

An interpretable machine learning framework for early-stage diabetes mellitus prediction using comparative classification models and SHAP

J Naveen, Deekshitha U and Kavya V *

Department of MCA (NBA Accredited), Surana College (Autonomous), Bangalore, India.

International Journal of Science and Research Archive, 2026, 20(01), 930–938

Publication history: Received on 13 June 2026; revised on 25 July 2026; accepted on 27 July 2026

Article DOI: <https://doi.org/10.30574/ijrsra.2026.20.1.1543>

Abstract

Diabetes mellitus is a long-term metabolic disorder that typically goes unrecognised until it has already caused significant harm, making timely identification a critical clinical priority. Unfortunately, conventional diagnostic approaches frequently fall short in detecting the disease before it progresses, particularly within busy healthcare settings. To tackle this gap, the current study introduces a machine learning-based framework enhanced with explainability, built around a structured data preparation process that handles categorical encoding, numerical scaling, and minority class oversampling through the SMOTE technique. A pair of classification models, Logistic Regression and Random Forest, are trained, tested, and directly compared to assess their diagnostic reliability. The findings reveal that Random Forest consistently delivers stronger results, reaching a classification accuracy of 98%, which reflects its capacity to learn intricate relationships within real-world clinical data. To move beyond raw performance, SHAP analysis is integrated to shed light on how individual patient attributes shape each prediction outcome. The resulting system strikes a meaningful balance between diagnostic accuracy and model interpretability, positioning it as a trustworthy tool for assisting medical professionals in data-driven clinical decision-making.

Keywords: Diabetes Mellitus; Explainable Artificial Intelligence (XAI); Machine Learning; SMOTE; Logistic Regression; Random Forest; SHAP; Medical Diagnosis.

1. Introduction

Diabetes Mellitus has become one of the most widespread metabolic conditions worldwide, placing a growing pressure on health systems around the world. The condition develops when the body either fails to produce enough insulin or loses its ability to use it effectively, eventually causing blood sugar levels to rise to harmful levels. When left unmanaged in its early phases, the disease can lead to a range of severe health complications, including cardiovascular problems, kidney failure, nerve damage, and vision loss [1]. Catching the disease early gives both patients and doctors a valuable opportunity to manage the condition better and maintain a good quality of life [10]. Yet routine diagnostic methods, which mostly rely on lab tests and physical examination, often fail to catch the disease before it advances. As electronic health records and patient monitoring systems produce vast amounts of patient data every day, machine learning has become a valuable tool for spotting disease risk patterns that doctors alone may overlook. Several learning algorithms, from Logistic Regression and Decision Trees to Support Vector Machines and methods like Random Forest, are now widely used in medical prediction tasks. Logistic Regression is still popular due to its simplicity and easy interpretation, while Random Forest usually performs better by merging outputs from many decision trees and reducing the chance of overfitting [2], [17]. However, high accuracy alone is insufficient for clinical practice, where safety, ethics, and transparency also matter. Many of these models work as black-box systems, producing results without explaining how the decision was made, which makes doctors hesitant to trust and use them in practice. XAI methods such as SHAP provide a useful solution to this problem by showing how each input variable affects the model output, bringing the transparency that clinical predictions truly require [5], [19].

* Corresponding author: Kavya V

2. Problem Statement

Detecting diabetes at an early stage is much harder than it seems, mainly because the symptoms it produces closely resemble those of several other conditions, and the patterns hidden within patient datasets are very complex to distinguish without advanced analysis [10]. This overlap in symptoms causes many patients to receive a proper diagnosis only after the disease has already progressed significantly, making treatment far more difficult and costly at that point [1]. While machine learning has clearly opened new doors in medical prediction, a serious limitation continues to affect most of these methods. Models that score well on performance metrics often do not explain how they arrived at a particular conclusion, leaving clinicians completely in the dark about the reasoning behind each result. In a healthcare environment where every clinical decision directly impacts a patient's life, this lack of explanation is not just inconvenient but becomes a real obstacle to practical use [3], [8]. What is truly required, therefore, is a combined framework that delivers both strong prediction capability and full transparency so that medical professionals can apply the results with genuine confidence and understanding [7], [21].

3. Proposed System

This study proposes an explainable approach developed for detecting diabetes mellitus at an early stage. Instead of assessing each element separately, two classification models are evaluated, and SHAP explanations are incorporated to better understand how the model behaves internally. The system brings together data processing, imbalance handling, predictive modelling, and output interpretation into one unified end-to-end workflow. Logistic Regression is used as the interpretable reference model for comparison, while Random Forest was chosen for its potential to capture non-linear patterns and feature interactions commonly found in clinical data, delivering consistently better predictive results [2], [9], [15].

A well-known issue in medical datasets is the uneven spread of class labels, where non-diabetic cases greatly outnumber diabetic ones. To address this, SMOTE is applied during the training phase to produce artificial samples representing the rare class, reducing majority-class dominance and improving the model's ability to correctly identify diabetic patients [4], [12]. After predictions are made, SHAP is used to examine each result and quantify exactly how much every feature influenced the outcome, transforming what would otherwise be a black-box decision into something clinicians can clearly understand and rely on [3], [5], [16]. The proposed framework therefore aims to deliver a system that is both highly accurate and fully transparent, providing a practical path toward real-world clinical integration [7].

4. Literature Review

The role of data-driven methods in healthcare has grown rapidly in recent years, driven by a surge in digitally recorded patient information and a growing clinical imperative to intervene before diseases reach advanced stages. Among the many conditions being studied through this lens, diabetes has attracted a particularly strong research focus. Machine learning has consistently demonstrated value in detecting subtle patterns hidden in complex patient datasets that could easily go unseen [1]. When comparing the range of algorithmic options available to researchers, ensemble methods tend to rise above conventional single-model approaches in terms of prediction reliability. Random Forest has proven itself as a leading choice in medical applications, functioning by combining the outputs of many individually grown decision trees into a single, more dependable verdict. This collective mechanism naturally improves how well predictions hold up on new data, reduces sensitivity to noise, and prevents the model from latching onto patterns that exist only in the training sample [2]. Among the most prevalent challenges in building medical prediction models is the unequal class distribution present in most health datasets, where patients without the condition of interest far outnumber those who have it. When a model is trained on such skewed data without correction, it tends to perform well on the majority group while struggling to correctly flag the minority cases that are clinically most important. SMOTE tackles this by creating artificial samples for the minority class, creating a more balanced environment for the learning algorithm [4].

Equally important to getting predictions right is being able to explain them in terms that are meaningful to the people who will act on them. In healthcare, particularly, a model that cannot justify its conclusions is unlikely to earn the confidence of practicing clinicians, regardless of how accurate it may be. SHAP has become a widely adopted tool for bridging the gap. It attributes a specific degree of influence to each input feature for every prediction the model makes and, in doing so, turns abstract outputs into traceable, clinically relevant reasoning [3], [5]. The most effective and trustworthy diabetes prediction systems are those that bring together careful data handling, strong modelling, and transparent explainability within a single integrated approach [12], [23].

5. Methodology

We designed our research across four connected phases, each tackling its own distinct problem, with the ultimate goal of creating a diabetes detection system that works reliably and can be understood by clinicians. Every phase solves a particular challenge—whether technical or clinical—so that our final system produces results grounded in statistics while remaining practically useful in real medical settings. First comes data preparation, where we clean up messy patient records, tackle the problem of imbalanced classes, and transform variables into formats that machine learning algorithms can actually work with [4], [12]. Then we move to building models, training and testing two algorithms that operate on fundamentally different principles, which gives us a meaningful way to see which approach truly outperforms the other [2], [9]. The testing phase doesn't stop at simple accuracy numbers; instead, we evaluate both classifiers using multiple complementary metrics, creating a comprehensive and realistic assessment of how well each model can be trusted for diagnosis [14]. The final stage focuses on interpretation, making certain that achieving high performance doesn't mean losing the transparency that clinicians need; we employ SHAP methods to unpack each prediction, revealing which patient factors drove the model's conclusion [3], [5].

5.1 Data Preparation

The experiments are conducted on the Early-Stage Diabetes Risk Prediction Dataset, which comprises 520 patient records and 17 attributes capturing a combination of demographic details and clinically observed diabetes symptoms. Before feeding the data into any learning algorithm, categorical variables are transformed through label encoding, and all numerical features are brought to a comparable scale using standard normalization techniques, both of which improve algorithmic compatibility and accelerate model convergence [11]. A particularly important concern with this dataset is the unequal distribution of class labels, where non-diabetic cases outnumber diabetic ones by a notable margin. SMOTE is applied during the training phase to produce synthetic yet statistically representative samples of the minority diabetic class, helping the model develop a more balanced and sensitive decision boundary [4], [12].

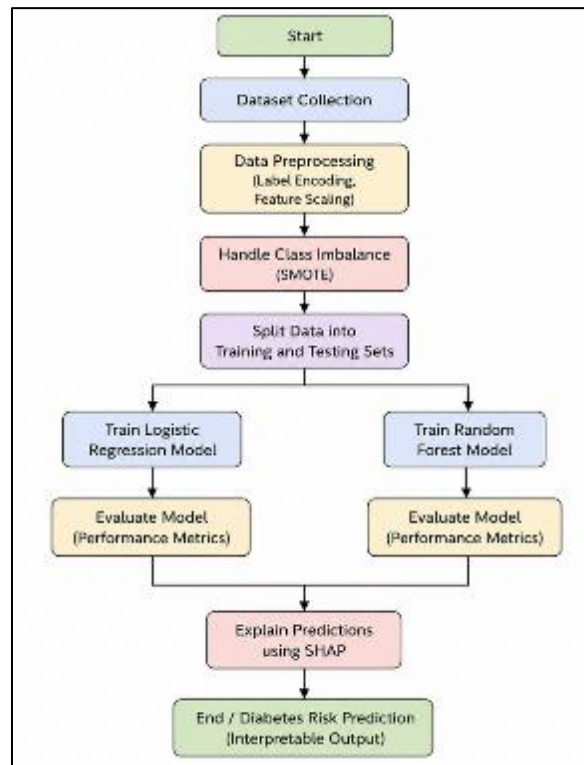


Figure 1 Proposed methodology pipeline for diabetes prediction using machine learning and XAI.

Fig. 1 presents the step-by-step workflow of the proposed diabetes prediction system. The process begins with raw patient data that is cleaned and prepared through encoding, normalisation, and SMOTE to fix the class imbalance in the training set. Predictive models are then trained on this balanced data and determined by performance metrics. The best model is further analysed using SHAP to know how individual features affect the final prediction. This pipeline ensures that the system delivers both strong predictive performance and clear, clinically meaningful explanations.

Table 1 Summary of the Early-Stage Diabetes Risk Prediction Dataset

Property	Value
Dataset Name	Early-Stage Diabetes Risk Prediction
Total Records	520
Input Features	16
Target Variable	Diabetes Class
Total Attributes	17
Numerical Features	1 (Age)
Categorical Features	15
Class Labels	Positive, Negative
Missing Values	None

Table I describes the data used for this study. It contains information from 520 patients, represented by 16 recorded attributes and one class label that specifies the diabetes status of each individual. Only Age is measured as a numerical value, whereas all other attributes are categorical and correspond to commonly observed diabetes-related symptoms. Because the dataset is complete and contains no missing entries, preprocessing is limited to label encoding, feature scaling, and SMOTE before training the prediction models.

5.1. Predictive Model Development

Two supervised classification algorithms are selected to enable a direct and meaningful performance comparison across different modelling philosophies. Logistic Regression is included as the interpretable baseline, valued for its straightforward mathematical structure and its well-established track record in binary medical classification tasks [17]. Random Forest is selected as the primary candidate owing to its ensemble design, which builds and aggregates predictions from a large number of individually trained decision trees, allowing it to capture complex and nonlinear relationships within clinical feature sets while remaining resistant to overfitting [2], [9].

5.2. Model Evaluation and Explainability

Both classifiers are assessed against five widely accepted performance metrics—accuracy, precision, recall, F1-score, and ROC-AUC—collectively offering a comprehensive and balanced picture of the individual algorithm's diagnostic strengths [14]. In a medical prediction context, increasing recall carries particular weight, as missed diabetic cases can lead to delayed treatment and worsening patient outcomes, making recall an especially critical measure in this evaluation. Once the superior model is identified, SHAP is applied to examine its internal decision logic, assigning a measurable contribution score to each clinical feature for every prediction made, thereby converting numerical model outputs into transparent and clinically interpretable explanations that practitioners can realistically act upon [3], [5], [16].

5.3. Proposed Algorithm

The proposed method preprocesses the dataset, applies SMOTE to address class imbalance, and develops suitable models for diabetes prediction. The selected model is then evaluated and interpreted using SHAP to provide understandable diabetes predictions [8].

- Input: Dataset (D) containing patient medical features and corresponding class labels.
- Apply preprocessing steps, including label encoding, feature scaling, and missing value treatment.
- Partition (D) into separate training and evaluating subsets.
- Implement SMOTE on the training subset to address class imbalance.
- Train Logistic Regression and Random Forest on the class-balanced training set.
- Evaluate both models using Accuracy, Recall, Precision, ROC-AUC, and F1-Score [8].
- Identify the best-performing model (best) and apply SHAP to generate feature-level explanations.
- Output: Diabetes prediction result accompanied by an interpretable feature contribution breakdown.

6. Results and Discussion

An Explainable AI (XAI) framework is introduced to monitor the progress of Random Forest and Logistic Regression classifiers in predicting early-stage diabetes [2], [9]. The performance is conducted using standard evaluation metrics [11], [14] and presented through tables and figures for clear interpretation. The results demonstrate the performance improvement achieved by the proposed model.

6.1. Performance Evaluation

Performance evaluation of the proposed algorithms was conducted using standard classification metrics [11], [14], as shown in the table below.

Table 2 Performance Comparison of Classification Models

Metric	Logistic Regression	Random Forest
Accuracy	93.33%	98.08%
Precision	91.52%	100.00%
Recall	93.10%	96.88%
F1-Score	92.30%	98.41%
ROC-AUC	98.90%	100.00%

Table II shows the comparative performance of Logistic Regression and Random Forest models. The Random Forest classifier demonstrated the best accuracy of 98.08%, in addition to the highest precision, F1-Score, and ROC-AUC values. This superior performance suggests that Random Forest better captures underlying data complexity, resulting in more reliable early-stage diabetes prediction.

6.2. Performance Metric Formulations

Accuracy is defined as:

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

Precision is defined as:

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

Recall is defined as:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

F1-score is defined as:

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

ROC-AUC describes the model's ability to distinguish between positive and negative classes, where a higher AUC value indicates better classification performance.

6.3. Accuracy Evaluation

Accuracy essentially answers a simple question: out of all the predictions a model makes, how many are actually correct? It is one of the most straightforward ways to check whether a classifier is doing its job, since it tallies up all the right guesses divided by the total number of guesses. While accuracy gives a quick snapshot of overall performance, it is particularly useful when comparing different models to determine which one correctly classifies more cases [14]. For the proposed diabetes detection framework, accuracy was used as one of the key evaluation metrics to compare the performance of Logistic Regression and Random Forest classifiers in identifying early-stage diabetes.

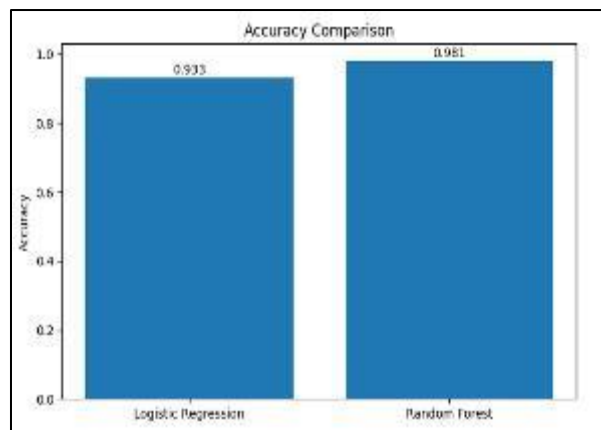


Figure 2 On comparing the accuracy analysis of logistic regression and random forest models.

Fig. 2 presents the accuracy comparison between the Logistic Regression and Random Forest classifiers. The Random Forest model achieved an accuracy of 98.08%, outperforming Logistic Regression, which achieved 93.33%. This improvement indicates that the Random Forest classifier is more effective at learning complex relationships within the dataset and provides better generalization on unseen data. Consequently, it demonstrates superior predictive performance for early-stage diabetes classification in the proposed framework.

6.4. Confusion Matrix Analysis

The Classification performance of a confusion matrix, which contrasts predicted labels against actual labels [14]. It delivers a detailed summary of correct and incorrect predictions, enabling better interpretation of model performance.

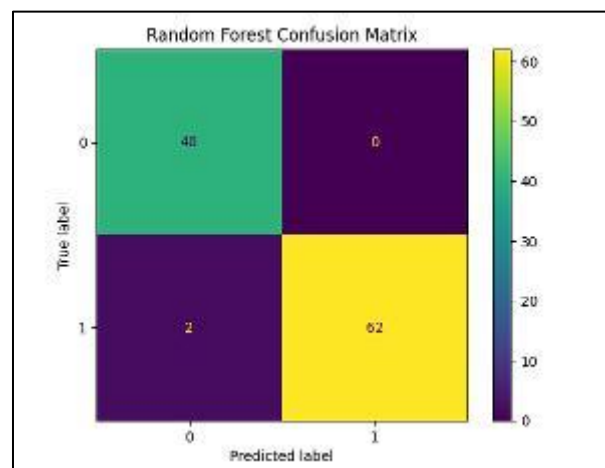


Figure 3 Confusion matrix showing TP, TN, FP, and FN values.

Fig. 3 The confusion matrix obtained from the Random Forest classifier shows that the model correctly identified 62 diabetic cases (True Positives) and 40 non-diabetic cases (True Negatives). No non-diabetic cases were wrongly classified as diabetic (False Positives = 0), while only 2 diabetic cases were missed (False Negatives = 2). These results highlight the strong classification efficiency of our model with very low prediction error.

6.5. ROC Curve Analysis

The ROC curve serves to assess the trade-off between sensitivity and specificity [14]. It facilitates the validation of the model's ability to differentiate between labels and quantify overall classification performance.

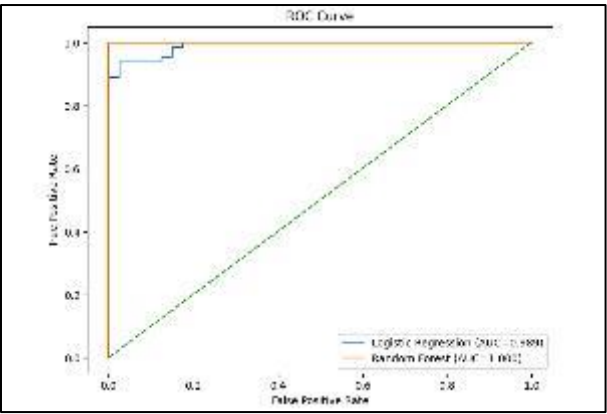


Figure 4 ROC curve illustrating AUC performance for Random Forest and Logistic Regression.

Fig. 4 illustrates the ROC curve analysis, where Random Forest achieves a higher AUC, indicating stronger class separability and better classification capability.

6.6. SHAP Explainability Analysis

SHAP explains the prediction process by quantifying how much each feature contributes to a prediction. It enhances transparency and builds trust in the model's decisions [16].

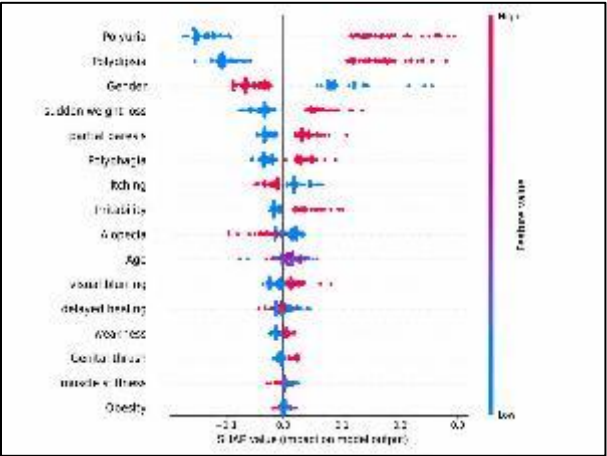


Figure 5 SHAP plot highlighting feature importance in diabetes prediction.

Fig. 5 The SHAP analysis results. Features such as polyuria, polydipsia, sudden weight loss, Age, and partial paresis have the highest impact on diabetes prediction. The SHAP explanations improve model transparency by clearly showing how individual features influence the prediction outcome.

6.7. Feature Importance Table

Table 3 Top features identified using SHAP analysis for diabetes prediction

Rank	Feature
1	Polyuria
2	Polydipsia
3	Sudden Weight Loss
4	Age
5	Partial Paresis

The ranking obtained from SHAP analysis aligns with common clinical symptoms associated with diabetes, supporting the trustworthiness of the model explanations [10].

Table 3 highlights the most significant features contributing to diabetes prediction. Polyuria and Polydipsia are the strongest indicators, followed by Sudden Weight Loss, Age, and Partial Paresis, confirming their strong clinical relevance.

6.8. Clinical Implications

The proposed developed model has the capability to aid healthcare professionals in the advanced diagnosis of individuals at risk of diabetes. The composition of high predictive performance of the SHAP-based explanations and Random Forest allows clinicians to better understand the reasoning behind model predictions [3], [7]. Important features such as polyuria, polydipsia, and sudden weight loss were identified as key indicators of diabetes risk, which is consistent with findings reported in previous clinical studies [10]. By providing both accurate predictions and interpretable insights, the proposed system can assist in timely decision-making, early intervention, and improved patient management [13].

Future Scope

Even though the presented framework exhibits high predictive accuracy, there are many possible ways to improve it. Advanced deep learning techniques, namely Artificial Neural Networks (ANNs) and hybrid architectures, could be investigated to achieve higher prediction accuracy and better model performance. Another way to improve the framework is to incorporate bigger and more diverse datasets of real-world patients' medical data. Furthermore, combining the framework with real-time hospital information systems and electronic health records could make it more applicable in clinical practice and capable of continuous patient monitoring. Exploring methods other than SHAP to analyse the model's behaviour and explainability would be one of the possible improvements to make. Finally, creating a web/mobile healthcare application based on the presented framework could allow healthcare specialists to perform early detection of diabetes and risk assessment.

7. Conclusion

We propose a novel set of XAI algorithms to predict early the development of diabetes by applying Logistic Regression and Random Forest classifiers. In particular, the suggested framework combines data preprocessing, class balancing using SMOTE, and SHAP explanations to provide improved predictions and better reliability of the models. As evidenced by the experimental results, the Random Forest classifier performed well compared to Logistic Regression. The model implemented a stronger ability to identify underlying data patterns, leading to more precise predictions. To improve the transparency of the prediction algorithm, the SHAP explanation was applied to reveal how the individual features affect the model. Thus, it was discovered that polyuria, polydipsia, and sudden weight loss are the features that have the strongest influence on the disease.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

- [1] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, no. 3, Art. no. 160, 2021.
- [2] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [3] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
- [5] C. Molnar, Interpretable Machine Learning, 2nd ed. Lulu.com, 2022.

- [6] M. Du, N. Liu, and X. Hu, "Techniques for Interpretable Machine Learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2020.
- [7] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
- [8] Z. C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.
- [9] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.
- [10] M. M. F. Islam, F. Karray, R. Alhajj, and J. Zeng, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, 2020, pp. 113–125.
- [11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [12] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.
- [13] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
- [14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [15] S. Sisodia and D. S. Sisodia, "Prediction of Diabetes Using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.
- [16] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *ITU Journal: ICT Discoveries*, vol. 1, no. 1, pp. 1–10, 2017.
- [17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [18] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
- [20] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Health Care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.
- [21] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and W. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine*, vol. 178, no. 11, pp. 1544–1547, 2018.
- [22] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Scalable and Accurate Deep Learning with Electronic Health Records," *Nature Partner Journals Digital Medicine*, vol. 1, no. 1, p. 18, 2018.
- [23] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [24] S. Nagpal, A. Das, R. Dey, C. Rana, J. Bhatnagar, and A. P. Reddy, "Machine Learning Models for Classification of Diabetes Progression," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2574–2581, 2020.
- [25] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is Not All You Need," *Information Fusion*, vol. 81, pp. 84–90, 2022.
- [26] B. Haibe-Kains et al., "Transparency and Reproducibility in Artificial Intelligence," *Nature*, vol. 586, no. 7829, pp. 313–316, 2020.
- [27] T. Yarkoni and J. Westfall, "Choosing Prediction Over Explanation in Psychology: Lessons from Machine Learning," *Perspectives on Psychological Science*, vol. 12, no. 6, pp. 1100–1122, 2017.
- [28] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Y. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. So, B. Repetowski, S. Afdel, and A. Y. Ng, "Deep Learning for Chest Radiograph Diagnosis: A Deep Neural Network Trained on CheXpert Dataset," *Radiology: Artificial Intelligence*, vol. 2, no. 1, p. e190065, 2020.