IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Explainable Machine Learning-Based Diabetes Risk Assessment Using Balanced CatBoost, Threshold Optimization and SHAP with Multilingual Accessibility

Geetanjali Manohar Shelke[1] and Dr. Kiran Tikar[2]
Student, Master of Computer Applications[1]
Assistant Professor, Master of Computer Applications[2]
Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra, India.

Abstract: *A significant example of using machine learning is diabetes risk assessment using routinely collected health indicators. The current screening methods may be ineffective due to class imbalance, complex interactions between risk factors, and limited availability. Explainability of predictive models. This study brings forward an explainable binary diabetes–risk assessment pipeline based on the 2015 Diabetes Health Indicators data are from the CDC Behavioral Risk Factor Surveillance System (BRFSS). The original this dataset consists of 22 columns (21 predictors plus the target column) and 253,680 records. Diabetes_binary. The workflow that was implemented considers female participants aged between 6-12 (Age codes 6-12). Approximately 45-79 years of the standard BRFSS age coding, eliminating duplicate records and assessing Logistic The models used are Regression, Random Forest, XGBoost, CatBoost, and a model with class weighted Balanced CatBoost. After duplicate If it is removed, the study cohort consists of 91981 observations,76643 are negative and 15338 are positive observations. Because the CatBoost is trained weights[1, 5] on the classes and the positive class is significantly smaller. Then threshold analysis of the range of 0.30 to 0.60 was performed to check the precision-recall trade-off. The resulting threshold value is 0.60, and Balanced is the classification for this classification. CatBoost obtains 77.66% accuracy, 39.75% precision, 65.87% recall, 49.58% F1-score, and 81.55% ROC-AUC. SHAP to interpret the final model, an analysis is performed on it, which involves the factors of General Health, BMI, High Blood Pressure, High Cholesterol, Age is one of the most important characteristics. A multilingual presentation layer is created offering English, Marathi and Hindi. Prediction messages without any change in the model. The study exemplifies an actual marriage of the practical and the possible. Imbalance aware boosting, threshold selection, explainability, and multilingual accessibility for research oriented diabetes risk assessment. Formal demographic fairness metrics, external validation, calibration analysis, and clinical no validation was undertaken and this is an important direction of future work..*

Key Words: Diabetes risk assessment, CatBoost, class imbalance, threshold optimization, SHAP, explainable AI, BRFSS, multilingual accessibility

1. INTRODUCTION

Diabetes is an important chronic illness and identifying individuals with a high risk of developing diabetes could allow for early preventive measures to be taken. With the increasing availability of health-related data, the relationships

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

between these data are often complex and nonlinear, making it an increasingly attractive target for diabetes risk assessment with machine learning (ML).

A major issue with the diabetes classification is the lack of balance in the number of diabetes instances and negative instances. The diabetes class has very few instances compared to the number of negative classes, which is a major problem in the diabetes classification. Therefore, accuracy may not be an appropriate assessment of model performance in picking up positive cases. Under these situations, other metrics such as precision, recall, f1 score, or roc-auc are important to consider for evaluating the model performance. Another aspect to take into account is the interpretability of the model. In particular, AI tools that can explain their predictions, such as SHApley Additive exPlanations(SHAP), can be very helpful when trying to be more precise of what information each component is contributing to a prediction, or when trying to make the process of risk assessment done by a machine-learning system more transparent.

This study suggests a combined pipeline using BRFSS 2015 data, model comparison, Balanced CatBoost model, threshold optimization, explanation using SHAP and presenting results in various languages (English, Marathi and Hindi). The present study only considers the female population for Age codes 6–12 (45–79 years). The contribution is a holistic and understandable prediction workflow, not a new machine-learning algorithm. Diabetes_binary is a binary classification of diabetes-related health status, and doesn't differentiate between Type 1 and Type 2. However, the proposed system is not meant for diagnosis in clinical settings, but for research purposes for risk assessment.

II. LITERATURE REVIEW

The diabetes prediction research has evolved from simple classifiers to ensemble learning, interpretable models. Personalized risk estimation, and fairness-aware evaluation. Elated et al. studied Extreme Learning Machine. Learning to predict diabetes at early stages based on questionnaire data[1] and Zhang and Wang studied cascade and ensemble Learning. Other studies have assessed learning for gestational diabetes with results of 80.3% accuracy, 79.3% recall[2]. The authors of the “Boosting” paper are Ganie et al.algorithms and reported the best accuracy of the evaluated algorithms[3]. Geo Jenefer and Deepa submitted their Proposed Firefly optimization based CatBoost Classifier on the Pima dataset which showed the growing trend. The application of CatBoost in the prediction of diabetes[4]. The current study is especially relevant to BRFSS-based research as it offers a big population-health dataset. Pias et al.focused on the resampling design for diabetes prediction for use of BRFSS 2015 and later years, focusing on sensitivity, balanced accuracy and subgroup[5]. In this study, Ahmed et al. have compared LIME and SHAP explanations on a large diabetes dataset of 253,680 responses and showed the role of interpretability in the medical prediction[6]. A self-explainable interface study was conducted that combined SHAP local explanations and machine learning prediction in a separate study to improve transparency[7]. Advanced boosting and XAI were used in combination in more recent work. Rafie et al.compared two different algorithms for Type 2: XGBoost and CatBoost. In 3,203 adults from the Dena Cohort, diabetes risk prediction was performed using SHAP for global and case-specific interpretation[8]. Hu et al. have proposed an interpretable CatBoost model for diabetic foot risk evaluation and obtained good performance. Internal and external performance, demonstrating the usefulness of CatBoost plus SHAP in a related diabetes-care task[9]. Mehrnoush et al. tested seven models for GD and determined that CatBoost and XGBoost models were among the best models. Predictive power was moderate, but stronger approaches were found[10]. Rossi et al. suggest the hybrid approach of D.R.E.A.M., which is a mix of the two. For instance, the probabilities of risk have been used for Type 2 diabetes risk estimation[11]. As well as the optimization of thresholds and SHAP. The fairness and generalizability of healthcare ML is becoming more of an accepted dimension. Pessach and Shmueli presented a literature review of definitions of fairness and mitigation strategies for fairness pre-processing, in-processing, and post processing[12]. Raza et al. discussed the bias and quantitative fairness metrics for equity-oriented public health ML[13]. Chakradeo et al. had discussed the participatory view on fairness assessment of clinical prediction models[14].

Pall et al. recently conducted a large-scale external validation and fairness assessment of Type 2 diabetes risk prediction with XGBoost demonstrates that internal performance can be different from external performance and that performance and that performance and that performance can vary significantly across age groups[15]. These studies

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

highlight the significance of fairness and external validation and this is not a formal fairness metric calculation in itself in the present implementation. The diabetes prediction literature matrix for this project has 25 entries between 2022-2026 and the topic is diabetes prediction. CatBoost, SHAP/XAI, fairness, personalization, calibration and external validation. Based on the evidence review, the following is listed as evidence that: CatBoost, imbalance handling and threshold selection, SHAP are independent techniques. The practical and unique contribution of this work is that these components have been brought together in a well-defined female population from BRFSS 2015. Add English, Marathi and Hindi accessibility. This is not to be seen as a contribution to the application, but an application-level contribution. The newness of the algorithms.

III. RESEARCH GAP

The literature shows that there are three recurring gaps that concern the implemented workflow. The literature reveals three common gaps that concern the workflow that has been implemented. First, many studies have shown that without equal treatment for positive-class recall in case of severe imbalance. Second, explainability is often Independent from threshold decision and imbalance treatment. Thirdly, multilingual accessibility is less. Often put forward as part of the end-user facing prediction process. The present study tackles these practical issues: gaps between one repeatable pipeline.

IV. OBJECTIVE

- To produce a female older-adult cohort from the Diabetes Health Indicators data for the BRFSS 2015..
- To clean the data and create a stratified training and testing set.
- To compare Logistic Regression, Random Forest, XGBoost,Standard CatBoost and Balanced CatBoost.
- To create a class weighted Balanced CatBoost model for better recall of positive classes.
- To test the various decision thresholds from 0.30 to 0.60.
- To choose the ultimate decision threshold according to the maximum F1-score obtained from the threshold tested. .
- To describe model prediction with SHAP and determine important risk factors.
- To make prediction messages accessible to people by providing them in English, Marathi and Hindi.

V. METHODOLOGY

5.1 Dataset

The data used in the study comes from the diabetes_binary_health_indicators_BRFSS2015. csv dataset, which is a commonly used cleaned version of the data. Diabetes health indicators from 2015 BRFSS. The full file has 22 columns and 253,680 observations, with a single binary column and 21 predictor variables[26]. The predictors include health conditions, lifestyle indicators, healthcare access, general, and environmental factors. Welfare, mental health and physical health, mobility, sex, age, education and income. The data file, Diabetes_binary, is a binary file containing a diagnosis of either Diabetes or NoDiabetes. Used the classification outcome as the dependent variable.

5.2 Population Selection and Data Cleaning

The dataset was filtered using Sex == 0 and age code 6-12, which correspond approximately to ages 45-79 under standard BRFSS coding. This produced 101,793 records, with 86,389 negative cases and 15,404 positive cases. No missing values were found. After removing 9,812 duplicate rows, the final dataset contained 91,981 records, including 76,643 negative cases and 15,338 positive cases. Since the positive class remained smaller, class weighting was used in Balanced CatBoost to help address the class imbalance. No synthetic oversampling or feature scaling was applied in the final CatBoost pipeline.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

5.3 Train–Test Split

Data is split into training (80%) and testing (20%) subsets with a stratified split and random_state=42. Both subsets contain approximately the same ratio of target classes as in the original data set,provided that stratification is used to maintain these ratios. The resulting training set will have 21 predictor variables, and 73,584 observations in the dataset, and 18,397 observations in the test set. The test set uses Model comparison, threshold analysis and model interpretation using SHAP.

5.4 Machine Learning Models

Five machine-learning models were compared in this study. The base model was Logistic Regression and the model with Random Forest was set up with 300 trees with class weights from balanced classes. The parameters used in XGBoost were 300 estimators, max depth = 6, and learning rate = 0.05. The parameters used for the Standard CatBoost were 500 iterations, depth of 6, learning rate of 0.05. Balanced CatBoost was the same as Standard CatBoost,but with class weights of [1,5] to emphasize the positive class. Balanced CatBoost was chosen due to its ability to enhance the positive-class recall, as this component of the model is able to detect more instances of the positive class, while the accuracy rate of the model itself was slightly lower.

5.5 Evaluation Metrics

Accuracy is the proportion of all the test observations that are classified correctly. The fraction is measured by precision. Made predictions that were correct and the observation was positive Recall is the ratio of the number of correct responses to the total number of positive ones. The model is able to detect the material. The harmonic mean of precision and recall is F1-score. ROC-AUC is a summary of the ranking properly handling ability of predicted probabilities over decision thresholds. The study concentrates on the imbalance nature of the data set since it is so striking. Recall, F1-score and ROC-AUC in addition to accuracy.

5.6 Threshold Optimization

The CatBoost probability output does not come with a default classification threshold of 0.50. The notebook evaluates threshold of 0.30, 0.35, 0.45, 0.55 and 0.60. For each threshold, the positive class is assigned if the probability predicted is at or above a threshold. For the values that were considered, 0.60 yield a maximum F1-score of 49.58% and it is thus used for final classification. This is an empirical operating-point selection on the PICE, the laptop PC, or the notebook PC. This is a hold-out test set that is available and should not be used as a clinically validated threshold.

5.7 Explainability with SHAP

The final Balanced CatBoost model was used to make predictions, which were then explained using SHAP with TreeExplainer. The most important components were General Heath, BMI, High Blood Pressure, High Cholesterol and Age. Some other factors were also important: income, alcohol consumption, difficulty walking, cholesterol check, and heart disease. The model's probability of prediction and class 0 for one test observation were 41.17% and 0, respectively, at a threshold of 0.60. SHAP is used to explain predictions made by the model, but it does not imply causation or make a clinical diagnosis.

5.8 Multilingual Accessibility

The final notebook contains a function named multilingual_result that has English, Marathi, and Hindi result messages. The underlying only part that's translated is the presentation layer, Probability and classification remain the same. The interface communicates the predicted class, diabetes-related risk probability, a short interpretation and a safety note stating that the output is not clinical, but machine learning prediction. This component is therefore an accessibility contribution as opposed to tweaking the prediction algorithm itself.

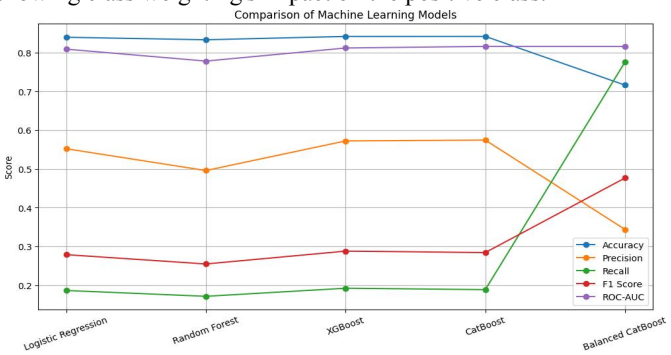
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

VI. EXPERIMENTAL RESULTS

6.1 Model Comparison

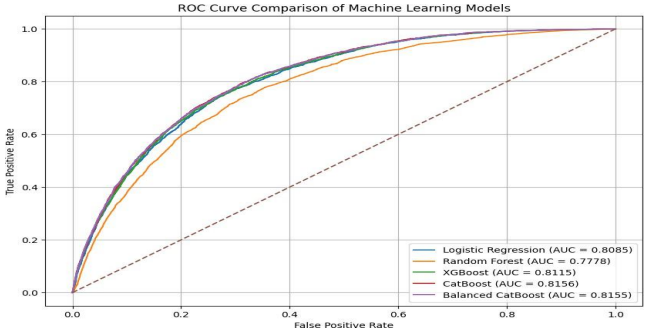
Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	83.91%	55.20%	18.68%	27.91%	80.85%
Random Forest	83.27%	49.58%	17.18%	25.51%	77.78%
XGBoost	84.13%	57.21%	19.26%	28.82%	81.15%
CatBoost	84.14%	57.43%	18.90%	28.45%	81.56%
Balanced CatBoost (0.50)	71.59%	34.40%	77.54%	47.66%	81.55%

Typical models have an accuracy score greater than 83% and a recall score less than 20% for the positive class. This disparity illustrates the need for accuracy beyond that the maximum ROC-AUC values of the standard-model are achieved by XCBoost (81.15%) and CatBoost (81.15%). However, The Balanced CatBoost model boosts the recall to 77.54% at threshold 0.50, showing class weighting's impact on the positive class.



6.2 ROC Curve Analysis

Comparing the result with ROC indicates that the overall performance of standard CatBoost and Balanced CatBoost are very similar. The discrimination performance of the two classifiers was 0.8156 and 0.8155 respectively. The significant difference here is not one of performance ranks but one of classification behavior resulting from class weighting and threshold selection. This helps to validate the use of a probability-based evaluation rather than the single accuracy value being the complete model evaluation.



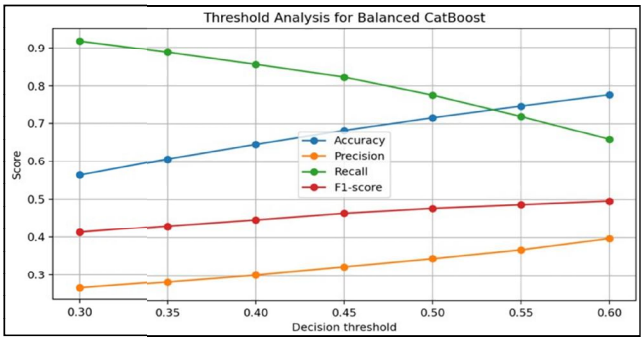
6.3 Threshold Analysis

Threshold	Accuracy	Precision	Recall	F1-score
0.30	56.33%	26.57%	91.79%	41.21%
0.35	60.60%	28.30%	88.85%	42.93%

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

0.40	64.49%	30.13%	85.66%	44.59%
0.45	68.22%	32.25%	82.30%	46.34%
0.50	71.59%	34.40%	77.54%	47.66%
0.55	74.66%	36.73%	71.90%	48.62%
0.60	77.66%	39.75%	65.87%	49.58%

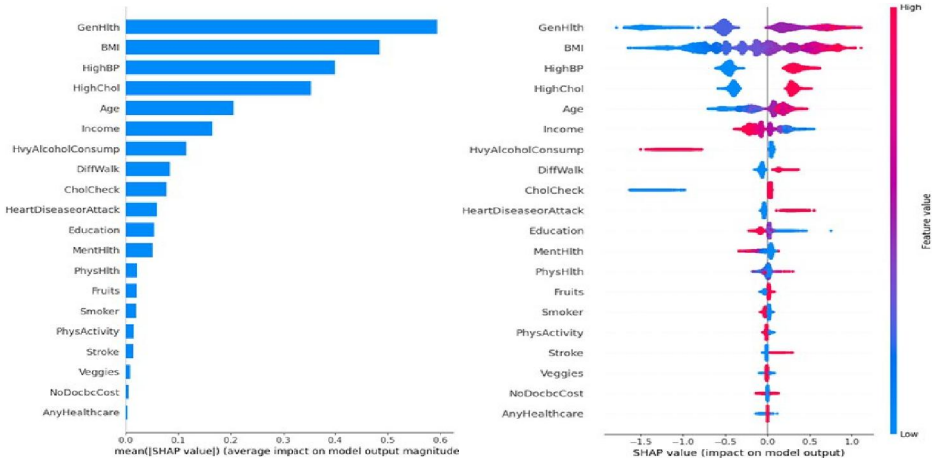
Having higher thresholds is more precise and more accurate, but more likely to miss the true case, higher thresholds have a higher sensitivity, but are more likely to yield false positives. The F1-score is maximum at threshold 0.60 in the tested range (49.58%). It also achieves the highest precision(39.75%) among the thresholds evaluated but with a recall of 65.87%. The decision is therefore a recorded operating point, rather than a universal medical threshold, determined by the study's F1 criteria.



6.4 Final Model and Confusion Matrix

The end result is a balanced CatBoost model with a decision threshold of 0.60. The final test-set metrics are 77.66% accuracy, 39.75% precision,65.87% recall, 49.58% F1-score and 81.55% ROC-AUC. The confusion matrix is [[12266, 3063], [1047, 2021]]. So 12,266 negative observations have been correctly classified, 3,063 negative observations are classified as positive, 1,047 positive observations are missed and 2,021 positive observations are correctly detected.

6.5 Explainable AI Results



The SHAP results offer an explanation at the model level as to which variables are the most important for the prediction. The General Health is most prominent, followed by BMI, HighBP, HighChol and Age. The results are consistent with the diabetes prediction literature in that health status, adiposity, blood pressure, cholesterol and age are among the characteristics that generally contribute to the prediction. However, SHAP is not definitive proof that a

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

feature, on its own, leads to diabetes, it is merely a description of the behaviour of the trained model. HeartDiseaseorAttack is also considered influential, so its presence should also be taken as a predictive association between the dataset. It could represent shared risk factors or health burden. The analysis doesn't provide causal pathways nor take the place of clinical interpretation.

VII. LIMITATIONS

- Cross sectional survey-derived data, may not be generalizable.
- Limited applicability to other age and gender groups, as the study does not involve other participants.
- The binary target does not differentiate between Type 1 and Type 2 diabetes.
- Instead of a separate validation set, the decision threshold of 0.60 was chosen based on the hold-out test results.
- External validation of the model was not performed.
- Probability calibration was not performed.
- There was no formal demographic fairness evaluation.
- A formal user study of the multilingual interface was not carried out.
- The proposed model should not be regarded as clinical as a diagnostic system, without further validation and appropriate governance, but as a research-based risk assessment system.

VIII. FUTURE WORK

The proposed system can be improved through future studies in the following ways: Model performance on different age and gender groups can be used to assess fairness. The model can also be evaluated using the independent data sets to assess its performance in other populations. Further, by using a separate validation set, the probability calibration and threshold can be optimized to increase the reliability of predictions. Moreover, counterfactual explanations might better equip users with understanding of options to mitigate risk. The multilingual interface should be tested with Marathi and Hindi speaking users to assess its clarity and usability. Lastly, data sets with specific diabetes types could be utilized to make a model that separates Type 1 diabetes from Type 2 diabetes.

IX. CONCLUSION

This work proposed a machine learning system that helps in performing diabetes risk assessment in a female population, extracted from the BRFSS 2015 database.Five models were tested and it was determined that Balanced CatBoost was the best model to make the Five models were tested and it was determined that identification of positive cases better. At a threshold of 0.60, the model achieved 77.66% accuracy, 39.75% precision, 65.87% recall, an F1-score of 49.58%, and a ROC-AUC of 81.55%. The messages supported SHAP in making the results more accessible and understandable in English, /marathi and Hindi. Most importantly, the class balancing, threshold selection, explainability and multilingual presentation are integrated in a single workflow. However, the system needs to be further assessed for fairness, validated with external patients and assessed clinically before it can be adopted for healthcare use in the real world.

REFERENCES

[1] N. Elsayed, Z. Elsayed, and M. Ozer, “Early Stage Diabetes Prediction via Extreme Learning Machine,” arXiv:2202.11216, 2022.
[2] J. Zhang and F. Wang, “Prediction of Gestational Diabetes Mellitus under Cascade and Ensemble Learning Algorithm,” Computational Intelligence and Neuroscience, Art. 3212738, 2022, doi: 10.1155/2022/3212738.
[3] A. A. Ganie, S. Pramanik, B. Malik, M. Mallik, and Z. Qin, “An Ensemble Learning Approach for Diabetes Prediction Using Boosting Techniques,” Frontiers in Genetics, 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

[4] G. Geo Jenefer and A. J. Deepa, "Diabetes Disease Prediction Using Firefly Optimization-Based CatBoost Classifier in Big Data Analytics," Journal of Intelligent & Fuzzy Systems, vol.44, no.6, pp.9943-9954, 2023, doi: 10.3233/JIFS-223105.
[5] T. S. Pias, Y. Su, X. Tang, H. Wang, S. Faghani, and D. Yao, "Enhancing Fairness and Accuracy in Type 2 Diabetes Prediction through Data Resampling," medRxiv, 2023, doi: 10.1101/2023.05.02.23289405.
[6] S. Ahmed, M. S. Kaiser, M. S. Hossain, and K.Andersson, "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions," IEEE Access, vol.13, pp.37370-37388, 2024, doi: 10.1109/ACCESS.2024.3422319.
[7] G. Dharmarathne et al., "A Novel Machine Learning Approach for Diagnosing Diabetes with a Self-Explainable Interface," Healthcare Analytics, vol.5, Art.100301, 2024, doi: 10.1016/j.health.2024.100301.
[8] Z. Rafie et al., "Leveraging XGBoost and Explainable AI for Accurate Prediction of Type 2 Diabetes," BMC Public Health, vol.25, Art.3688, 2025, doi: 10.1186/s12889-025-24953-w.
[9] X. Hu et al., "An Interpretable Machine Learning Model for Diabetic Foot Risk Classification in Patients with Diabetes," Scientific Reports, 2026, doi:10.1038/s41598-026-58612-w.
[10] V. Mehrnoush et al., "The Power of Machine Learning Models in Predicting Gestational Diabetes Mellitus," BMC Pregnancy and Childbirth, vol.26, Art.348, 2026, doi: 10.1186/s12884-026-08856-1.
[11] D. Rossi et al., "D.R.E.A.M: Diabetes Risk via Explainable AI Modeling," Multimedia Tools and Applications, vol. 85 , Art.145, 2026, doi:10.1007/s11042-026-21240-7.
[12] D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning," ACM Computing Surveys, vol.55, no. 3, Art. 51, 2022, doi:10.1145/3494672.
[13] S. Raza, A. Shaban-Nejad, E. Dolatabadi, and H. Mamiya, "Exploring Bias and Prediction Metrics to Characterise the Fairness of Machine Learning for Equity-Centered Public Health Decision-Making: A Narrative Review," arXiv:2408.13295, 2024.
[14] K. Chakradeo et al., "Navigating Fairness Aspects of Clinical Prediction Models," BMC Medicine, vol.23, Art.567, 2025, doi: 10.1186/s12916-025-04340-3.
[15] R. S. Pall et al., "Comprehensive Evaluation of Machine Learning for Type 2 Diabetes Risk Prediction: Large-Scale External Validation and Fairness Analysis," arXiv:2607.16253, 2026