2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON) | 978-1-7281-7663-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ODICON50556.2021.9428943

CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages

P. Suresh Kumar[1]
Deptartment of CSE, Dr.LankapalliBullayya College of Engineering (W), Visakhapatnam, India, 530013
reshu.suri@gmail.com

Anisha Kumari K[2]
Deptartment of CSE, Dr.LankapalliBullayya College of Engineering (W), Visakhapatnam, India 530013
anishakushwaha14@gmail.com

Subhashree Mohapatra[3]
Department of Computer Scienece and Engineering, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India, 751030,
subhashreemohapatra68@gmail.com

Bighnaraj Naik[4]
Department of Computer Application, Veer Surendra Sai University of Technology, Burla, 768018, India
mailtobnaik@gmail.com

Janmenjoy Nayak[5]
Department of CSE, Aditya Institute of Technology and Management (AITAM), Tekkali, AP-532201, India
mailforjnayak@gmail.com

Manohar Mishra[6,*]
Department of Electrical and Electronics, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India, 751030
*manohar2006mishra@gmail.com
(Corresponding Author)

Abstract—**Diabetes prediction at the early stage is an important issue in the healthcare field and helps an individual to avoid dangerous situations by initiating treatment. For the prediction of diabetes at the early stages, many techniques in the area of machine learning and ensemble learning have been used. In this paper, we propose an ensemble technique CatBoost which is a Gradient Boosting Decision Tree (GBDT) for diabetes prediction at early stages. The experiment is conducted by comparing the performance of CatBoost with other machine learning methods such as K-Nearest neighbor, Multi-layer perceptron, Logistic regression, Gaussian Naive Bayes, and Stochastic gradient descent and the result is evaluated using accuracy, precision, recall, f1-score, and AUC-ROC curve. Experimentation is conducted using the dataset available in the UCI machine learning repository named "Early stage diabetes risk prediction". The results prove that CatBoost outperforms compared to the other machine learning methods.**

I. Introduction

Diabetes Mellitus also named as diabetes is a metabolic disease caused when the blood glucose or sugar level increases, mainly caused due to less production of insulin or the insulin produced is not consumed by the body cells[1]. Our body converts the food we eat into glucose which is required by body cells to do daily activities. The glucose in the blood is provided to the body cells by the hormone insulin. So,a decrease in the production of insulin increases the amount of glucose in the blood causing diabetes [2]. Diabetes is of four types: (i) Type 1 - caused due to less or no production of insulin, (ii) Type 2 – caused when the body cells could not use the insulin produced by the pancreas, (iii) Gestational diabetes – caused in pregnant women with high glucose level in the blood, (iv) Pre-diabetes - caused when the glucose level in blood is high but not enough to be diagnosis diabetes[3]. According to the estimation of the International Diabetes Federation, there are about 382 million people with diabetes around the world while would be doubled to 592 million people by 2035 [4]. Diabetes increases the chances of cardiovascular disease like stroke and heart attack and also affects other organs like nerves, eye, kidney, foot causing multi-organ damage [5]. However, detecting diabetes in early stages can help to control glucose levels and prevent these complications with treatment.

A huge amount of data regarding diabetes, reasons for illness, symptoms, and effect on health is available today. Analysis of this diabetic data is a challenging issue while diagnosis and treatment initiation of diabetes. Data analysis fields like data mining and machine learning have been applied for diabetes prediction. Machine Learning (ML) using statistical analysis on the given diabetes dataset has produced remarkable results in the classification and prediction of diabetes at the early stages. ML approaches such as artificial neural networks (ANN), support vector machine (SVM), naive Bayes (NB), k-nearest neighbour (KNN), decision tree (DT), Random Forest (RF) etc. have been used by many researchers for this process. The ability of these algorithms such as learning from the previous data and using that information for further classification and detection of diabetes is the main strength of ML. [6][7]. But these ML methods have their own drawbacks as well. The advantages and disadvantages of each ML method used for odiabetes prediction using the diabetic dataset are as follows: (i) SVM: It easily handles non-linear complex data and provides accuracy. It removes over-fitting in the dataset. But it is difficult to handle large datasets and execution is slow comparatively. (ii) KNN: It is easy to implement and training is faster. But it requires large space and knowledge transparency is low. (iii) DT: It does not require domain knowledge, easy interpretation, and can handle both numerical and categorical data. But it is an unstable classifier and generates categorical output. (iv) RF: It works well on data with several input parameters and has improved classification accuracy. But the interpretation is difficult and evaluation is slow. (v) ANN: it has the capability to develop a model that can handle non-linear complex data. It can also generalize the model to predict unseen. But the optimization of a model for large networks can be challenging due to the number of parameters and high processing time (vi) Logistic Regression (LR): Easily implemented and efficient to train. But the prediction of continuous output cannot be done and requires large sample data for stable outcome. (vii) Gaussian Naïve Bayes (GNB): Easy to implement and has more accuracy due

978-1-7281-7663-5/21/$31.00 ©2021 IEEE

to higher probability. But makes a strong assumption on the shape of data distribution and data is lost while making continuous features to discrete. (viii) Stochastic Gradient Descent (SGD): It is easy to implement, efficient, and gives good accuracy if input variables are selected sensibly. But it requires high memory and has complexity [8].

Though the performance of ML methods is superior, it is not widely used as it should be. These methods are complex and hard to understand the reason for the outcome. The outcomes should be understandable to trust the result in the healthcare system as the life of the person is depending upon the outcome of the methods. Understanding the outcome helps in better treatment and also helps in the new discovery of knowledge in this field [9]. The ensemble learning approach is the process of combining individual classifier methods to form a hybrid model to increase performance and improve the accuracy of the model [10]. It combines the classification ability of each classifier to increase the performance of the model by reducing the chances of misclassification of the dataset. The outcome is based on final ensemble classification rather than predictive analysis of an individual classifier [11]. CatBoost, a new gradient enhancement technique is also used for diabetes prediction and in other fields liken short-term weather forecast, driving style recognition, and kick-starter campaign prediction. CatBoost technique requires lower computational cost and shows better accuracy compared to individual classifier techniques like RF, SVM, and ANN techniques [12].

Sarwar et al., 2020[13] proposed an ensemble network for diagnosing type-II diabetes. A comparative experiment is done using a SVM, ANN, NB, KNN, and Ensemble models. Experimentation is done by using a dataset having 400 people recorded from wide geographical regions and the attributes are: age, gender, smoking, drinking, urination, Thirst, height, weight, fatigue, and family history. The experiment results provide that the Ensemble technique outperforms and shows efficient results with 98.60% accuracy compared to individual techniques.

Sisodia & Sisodia, 2018[7] compared three models NB, DT, and SVM method for diabetes prediction at an early stage. Pima Indians Diabetes Dataset (PIDD) consisting of 768 female patients' records used for the experiment and the results are evaluated based on accuracy, recall, precision, and f-measure. The results proved that the NB method has the highest accuracy with 76.30% compared to other methods.

Some more literature in diabetes risk prediction at early stages is presented in TABLE I.

TABLE I
The literature of Early prediction of diabetes

Author& Year	Technique	Database	Evaluation	Ref
Islam et al., 2020	Naive Bayes, J48 Decision Tree, Logistic Regression, Random Forest.	Data provided by Sylhet Diabetic Hospital	Recall, Precision, F-measure, FP rate, and TP rate.	[14]
Reddy et al., 2020	Ensemble learning using K-NN,Decision tree, Logistic Regression, Random Forest, Adaboost classifier.	Diabetic retinopathy dataset.	Precision, Recall, f1-score, support	[11]
Nguyen et al., 2019	Deep feedforward neural network	Data provided by Practice Fusion EHRs	AUC, Sensitivity, Specificity.	[15]
Dinh et al., 2019	Ensemble model using Random Forest, SVM, gradient boosting, Logistic regression.	NHANES dataset	AUC, Precision, Recall, f1-score	[16]
Fitriyani et al., 2019	Ensemble model (DPM)	4 different datasets	Precision, Recall, F-measure, accuracy, AUC	[17]

In this article, we proposed the ensemble learning algorithm CatBoost to predict diabetes at early stages by using a dataset available in the UCI ML repository. A comparive study has been performed amongst the proposed method and various other ML algorithms such as KNN, MLP, LR, GNB, and SGD to evident the efficiency.

The rest of the work is reported as follows: proposed CatBoost method is presented in Section-II. Section-III discusses the experiment, dataset features, parameter setting, results, and analysis. Section-IV summarizes the conclusion of the paper andthe future research direction is discussed.

II. Proposed Methodology

CatBoost is an efficient classifier algorithm that uses gradient boosting on decision trees and handles categorical features in the data [18]. It automatically handles categorical data using statistical methods; unlike other methods require to fit the categorical data beforehand. CatBoost can avoid the over-fitting of data by optimizing the extensive input parameters. It does not deal with the categorical features during processing time; it deals with them during training time. Instead of using binary substitution on categorical data it performs random permutation and computes a mean label value. For example, the value corepond to similar class value is placed before the known one in the permutation. This approach avoids the over-fitting of categorical data [19]. It performs well on small-sized data.

The categorical features present in the dataset can be combined to form a single new feature and then converted into a numerical value. After splitting the tress, it utilises a craving path to syndicate the unconditional features. First split

includes non-combination of categorical features. Combination starts from the second and subsequent splits and combining all the categorical features in the dataset forms a new feature. The target statistic method converts the categorical feature to the numerical value. The process of unbiased boosting of categorical features is expressed by ordered boosting [20]. The overall system process is shown in Fig.1.

Algorithm: CatBoost Algorithm

1. Dataset observations are given $\boldsymbol{D\{X_i, Y_i\}}$, where $\boldsymbol{i = 1, \ldots n}$
2. Random permutation is applied and it adds its prior value. If the permutation is $\boldsymbol{\sigma = (\sigma_1, \ldots, \sigma_n)}$, p is the prior value, and w is the weight correspond to p then, $\boldsymbol{x_{\propto_{P,k}}}$ is substituted by Eq 1.

$$x_{\propto_{P,k}} = \frac{\sum_{j=1}^{P-1}\left[x_{\propto_{(j,k)}} = x_{\propto_{(P,k)}}\right].Y_{\propto_j} + w.P}{\sum_{j=1}^{P-1}\left[x_{\propto_{(j,k)}} = x_{\propto_{(P,k)}}\right] + w} \quad (1)$$

3. Combine categorical features by using greedy target-based statistics
4. Ordered boosting is used to overcome gradient bias
5. Train each model $\boldsymbol{L_i}$ for each sample $\boldsymbol{x_i}$
6. These gradients $(\boldsymbol{R_1, \ldots R_n})$ can be considered as base learners for each model
7. The result is averaged to get the final outcome.

Algorithm: Ordered Boosting

Input: $\boldsymbol{(X_k, Y_k)_{k=1}^n}$, number of iterations N;
$\boldsymbol{\sigma}$ ← random permutation of [1, n];
$\boldsymbol{L_i}$ ← 0 for i=1, 2, …, n;
for s ← 1 to N **do**
  for i ← 1 to n **do**
    $\boldsymbol{r_i \leftarrow y_i - L_{\sigma(i)-1}(x_i)}$;
  end for
  for i ← 1 to n **do**
    $\boldsymbol{\Delta L \leftarrow LearnModel\left[(x_i, r_j) : \sigma(j) \le i\right]}$;
    $\boldsymbol{L_i \leftarrow L_i + \Delta L_j}$;
  end for
end for
return $\boldsymbol{L_n}$

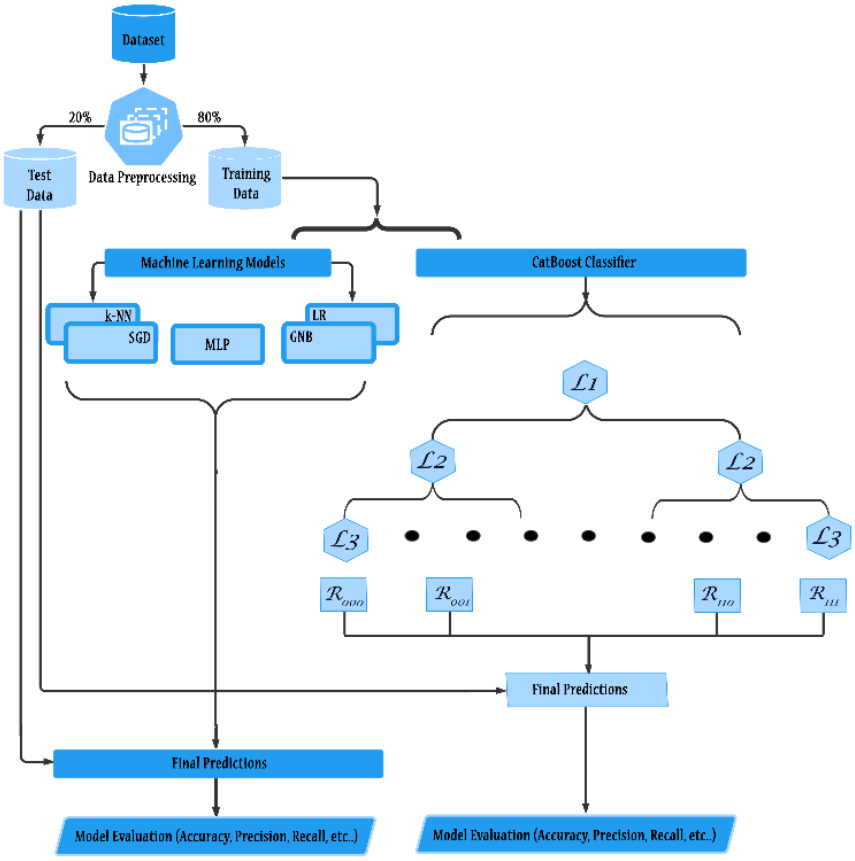

Fig 1. The framework of the proposed system

III. RESULT ANALYSIS

In this paper, the prediction of diabetes at the early stages is done using CatBoost ensemble learning algorithm. We compared the proposed method with different ML methods such as KNN, LR, MLP, GNB, and SGD. This comparison is carried out based on different performance measures such as recall, precision, confusion matrix, f1-score, accuracy, and AUC-ROC.

Experimentation is conducted by using a dataset "Early stage diabetes risk prediction" that is accessible from UCI ML repository [21].The dataset contains various details of a patient such as Age, Gender, Polydipsia, Polyuria, weakness, sudden weight loss, Polyphagia, visual blurring, Itching, Genital thrush, delayed healing, Irritability, muscle stiffness, partial paresis, Obesity, and Alopecia. There is a total of 520 instances in the dataset, among them 320 are positive and 200 are negative instances and is shown in the Fig.2.

A. Accuracy

Classification accuracy is well-defined as the ratio of correctly classified or predicted instances to the total number of instances in the data. It is represented in equation (2).

$$Accuracy = \frac{Correctly\ classified\ instances}{Total\ Number\ of\ Instances} \quad \textbf{(2)}$$

Fig.2. The class balance of the dataset

B. Precision

It is the ratio of correctly classified instances to the sum of both true and false positive instances. It is expressed by equation (3).

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(3)}$$

C. Recall

Recall endeavours to answer what proportion of actual true positives are correctly classified. It is given in equation (4)

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(4)}$$

D. F1-score

It is a mean of sensitivity and precision. It is expressed in equation (5).

$$f1 - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad \textbf{(5)}$$

Here, 8:2 ratio of training and testing data is considered for the analysis. The parameter setting of every ML algorithm and the suggested one is presented in TABLE II. The Confusion matrix of the proposed method with respect to the testing data is presented in Fig.3. From the confusion matrix it is clearly seen that 64 true positive instances, 40 true negative instances out of 104 instances in the testing dataset. Accuracy of the proposed method and various machine learning algorithms are shown in the TABLE III. The proposed method's accuracy is significantly better at 100percent compared to other ML algorithms. MLP, LR, KNN, SGD followed by the proposed method with 95, 93, 84, 84, and 62 percent respectively.

TABLE II
Parameter setting of the proposed method and various machine learning algorithms

Technique	Parameter Setting
CatBoost	iterations=10, learning rate=3
KNN	n_neighbors=7, algorithm='kd_tree'
MLP	activation='relu', batch_size=500, random_state=0, max_iter=100
LR	solver='newton-cg'
GNB	var_smoothing=4.5
SGD	loss='modified_huber',random_state=1, max_iter=600

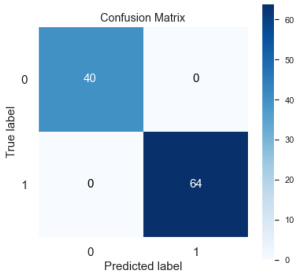

Fig.3. Confusion matrix of the proposed method

TABLE III
ACCURACY OF PROPOSED METHOD AND VARIOUS MACHINE LEARNING ALGORITHMS

Technique	Accuracy	Precision	Recall	f1-score	AUC-ROC
CatBoost	1.0	1.00	1.00	1.00	1.00
KNN	0.84	0.94	0.78	0.85	0.85
MLP	0.93	0.94	0.95	0.95	0.93
LR	0.95	0.95	0.97	0.96	0.95
GNB	0.62	0.62	1.00	0.76	0.50
SGD	0.84	0.79	1.00	0.88	0.79

Evaluation measures such as precision, recall, and f1-score are used to calculate the performance of the classification models rather than accuracy. The proposed method produced the better precision 1.0, followed by LR, KNN, MLP, SGD, GNB with 0.95, 0.94, 0.94, 0.79, and 0.62 respectively. In the case of a recall, SGD, GNB performed equally as the proposed method with 1.0, followed by MLP, LR, KNN with 0.97, 0.95, and 0.78.The proposed method performed well in terms of f1-score with 1.0 followed by LR, MLP, SGD, KNN, GNB with 0.96, 0.95, 0.88, 0.85, 0.78.The area under the curve of the proposed method is 1.0, logistic regression and multi-layer perceptron achieved 0.95, and 0.93 followed by KNN, SGD, and GNB obtained 0.85, 0.79, and 0.50 respectively.

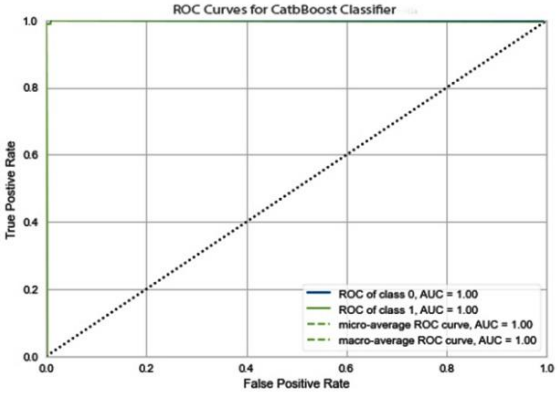

Fig 4. AUC-ROC of proposed method

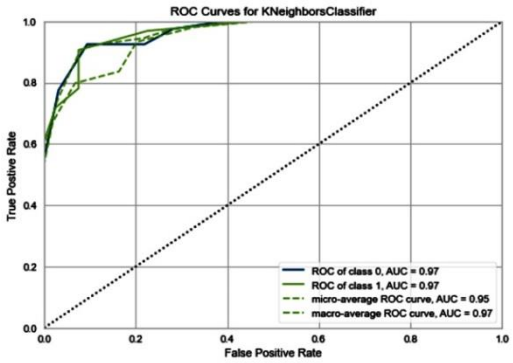

Fig 5. AUC-ROC of K-Nearest Neighbour Classifier

Fig.4 shows the plots of ROC-AUC curve analysis for the proposed method and machine learning classifiers. The performance of CatBoost is outstanding in terms of AUC-ROC. Fig 4 shows the proposed method's micro average ROC curve area 1.0, macro average ROC curve area 1.0, ROC curve areas are 1.0 for both classes.

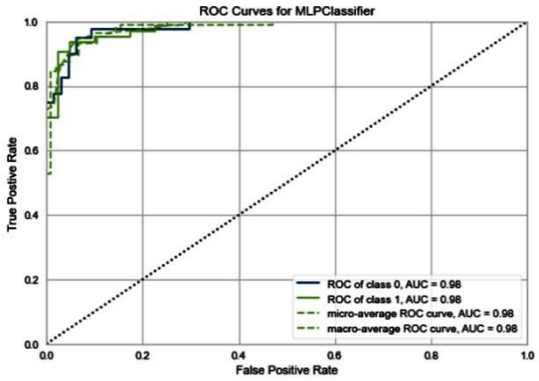

Fig 6. AUC-ROC of MLPClassifier

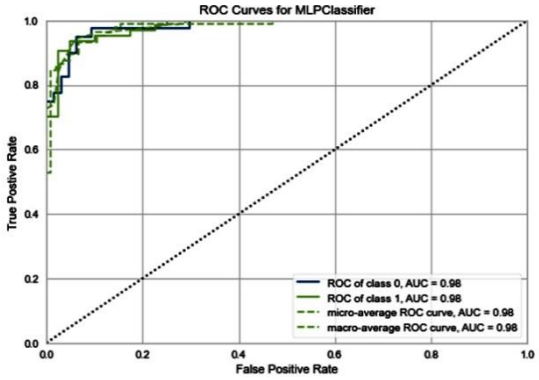

Fig 7. AUC-ROC of Logistic regression

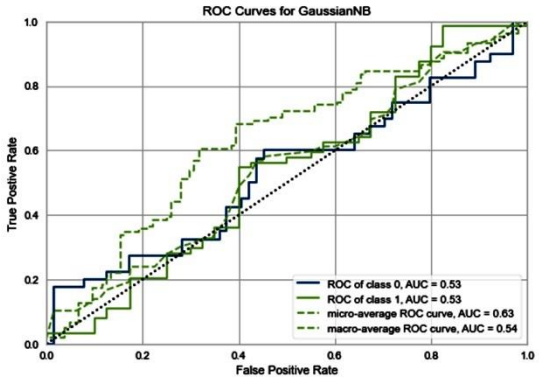

Fig 8. AUC-ROC of Gaussian Naïve Bayes

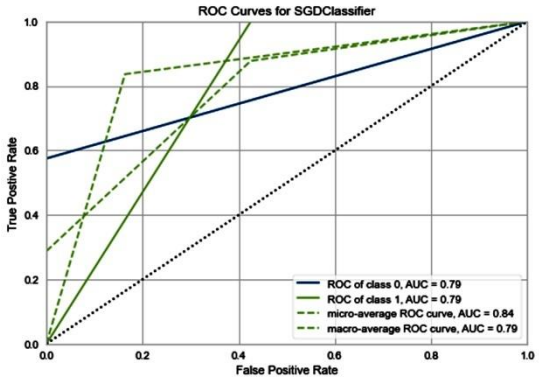

Fig 9. AUC-ROC of SGD Classifier

IV. CONCLUSION

Diabetes is increasing rapidly for all ages. So, the prediction of diabetes at the early stages is very curious and important. The research endeavors presented in this article

are focused on predicting diabetes at an early stage using the CatBoost algorithm and the performance of CatBoost is compared to the performance of various machine learning algorithmssuch as KNN, MLP, LR, GNB, and SGD.The proposed CatBoost ensemble technique outperformed very well compared to different ML algorithms based on accuracy, precision, recall, f1-score, and AUC-ROC. The impediments of the present research work include the proposed model being tried on a dataset with sparse size. As future work, we will test the proposed model and various ensemble learning algorithms by considering he dataset with huge records.

REFERENCES

[1] K. Sowjanya, A. Singhal, and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," in *2015 IEEE International Advance Computing Conference (IACC)*, Jun. 2015, pp. 397–402, doi: 10.1109/IADCC.2015.7154738.

[2] N. Jayanthi, B. V. Babu, and N. S. Rao, "Survey on clinical prediction models for diabetes prediction," *J. Big Data*, vol. 4, no. 1, p. 26, Dec. 2017, doi: 10.1186/s40537-017-0082-7.

[3] M. Maniruzzaman *et al.*, "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers," *J. Med. Syst.*, vol. 42, no. 5, p. 92, May 2018, doi: 10.1007/s10916-018-0940-7.

[4] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images," *Comput. Methods Programs Biomed.*, vol. 157, pp. 121–128, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.004.

[5] M. Komi, Jun Li, Yongxin Zhai, and Xianguo Zhang, "Application of data mining methods in diabetes prediction," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Jun. 2017, no. S Ix, pp. 1006–1010, doi: 10.1109/ICIVC.2017.7984706.

[6] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," in *2015 4th Mediterranean Conference on Embedded Computing (MECO)*, Jun. 2015, pp. 378–382, doi: 10.1109/MECO.2015.7181948.

[7] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[8] A. Choudhury and D. Gupta, "A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques," in *Advances in Intelligent Systems and Computing*, vol. 740, 2019, pp. 67–78.

[9] G. Luo, "Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction," *Heal. Inf. Sci. Syst.*, vol. 4, no. 1, p. 2, Dec. 2016, doi: 10.1186/s13755-016-0015-4.

[10] K. Bhatia, S. Arora, and R. Tomar, "Diagnosis of diabetic retinopathy using machine learning classification algorithm," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Oct. 2016, no. October, pp. 347–351, doi: 10.1109/NGCT.2016.7877439.

[11] G. T. Reddy *et al.*, "An Ensemble based Machine Learning model for Diabetic Retinopathy Classification," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Feb. 2020, pp. 1–6, doi: 10.1109/ic-ETITE47903.2020.235.

[12] A. Mancini *et al.*, "Machine learning models predicting multidrug resistant urinary tract infections using 'DsaaS,'" *BMC Bioinformatics*, vol. 21, no. S10, p. 347, Aug. 2020, doi: 10.1186/s12859-020-03566-7.

[13] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 419–428, Jun. 2020, doi: 10.1007/s41870-018-0270-5.

[14] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in *Advances in Intelligent Systems and Computing*, vol. 992, 2020, pp. 113–125.

[15] B. P. Nguyen *et al.*, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Comput. Methods Programs Biomed.*, vol. 182, p. 105055, Dec. 2019, doi: 10.1016/j.cmpb.2019.105055.

[16] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 211, Dec. 2019, doi: 10.1186/s12911-019-0918-5.

[17] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," *IEEE Access*, vol. 7, pp. 144777–144789, 2019, doi: 10.1109/ACCESS.2019.2945129.

[18] X. Fei, Y. Fang, and Q. Ling, "Discrimination of Excessive Exhaust Emissions of Vehicles based on Catboost Algorithm," in *2020 Chinese Control And Decision Conference (CCDC)*, Aug. 2020, pp. 4396–4401, doi: 10.1109/CCDC49329.2020.9164224.

[19] G. Huang *et al.*, "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *J. Hydrol.*, vol. 574, no. December 2018, pp. 1029–1041, Jul. 2019, doi: 10.1016/j.jhydrol.2019.04.085.

[20] Y. Zhang, Z. Zhao, and J. Zheng, "CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China," *J. Hydrol.*, vol. 588, no. April, p. 125087, Sep. 2020, doi: 10.1016/j.jhydrol.2020.125087.

[21] F. I. M M, F. Rahatara, R. Sadikur, and B. Yasmin, "UCI Machine Learning Repository: Early stage diabetes risk prediction,"2020. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.