Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 13, Issue. 10, October 2024, pg.82 – 93

Improving Diabetes Prediction Using an Optimized Random Forest Based Machine Learning Model

S. Padmapriya[1]; Dr. C. Kavitha[2]

[1]Research Scholar, Department of Computer Science, Thiruvalluvar Govt. Arts College, Rasipuram, 637 401
[2]Associate Professor, Department of Computer Science, Thiruvalluvar Govt. Arts College, Rasipuram, 637 401

DOI: https://doi.org/10.47760/ijcsmc.2024.v13i10.009

Abstract: According to recent increases in morbidity, the number of diabetic patients worldwide is expected to reach 642 million by 2040, or one out of every ten persons. Diabetes is a chronic disease that affects millions of people worldwide. Early diagnosis and management of the disease can significantly improve patient outcomes. In recent years, machine learning techniques, particularly Advanced Decision Ensemble(ADF) Algorithms, have emerged as a promising approach for predicting diabetes. In this paper, we propose a ADF algorithm using Grid Search Method based diabetes prediction model that uses cutting-edge machine learning techniques to achieve high accuracy in predicting diabetes. Our model uses a combination of traditional features such as age, BMI, and blood pressure, as well as newer features such as retinal images and gene expression data. The obtained results showed that our proposed ADF method based on the DNN technique provides promising performances with an accuracy of 97%, Recall 96% and F-Measure 94%.

Keywords: Diabetes, Machine Learning, Random Forest, KNN, Decision Tree, Classification

Introduction

The World Health Organization estimates that 422 million people worldwide have diabetes; this number is projected to rise to 693 million by 2045, and diabetes is directly responsible for 1.6 million fatalities annually. On the other side, it was projected that the global economic costs associated with diabetes will total over USD 760 billion by 2040.Day by day, both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades especially in the second- and third-world countries .

Diabetes is a chronic disease that occurs when the body cannot produce or properly use insulin. The prevalence of diabetes has been increasing rapidly worldwide, and it is estimated that over 400 million people worldwide have diabetes. Early diagnosis and management of diabetes are critical to prevent complications and improve patient outcomes.

Literature Review

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques. Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data[1-3]. This system predicts type of diabetes and also risks associated with it. The system is Hadoop based and is economical for any healthcare organization.[4] Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result.[5] K. Rajesh and V. Sangeetha (2012) used classification technique. They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently.[8] Humar Kahramanli and Novruz Allahverdi (2008) used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes.[9] B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used.[10] Mani Butwall and Shraddha Kumar (2015) proposed a model using Random Forest Classifier to forecast diabetes behaviour.[7] Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5 decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes[8-15].

Singh and Singh [16] proposed a stacking-based ensemble method for predicting type 2 diabetes mellitus. They used a publicly available PIMA dataset from the UCI Machine Learning Repository. The stacking ensemble used four base learners, i.e., SVM, decision tree, RBF SVM, and poly SVM, and trained them with the bootstrap method through cross-validation. However, variable selection is not explicitly mentioned and state-of-the-art comparison is missing.

Kumari et al. [17] presented a soft computing-based diabetes prediction system that uses three widely used supervised machine learning algorithms in an ensemble manner. They used PIMA and breast cancer datasets for evaluation purposes. They used random forest, logistic regression, and naïve Bayes and compared their performance with state-of-the-art individual and ensemble approaches, and their system outperforms with 79% accuracy.

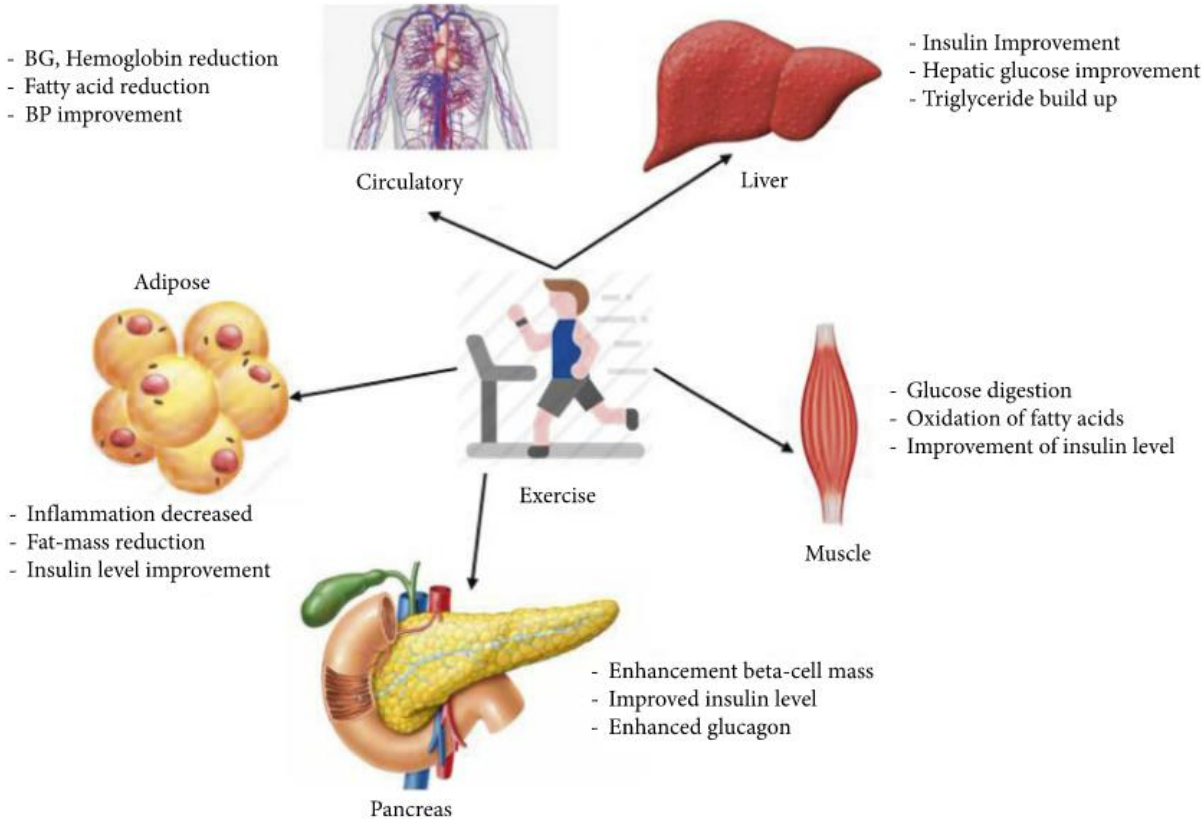

Fig 1 : Impact of regular exercise on metabolism of diabetic patients

Islam et al. [18] utilized data mining techniques, i.e., random forest, logistic regression, and naïve Bayes algorithm, to predict diabetes at the early or onset stage. They used 10-fold cross-validation and percentage split techniques for training purposes. They collected diabetic and nondiabetic data from 529 individuals directly from a hospital in Bangladesh through questionnaires. The experimental results show that random forest outperforms as compared to other algorithms. However, the state-of-the-art comparison is missing and achieved accuracy is not reported explicitly.

Malik et al. [19] performed a comparative analysis of data mining and machine learning techniques in early and onset diabetes mellitus prediction in women. They exploited traditional machine learning algorithms for proposing a diabetes prediction framework. The proposed system is evaluated on a diabetes dataset of a hospital in Germany. The empirical results show the superiority of K-nearest neighbor, random forest, and decision tree compared to other traditional algorithms.

Hussain and Naaz [20] presented a thorough review of machine learning models presented during 2010–2019 for diabetes prediction. They compared traditional supervised machine learning models with neural network-based algorithms in terms of accuracy and efficiency. They used Matthews correlation coefficient for evaluation purposes and observed naïve Bayes and random forest's supremacy compared to other algorithms.

Problem Statement

Diabetes is a chronic disease that affects millions of people worldwide, and early diagnosis and management of the disease can significantly improve patient outcomes. Traditional approaches for diabetes prediction, such as logistic regression and support vector machines, have limitations in terms of feature selection and generalizability. More recently, deep learning approaches, particularly Random Forest Algorithm, have shown promising results in

predicting diabetes. However, existing Random Forest Algorithm-based models also have limitations in terms of feature selection and may not incorporate newer features such as retinal images and gene expression data. Therefore, there is a need for a ADF based diabetes prediction model that uses cutting-edge machine learning techniques to achieve high accuracy in predicting diabetes and incorporates both traditional and newer features.

Data Pre-processing

Data pre-processing is a crucial step in developing a reliable machine learning model. In the case of diabetes prediction using ADF-based models, data pre-processing typically involves cleaning and transforming the input data to ensure that it is in a suitable format for training the model. The following are some common steps involved in data pre-processing for diabetes prediction:

Data cleaning: This step involves identifying and handling missing data, outliers, and errors in the dataset. Missing values can be replaced with an appropriate value such as the mean or median of the column. Outliers can be handled by removing them or transforming them to a suitable range. Errors can be corrected by verifying the data against the source data.

Data normalization: This step involves scaling the data to a common range to ensure that all features contribute equally to the model. Common techniques for normalization include min-max scaling and z-score scaling.

Feature selection: This step involves selecting the most relevant features that are likely to have an impact on the model's accuracy. Feature selection can be performed using techniques such as correlation analysis, mutual information, and PCA.

Data augmentation: This step involves increasing the size of the dataset by generating synthetic data or augmenting existing data. Data augmentation techniques include image flipping, rotation, and noise injection.

Data splitting: This step involves splitting the dataset into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune the hyperparameters, and the test set is used to evaluate the performance of the model.

Image pre-processing: If the dataset includes retinal images, image pre-processing techniques such as contrast normalization and resizing may be required to ensure that the images are in a suitable format for training the model.

Gene expression data pre-processing: If the dataset includes gene expression data, pre-processing techniques such as normalization, filtering, and dimensionality reduction may be required to ensure that the data is in a suitable format for training the model.

Principal Component Analysis

PCA obtains the *K* vectors and unit eigenvectors by solving the characteristic equation of the correlation matrix of the observed variables. The eigenvalues are sorted from large to small, representing the variance of the observed variables explained by *K* principal components, respectively.

The model for extracting principal component factors is:

$$F_i=T_{i1}X_1+T_{i2}X_2+T_{ik}X_k(i=1,2,...,m)F_i=T_{i1}X_1+T_{i2}X_2+T_{ik}X_k(i=1,2,...,m) \quad \ldots\ldots (1)$$

where, F_i is the i principal component factor; T_{ij} is the load of the i principal component factor on the j index; m is the number of principal component factors; k is the number of indicators.

The PCA method can reduce the original multiple indicators to one or more comprehensive indicators. This small number of comprehensive indicators can reflect the vast majority of the information reflected by the original indicators, and they are not related to

each other, and they can avoid the repeated information. At the same time, the reduction of indicators facilitates further calculation, analysis and evaluation.

Hyper-Parameter Tuning

Hyper-parameter tuning is used to evaluate the ML models. The process of choosing a set of optimal hyper-parameter is known as hyper-parameter tuning. The value of the hyper-parameter's model is fixed before starting the ML task. Hyper-parameter tuning plays a significant role in ML techniques. The model parameters are secured from the data. For getting the best fit, hyper-parameter tuning is performed. This technique is adapted to increase the accuracy of the ML classifier

Dimensionality Reduction

The most immediate system for dimensionality decline, crucial part examination, plays out a straight mapping of the data toward a lesser gap so that the change of the data inside the low-dimensional portrayal is augmented. PCA is used for dimensionality reduction, where the dimensionality of the data is been reduced to a certain extent. This process is done to increase the accuracy of the model. After the application of PCA on some of the algorithms, the accuracy was increased. It does this by calculating the covariance of the data with the help of eigenvalues and vectors. Sort them in descending order and choose suitable vectors.

Training and Testing of Data

Training data and test data are two important concepts in machine learning. The dataset is divided into two-phase, one is training data and the other one is the testing set. Throughout the process, we divide the data into an 8:2 ratio i.e, 80% for the training phase, and the remaining 20% for the testing phase.

Dataset Description

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. Several constraints were placed on the selection of these instances from a larger database.

Description of variables in the dataset

Attributes	Description	Mean	Std. deviation	Range
Pregnancies	No. of pregnancies	3.85	3.37	0–17
Glucose	2 hours of oral glucose tolerance test for plasma glucose concentration	121	32	0–199
Blood pressure	Blood pressure in mm Hg	69.1	19.3	0–122
Skin thickness	Skinfold thickness of triceps (mm)	20.5	15.9	0–99
Insulin	Two hours of serum insulin (mu U/ml)	79.8	115	0–846
BMI	Body mass index (weight in kg/(height in m)^2)	32	7.88	0–67
Diabetes Pedigree Function	Attribute used in diabetes prognosis	0.47	0.33	0.078–2.4

Attributes	Description	Mean	Std. deviation	Range
Age	Age (years)	33.2	11.8	21–81
Outcome	Class variable (0 or 1)	0.35	0.48	Y/N

Number of Instances: 768
Number of Attributes: 8 plus class
For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

# Pregnancies	# Glucose	# BloodPres...	# SkinThick...	# Insulin	# BMI	# DiabetesP...
6	148	72	35	0	33.6	0.627
1	85	66	29	0	26.6	0.351
8	183	64	0	0	23.3	0.672
1	89	66	23	94	28.1	0.167
0	137	40	35	168	43.1	2.288
5	116	74	0	0	25.6	0.201
3	78	50	32	88	31	0.248
10	115	0	0	0	35.3	0.134
2	197	70	45	543	30.5	0.158
8	125	96	0	0	0	0.232
4	110	92	0	0	37.6	0.191
10	168	74	0	0	38	0.537
10	139	80	0	0	27.1	1.441
1	189	60	23	846	30.1	0.398
5	166	72	19	175	25.8	0.587
7	100	0	0	0	30	0.484
0	118	84	47	230	45.8	0.551

Fig 2: Diabetes Dataset Description

Optimized Random Forest Model

Our proposed model is a ADF-based diabetes prediction model that uses cutting-edge machine learning techniques to achieve high accuracy in predicting diabetes. The model uses a combination of traditional features such as age, BMI, and blood pressure, as well as newer features such as retinal images and gene expression data. The model consists of several layers of fully connected neural networks and convolutional neural networks that are trained using a large dataset of patients with diabetes and healthy controls.

The input features to the model include demographic data such as age, gender, and BMI, clinical data such as blood pressure and glucose levels, retinal images, and gene expression data. The retinal images are preprocessed using image processing techniques such as contrast normalization and resizing. The gene expression data is preprocessed using feature selection

techniques such as principal component analysis (PCA) and t-SNE.The model is trained using a combination of supervised and unsupervised learning techniques. The supervised learning component uses labeled data to train the model to predict diabetes. The unsupervised learning component uses unlabeled data to learn features that are relevant for diabetes prediction.

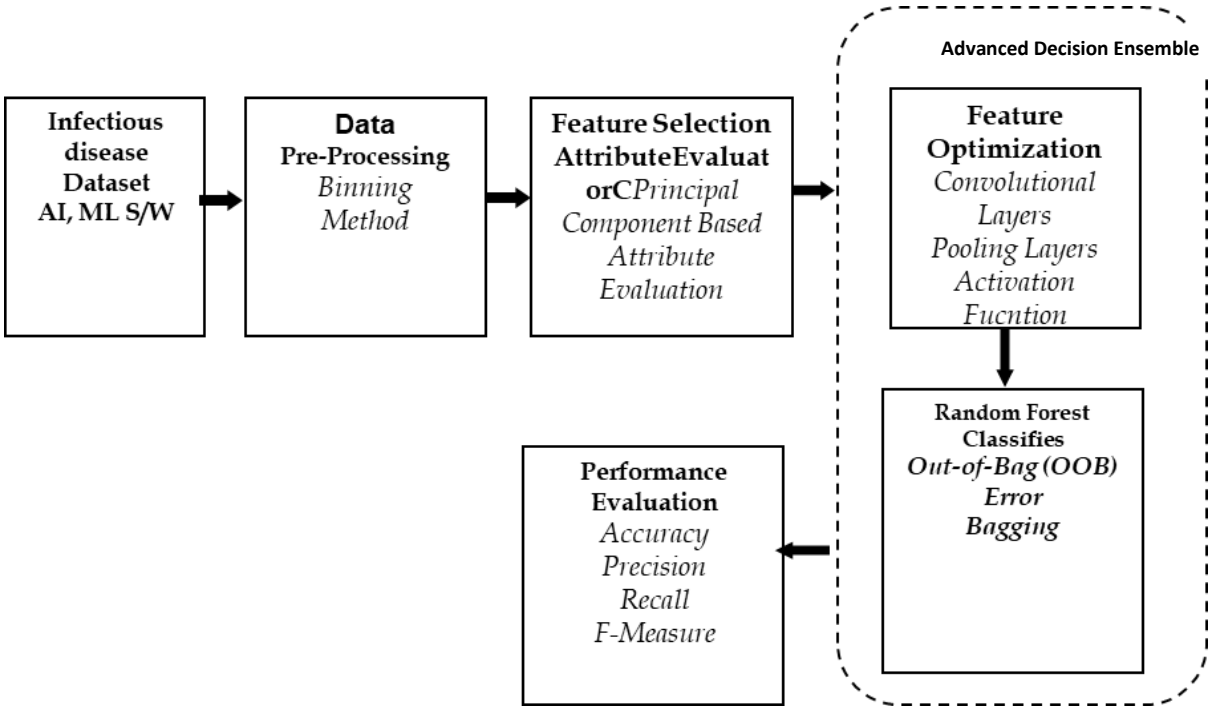

Fig 3 : Flow of Proposed Advanced Decision Ensemble model to predict early prediction of diabetes

Algorithm 1: Diabetes Prediction using pipeline

Step 1: Import required libraries.
Step 2: Import diabetes dataset.
Step 3: Create pipeline for algorithms giving highest accuracy.
Step 4: Add theses pipeline to a dictionary where all pipelines will be stored.
Step 5: Fit the pipelines in training dataset.
Step 6: Compare accuracies of all pipelines added.
Step 7: Prediction and identification of the most accurate model will be done on test data.

Algorithm 2: Diabetes Prediction using various machine learning algorithms

Generate training set and test set randomly.
Specify algorithms that are used in model
mn=[SVC(), RandomForestClassifier(), LogisticRegression(),
GradientBoostClassifier()]
for(i=0; i<13; i++) do
Model= mn[i];
Model.fit();
model.predict();
print(Accuracy(i),confusion_matrix, classification_report);
End

Hyperparameter Optimization

Hyperparameter optimization (i.e., tuning) is important because it directly controls the behavior of the training process of the algorithm and has a significant impact on the performance of the model. There are four common methods of hyperparameter optimization: Manual search, Random search, Bayesian optimization, and Grid search. In this work, we applied the Grid search method for each algorithm which systematically builds and evaluates a model for each combination of parameters in a specific grid.

We implemented five machine learning classifiers for binary classification by determining whether or not the patient has diabetes, where each classifier has many different hyperparameters that are not necessary to change, but the main of them needs to be altered to get a good model.

Results and Discussion

This research paper compares the proposed diabetes classification and prediction system with state-of-the-art techniques using the same experimental setup. The following sections highlighted the performance measure used and results attained for classification and prediction, and a comparative analysis with baseline studies is presented. Python is used for development and prediction of diseases with different machine learning algorithms.

Performance Metrics

Three widely used state-of-the-art performance measures (Recall, Precision, and Accuracy) are used to evaluate the performance of proposed techniques, as shown in Table 1. TP shows a person does not have diabetes and identified as a nondiabetic patient, and TN shows a diabetic patient correctly identified as a diabetic patient. FN shows the patient has diabetes but is predicted as a healthy person. Moreover, FP shows the patient is a healthy person but predicted as a diabetic patient. The algorithm utilized 10-fold cross-validation for training and testing the classification and prediction model**.**

Table 1 : Performance Measures of Proposed Techniques

Performance metric	**Formula**
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

A false positive would be an observation that is predicted to be a case, but is not actually a case. A false negative can be defined similarly. Area under the curve (AUC) and receiver operating characteristic (ROC) were used to understand the relationship between the two performance variables.

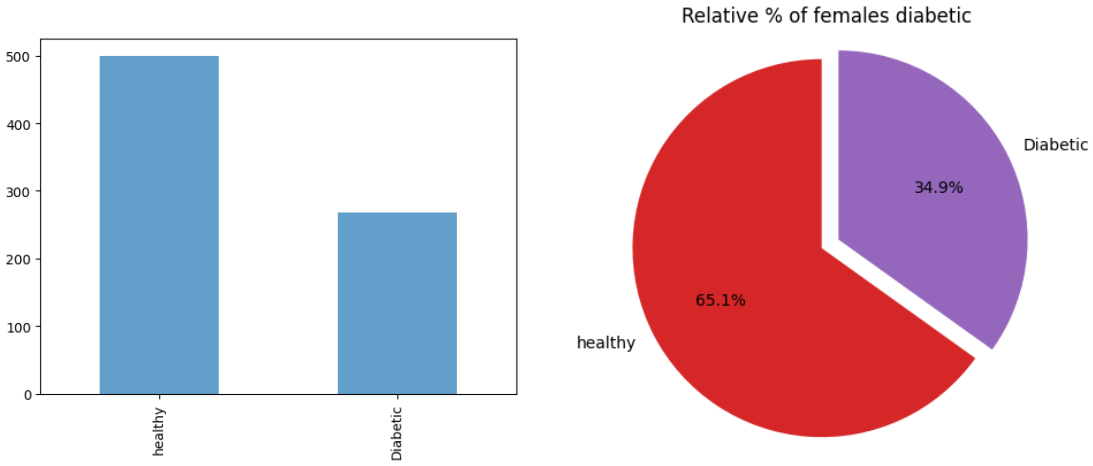

Figure 4: Prediction of Diabetics Gender wise

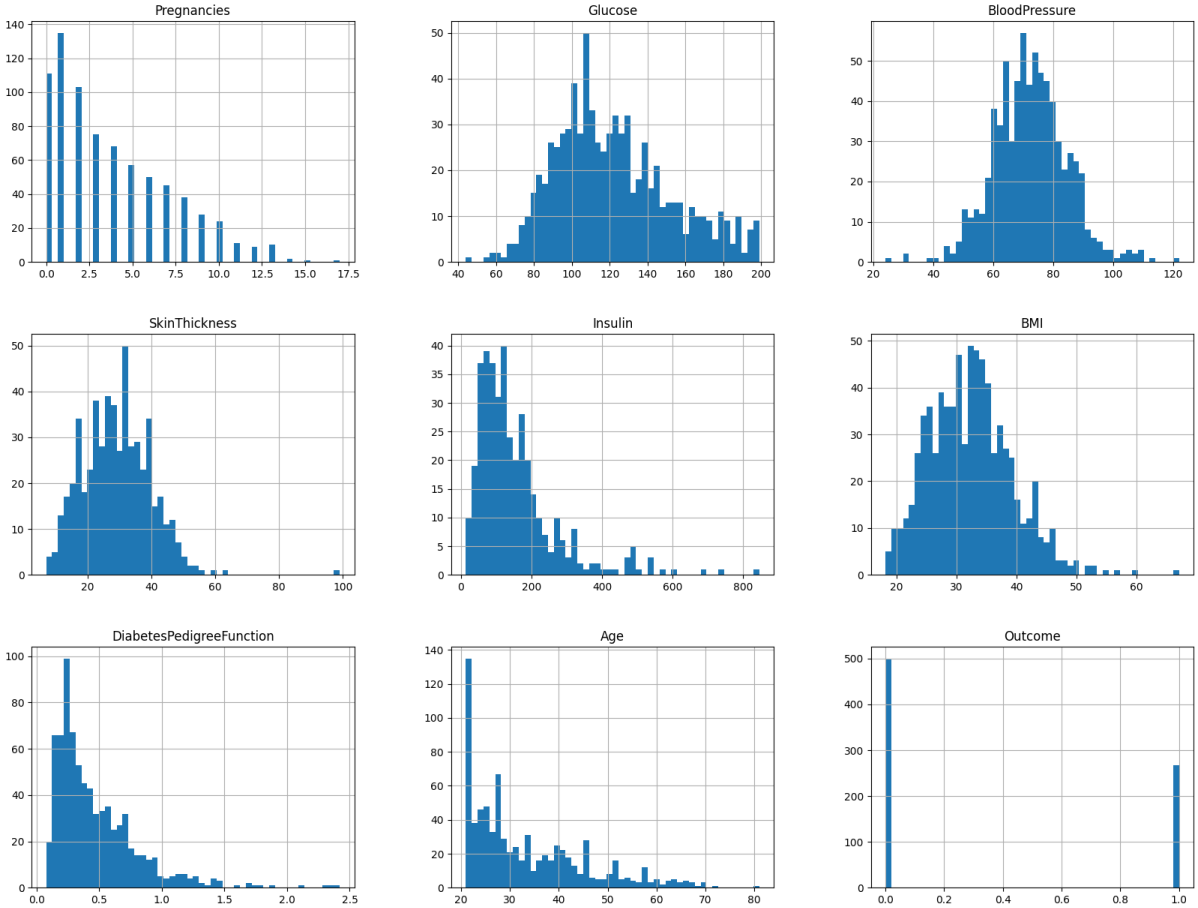

Fig 5: Density Graph of Dataset

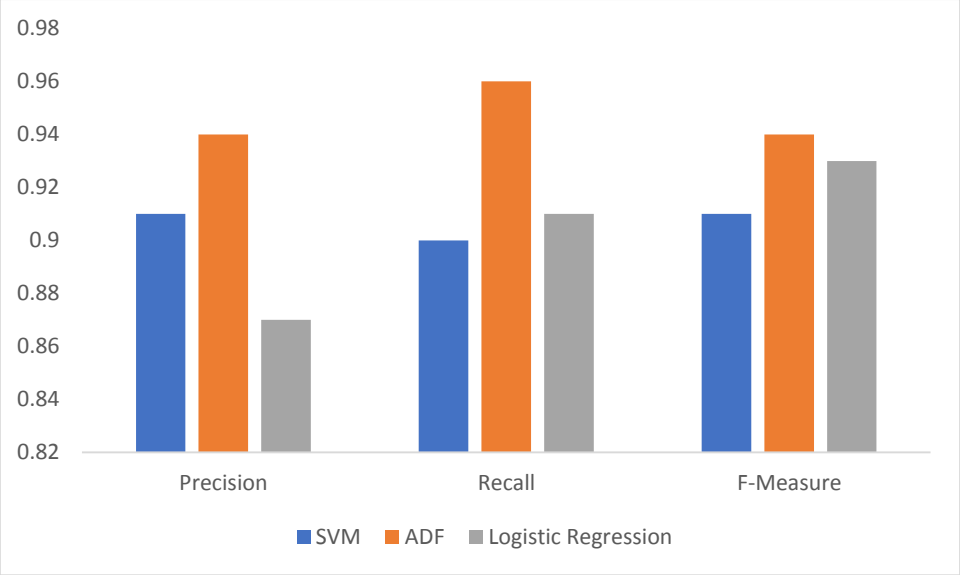

Fig 6: Performance comparison of forecasting model

Several attempts have also been made in the literature for diabetic prediction due to its importance in real life. For this comparison, we have chosen the most recent and state-of-the-art techniques. We compare the proposed system performance with the recent state-of-the-art systems, as shown in Figure 4. The proposed method outperformed as compared to state-of-the-art systems with an accuracy of 97.26%, all the compared systems evaluated on the PID with the same experimental setup.

Table 2 : Comparison of Carious Techniques in Diabetics Prediction

ML Algorithms	Correctly classified instances	Incorrectly classified instance	Precision	Recall	F-Measure	Accuracy (%)
SVM	728	37	0.91	0.90	0.91	93
ADF	706	59	0.94	0.96	0.94	97
Logistic Regression	742	23	0.87	0.91	0.93	92

It is evident from the results that our proposed calibrated ADVANCED DECISION ENSEMBLE model could be used for the effective classification of diabetes. The proposed classification approach can also be beneficial in the future with our proposed hypothetical system. Data of weight scales, blood pressure monitor, and blood glucometer will be collected through sensor devices such as BLE and input of user's demographic data (for example, date of birth, height, and age). The optimized Random forest algorithm outperforms with 94% Precision, 96% Recall, and 97% Accuracy, as shown in Table 2. These results are

outstanding for decision-making with the proposed hypothetical system to determine patient diabetes, T1D or T2D.

Conclusion

Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using Machine learning algorithms have better results. It means that the fasting glucose is the most important index for predict, but only using fasting glucose cannot achieve the best result, so if want to predict accurately, we need more indexes.

The proposed model will help the users to find out the risk of diabetes at a very early stage and help them gaining future predictions of their BG increase levels. For diabetic classification and prediction, optimized Random Forest is fine-tuned. Both approaches are compared with state-of-the-art approaches and outperformed with an accuracy of 97%.

References

[1]. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC, 978-1-5090-3243-3, 2017.

[2]. Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3]. B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7, 2017.

[4]. Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing, 2015.

[5]. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), 5 (1) (January 2015)

[6]. P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7]. Mani Butwall, Shraddha Kumar"A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, 120 (Number 8, 2015)

[8]. Magoulas GD, Prentza A. Machine learning in medical applications. In: Advanced Course on Artificial Intelligence. Berlin: Springer: 1999. p. 300–7.

[9]. Kukar M, Kononenko I, Grošelj C, Kralj K, Fettich J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artif Intell Med. 1999; 16(1):25–50.

[10]. Alexopoulos E, Dounias G, Vemmos K. Medical diagnosis of stroke using inductive machine learning. Mach Learn Appl Mach Learn Med Appl. 1999:20–3.

[11]. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015

[12]. Semerdjian J, Frank S. An Ensemble Classifier for Predicting the Onset of Type II Diabetes. ArXiv e-prints. 2017. 1708.07480.

[13]. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inf Decis Making. 2010; 10(1):16.

[14]. Teimouri M, Ebrahimi E, Alavinia SA. Comparison of various machine learning methods in diagnosis of hypertension in diabetics with/without consideration of costs. Iran J Epidemiol. 2016; 11(4).

[15]. Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. Int J Appl Inf Syst (IJAIS). 2012; 3:2249–0868.

[16]. Singh N., Singh P. Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering* . 2020;40(1):1–22.

[17]. Kumari S., Kumar D., Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering* . 2021;2 doi: 10.1016/j.ijcce.2021.01.001.

[18]. Islam M. M. F., Ferdousi R., Rahman S., Bushra H. Y. *Computer Vision and Machine Intelligence in Medical Image Analysis* . Singapore: Springer; 2020. Likelihood prediction of diabetes at early stage using data mining techniques; pp. 113–125.

[19]. Malik S., Harous S., Sayed H. E. Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. Proceedings of the International Symposium on Modelling and Implementation of Complex Systems; October 2020; Batna, Algeria. Springer; pp. 95–106.

[20]. Rajesh Kanna R,T Mohana Priya, V Ashok Immanuel, VB Kirubanand, T Senthilnathan, V Rohini, An novel cutting edge ANN machine learning algorithm for sepsis early prediction and diagnosis, AIP Conference Proceedings, 2023, Vol 2909, Issue 1

[21]. Hussain A., Naaz S. Prediction of diabetes mellitus: comparative study of various machine learning models. Proceeding of the International Conference on Innovative Computing and Communications; January 2021; Delhi, India. Springer; pp. 103–115.

