Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 3172–3180, DOI 10.12694/scpe.v25i4.2873

CLASSIFICATION OF DIABETES USING ENSEMBLE MACHINE LEARNING TECHNIQUES

ASHISHA G R,* ANITHA MARY X† AND MAHIMAI RAJA J‡

Abstract. Diabetes is a widespread chronic condition that impacts people all over the globe and requires a clear and timely diagnosis. Untreated diabetes leads to retinopathy, nephropathy, and damage to the nervous system. In this context, Machine Learning (ML) might be used to detect health problems early, diagnose them, and track their progress. Ensemble techniques are a promising approach that combines many classifiers to improve forecast accuracy and resilience. This study investigates the categorization of diabetes using an ensemble machine learning technique known as a voting classifier. Using a variety of classifiers, including Light Gradient Boosting Machine (LightGBM), Gradient Boost classifier (GBC), and Random Forest (RF). The predictions are aggregated using voting methods to get a final classification result. The research is carried out using two benchmarking datasets: the Pima Indian Diabetes Dataset (PIDD) and the German Dataset. The Boruta technique is used to choose the best attributes from the datasets, while the Random Over Sampling approach balances the range of classes and eliminates abnormal data using the interquartile range approach. The findings showed that the combination of the Boruta feature selection algorithm and ensemble Voting Classifier performed better for both PIDD and German datasets with an accuracy of 93% and 90% respectively. These algorithms are evaluated and the maximum accuracy is produced using the combination of the Boruta feature selection algorithm and ensemble Voting Classifier. This research helps medical professionals in the early prediction of diabetes, reducing physician's time.

Key words: Machine Learning, Ensemble Voting Classifier, Random Over Sampling, Diabetes, Gradient Boost.

1. Introduction. Diabetes is a metabolic disorder characterized by elevated blood glucose levels. Insulin carries glucose from blood arteries to tissues, where it is converted into energy. The body of a diabetic patient is unable to produce enough insulin. Pre-diabetes, gestational diabetes, Type 1 diabetes (T1D), and Type 2 diabetes (T2D) are the four forms of the disease. In type 1 diabetes (T1D), the immune system of the patient targets and kills the insulin-secreting beta cells in the pancreas [1].

To manage these diseases, additional investigation is necessary. This research work discusses diabetes, one of the long-term diseases. In T2D, the patient's insulin production minimizes, causing high blood glucose levels. Recent researches indicate that 80 percent of T2D can be prevented with early detection.

Pre-diabetes is a term used to describe a condition in which blood glucose levels are elevated but not adequate to be classified as Type 2 Diabetes. Women with elevated blood glucose levels during pregnancy are referred to as having gestational diabetes. According to the International Diabetes Federation (IDF), India has the second-highest number of diabetic patients worldwide, behind China. There are currently an estimated 77 million diabetic patients in India, and that figure is predicted to rise to 134 million by 2045. According to statistics, 17.5% of world diabetic patients were in the Indian population. Diabetes must be treated as soon as quickly as possible to avoid its negative effects. Diabetes is an unsafe medical condition, hence it is critical to diagnose it using an automated technique.

The use of machine learning techniques to improve illness categorization and diagnosis has gained popularity in the medical business [2]. Many machine learning (ML) algorithms have been developed in recent studies for categorizing diabetes data, including Support Vector Machine (SVM) [3], Decision Tree (DT) [4], XGBoost classifier (XGB) [5], etc. But DT achieves lower accuracy due to imbalances in data and the absence of choosing

*Electronics and Instrumentation Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India (ashisha@karunya.edu, ashishaagr@gmail.com, Corresponding Author)
†Robotics Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India (anithamary@karunya.edu)
‡UG Scholar, Computer Science Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India (mahimairaja@karunya.edu.in)

features method. The SVM algorithm is prone to data imbalance, which results in low data accuracy when one group has greater samples than another. If the data includes outliers and missing values, the SVM approach may produce incorrect classification. The poor quality of data obtained from various sources reduced the effectiveness of the XGBoost classifier, resulting in the low accuracy of the model.

The primary purpose of this research is to identify efficient prediction strategies among the many machine learning approaches that can benefit healthcare professionals and hospitals. This work could act as a foundation for researchers interested in learning more about diabetes research. Presenting the right approach for handling missing values, outliers, and unbalanced data sets is the main goal of the proposed approach. It also concentrated mainly on selecting the best feature extraction approach 3 for identifying the most essential features, as well as introducing the suitable ensemble method for diabetes prediction. Light Gradient Boosting Model, Gradient Boosting Classifier, Random Forest, and ensemble voting classifier are utilized in this study to examine PIDD and German datasets. The research also includes an effectively structured assessment of the models used.

The structure of the article is as follows. Section 4 presents the proposed method, which comprises Interquartile Range (IQR), random oversampling, and ensemble ML classifier techniques. Section 2 contains the relevant study based on ML methods in diabetes classification. Section 3 explains the benchmarking PIMA and German datasets. Section 5 shows the performance evaluation of selected attributes, with a comparison to past approaches, and Section 6 concludes the research.

2. Literature Review. Several studies applied the ML approach to classifying diabetes using the PIDD dataset [6]. Mushtaq et al introduced a novel method to classify DM using ML techniques. Initially, the dataset was balanced using the oversampling strategy SMOTE, and outliers were identified using the Interquartile Range (IQR) technique. The accuracy and other performance indicators were then tested using several classification algorithms, including NB (Naïve Bayes), KNN (K closest neighbour), RF (Random Forest), and GB (Gradient Boost). In this research, the maximum accuracy of 81.5% was achieved for the PIDD dataset [7]. Rajeswari et al proposed the Logistic Regression (LR) and SVM to classify diabetes. Data for this research was collected from NC University. SVM achieved 82% accuracy for the training dataset and for the testing dataset, and it was 75% accurate. [8].

A technique using SDKNN for diabetes categorization was proposed by Patra et al. In this investigation, the PIMA dataset was used and they split the dataset into 90% training and 10% testing. An accuracy of 83.2% was demonstrated by the suggested SDKNN method [9]. Kumari et al proposed a diabetes prediction paradigm based on Decision Trees (DT). The author used 200 patient data from a medical health center and compared the DT with SVM, NB, and KNN. The experiment's findings reveal that the results of the DT are better than the other methods [10]. An ensemble approach was proposed by Saloni [11] to predict diabetes. This work uses the PIMA diabetes dataset. The dataset's data were divided into 70% training and 30% testing categories. The execution was done in the Python platform. The work utilized Gradient Boost, XGBoost, CatBoost, AdaBoost, SVM, NB, LR, and Random Forest. The result of the work shows that LR outperforms the other techniques.

Joshi et al. predicted diabetes using DT and LR. The experiment made use of the PIMA diabetes dataset. In this research, the significant features were selected using a classification tree and obtained a 78.26% accuracy rate [12]. A novel approach to diabetes prediction based on XGBoost and RF ML algorithms was presented by Barik et al. The PIMA dataset was used in the course of the research. The execution of the research was done in the Python platform. XGBoost algorithm performs better than the RF algorithm and achieves 74% accuracy [13]. GA hybrid learning method was proposed by Tan et al., [14]. This research used the collected dataset for diabetes prediction. The important features of the study were extracted using the DT method.

Azad et al suggested a diabetes classification method based on the PMSGD algorithm. The PIMA dataset was used in this research. The suggested model achieved 80.70% accuracy [15]. Singh et al introduced an evolutionary NSGA-II model for the classification of T2D. The research made use of the PIMA dataset and its implementation made use of MATLAB. The suggested method achieves an accuracy of 83.8% [16]. Numerous machine learning approaches and ensemble methods have been employed to predict diabetes, as the publications on this subject show, but none of these systems have been able to attain an accuracy of greater than 85% [17,18].

ML has currently endured extensive research to properly detect the presence of diabetes in its early stages. Despite improved accuracy, they failed to tackle bias handling, database balancing, outlier removal, and choosing

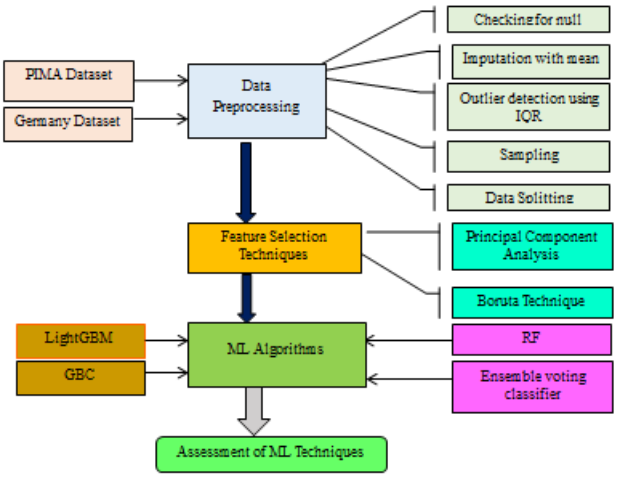

Fig. 4.1: Visibility approximation algorithm

essential diabetes attributes. As a result, it was evident that the result had not been verified because the presence of bias, an unbalanced dataset, and the lack of an attribute selection process would result in incorrect accuracy. Dealing values that are missing, concentrating on finding outliers, dataset balancing, and emphasizing selecting features approaches have been recognized as critical phases for diabetes classification investigation in the literature. ML models usually struggle to produce better results during the training phase when there are more missing values in the dataset. Detecting and working with missing data for every input attribute is an essential phase in the pre-processing step. Many survey studies on diabetes research reveal that ensemble models will produce better accuracy than common ML models.

The primary focus of the proposed approach is to provide a suitable method for handling missing values, outliers, and unbalanced data sets. It also concentrated mainly on selecting the best feature extraction approach for identifying the most essential features, as well as introducing the suitable ensemble method for diabetes prediction. Light Gradient Boosting Model, Gradient Boosting Classifier, Random Forest, and ensemble voting classifier are utilized in this study to examine PIDD and German datasets. The research also includes an effectively structured assessment of the models used.

3. Datasets. There are two different datasets used in the proposed research. They are Germany Dataset [19] and PIDD (PIMA Indian Diabetes Dataset) [20]. The NIDDK Center generates this dataset. It includes medical markers for blood pressure, body mass index (BMI), pregnancy, skin thickness, parent's history of diabetes, insulin, age, glucose levels, and results. The dataset comprises 769 imbalanced data values. Diabetes data was obtained by the hospital in the German city of Frankfurt and acquired via Kaggle.

4. Methodology. As shown in Figure 4.1, the model stream is divided into six phases. The entire model is implemented in Google Colab with the programming language Python. The proposed work is an appropriate technique for dealing with values that are missing, outliers, and data sets that are imbalanced. Also, selecting the best feature extraction approach for identifying the most essential features, as well as introducing the suitable ensemble method for diabetes prediction.

4.1. Data Preprocessing. The procedure includes loading and organizing data to perform training, validating, and testing. After importing the dataset, the shape of the dataset was verified. Duplicate rows from the datasets were identified and dropped. Through a procedure known as "imputation with mean," all of the datasets' null data were replaced with the mean values of the respective characteristics. To find the outlier, Inter Quartile Range (IQR) algorithm was applied to the data. IQR is a statistically dispersed measure derived from the difference between Q3 and Q1 in a dataset. It is less susceptible to outliers because it shows the distribution of the middle 50% of the data. Outliers are detected in the IQR method by taking values greater

than Q1-1.5IQR and Q3+1.5IQR. When there are disparities in the dataset between the classes, the model may be biased during training and perform well for the majority class but badly for the minority class. For this study, the Random oversampling strategy is chosen to solve the problem of class imbalance. Since random oversampling involves reproducing samples from the minority group. The model obtains its knowledge directly from the data and it is suitable for a wide range of ML techniques. The dataset has two sets of data: training and testing. 20% of the datasets are chosen for testing and 80% for training using the train test split function.

4.2. Feature Selection Techniques. It entails selecting the most relevant attributes from the dataset that have a positive relationship with the prevalence of diabetes. In this system, two feature extraction strategies are proposed: Boruta Algorithm and PCA (Principal Component Analysis).

4.3. Boruta Algorithm. Boruta creates shadow features that are compared to the importance of the initial features in the dataset. Utilized as a comparative tool is the importance of the shadow features, which stand in for data noise. The steps below are the process of the Boruta method.

Step 1: Create a new feature matrix. Each matrix feature, B, is used to generate the shadow feature matrix B_D. Merge the shaded matrix B_D to the original matrix B, to construct the new matrix Bn. Equation (4.1) gives the expression for B_n.

$$B_n = [B, B_D] \tag{4.1}$$

Step 2: Train the model using the newly generated matrix, B_n.

Step 3: Determine the Z-score of the largest shadow feature, D_{max} for the matrix N and the attribute matrix, N_D.

Step 4: Check the state to determine the important and unimportant features. If D_{max} is greater than the Z-score, then it is considered an unimportant feature, and the features in which the Z score is greater than D_{max} is considered an important feature.

Step 5: Remove all shadow attributes.

Step 6: Carry out the above steps until all of the important features have been chosen.

4.4. Principal Component Analysis. The PCA approach is used for dimension reduction and to generate related features to describe the type of data. All these elements are linear mixtures of the primary attributes, ordered by their capacity to incorporate the greatest variance. To use PCA in a database, the data must first be standardized for all features to have an even scale.

Mean centered, M_{cen} is estimated as (4.2),

$$M_{cen} = M - mean(M) \tag{4.2}$$

where M represents the dataset matrix.

The normalized data's covariance matrix (4.3) is then determined, and it indicates how distinct attributes vary from the other.

$$CV = \frac{1}{(m-1)} M_{cen}^T M_{cen} \tag{4.3}$$

where CV indicates the covariance matrix, and m represents the data.

The dimension of the dataset is reduced from m to l by maintaining only the top l components, while l is frequently much less than m.

4.5. ML Techniques. After being processed with feature selection methods, the dataset is fed to ML models. The following classifiers are employed in the proposed work: LightGBM, GBC, RF, and ensemble voting classifier.

4.6. Light Gradient Boosting Machine. LightGBM algorithm determines a function that maps the input features to the target feature. Gradient-based One-Side Sampling is an advanced method that is used to

reduce the amount of data of the gradient boosting method. The objective function of LightGBM is expressed in (4.4).

$$OJ(V) = \Sigma_{j=1}^{m} n(z_j, z_j^{1^{u-1}}) + \Sigma_{i=1}^{D} f_i(x_j; V_i) + \Sigma_{j=1}^{D}\ \Omega(f_i) \tag{4.4}$$

where V represents the set of parameters to be learned by the ML algorithm, D denotes the number of DT in total, x is the input attributes and z represents the target attribute, n is the loss function. This method builds an integrated DT to estimate f(x) using a gradient boosting technique. To carry out the classification, the combination of DT is trained with x. Estimation of DT P_D and estimated values of DT P_p is presented mathematically in equations (4.5)-(4.6).

$$P_D = \Sigma_{i=1}^{D} \Omega(f_i) \tag{4.5}$$

$$P_p = \Sigma_{j=1}^{m} n(z_j, z_j^{1^{u-1}}) + \Sigma_{i=1}^{D} f_i(x_j; V_i) \tag{4.6}$$

Equation (4.7) expresses the loss function of LightGBM.

$$f(n) = \Sigma_{j=1}^{m} n(z_j, z_j^{1^{u-1}}) + \Sigma_{i=1}^{D} f_i(x_j; V_i) - Z_j \tag{4.7}$$

The loss function of the LightGBM classifier allows the method to find the variation between the predicted values and target values during learning.

4.7. Gradient Boost Classifier. Gradient boosting classifier (GBC) operates by building a collection of DTs, with each DT attempting to correct the faults made by the previous DT in the sequence. This algorithm's result is a sum of all the DT's predictions. Data taken for training is given in equation (4.8).

$$T = \Sigma_{j=1}^{N} (x_j, z_j) \tag{4.8}$$

An approximation function of GBC is expressed in the below equation (4.9).

$$A_r(x) = A_{r-1} + R_r H_r(x) \tag{4.9}$$

where R_r represents the weight of the r^{th} approximation function, $H_r(x)$. Pseudo value (4.10)-(4.11) is used to train $H_r(x)$.

$$T = \Sigma_{j=1}^{N} (x_j, S_o) \tag{4.10}$$

$$S_o = \frac{\partial f(z_j, F(x))}{\partial F(x)} \tag{4.11}$$

GBC constructs a predictive model in the form of a sequence of weak DTs and then incorporates their predictions to produce a more powerful overall model.

4.8. Random Forest. A hybrid machine learning method called Random Forest (RF) combines many DTs to generate predictions. The number of random DTs (4.12)-(4.14) is represented as follows.

$$R_r(x, T_r) = EO'[R_r(x, O', T_r] \tag{4.12}$$

$$R_r(x, O') = \frac{\Sigma_{j=1}^{m} Z_j(x_j \epsilon B_m(x, O'))}{\Sigma_{j=1}^{m} (x_j \epsilon B_m(x, O'))} 1E_r(x, O') \tag{4.13}$$

Here, the randomizing element is represented by O' and $B_m(x, O')$ denotes the rectangular cell of $R_r(x, O')$.

$$EO'[R_r(x, O')] = EO' \frac{\Sigma_{j=1}^{m} Z_j(x_j \epsilon B_m(x, O'))}{\Sigma_{j=1}^{m} (x_j \epsilon B_m(x, O'))} 1E_r(x, O') \tag{4.14}$$

The approach begins by randomly selecting a set of features and sampling the data. This classifier's DT uses randomly picked features to train every single tree. The Random Forest method generates an estimated output by gathering votes from all the trees.

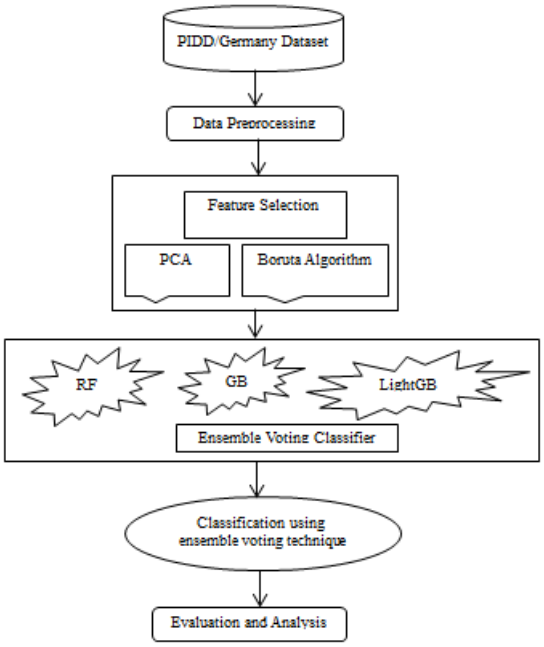

Fig. 4.2: Proposed ensemble voting method for diabetes prediction

Table 4.1: Parameters used for Experimental Analysis

Parameters	LGBM	GBC	RF
Step size	0.01, 0.05, 0.1, 0.2, 0.5	50, 100, 200	-
Maximum depth	3, 5, 7, 10	3, 5, 7	None, 10, 20
Number of estimators	50, 100, 200, 300, 500	50, 100, 200	50, 100, 200
Random state	-	-	42

4.9. Ensemble Voting Classifier. The voting technique employs a pair of voting methods: either soft or hard voting. In hard voting, the final prediction is determined by a majority vote in which the combiner selects the class estimate that appears most frequently from the models that comprise the foundation.

$$y = argmax_e(\Sigma_{j=1}^{M}||(I_j(Z) = e)) \tag{4.15}$$

where y is the voting algorithm's prediction, M indicates the total number of models, and e signifies the class label. $I_j(z)$ is the prediction of the j^{th} model for the input z.

Figure 4.2 shows the proposed model, which incorporates the Random Forest, GB, and LightGBM algorithms. A voting technique has been used to provide the probability of every targeted parameter. The initial training data and data points are then shuffled, and these data points are delivered to Random Forest, GB, and LightGBM techniques. The mathematical expressions of the performance measures are as follows. Table 4.1 shows the experimental parameters.

5. Result and Performance Evaluation. The proposed methodology combines three ML classifiers: RF, XGB, and LR with an ensemble voting technique. Experiments were conducted using the PIDD and Germany Diabetes datasets. To assess the ensemble voting technique's reliability and efficacy, accuracy, precision, and recall were assessed.

Table 5.1: Comparison of ML models for PIDD

Models	Algorithm	Accuracy	Precision	Recall
Light GBM	PCA	87	84	92
GBC	PCA	85	85	84
RF	PCA	87	85	89
Ensemble voting model	PCA	90	91	96
Light GBM	Boruta Algorithm	92	90	91
GBC	Boruta Algorithm	89	88	90
RF	Boruta Algorithm	90	91	92
Ensemble voting model	Boruta Algorithm	93	91	96

Table 5.2: ML model comparison for the Germany Diabetes Dataset

Models	Algorithm	Accuracy	Precision	Recall
Light GBM	PCA	87	80	93
GBC	PCA	84	82	88
RF	PCA	86	82	93
Ensemble voting model	PCA	88	83	95
Light GBM	Boruta Algorithm	89	85	95
GBC	Boruta Algorithm	88	88	95
RF	Boruta Algorithm	88	87	96
Ensemble voting model	Boruta Algorithm	90	85	97

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositivies + FalseNegatives} \tag{5.1}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositivies} \tag{5.2}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5.3}$$

Table 5.1 describes the comparison of ML algorithm results for the PIDD dataset. It shows that the Boruta algorithm and ensemble voting approach outperformed all other machine learning models with an accuracy of 93%. Figure 4.2 shows a comparison of ML algorithm outcomes for the German diabetes dataset.

Table 5.2 shows that the Boruta algorithm and an ensemble voting approach beat the German diabetes dataset. It achieved the highest accuracy of 90% when compared to other machine learning models.

In this study, diabetes was predicted by handling missing values and outliers, rebalancing the data that was out of balance, and selecting the best feature selection technique to extract the important characteristics (BMI, blood pressure, etc.) from the dataset. Next, by evaluating the performance of suitable ML techniques the reliability of models will be tested. The proposed model was compared with the existing research works. Table 5.3 presents the comparison, which shows that the proposed method outperformed prior diabetes prediction studies in terms of performance.

6. Conclusion. The proposed research determines the greatest accuracy and methodology for predicting patients with diabetes. The ensemble voting classifier technique is used in the proposed system. It is based on the combination of three ML models: Random Forest, XGB, and Logistic Regression. The PIDD was used for testing a proposed method and the developed model was also applied to the Germany diabetes dataset.

Table 5.3: Comparison of other diabetes prediction models

ML Classifier	Accuracy
SVM [8]	75%
KNN [9]	83.2%
DT [10]	73%
DT [12]	78.26%
XG Boost [13]	74%
DT [14]	85%
PMSGD [15]	80.70%
NSGA – II [16]	83.8%
Soft Classifier [11]	79.08%
Proposed model - PIDD	93%
Proposed model - German Dataset	90%

Comparing the Precision-recall curve with the ROC curve, the proposed method performs better. The ensemble voting Method has achieved 93% accuracy on the PIDD and 90% accuracy on the Germany Diabetes Dataset. The overall investigation of this research shows that the proposed ensemble voting classifier can be used for both PIMA and German Diabetes Datasets. This proposed system can be improved by introducing various deep-learning techniques.

REFERENCES

[1] K. Kangra and J. Singh, *"Comparative analysis of predictive machine learning algorithms for diabetes mellitus," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 3, pp. 1728–1737, Jun. 2023.*

[2] J. J. Khanam and S. Y. Foo, *"A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021*

[3] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," in Procedia Computer Science, Elsevier B.V., 2018, pp. 1578–1585.

[4] H. M. Deberneh and I. Kim, "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm," International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 3317, vol. 18, no. 6, p. 3317, Mar. 2021.

[5] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Comput Appl, pp. 1–17, Mar. 2022.

[6] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front Genet, vol. 9, Nov. 2018.

[7] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," Mobile Information Systems, vol. 2022, 2022.

[8] S. V. K. R. Rajeswari and Vijayakumarponnusamy, "Prediction of Diabetes Mellitus Using Machine Learning Algorithm," 2021. [Online]. Available: http://annalsofrscb.ro

[9] R. Patra and B. khuntia, "Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique," IOP Conf Ser Mater Sci Eng, vol. 1070, no. 1, p. 012059, Feb. 2021.

[10] K. Srujana Kumari, "Performance Analysis of Diabetes Mellitus Using Machine Learning Techniques," 2021.

[11] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40–46, Jun. 2021

[12] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," Int J Environ Res Public Health, vol. 18, no. 14, Jul. 2021

[13] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," in Smart Innovation, Systems and Technologies, Springer Science and Business Media Deutschland GmbH, 2021, pp. 399–409.

[14] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early Risk Prediction of Diabetes Based on GA-Stacking," Applied Sciences (Switzerland), vol. 12, no. 2, Jan. 2022

[15] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," in Multimedia Systems, Springer Science and Business Media Deutschland GmbH, Aug. 2022, pp. 1289–1307.

[16] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of

diabetes mellitus," Biocybern Biomed Eng, vol. 40, no. 1, pp. 1–22, Jan. 2020.
[17] M. Fatima, S. Srivastav, and A. C. Mondal, "Prenatal stress and depression associated neuronal development in neonates," International Journal of Developmental Neuroscience, vol. 60. Elsevier Ltd, pp. 1–7, Aug. 01, 2017.
[18] A. Husain and M. H. Khan, "Early diabetes prediction using voting based ensemble learning," in Communications in Computer and Information Science, Springer Verlag, 2018, pp. 95–103.
[19] Germany Dataset" Accessed: Feb. 22, 2024. [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes?resource=download
[20] "Pima Indians Diabetes dataset" Accessed: Feb. 22, 2024. [Online]. Available: https://data.world/uci/pima-indians-diabetes

Edited by: Dhilip Kumar V
Special issue on: Unleashing the Power of Edge AI for Scalable Image and Video Processing
Received: Nov 17, 2023
Accepted: Apr 1, 2024