

A Novel Ensemble Learning Approach for Diabetes Prediction in Imbalanced Datasets

Djalila Boughareb^{1*}, Said Bouteldja² and Hamid Seridi¹

¹Department of Computer Science, LabStic Laboratory, University 8 May 1945, Guelma-Algeria

²Department of Computer Science, University of 8 May 1945, Guelma-Algeria

Abstract

The incapacity of the body to effectively make or use insulin results in diabetes, a chronic illness. Over time, this illness may cause harm to the kidneys, blood vessels, heart, eyes, nerves, and kidneys. Timely treatment is essential to stop the progression of diabetes and requires early detection. We provide a hybrid machine learning strategy in this work that predicts diabetes by combining two strong algorithms. XGBoost (eXtreme Gradient Boosting)-based voting classifier and bagging classifier are the two main components of our system. We evaluated our model using three distinct datasets: the Pima Indian diabetes dataset (PIDD), its extended version, and the Frankfurt Hospital Germany Diabetes Dataset (FHGDD). In comparison to individual algorithms (XGBoost, Bagging with Decision Tree) and other ensemble methods (Voting Classifier, HM-Bag Moov Voting Classifier, XGBoost+ Data feature stitching, and Soft), our experimental results show that our proposed approach achieved higher accuracy of 92.7%, precision of 97.1%, recall of 81.7%, and an F1 score of 88.7%. Therefore, our results imply that the hybrid machine learning approach that has been suggested can be a dependable tool for the early diagnosis of diabetes, resulting in better patient outcomes and more prompt and efficient therapies.

Keywords

Diabetes Prediction, Bagging, XGBoost, Decision Tree, Pima, FHGDD, Machine Learning, Ensemble Learning.

1. Introduction


Diabetes is becoming more commonplace worldwide, which presents a serious public health risk. The early identification and treatment of diabetes can improve the quality of life for persons with the disease by preventing or delaying the onset of complications. Conventional techniques for identifying diabetes, such blood glucose monitoring, can be costly, intrusive, and time-consuming. Because machine learning techniques automate the process and enable more precise and effective disease identification, they hold the potential to completely transform the diagnosis and management of diabetes. Large-scale datasets and sophisticated algorithms are used to find patterns and risk variables that could be hard for human experts to find.

The 13th International Conference on Research in computing at Feminine, May 20-21 2024, Constantine, Algeria

* Corresponding author.

† These authors contributed equally.

✉ boughareb.djalila@univ-guelma.dz (D. Boughareb); bouteldja.said@univ-guelma.dz (S. Bouteldja); seridi.hamid@univ-guelma.dz (H. Seridi)

 0000-0002-3432-2548 (D. Boughareb); 0000-0002-0236-8541 (H. Seridi)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

There are two main forms of diabetes: Type 1: Usually identified in young people, it is caused by the immune system targeting cells that produce insulin. It requires daily insulin pumps or injections to control. If left untreated, it can result in heart disease, retinopathy, neuropathy, and nephropathy. Type 2: More frequent, usually affecting older persons, it is characterized by either insufficient or inefficient insulin production or utilization by the body. controlled via dietary adjustments, prescription drugs, and occasionally insulin. If left unchecked, the hazards are comparable to those of Type 1. Because gestational diabetes during pregnancy results in decreased insulin sensitivity and elevated blood sugar levels, it raises the risk of Type 2 diabetes later in life for both the mother and the child.

Numerous researches have examined the prediction of diabetes by taking into account a range of characteristics, including lifestyle, Electronic Health Records (EHRs), environment, and molecular attributes. Based on prior experience and medical records that contain patient conditions and vital signs. The most widely used dataset in these studies is the Pima Indians Diabetes Database [8; 1;6;4], which has 768 samples, 268 of which are patients with diabetes, and 8 independent factors that are used to identify whether a patient has diabetes.

Using the PIMA Indian Dataset, authors in Ref. [4] used Decision Tree, SVM, and Naive Bayes classifier techniques to predict diabetes. They discovered by 10-fold cross-validation that Naive Bayes had the best accuracy, at 76.30%. On the basis of the same dataset, authors in Ref. [6] created a predictive model with XGBoost and feature stitching, which produced an astounding 80.2% accuracy and identified important predictive factors like diabetes pedigree function, glucose, age, pregnancies, and BMI. A new method for classifying diabetes called HMBag Moov was introduced by Bashir et al. [1]. It was compared with a number of other approaches, including as Naive Bayes, SVM, Logistic Regression, etc. The accuracy of the HMBag Moov Voting Classifier was 77.21%, even though it did not employ hyperparameter tweaking or cross-validation and only evaluated a small number of ensembling techniques.

Despite the extensive utilization of the Pima dataset and the notable success in prediction outcomes derived from it, a significant challenge arises due to class imbalance within the dataset. Specifically, there exists a prevalence of healthier patients compared to those afflicted. This inherent imbalance poses a substantial hurdle for classification algorithms, as the minority classes are overshadowed. Consequently, even if misclassifying every minority instance, the algorithm could still exhibit low error rates [11].

One potential remedy for this problem is data augmentation [2], which entails raising the minority class's representation in order to avoid overfitting. Also, previous studies on imbalanced datasets, including those focused on biomedical data, affirm the efficacy and reliability of ensemble learning methods in alleviating the challenges posed by class imbalance [14-16]. For instance, in order to mitigate the impact of class imbalance [15] present Sample and Feature Selection Hybrid Ensemble Learning (SFSHEL), a novel approach aimed at tackling the complexities posed by imbalanced datasets in classification tasks. Base learner weights are assigned through validation, enabling weighted voting for predictions. SFSHEL-RF, based on random forest, shows superior performance on clinical datasets, validating its effectiveness. In response to the limitations of traditional classification methods, [14] introduced an innovative ensemble learning framework tailored for medical diagnosis with imbalanced data. Comprising

three phases—data pre-processing, base classifier training, and final ensemble—the proposed approach was evaluated across nine imbalanced medical datasets. Results demonstrate its superiority over other state-of-the-art classification techniques. Furthermore, [16] introduced a multi-criteria ensemble training method tailored for imbalanced datasets, simultaneously optimizing precision and recall. It presents a set of Pareto optimal solutions, allowing the end-user to select the most suitable solution based on their preferences. Results confirmed the method's utility, ensuring high-quality outcomes comparable to single-criterion optimization.

The primary aim of this project is to address the challenge of imbalanced datasets in diabetes diagnosis by leveraging data augmentation techniques and ensemble learning models. The ultimate goal is to enhance patient outcomes and alleviate the strain on healthcare systems caused by diabetes. To achieve this objective, the project proposes an effective and efficient system capable of analyzing clinical data to accurately identify diabetes or determine if an individual is in the pre-diabetic stage. The project involves the utilization of both bagging and XGBoost (eXtreme Gradient Boosting) algorithms for diabetes prediction, using an expanded version of the Pima dataset previously generated via a GAN (Generative Adversarial Network) algorithm [2], in addition to the Frankfurt Hospital Germany Diabetes Dataset (FHGDD) [17].

Bagging also known as bootstrapping is a prominent ensemble learning technique that combines predictions from multiple decision trees trained on different subsets of the same dataset [5]. Another variant, ensemble bagging, involves constructing a collection of classifiers that iteratively apply a specific algorithm to various versions of the training dataset [9]. These ensemble methods are valuable tools for enhancing predictive performance and addressing overfitting in classification tasks.

The proposed combined approach is designed to improve the accuracy and robustness of the predictive model. Furthermore, the project seeks to advance the field by comparing its proposed methodology with various state-of-the-art research studies, thereby providing insights into the efficacy and superiority of the suggested approach.

The remaining sections of the paper are arranged as follows: In Section 2, the study technique is explained; in Section 3, the evaluation's specifics are outlined and the results are discussed; and in Section 4, the article is concluded and potential future directions are discussed.

2. Materials and Method

This section will provide an explanation of the research technique.

2.1. Dataset

Our research utilized the PIMA Indian Diabetes dataset [8], originally gathered by the National Institute of Diabetes and Digestive and Kidney Diseases. Widely recognized as a benchmark dataset, it has been extensively employed in machine learning studies to assess the efficacy of various classification and prediction algorithms in diabetes prediction. We utilized two versions of this dataset: one with 768 instances and an extended version containing 1602 instances [2]. Additionally, we incorporated the Frankfurt Hospital Germany Diabetes Dataset (FHGDD) [17] into our analysis. Each dataset consists of exactly eight attributes, including:

- Pregnancies: The total number of pregnancies.
- Blood glucose level measured in an oral glucose tolerance test after two hours.
- Blood Pressure: Miligrams of Hg for the diastolic blood pressure.
- SkinThickness: Skin fold thickness (mm) of the triceps.
- Insulin: Two-hour serum level.
- Body Mass Index (BMI).
- DiabetesPedigreeFunction: The function of the diabetes pedigree.
- Age: The number of years.

"Outcome," the only dependent variable in the dataset, had binary values of either 0 or 1. The dataset was split into two sets, a training set and a testing set, using a 70:30 ratio in order to assess the performance of the model. Using four folds of cross-validation on the testing set, the model was assessed after being trained on the training set.

The used extension of Pima was generated recently in a previous work [2] where we used Generative Adversarial Networks (GANs) for data imputation, a technique introduced by Goodfellow et al. in 2014 [13]. GANs employ a game-theoretic framework wherein a generator network competes against a discriminator network. The generator's objective is to create synthetic data samples resembling those from the training set, while the discriminator aims to distinguish between real and synthetic samples. The experiment generated 1602 data lines, including 602 authentic and 1000 synthetic lines.

The Frankfurt Hospital Germany Diabetes Dataset (FHGDD) serves as another resource in diabetes prediction and classification research. Comprising the same attributes as the PIDD but with an expanded size of 2000 instances, it provides a rich data source for analyzing diabetes-related factors. Table 1 illustrates the distribution of instances across each class, delineating the counts for Class 1 (diabetic) and Class 2 (non-diabetic).

Table 1

The distribution of instances among the classes within the three datasets.

Dataset	# Class 1 (diabetic)	#Class 2 (non- diabetic)
PIDD	268	500
Extended PIDD	801	801

3. Proposed Approach

Our methodology combines XGBoost (eXtreme Gradient Boosting) and bagging techniques. XGBoost, a gradient boosting decision tree (GBDT) algorithm, efficiently iterates weak models to create a strong one, proven effective in prediction tasks [3], [7], [10]. It optimizes model parameters by merging regression trees and gradient descent. Initially, the model is initialized with weak learners, typically shallow decision trees. Iteratively, the algorithm fits the gradient of the loss function to the predictions of current weak learners, then trains new weak learners based on this gradient information, adding them to the model. This process repeats until a stopping criterion is met. Predictions are made by aggregating the predictions of all weak learners. The optimization problem in XGBoost combines a loss function, measuring prediction error, and a regularization term, penalizing model complexity to prevent overfitting.

Decision trees, particularly CART (Classification and Regression Tree), are pivotal in XGBoost, where their shallow structure mitigates overfitting risks. XGBoost defines an objective function to optimize during training, comprising a regularization term controlling model complexity and a loss function quantifying prediction accuracy against actual values. For binary classification, the logistic loss function, also called log loss or cross-entropy loss, is employed as the objective function in XGBoost, denoted by equation (1) [12].

$$L = I^n (y_i - \hat{y}_i)^2 + (f) \quad (1)$$

Such as,

L : the objective function.

n : the number of samples.

y_i : the true label.

\hat{y}_i : the predicted label.

(f) : the regularization term, which is a function of the model parameters f.

XGBoost is a powerful machine learning algorithm that trains an ensemble of decision trees iteratively. It utilizes gradients to understand instance deviations and constructs trees to identify patterns efficiently. Weighted updates adjust instance weights based on prediction errors, while ensemble building combines individual predictions using their importance. Regularization techniques control model complexity, preventing overfitting, and control parameters fine-tune its behavior. By aggregating ensemble predictions weighted by

importance, XGBoost produces accurate predictions, making it effective for various machine learning tasks.

In our study, we aimed to boost the accuracy and robustness of our model by integrating Bagging Classifier with XGBoost through a Voting classifier. The Voting Classifier, a form of ensemble classifier, combines predictions from multiple base classifiers via a majority or weighted vote. We adopted hard voting, where the final prediction for an input is determined by the majority vote of individual model predictions. Mathematically, for N individual models (f_1, f_2, \dots, f_N), the final prediction y_{pred} for input x is obtained using equation (2), where argmax selects the class label with the highest number of votes. Figure 1 illustrates the flowchart of the proposed model.

$$y_{pred} = \text{argmax}(\text{sum}(f_i(x))) \text{ for } i = 1 \text{ to } N \quad (2)$$

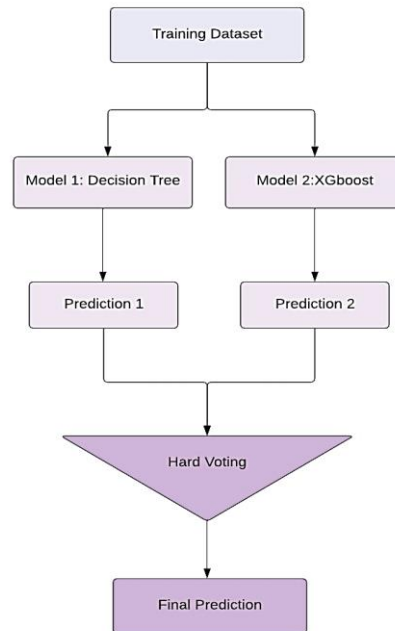


Figure 1: The ensemble learning model architecture

4. Results and discussion

The experimental hardware setup comprises an Intel Core i3-3110M CPU clocked at 2.40GHz, paired with 8 GB of RAM and a capacious 1 TB HDD for storage needs. Python 3.7 was utilized to develop the machine learning model, with the following libraries employed: NumPy as a fundamental tool for mathematical operations, pandas for efficient data loading, and scikit-learn providing a suite of base classifiers.

In the realm of binary classification evaluation for this task, four pivotal terms emerge—True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)—are fundamental for evaluating classifier model performance. TP denotes correctly classified positive instances, TN signifies correctly classified negative instances, FP represents incorrectly classified positive instances, and FN indicates incorrectly classified negative instances. These terms are pivotal in calculating various evaluation metrics such as precision, recall, F1 score, and accuracy, essential for assessing classifier model performance. Precision measures the proportion of correctly identified positives, recall gauges the proportion of actual positives correctly identified, and F1 score offers a balanced assessment of both precision and recall. Accuracy quantifies the overall correctness of the classifier's predictions. Understanding these terms and associated evaluation metrics is crucial for evaluating and enhancing the performance of binary classification models. The formulas for these metrics are referenced in equations (3-6) as follows:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (3)$$

$$\text{Rappel} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

$$\text{F1-score} = \text{TP} / (\text{TP} + 1/2(\text{FP} + \text{FN})) \quad (5)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (6)$$

As demonstrated in Table 2, our approach exhibited strong predictive capability across multiple datasets. Specifically, high accuracy rates of 91% and 92.7% were achieved using the Pima and Pima extended datasets, respectively, indicating the robustness of our model. The precision values for the Pima and Pima extended datasets were notably high at 89.55% and 97.1%, respectively, indicating a high proportion of true positive predictions. Similarly, solid recall values of 81.08% and 81.7% were obtained for the Pima and Pima extended datasets, respectively, demonstrating the model's ability to accurately identify actual positives. Maintaining a balance between recall and precision, crucial in medical diagnostic models, our algorithm delivered F1-scores of 85% and 88.7% for the Pima and Pima extended datasets, respectively, indicating a suitable balance between the two metrics. Furthermore, when applied to the FHGDD dataset, our approach achieved an accuracy rate of 85.6% and a precision of 87.8%, albeit with a slightly lower recall of 68.7%. Nevertheless, the F1-score of 77.1% demonstrates a reasonable balance between precision and recall. These results underscore the effectiveness of our approach in accurately predicting the presence of diabetes across different datasets. For further context and comparison, detailed performance metrics relative to other state-of-the-art methods are provided in Table 3, and Figure 2 respectively.

Let's explore two contrasting scenarios: one featuring a diabetic patient and the other a non-diabetic individual. In the first case, the model identifies the patient, with attributes such as one pregnancy, blood glucose level of 119, blood pressure of 78, skin thickness of 29, insulin level of 180, BMI of 38.19, diabetes pedigree function of 0.53, and age of 25, as non-diabetic. Conversely, the second scenario portrays a patient with four pregnancies, blood glucose level of 129, blood pressure of 70, skin thickness of 18, insulin level of 122, BMI of 29.43, diabetes pedigree function of 1.17, and age of 41, classified as diabetic. This classification suggests that the combined characteristics in the latter set imply a higher likelihood of diabetes, according to the model's interpretation.

Table 2

Outcomes of the proposed method.

Approach	Accuracy	Precision	Recall
XGBoost+ DT on pima	91%	89.5%	81%
XGBoost+ DT on pima extended	92.7%	97.1%	81.7%
XGBoost+ DT on FHGDD	85.6%	87.8%	68.7%

Table 3

Comparison of accuracy achieved with other state of the art works.

Techniques	Research work	Accuracy
XGBoost+ DT	Our model on pima	91%
XGBoost+ DT	Our model on extended pima	92.7%
XGBoost+ DT	Our model on FHGDD	85.6%
Soft voting classifier	[3]	79.08%
XGBoost	[3]	75.75%
Bagging	[3]	74.89%
Random Forest	[3]	77.48%
XGBoost+ Data feature stitching	[4]	80.2%
HM-Bag Moov Voting Classifier	[1]	77.21%
Voting Classifier	[6]	86%

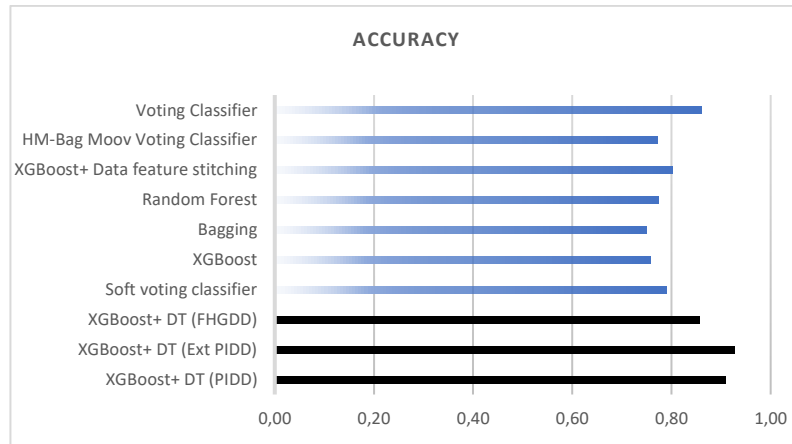


Figure 1: Comparison of the accuracy achieved by the proposed model against other state-of-the-art models.

Combining XGBoost with decision trees for binary classification tasks leverages the strengths of both algorithms. Decision trees offer intuitive interpretations, robustness to noisy data, and the ability to handle both numerical and categorical features effectively. XGBoost enhances performance through boosting, regularization techniques, and scalability, making it suitable for large datasets.

The results obtained from our model using different datasets showcase its effectiveness in predicting diabetes compared to various existing techniques. Our model achieved high accuracy rates when applied to the Pima and extended Pima datasets, respectively. Additionally, when our model was applied to the FHGDD dataset, it achieved a respectable accuracy rate of 85.6%, indicating its applicability across different datasets.

Comparing our results to those of other techniques, we observe that our model outperforms several state-of-the-art methods in terms of accuracy. For instance, the soft voting classifier, XGBoost, Bagging, Random Forest, and XGBoost+ Data feature stitching techniques achieved accuracy rates of 79.08%, 75.75%, 74.89%, 77.48%, and 80.2%, respectively. Our model's accuracy surpasses these benchmarks, highlighting its superior predictive performance. Moreover, our approach also compares favorably to other ensemble methods, such as HM-Bag Moov Voting Classifier and Voting Classifier, which achieved accuracy rates of 77.21% and 86%, respectively.

In this study, we compare the accuracy obtained by our proposed model with other state-of-the-art models. While ensemble learning, which combines multiple algorithms, often demands significant computational resources, our approach mitigates this challenge by ensembling just two robust algorithms, Decision Trees (DT) and XGBoost. Additionally, some methods in related works utilize more intricate techniques such as deep learning, which may enhance performance but at the expense of increased computational demands and interpretability challenges. Moreover, disparities in data preprocessing techniques and evaluation metrics further complicate direct comparisons between different models.

5. Conclusion

This study aimed to predict diabetes using machine learning techniques on the Pima Indian Diabetes dataset. XGBoost and Bagging with Decision Trees were among the techniques used, along with data pretreatment techniques like median imputation. For diabetes prediction, our method produced remarkable accuracy rates of 92.7% with the extended Pima dataset and 91% with the Pima dataset. With a precision of 89.55% for Pima and 97.1% for extended Pima, our model demonstrated a high percentage of accurate positive predictions. The model's recall values of 81.08% for Pima and 81.7% for extended Pima showed that it could recognize real positive cases.

These findings highlight the efficacy of our approach in predicting diabetes, although with acknowledgment of potential biases and incomplete data. Future research should prioritize addressing these limitations by diversifying datasets and incorporating more comprehensive medical information to reinforce the model's accuracy and robustness. Moreover, ensuring the model's applicability across diverse demographic groups is imperative for its generalizability.

In addition, while this study showcased promising outcomes with XGBoost and Bagging with Decision Trees, it's essential to explore additional algorithms and ensemble methods, such as stacking, which may offer complementary benefits.

References

- [1] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *Journal of Biomedical Informatics*, vol. 59, pp. 185-200, 2016.
- [2] D. Boughareb, H. Bensalah, and H. Seridi, "A Hybrid GAN-ANN-Based Model for Diabetes Prediction," *International Journal of Scientific Research in Science and Technology*, vol. 10, no. 14, pp. 30-41, 2022.
- [3] 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, 2021.
- [4] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: Review and case study," *Applied Sciences*, vol. 14, pp. 2519-2528, 2019.
- [5] B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inform. Med. Unlocked*, vol. 16, p. 100203, 2019.
- [6] M. Li, X. Fu, and D. Li, "Diabetes prediction based on XGBoost algorithm," in *IOP Conference Series: Materials Science and Engineering*, vol. 768, no. 7, p. 072093, Mar. 2020.

- [7] R. Paranjape et al., "An agent-based simulation system for modeling a diabetic patient," *International Journal of Intelligent Information and Database Systems*, vol. 4, no. 3, pp. 264, 2010.
- [8] Pima Indian Diabetes. Retrieved from <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- [9] D. Ramesh and Y. S. Katheria, "Ensemble method based predictive model for analyzing disease datasets: A predictive analysis approach," *Health Technol.*, vol. 9, pp. 533-545, 2019.
- [10] G. N. Sundar et al., "Intelligent computational techniques of machine learning models for demand analysis and prediction," *International Journal of Intelligent Information and Database Systems*, vol. 16, no. 1, pp. 39, 2023.
- [11] F. H. K. Tanaka and C. Aranha, "Data augmentation using GANs," in *2020 IEEE International Conference on Big Data*, pp. 5048-5053.
- [12] XGBoost. Retrieved from <https://xgboost.readthedocs.io/en/latest/>
- [13] I. J. Goodfellow et al., "Generative adversarial networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Z. Ghahramani, M. Welling, and C. Cortes, Eds. Massachusetts: MIT Press, 2014, pp. 2672–2680.
- [14] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data," *IEEE Access*, vol. 8, pp. 171263-171280, 2020.
- [15] Z. Wang, P. Jia, X. Xu, B. Wang, Y. Zhu, and D. Li, "Sample and feature selecting based ensemble learning for imbalanced problems," *Applied Soft Computing*, vol. 113, p. 107884, 2021.
- [16] W. Węgiel, M. Koziarski, and M. Woźniak, "Multicriteria classifier ensemble learning for imbalanced data," *IEEE Access*, vol. 10, pp. 16807-16818, 2022.
- [17] <https://www.kaggle.com/datasets/johndasilva/diabetes>