selected for the testing set, and Random-Forest was applied for prediction, feature selection techniques, and feature importance techniques have not been used to select the best features that would enhance the overall results. Vigneswari et. al. [8] proposed the diabetic prediction system using the logistic model tree based on the blood pressure, insulin level, BMI, glucose, and other parameters. The diabetic dataset processed feature selection using the Weka and attained the accuracy of 79.31%, but still, much more analysis is necessary to learn about the dataset on which accuracy can be improved. Yahya et. al. [9] compare the different classifiers for the diabetes prediction based on the provided-PIMA Indian dataset by the usage of the Random-Forest that accuracy was 83.67%, but still, there is a gap in the features extraction methods to best fit the model using the important features for prediction.

Hassan et. al. [10] proposed the diabetic prediction system using the PIMA Indian dataset by the outlier detection techniques using the IQR, then compute the mean of values to fill the null values for accurate classification, and then used the correlation method for feature selection. The ensemble classifier of adaboost and xgboost achieved the highest AUC rate among other classifiers and attained an accuracy rate of 88.84%. Dutta et. al [11] experimented with the PIMA Indian dataset using the three classifiers namely Logistic regression, Support vector machine, and Random-Forest. In this study, Random-Forest achieved the best prediction results with an accurate rate of 83.81% with less detailed analysis results. Tigga et. al [12] discussed

the Random-Forest for diabetes prediction using the questionnaire-based dataset. Sonar et. al [13] proposed the diabetes prediction system from the PIMA Indian dataset by Support Vector Machine and Artificial Neural Networks that achieved the same metric results. Mujumdar et al. [14] proposed the diabetic prediction using the Logistic regression approach which proved the best classifier and achieves an accuracy rate of 96% by the PIMA Indian dataset. Naz et. al. [15] proposed the diabetic prediction system from the PIMA Indian dataset using deep learning with the help of a rapid miner tool.

Mushtaq et. al. [16] proposed the diabetic prediction system from the PIMA Indian dataset using different preprocessing techniques. Ensemble classifier achieved the highest accuracy rate of 81.7% among other implemented classifiers. Kumari et al. [17] developed a diabetes prediction system based on soft computing that utilized the ensemble classifier of Random-Forest, naïve Bayes, and logistic regression. Malik et al. [18] compared data mining and machine learning algorithms for predicting diabetes mellitus in women. The empirical results show that K-Nearest Neighbors, Random-Forest, and Decision-Tree outperform rather than other techniques. Yahyaoui et. al. [19] proposed the diabetic prediction system using the Random-Forest with the approach of grid search cv technique with possible hyperparameters and reached the accuracy rate of 79.26%. Sisodia et. al. [20] used the Weka tool for diabetes prediction by the Naïve Bayes that achieved the higher metrics results. Birjais et. al. [21] used the gradient boosting classifier for the PIMA Indian dataset. Ahmed [22] has used MATLAB R2020a for simulation purposes using the fused model with the embedment of artificial neural networks and a support vector machine. Md. Maniruzzaman [23] proposed the diabetes prediction system, using the dataset derived from the National Health and Nutrition Examination Survey (NHANES).

In this study, the authors conducted the Random-Forest classifier using the k-fold cross validation. Prasanth et. al. [24] addressed the diabetes issues by the proposed voting classifier using the local factor outlier detection. Benbelkacem [25] proposed the diabetic recommender system for prediction purposes by deriving the PIMA Indian dataset using the Random-Forest that achieved a low error rate. Islam Ayon [26] proposed a strategy to diagnose diabetes by deriving the PIMA Indian dataset using the deep learning classifier applied with cross-validation techniques. Zou et. al. [27] proposed a method for diabetes prediction using the Random-Forest classifier with the help of the Weka tool. Random-Forest proves to be the best classifier both for the Luzhou dataset the accuracy achieved 0.8084, and the accuracy for Pima Indians is 0.7721. Kaur [28] utilized the machine learning technique for Pima Indian diabetes dataset using the R data manipulation tool and performed different preprocessing techniques. In this study, Linear Kernel Support Vector Machine achieved a higher accuracy result of the 89%. Butt et. al. [29] utilized the Multilayer Perceptron Network by the inducement of sigmoid activation function and analyzed evaluation metrics such as precision, recall, and accuracy. Song [30] addressed the diabetic prediction using the multilayer perceptron and Adam gradient descent for weight optimization. Paul [31] proposed the diabetic prediction system by the implementation of the Scaled Conjugate Gradient Algorithm on the PIMA Indian dataset.

In this study, the authors conducted normalization on the entire dataset and performed K-Fold validation of the feed neural network with line search and Broyden Fletcher Goldfarb Shanno (BFGS) algorithm. Nahzat et. al. [32] proposed the diabetic prediction system using the PIMA Indian dataset. Random-Forest proved as the best classifier in this study. Soni et. al. [33] conducted the diabetes prediction by the Random-Forest and reached the highest accuracy of 77%. Alshamlan et. al. [34] proposed the diabetic prediction system by obtaining the public datasets GSE38642 and GSE13760 using the Support Vector Machine Classifier. Dunbray et. al. [35] addressed the diabetic issues by adapting an ensemble classifier using the light gradient boosting, k-nearest neighbor for the PIMA Indian dataset. Diab et. al. [36] proposed the diabetic prediction system using the neural networks by obtaining the PIMA Indian dataset in the MATLAB tool. Yadav et. al. [37] analyzed the PIMA Indian dataset and proposed a solution to encounter the diabetic issues, using different classifiers by the inducement of the hyper-tuning parameters. Costa et. al. [38] analyzed the PIMA Indian dataset and diabetes dataset 2019 to predict diabetics using the Random-Forest and achieved the accuracy results 0f 90.5%.

Saxena et. al. [39] proposed a solution to encounter the diabetic prediction by firstly preprocessing the dataset using the mean imputation method and then, various classifiers applied on the preprocessed dataset using k-nearest neighbor by the utilization of Minkowski distance, p value has 1 and number of neighbors (k-value) has 9. Other classifiers such as gradient boosting classifiers by the utilization of learning rate, shrinkage, and depth of the tree were used. In this study, the gradient boosting classifier has got accuracy of 91.63%. Prabhu et. al. [40] aimed to compare the results of deep belief networks with other different classifiers including familiar classifiers namely naïve Bayes, Decision Tree, Logistic Regression, Random-Forest, and Support Vector Machine. In this case, a comparative analysis was carried out, and deep belief networks proved a best neural network than other compared classifiers in terms of precision, recall, and F1 scores. Agarwal et. al. [41] diabetic prediction in women by obtaining the PIMA Indian dataset the k-nearest neighbor using linear kernel and K-Nearest Neighbor and voting classifier. By adding the cross-validation on the top of selected classifiers, accuracy reached 81.17% with the logistic regression. Aboalnaser et. al. [42] proposed the diabetic prediction

.

system using the orange data mining tool by accessing the public dataset 'PIMA Indian dataset' by the K-Nearest neighbor that outperformed and achieved the accuracy by 99.0% compared to other classifiers.

Khanam et. al. [43] proposed a strategy to apply the seven different machine learning algorithms by the approach of neural networks. Bettini et al. [44] suggested a novel approach for predicting type 2 diabetes and employed cross-validation techniques to develop an optimal Machine Learning model. Gupta et. al. [45] proposed a solution to encounter the diabetic issue using the prediction of different diabetic symptoms using the utilization of different machine learning models such as deep learning and the quantum machine learning algorithms by accessing the public dataset. Alam et. al. [46] proposed the diabetic prediction system using the artificial neural network with the inducement of possible hyper-tuning parameters.

Materials and Methods

PIMS Indian dataset was initiated by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset is mainly for women and the dataset has 9 columns consisting of eight attributes and other one has a target label as an outcome, and the 768 records have 268 diabetic patients and 500 unhealthy patients. This dataset consists of pregnancies, glucose, blood pressure, skin thickness, BMI, diabetes pedigree function, and insulin. In the dataset, there are zeros values handled differently by different authors including the mean imputation method. Also, the dataset is unbalanced, on which different sampling techniques have been used in different articles. In different papers, the authors did not pay heed to the unbalanced dataset.

The symptom dataset was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and approved by a doctor. This Symptoms dataset is mainly for both genders and the dataset have 17 columns consisting of 16 attributes, and another one has a target label 'class and this dataset contains 520 records. This dataset consists of symptoms including age, gender, polydipsia, polyuria, polyphagia, weakness, sudden weight loss, genital thrush, visual blurring, irritability, delayed healing, partial paresis, muscle stiffness, Itching, and alopecia. In the dataset, there are no missing values, and the dataset is balanced but contains outliers that are not addressed in papers. Another dataset has the Luzhou dataset which has 14 attributes, and some others have a survey-based dataset. The other two datasets that have used GSE38642 (54 healthy vs. 9 unhealthy) [19], and GSE13760 (11 healthy vs. 10 unhealthy) were obtained from the Hematology Department in Roskilde Hospital. These datasets have been taken from the Gene Expression Omnibus database, in this study, all the people were of age 20.

2. RESEARCH METHOD

The Xingchen Xu method

In this study, the diabetic prediction system was proposed by the analysis of correlated symptoms. Different machine learning algorithms have been used such as Stochastic Gradient Descent, K- Nearest neighbor, Decision- Tree, Random-Forest Bagging, Fully-connected Neural-Network, Ada- boost Decision-Tree, and Ada-boost Multilayer Perceptron Network. With the usage of hyperparameters tuning, Exhaustive grid search techniques applied on the Random-Forest and Decision-Tree, both achieve the performance accuracy of 98%. Correlation removal techniques have not been used to balance the dataset in this proposed study.

Jingyu Xue's Method

In this study, the diabetic prediction system proposed to classify early diabetes by the usage of supervised machine learning algorithms like Support Vector Machine, Naive Bayes classifier, and Light gradient boosting machine. In all these supervised machine learning algorithms, the Supportvector machine performs the best prediction performance in achieving accuracy at 96.54%. Limited recommendation techniques for diabetes prediction have been used in this paper.

Minhaz Uddin Emon's Method

In this study, the diabetic prediction system was proposed to lower the health risks. For this purpose, Logistic-Regression, Gaussian Process, Adaptive-Boosting, Decision-Tree, K-Nearest Neighbors, Multilayer Perceptron, Support Vector Machine, Bernoulli Naive Bayes, Bagging Classifier, Random-Forest, and Quadratic Discriminant Analysis have used in this study. In this study, 25% of the dataset has been selected for the testing set. In this study, Random-Forest has achieved the overall best accuracy metrics results at 98%, but feature selection techniques and feature importance techniques have not been used to select the best features that would enhance the overallresults.

Vigneswari's Method

In this study, the diabetic prediction system was proposed based on the blood pressure, insulinlevel, BMI, glucose, and other parameters. The diabetic dataset processed feature selection using the Weka tool. To achieve the goal of diabetes prediction, different classifiers used such as Random-Forest, C4.5, Random-Tree, REPTree, and Logistic-Model Tree were evaluated by their accuracy and True Positive Rate. In such analysis, the logistic model tree accomplishes the best prediction method and attained the accuracy of 79.31%, but still, much more analysis is necessary to learn about the dataset on which accuracy can be improved.

Yahyaoui's Method

In this study, the author compared the different classifiers for diabetes prediction based on the provided PIMA Indian dataset. In this study, the authors used Random-Forest, a Support vector machine, and fully connected neural networks by inducing the complete dataset records for prediction. Theperformance of these classifiers was analyzed and get the prediction accuracy by the Random-Forest was 83.67%, but still, there is a gap in the features extraction methods to best fit the modelusing the important features for prediction.

Md. Kamrul Hassan's Method

In this study, the diabetic prediction system was proposed using the PIMA Indian dataset. Different classifiers have predicted the diabetic results through the preprocessing step. Outlier detection using the IQR, then compute the mean of values to fill the null values for accurate classification and then used correlation method for feature selection. Different Machine Learning classifiers havebeen used such as K-Nearest Neighbor, Decision-Trees, Random-Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron Network. Ensemble classifiers including AdaBoost and XGBoost were utilized for the diabetes prediction. In this study, the utilization of these mentioned classifiers was analyzed. As a result, the ensemble classifier proved the best performer for diabetes predictionfrom the PIMS Indian preprocessed dataset, achieved the highest AUC rate among other classifiers, and attained an accuracy rate of 88.84%.

Debadri Dutta's Method

In this study, the author experimented with the PIMA Indian dataset using the three different machine learning algorithms namely Logistic-Regression, Support vector machine, and Random-Forest to predict which conditions can cause the development of diabetes in women. In this study, Random-Forest achieved the best prediction results with an accurate rate of 83.81%. This paper has not presented the detailed analysis results from the dataset and preprocessing techniques for outlier detection, some imputation methods have not been used for accurate classification.

Neha Prerna Tigga's Method

In this study, the author discussed different machine learning approaches to predict diabetes from the questionnaire-based dataset. In this study, the authors carried the comparative analysis from the PIMS Indian dataset and the questionnaire-based dataset through different classification techniques namely Logistic Regression, K- Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision-Tree, and Random-Forest. By the comparative analysis among these applied classifiers, Random-Forest performs the excellent classification operations, achieved the overall best results metrics, and attained an accuracy rate of 94%. The possible diabetic symptoms have been neglected from the questionnaire-based dataset that has a high impact on the diabetic level.

Priyanka Sonar's Method

In this study, the author proposed the diabetes prediction system from the PIMA Indian dataset by different machine learning classifiers including the Artificial neural networks, Decision tree, Support VectorMachine, and Naïve Bayes. In this study, Support Vector Machine, and Artificial Neural Networksachieved the same metric results, both attained the same accuracy rate of 82%, there are outlier detection techniques and sampling techniques have not been used to detect the outlier and balancethe dataset for the accurate classification.

Aishwarya Mujumdar's Method

In this study, the author proposed the diabetic prediction system based on the provided dataset whichhas features such as the number of pregnancies, glucose level, blood pressure, skin thickness, BMI,Age, and Job type). In this study, the authors performed the preprocessing step by the imputation method to fill the missing values, and then normalize the dataset. Different classifiers applied on the above-mentioned dataset records, Decision tree, Gaussian

.

naïve Bayes, Linear Discriminant Analysis, Support vector machine, Random-Forest, Extra trees, Ada-boost, Multilayer perceptronnetwork, Logistic-Regression, gradient boosting classifier, bagging and K-Nearest Neighbor used.By the comparative analysis of the PIMA Indian dataset and the other used dataset, Logistic-Regression proved the best classifier for the diabetes prediction and achieves the accuracy rate of 96% by the provided dataset, but still, feature selection techniques such as correlation-based, chi-square have not been used and feature importance for the provided dataset has also not been studied.

Huma Naz's Method

In this study, the author proposed the diabetic prediction system from the PIMA Indian dataset using different classification approaches such as Artificial Neural Network, Naive Bayes, Decision Tree,and Deep Learning. In this study, the authors used the rapid miner tool for classification and Deeplearning proved the best classifier in diabetes prediction based on the preprocessed PIMS Indian dataset (perform the imputation method and shuffled sampling technique to balance the dataset), has been achieved the accuracy rate of 98.07%. This paper has not the feature selection and analysis techniques for the dataset exploration.

Zaigham Mushtaq's Method

In this article, the author proposed the diabetic prediction system from the PIMA Indian dataset. In this study, the authors used the preprocessing techniques including the outlier detection using the interquartile range, imputation methods for missing values, and data sampling techniques such as oversampling and Tomek links sampling have used. Also there, a correlation-based featureselection technique has been used. Then, the cross- validation techniques were performed on the preprocessed dataset. To perform the classification, Logistic-Regression, Support Vector Machine,Naïve Bayes, K-Nearest neighbor, Gradient boosting classifier, and ensemble classifier have beenused. Ensemble classifier achieved the highest accuracy rate of 81.7% among other implemented classifiers.

Saloni Kumari's Method

In this article, the author developed a diabetes prediction system based on soft computing that utilized the ensemble of three machine learning algorithms. They assessed using PIMA and breast cancer databases. They used Random-Forest, Logistic- Regression, and Naive Bayes to compare their resultsagainst state-of-the-art individual and ensemble methods, and their system outperforms by 79 percent.

Sumbal Malik's Method

In this article, the author compared data mining and machine learning algorithms for predicting diabetes mellitus in women. They suggested a diabetes prediction system based on standard machine learning techniques. The proposed method is validated using a diabetes dataset from a German hospital. The empirical results show that K- nearest neighbors, Random-Forest, and Decision-Tree outperform other traditional techniques.

Amani Yahyaoui's Method

In this article, the author proposed the diabetic prediction system by using the Support Vector Machine, Random-Forest, and the deep learning models were utilized using the fully convolutional neural network to predict diabetes by obtaining the PIMA Indian dataset. In this study, the train test split option is used to split the dataset containing the 60% training dataset and 20% testing set, and 10% validation set to estimate the model functions. To conduct this study, the Support vector machine by selecting the kernel of the Radial Basis Function, then the accuracy reached 65.38%. The Random-Forest proved to be effective using the grid search cv technique with hyperparameters twenty number of trees and seven maximum-depth, then accuracy reached 79.26%. The other third algorithm used as a convolutional neural network using the fully connected layer and the pooling layers and soft-max activation function is used for the classification of output layers and reached the accuracy of 76.81%.

Deepti Sisodia's Method

In this study, the author used three machine learning classification algorithms namely Decision Tree, SVM, and Naive Bayes are used in this experiment to detect diabetes at an early stage. The author used the Weka tool for diabetes prediction using the different classifiers such as Decision-Tree, Support Vector Machine, and Naive Bayes. In this study, Naïve Bayes achieved the higher results in the perspective of accuracy (76.30%), precision (0.759), recall (76.3), and f-score (0.760).

Roshan Birgais's Method

In this article, the author used machine learning classifiers namely Gradient boosting, logistic regression, and Naive Bayes to identify the diabetic disease using various symptoms of Pima Indian dataset. Data preprocessing techniques such as Correlation variable selection in R language using the Boruta package that works as a wrapper around the Random-Forest. For the missing values imputation method, the K-nearest neighbor was computed by inducing the Euclidean distance. Different classifiers are used for diabetes prediction such as gradient boosting classifier by passing three number of folds and three iterations. The other one, Logistic-Regression used by passing the Logit parameter to get the probability for each target class. The highest accuracy achieved by the gradient boosting classifier is 86%.

Usama Ahmed's Method

In this article, the author has used MATLAB R2020a, incorporated the Artificial Neural Network and Support Vector Machine using the 5-fold Cross-Validation were incorporated for diabetes detection, while fuzzy logic was used in decision processes. The UCI data repository symptoms dataset is divided into training and testing datasets with a ratio of 70:30. The artificial neural network was used with Bayesian regularization (with 5% training set and 5% validation set) and tested on the remaining 30% test sets and achieved an accuracy of 92.31%. The model outputs become the input method for the fuzzy model, whereas the fuzzy logic determined the diabetes prediction inducement of the centroid method de-fuzzier. To store the fused models, a cloud storge method made effective in this study. The fused model diagnosed the diabetes using the inputs provided by the patients in real-time mode. By these such efforts, proposed fused model got the accuracy of 94.87% rather than previous conducted studies.

Maniruzzaman's Method

In this article, the author proposed the diabetes prediction system, got from the National Health and Nutrition Examination Survey (NHANES). In this study, the authors conducted the Feature selection technique by using logistic regression to detect the health risk factors based on a p-value less than 0.05 and regression coefficients. To conduct the validation, an Indian liver patient's dataset was used. Different classifiers were used by the authors such as Naïve Bayes, Decision-Tree, Ada-boost, and Random-forest. The results concluded that the overall best results were achieved by the Random-Forest with an accuracy rate of 95% and 0.95 AUC by tens number of k-folds.

S. Prasanth's Method

In this article, the author addressed the diabetes issues by proposing classification methods with increased precision for predicting people with diabetes using the PIMA Indian dataset. Random-Forest, Decision-Tree, Ada boost, Naïve-Bayes, Voting-classifiers, K-Nearest Neighbor, Light Gradient Boosting Machine, and Logistic Regression. In this study, the preprocessing technique was used to fill the null values using the median imputation method. Outliers can cause to develop inaccurate results, so those outliers were classified and eradicated using the Local outlier factor. In this article, Light gradient boosting analyzed that the Glucose attribute has a high importance feature and achieved a higher accuracy result by 90.01% and 0.90 validation score, voting classifier reached 94% of accuracy among others.

Sofia Benbelkacem's Method

In this article, the author proposed the diabetic recommender system for prediction purposes by deriving the PIMA Indian dataset. The main aim of this study, conduct comparisons among the tree-based variant classifiers including the Random-Forest, C4.5, REP-Tree, Simple-Cart, BF-Tree, and Support Vector Machine. By the comparison results, it is concluded that the Random-Forest achieved a higher accuracy (almost at 75%) and a low error rate of 0.21 among other tree-based classifiers.

.

Safial Islam Ayon's Method

In this article, the author proposed a strategy to diagnose diabetes by deriving the PIMA Indian dataset. In this study, cross-validation was specified as 5-fold and 10-fold. The deep learning approach used by the inducement of one input layer, four sizes of hidden layers, and the number of neurons in those layers are 12,16,16,14 respectively and specified the ReLU activation function using the Spyder tool. In our study, the deep learning approach by using the 5-fold cross-validation attained the best prediction accuracy results of 98.35%, F1 score of 98, and MCC of 97 by the evaluation of five-fold cross-validation.

Quan Zuo's Method

In this article, the author proposed the diabetes prediction system by deriving the dataset from hospital physical examination data in Luzhou, China, and the PIMA Indian dataset. For preprocessing step, deletion of missing values was performed in both datasets. Different classifiers are used such as J48-Decision-Tree and Random- Forest by using the Weka tool, and neural networks implemented in MATLAB. Hyper-tuning of neural networks settled by the size of the hidden layers was set to 10, set the value of the input layer and output layer to one, and specified the sigmoid activation function. To evaluate the model, 5-fold cross validation is used. In this article, Random-Forest proves to be the best classifier both for the Luzhou dataset and the accuracy achieved is 0.8084, and the best performance for Pima Indians is 0.7721.

Harleen Kaur's Method

In the current research article, the author utilized the machine learning technique in Pima Indian diabetes dataset to detect patterns with risk factors using the R data manipulation tool. Using K-Nearest Neighbor-based imputation method, null values get purified and the Boruta wrapper algorithm was used for feature selection. To classify the patients into diabetic and non-diabetic, the authors developed and analyzed five different predictive models namely linear kernel support vector machine by the specification of the hyper-tuning parameter as C=1, radial basis kernel of support vector machine by the specification of the hyper-tuning parameter as C=1 and sigma=0.107, k-nearest neighbor by the specification of K parameter size 13, artificial neural network by the sigmoid activation function and the hidden layer sizes=10, and multifactor dimensionality reduction by the recode activation function specification. In this study, Linear Kernel Support Vector Machine achieved a higher accuracy result of the 89%.

Umair Muneeb Butt's Method

In this article, the author utilized logistic regression by inducing the sigmoid function, Random-Forest, and Multilayer Perceptron Network by the inducement of the sigmoid activation function by neglecting the noisy dataset from sensor devices. Second, we implement three widely used machine learning algorithms for diabetes prediction, i.e., moving averages, Linear-Regression, and Long Short-Term Memory. Three performance measures (precision, recall, accuracy, root mean square error) were used to evaluate the performance. The results indicated that the Multilayer perceptron network outperforms with 86.6% Precision, 85.1% Recall, and 86.083% Accuracy.

Hwapyeong Song's Method

In this article, the author addressed the diabetes issues by finding out the different symptoms of diabetes on the PIMA Indian dataset using the multilayer perceptron network (activation function = 'Relu'), dense and sequential layers, output layer (activation function= 'sigmoid'), loss function as 'cross-entropy', and adam gradient descent used for weight optimization. To conduct this study, none of the preprocessing techniques were applied which is the main limitation of this study. In this study, the evaluation metric Area Under Curve (AUC) is used and the AUC result obtained was 0.915.

Banibrata Paul's Method

In this article, the author proposed the diabetic prediction system by the implementation of the Scaled Conjugate Gradient Algorithm on the PIMA Indian dataset. In this study, the authors conducted normalization on the entire dataset and performed K-Fold validation of the feed neural network with line search and Broyden Fletcher Goldfarb Shanno (BFGS) algorithm. It is observed that when the minimum accuracy is 77%, and the Mean Squared Error was achieved 0.1578 when the number of input neurons=8, number of hidden neurons=1, learning Rate= 0.6, and output neurons=2., maximum accuracy achieved 100% and Mean Squared Error is 0.0011 by the usage of eight input neurons=8, sixteen hidden neurons, 0.6 learning rate and two output neurons.

Shamriz Nahza's Method

In this paper, the author proposed the diabetic prediction system using the PIMA Indian dataset by the replacement of zeros values with the median, and correlation-based feature selection. Then, different algorithms were used for Classification K-Nearest Neighbors, Random-Forest, Support Vector Machine, Artificial Neural Network, and Decision Tree. As a result, Random-Forest proved as the best classifier with 88.31% accuracy, 88% precision, 86 % recall, and 87% F1 score.

Mitushi Soni's Method

In this article, the author addressed the diabetic issues and proposed a solution to encounter those issues and conduct the prediction on the PIMA Indian dataset. In this study, the author conducted the diabetes prediction by the removal of zeros values, and perform the prediction by using the different classifiers such as Random-Forest, support vector machine, k-nearest neighbor, decision tree, logistic regression, gradient boosting, and then ensemble classifier. By the comparative analysis, the Random - Forest attained an accuracy of 77%.

Hala Alshamlan's Method

In this article, the author proposed the diabetic prediction system by obtaining the public datasets GSE38642 and GSE13760 from the Gene Expression Omnibus database. In this study, the authors conducted the feature selection methods using the chi-square and Fisher score. Support vector machine and logistic regression were used to compute the diabetes prediction. The accuracy results of the two data were 90.23% and 61.90% respectively by the fisher score other than chi-square.

Nachiket Dunbray's Method

In this article, the author addressed the diabetic issues by adapting an ensemble classifier using the light gradient boosting, and k-nearest neighbor using the PIMA Indian dataset. In this study, an ensemble classifier by the grouping of light gradient boosting and k nearest neighbor was used with the adaption of the Grid Search CV technique to tune the hyperparameters such as the selection of Euclidean distance and k=7. In this study, the ensemble classifier achieved the higher accuracy results by using the cross-validation with three folds at 90.1%.

Maha S. Diab's Method

In this study, the author proposed the diabetic prediction system using the neural networks by obtaining the PIMA Indian dataset in the Matlab tool. Outlier detection using interquartile range and filled null values using mean was adapted in this proposed approach. In this study, classification models were conducted by splitting the dataset into a training set, validation set, and testing set in the ratios of 70:15:15, respectively using a random divide function in MATLAB by three main neural network functions including the feedforward, pattern, and cascade forward network. On the other hand, the feedforward network got 87.5% accuracy using the three hidden layers with 50, 30, and 20 neurons, eight epochs, and adjustment of lm function. In the case of the pattern network, the accuracy got 86.7% using the two hidden layers 50 and 20 neurons, four epochs, and adjustment of lm function. On the other hand, cascade forward network, the accuracy reached 91.1% using the two hidden layers 80 and 50 neurons, five epochs, and adjustment of lm function.

Shivani Yadav's Method

In this article, the author analyzed the PIMA Indian dataset and proposed a solution to encounter the diabetic issues. In this study, handling the outliers using the interquartile range, handling the null values by the imputation of the mean method, handling the feature selection technique by the ANOVA F-test, balanced the dataset by the oversampling technique (ADASYN method). Then, the Multilayer perceptron network using 5-fold cross validation has been used with hyper-tuned parameters such as using the five hidden layers induced with five neurons, the relu activation function, and 100 maximum iterations. By such techniques, the multilayer perceptron network achieved an accuracy of 80.4%.

Naomi Estera Costea's Method

In this article, the author analyzed the PIMA Indian dataset and diabetes dataset 2019 to predict diabetics. In this study, Support vector classification, Gaussian naïve Bayes, and Random-Forest were used on both datasets. In this study, missing values by the mean imputation method, label encoding used to preprocess the dataset, and Random-Forest achieved the accuracy results 0f 90.5%

Prakhar Saxena's Method

In this research article, the author proposed a solution to encounter the diabetic prediction by firstly

preprocessing the dataset using the mean imputation method and then, various classifiers applied on the preprocessed dataset such as naïve Bayes, support vector machine, k-nearest neighbor by the utilization of Minkowski distance, p value has 1 and number of neighbors (k-value) has 9. Other classifiers such as gradient boosting classifiers by the utilization of learning rate, shrinkage, and depth of the tree were used. In this study, the gradient boosting classifier has achieved a higher accuracy rate of 91.63.

P. Prabhu's Method

The author aimed to compare the results of deep belief networks with other different classifiers including familiar classifiers namely naïve Bayes, Decision Tree, Logistic Regression, Random-Forest, and Support Vector Machine. In this study, min-max normalization was applied for preprocessing technique and utilization of principal component analysis for feature selection. Then, the pretraining process using the deep belief neural network by the utilization of various parameters including epochs of Boltzmann's machine set to 10, input activation function set as sigmoid, hidden activation function set as Relu, number of hidden layers set to 3, hidden layers units [500,500,1000]. Fine-tuned the parameters of deep belief neural networks conducted for the testing purpose using the learning rate, Weight Initialization parameter type, both softmax, and the relu activation function set the four layers and the alike other parameters. In this case, a comparative analysis was carried out, and deep belief networks proved a best neural network than other compared classifiers in terms of precision, recall, and F1 scores.

Arushi Agarwal's Method

This article recommended a strategy to resolve the diabetic issues in women by obtaining the PIMA Indian dataset, utilized various classifiers including Decision Trees, Logistic Regression, Naïve Bayes, Support Vector Machine using linear kernel, and K-Nearest Neighbor that operated by two neighbors. By adding the cross-validation on the top of selected classifiers, accuracy reached 81.17% with the logistic regression.

Sara. A. Aboalnaser's Method

In this article, the author proposed the diabetic prediction system using the orange data mining tool by accessing the public dataset 'PIMA Indian dataset'. In this study, null values were replaced by the global mean value. Using the Anova, unimportant features were removed and the scalarization method was performed on the selected dataset. To verify the performance of models, 10-Fold cross validation was performed. Different classifiers have utilized and fine-tuned their parameters for this selected dataset such as Naïve Bayes, K-Nearest Neighbors by setting the neighbors to five and utilized the Euclidean distance, Artificial Neural Network, Decision Tree by utilization of 100 maximum depth, Random-Forest, Support Vector Machine by the selection of RBF kernel, Logistic Regression by the selection of ridge L2 and cost strength set to the default value. As a result, the K-Nearest neighbor outperformed and achieved the accuracy by 99.0% compared to other classifiers.

Jobeda Jamal Khanam's Method

In this article, the author proposed a strategy to apply the seven different machine learning algorithms including Decision Tree, K-Nearest Neighbor, Random-Forest, Naïve Bayes, Adaboost, Logistic Regression, and Support Vector Machine to the PIMA Indian dataset to predict diabetes and different performance metrics analyzed using the Weka and Jupyter tool. In this study, the preprocessing stage implemented different functions: outlier detection using the interquartile range, replacement of missing values with their mean value, correlation-based feature selection with a 0.2 cutoff, and data normalization, to improve the quality of data. In this article, the inducement of two hidden layers in neural networks proved as the most effective and provide accurate prediction with an accuracy rate of almost 86% for all epochs variations (200, 400, 800).

Gopi Bettini's Method's Method

This article suggested a novel approach for predicting type 2 diabetes. In this work, the authors used Machine learning classifiers to conduct trials for diabetes prediction using the pubic dataset of PIMA Indians. In this study, the authors' employed cross-validation techniques to develop an optimal ML model. The AUC for LR was 0.83, for RF it was 0.82, and for NB it was 0.81. All three models have been regarded as the most accurate for assessing whether diabetic patients or non-diabetic patients.

Himanshu Gupta's Method

This paper has proposed a solution to encounter the diabetic issue using the prediction of different diabetic symptoms using the utilization of different machine learning models such as deep learning and the

quantum machine learning algorithms by accessing the public dataset (PIMA Indian dataset). In this study, outlier detection using the interquartile range, replaced the null values by their target median value, multilayer feed-forward perceptron network utilized by the sequential layer, L2 regularization, and RMS-prop optimization technique was used to get the specified target level. On the other hand, the quantum machine learning approach was utilized by the hyper-tuned parameters using the adam optimizer with a learning rate of 0.01 and batch size of 10. As a result, the deep learning multilayer feed-forward perceptron network achieved a higher accuracy rate of 95%.

Talha Mahboob Alam's Method

In this article, the author used the three machine learning techniques namely Gradient boosting, logistic regression, and Naive Bayes on the PIMA Indian dataset. In this study, the authors carried out the preprocessing technique by replacement of null values with the median value, and the principal component analysis was carried out for feature selection. Further, the binning method set the association rules using the Apriori. The experimental results showed that the highest accuracy was achieved by the artificial neural network, the highest accuracy rate of 75.7, and the AUROC curve of 0.816 achieved by the inducement of learning rate to 10, number of hidden neurons to 5, and initial learning weight were set to 0.4.

Comparison among updated diabetic prediction analysis.

Table 1. Comparison analysis

S.NO	Year	Balancing Technique	Features used	Dataset	Classifiers used	Other parameters used	Accuracy
[16]	2022	Data sampling techniques such as oversampling and Tomeks links sampling	Correlation-based feature selection, and dropped the unnecessary features	PIMA Indian Dataset	Logistic-Regression, Support Vector Machine, Naïve Bayes, K-Nearest neighbor, Gradient boosting classifier, and ensemble voting classifier.	Outlier detection using the interquartile range, imputation methods for missing values, and Cross validationtechniques	81.7% with the voting classifier
[22]	2022	None	All features used	UCI Machine Learning Repository	Artificial neural network by the Bayesian regularization and support vector machine	5-fold cross validation	accuracy 94.87 by the artificial neural network.
[28]	2022	None	All features used	Luzhou and PIMA Indian dataset	J48 and Random-Forest by using the Weka tool, and neural networks	5-fold cross validation	Accuracy of Random-Forest achieved 80.4% for the Luzuho dataset and other has reached 77.7%
[25]	2022	None	Feature selected using the Boruta wrapper	PIMA Indian Dataset	Linear and RBF support vector machine, k- nearest neighbor, artificial neural network by the sigmoid activation, and multifactor dimensionality reduction by the recode activation function.	None	Linear Kernel Support Vector Machine achieved a higher accuracy result of the 89%.
[39]	2022	None	None	PIMA Indian Dataset	Naïve Bayes, support vector machine, k-nearest neighbor, and gradient boosting classifier	mean imputation method	gradient boosting classifier has achieved an accuracy rate of 91.63%.

.

3. RESULTS AND DISCUSSION

The In this article, a literature review has been done by studying the existing studies regarding machine learning models and algorithms also different techniques or methods have been used for the prediction of diabetes disease. Multiple classifiers have been used for the prediction of disease such as Random-Forest Tree, Naïve Bayes, Support Vector Machine, AdaBoost Decision Tree, Voting Classifier, and so on. The comparison of existing or recent studies has been done according to some specific criteria. Some limitations in the previous literature, they do not study or deep review newly used classifiers for prediction purposes, but this article will cover such gaps and facilitate multiple researchers and developers to study more about methods used for prediction in one frame. While studying multiple articles we have faced some challenges such as: Most of the authors have dropped missing qualities from the given dataset, which can influence the outcomes as the size of the dataset diminishes. ML algos are put in to the dataset; just a single creator has utilized AdaBoost and inclination support strategy. None of the creators has utilized the repetitive brain organization or profound learning innovation, which can help in expanding productivity. This strategy is considered to be more effective for prediction purposes.

CONCLUSION

We have determined from the above studies that the Mushtaq method is considered to be the best approach for diabetes prediction. To balance the dataset, SMOTE and Tomek have been used. Outliers have been eradicated from the dataset for equalization. Furthermore, this review looks at different ML-based models for predicting the diabetic state of patients at the earliest possible stage. After adjusting the dataset, the precision of classifiers was analyzed. Support Vector Machine, Naïve Bayes, Random-Forest, k-nearest neighbors, Gradient Boosting Classifier algorithm used by the Mushtaq Ahmed, and Random Forest achieved got 80.7% accurate prediction. Further, Mushtaq's method used a voting classifier and achieved 81.7% accurate prediction. In the future, accuracy can be increased by utilizing deep learning methods. So, the majority of the classifier has been utilized in Mushtaq's article that was considered to be good results.

REFERENCES

[1] A. Hussien, S. Saleh, and H. R. Tantawi, "Mothers' knowledge and practices toward their children suffering from juvenile diabetes: An assessment study," Egyptian Journal of Healthcare, vol. 10, no. 2, 2019. Available: https://doi.org/10.21608/EJHC.2019.36696.

[2] S. Malik, S. Harous, and H. El-Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in Proc. Int. Symp. Modeling and Implementation of Complex Systems, pp. 95–106, Springer, Algeria, Oct. 2020.

[3] J. Han, J. C. Rodriguez, and M. Behesti, "Discovering decision tree-based diabetes prediction model," in Proc. Int. Conf. Advanced Software Engineering and its Applications, pp. 99–109, Springer, Jeju Island, Korea, Dec. 2018.

[4] M. Batta, "Machine learning algorithms: A review," 2019. Available: https://doi.org/10.21275/ART20203995.

[5] X. Xu, X. Huang, J. Ma, and X. Luo, "Prediction of diabetes with its symptoms based on machine learning," in Proc. IEEE Int. Conf. Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), 2021, pp. 147–156. Available: https://doi.org/10.1109/CSAIEE54046.2021.9543343.

[6] IOP Publishing, "ShieldSquare Captcha," Journal of Physics: Conference Series, 2022. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012062/pdf.

[7] M. U. Emon, M. S. Keya, M. S. Kaiser, M.A. Islam, T. Tanha, and M. S. Zulfiker, "Primary stage of diabetes prediction using machine learning approaches," in Proc. Int. Conf. Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 364–367. Available: https://doi.org/10.1109/ICAIS50930.2021.9395968.

[8] D. Vigneswari et al., "Machine learning tree classifiers in predicting diabetes mellitus," in Proc. Int. Conf. Advanced Computing & Communication Systems (ICACCS), 2019, pp. 84–87. Available: https://doi.org/10.1109/ICACCS.2019.8728388.

[9] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in Proc. Int. Conf. Informatics and Software Engineering (UBMYK), 2019, pp. 1–4. Available: https://doi.org/10.1109/UBMYK48245.2019.8965556.

[10] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020. Available: https://doi.org/10.1109/ACCESS.2020.2989857.

[11] D. Dutta, D. Paul, and P. Ghosh, "Analyzing feature importances for diabetes prediction using machine learning," in Proc. IEEE 9th Ann. Information Technology, Electronics and Mobile Communication Conf. (IEMCON), 2018, pp. 924–928. Available: https://doi.org/10.1109/IEMCON.2018.8614871.

[12] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," Procedia Computer Science, vol. 167, pp. 706–716, 2020. Available: https://doi.org/10.1016/j.procs.2020.03.336.

[13] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in Proc. Int. Conf. Computing Methodologies and Communication (ICCMC), 2019, pp. 367–371. Available: https://doi.org/10.1109/ICCMC.2019.8819841.

[14] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, vol. 165, pp. 292–299, 2019. Available: https://doi.org/10.1016/j.procs.2020.01.047.

[15] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," J. Diabetes Metab. Disord., vol. 19, pp. 391–403, 2020. Available: https://doi.org/10.1007/s40200-020-00520-5.

[16] Z. Mushtaq et al., "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques," Mobile Information Systems, vol. 2022, Art. no. 6521532, pp. 1–16. Available: https://doi.org/10.1155/2022/6521532.

[17] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," Int. J. Cognitive Computing in Engineering, vol. 2, pp. 40–47, 2021.

[18] S. Malik, S. Harous, and H. E. Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in Proc. Int. Symp. Modelling and Implementation of Complex Systems, pp. 95–106, Springer, Batna, Algeria, Oct. 2020.

[19] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in Proc. Int. Conf. Informatics and Software Engineering (UBMYK), 2019, pp. 1–4. Available: https://doi.org/10.1109/UBMYK48245.2019.8965556.

[20] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018. Available: https://doi.org/10.1016/j.procs.2018.05.122.

[21] R. Birjais, A. K. Mourya, and R. Chauhan, "Prediction and diagnosis of future diabetes risk: A machine learning approach," SN Applied Sciences, vol. 1, art. 1112, 2019. Available: https://doi.org/10.1007/s42452-019-1117-9.

[22] U. Ahmed et al., "Prediction of diabetes empowered with fused machine learning," IEEE Access, vol. 10, pp. 8529–8538, 2022. Available: https://doi.org/10.1109/ACCESS.2022.3142097.

[23] M. Maniruzzaman et al., "Classification and prediction of diabetes disease using machine learning paradigm," Health Information Science and Systems, vol. 8, art. 7, 2020. Available: https://doi.org/10.1007/s13755-019-0095-z.

[24] S. Prasanth, M. Roshni Thanka, E. Bijolin Edwin, and V. Ebenezer, "Prognostication of diabetes diagnosis based on different machine learning classification algorithms," Annals of the Romanian Society for Cell Biology, vol. 25, pp. 372–395, 2021. Available: https://www.annalsofrscb.ro/index.php/journal/article/view/4299.

[25] B. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," in Proc. Int. Conf. Computer and Information Sciences (ICCIS), 2019, pp. 1–4. Available: https://doi.org/10.1109/ICCISci.2019.8716405.

[26] S. I. Ayon and M. Milon Islam, “Diabetes prediction: A deep learning approach,” Int. J. Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21–27, 2019.

[27] Q. Zou et al., “Predicting diabetes mellitus with machine learning techniques,” Journal of Biomedical Informatics, vol. 120, 2021.

[28] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” Applied Computing and Informatics, vol. 18, no. 1/2, pp. 90–100, 2022. Available: https://doi.org/10.1016/j.aci.2018.12.004.

[29] U. M. Butt et al., “Machine learning based diabetes classification and prediction for healthcare applications,” Journal of Healthcare Engineering, vol. 2021, Art. no. 9930985, pp. 1–17, 2021. Available: https://doi.org/10.1155/2021/9930985.

[30] H. Song and S. Lee, “Implementation of diabetes incidence prediction using a multilayer perceptron neural network,” in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), 2021, pp. 3089–3091. Available: https://doi.org/10.1109/BIBM52615.2021.9669583.

[31] B. Paul and B. Karn, “Diabetes mellitus prediction using hybrid artificial neural network,” in Proc. IEEE Bombay Section Signature Conf. (IBSSC), 2021, pp. 1–5. Available: https://doi.org/10.1109/IBSSC53889.2021.9673397.

[32] D. S. Diab, S. Husain, and A. Jarndal, “On diabetes classification and prediction using artificial neural networks,” in Proc. Int. Conf. Communications, Computing, Cybersecurity, and Informatics (CCCI), 2020, pp. 1–5. Available: https://doi.org/10.1109/CCCI49893.2020.9256621.

[33] S. Yadav et al., “A neural network based diabetes prediction on imbalanced data,” in Proc. IEEE Int. Conf. Communication Systems and Network Technologies (CSNT), 2021, pp. 515–521. Available: https://doi.org/10.1109/CSNT51715.2021.9509732.

[34] N. E. Costea, E. V. Moisi, and D. E. Popescu, “Comparison of machine learning algorithms for prediction of diabetes,” in Proc. Int. Conf. Engineering of Modern Electric Systems (EMES), 2021, pp. 1–4. Available: https://doi.org/10.1109/EMES52337.2021.9484116.

[35] P. Saxena, S. Saha, and S. K. Devi, “Diabetes prediction using support vector machine and random forest classifiers,” Proc. IEEE Int. Conf. Recent Advances in Computing and Software Systems (RACSS), 2022, pp. 315–319. Available: https://doi.org/10.1109/RACSS53876.2022.9751854.

[36] P. Prabhu and S. Selvabharathi, “Deep belief neural network model for prediction of diabetes mellitus,” in Proc. Int. Conf. Imaging, Signal Processing and Communication (ICISPC), 2019, pp. 138–142. Available: https://doi.org/10.1109/ICISPC.2019.8935838.

[37] A. Agarwal and A. Saxena, “Analysis of machine learning algorithms and obtaining highest accuracy for prediction of diabetes in women,” in Proc. Int. Conf. Computing for Sustainable Global Development (INDIACom), 2019, pp. 686–690.

[38] S. A. Aboalnaser and H. R. Almohammadi, “Comprehensive study of diabetes mellitus prediction using different classification algorithms,” in Proc. Int. Conf. Developments in eSystems Engineering (DeSE), 2019, pp. 128–133. Available: https://doi.org/10.1109/DeSE.2019.00033.

[39] J. Khanam and S. Foo, “A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 4, pp. 432–439, 2021.

[40] G. Battineni et al., “Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods,” Machines, vol. 7, no. 4, 2019. Available: https://doi.org/10.3390/machines7040074.

[41] H. Gupta et al., “Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction,” Complex & Intelligent Systems, 2021. Available: https://doi.org/10.1007/s40747-021-00398-7.

[42] M. T. Alam et al., “A model for early prediction of diabetes,” International Journal of Information and Computer Security, vol. 11, no. 2, pp. 32–41, 2022.

.