

# Machine Learning and Explainable Artificial Intelligence for Diabetes Prediction: A Review of Pre-processing Feature Optimization, Model Performance, Risk Assessment, and Personalized Healthcare

Punit C. Trivedi<sup>1</sup>, Ritesh B. Vaghasiya<sup>2</sup>

<sup>1</sup>*Department of Electronics and Communication, Dr. Subhash University, Junagadh, India*

<sup>2</sup>*Assistant. Professor., Department of Electronics and Communication, Dr. Subhash University Junagadh India*

doi.org/10.64643/IJIRTV13I5-208952-459

**Abstract**—Diabetes Mellitus is a major chronic metabolic disorder for which early prediction can support preventive healthcare. Machine Learning (ML) has been widely applied to diabetes prediction using clinical and patient health data. This paper reviews recent ML-based diabetes prediction studies with emphasis on datasets, pre-processing, feature optimization, classification algorithms, ensemble learning, model optimization, Explainable Artificial Intelligence (XAI), risk assessment, and healthcare recommendations. The review indicates that optimized and ensemble ML models generally achieve competitive prediction performance, while SHAP and LIME are increasingly used for model interpretation. However, most existing studies focus mainly on prediction performance, with limited integration of explainability, multi-level risk assessment, and personalized preventive recommendations. The review highlights the need for integrated and interpretable ML-based diabetes decision-support frameworks.

**Index Terms**—Diabetes Mellitus, Machine Learning, Diabetes Prediction, Explainable Artificial Intelligence, SHAP, LIME, Ensemble Learning, Risk Assessment, Healthcare Recommendations.

## I. INTRODUCTION

Diabetes Mellitus is a chronic metabolic disorder associated with elevated blood glucose levels and serious long-term complications. Early identification of individuals at increased risk can support timely evaluation and preventive interventions. The availability of clinical datasets has enabled the use of

computational approaches for diabetes screening and prediction.

Machine Learning (ML) has been widely applied to diabetes prediction using algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gradient Boosting, AdaBoost, XGBoost, and Extra Trees. Ensemble and hybrid approaches have also been investigated to improve predictive performance. However, reliable diabetes prediction depends not only on classifier selection but also on data quality, pre-processing, feature optimization, class balancing, and model parameter selection. Furthermore, high predictive performance does not necessarily provide sufficient insight into why a prediction is generated. Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME can improve model transparency by identifying influential features.

This review examines recent developments in ML-based diabetes prediction, focusing on pre-processing, feature optimization, model optimization, ensemble learning, explainability, risk assessment, and personalized healthcare recommendations.

## II. REVIEW METHODOLOGY

A structured review of 22 relevant research studies was conducted on ML-based diabetes prediction. The studies were compared based on datasets, pre-processing, feature selection, ML and ensemble

techniques, optimization, evaluation, XAI, risk assessment, and healthcare-related functionality.

The reviewed studies include research using the PIMA Indians Diabetes Dataset and other clinical and population-level datasets. The comparison was used to identify commonly adopted approaches, limitations, and research opportunities.

### III. DATASETS USED FOR DIABETES PREDICTION

The PIMA Indians Diabetes Dataset is a widely used benchmark for diabetes prediction. It contains 768 records, eight clinical input attributes, and a binary diabetes outcome. The features include pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.

Its frequent use enables comparison of different ML approaches under a common benchmark. However, its relatively small size and specific population characteristics may limit generalization.

Other reviewed studies use datasets such as BRFS, cohort-based datasets, and privately collected clinical data, providing larger or more diverse patient populations. The literature therefore indicates that while PIMA is useful for benchmarking, dataset diversity and external validation remain important considerations for developing more generalizable diabetes prediction models.

### IV. DATA PRE-PROCESSING AND FEATURE OPTIMIZATION

Data pre-processing is essential for diabetes prediction because clinical datasets may contain missing or invalid values, outliers, different feature scales, and class imbalance. In the PIMA dataset, zero values in features such as glucose, blood pressure, skin thickness, insulin, and BMI may represent missing or unrealistic measurements and therefore require appropriate handling.

Reviewed studies have investigated imputation, outlier detection using methods such as IQR and Isolation Forest, and feature scaling through standardization or normalization. Class imbalance has also been addressed using techniques including SMOTE, ADASYN, SMOTEENN, and SMOTETomek. Such balancing methods should be

applied only to training data to avoid information leakage.

Feature optimization methods including correlation analysis, Mutual Information, PCA, RFE, Boruta, and model-based feature importance have been used to identify relevant attributes, reduce redundancy, and potentially improve model performance.

### V. MACHINE LEARNING AND ENSEMBLE APPROACHES

Various supervised ML algorithms have been applied to diabetes prediction, including Logistic Regression, Decision Tree, SVM, KNN, and Naïve Bayes. Ensemble methods have received particular attention, including Random Forest, Extra Trees, AdaBoost, Gradient Boosting, and XGBoost.

Random Forest and Extra Trees combine multiple randomized decision trees, while boosting methods progressively improve weak learners. XGBoost further incorporates regularization and computational improvements for efficient gradient boosting.

The reviewed literature indicates that ensemble approaches often provide competitive performance on structured diabetes datasets. However, no single algorithm is universally optimal because performance depends on the dataset, pre-processing, feature configuration, class-balancing strategy, and parameter settings. Therefore, comparative evaluation of multiple algorithms is preferable.

### VI. MODEL OPTIMIZATION AND PERFORMANCE EVALUATION

Model optimization is important because default parameters may not provide optimal performance for a specific dataset. Reviewed studies have investigated feature optimization, ensemble methods, grid search, and hyperparameter tuning. Methods such as RandomizedSearchCV can efficiently evaluate parameter combinations and identify suitable model configurations.

Model performance should be evaluated using multiple metrics rather than accuracy alone, particularly when class imbalance exists. Common measures include Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. Comprehensive evaluation provides a better assessment of model performance and reliability,

with Recall being particularly relevant when identifying potentially at-risk individuals is important.

## VII. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable Artificial Intelligence (XAI) is increasingly important in healthcare because complex ML models may provide predictions without clearly indicating the factors responsible for them. SHAP (SHapley Additive exPlanations) estimates feature contributions using Shapley values and can provide both global and patient-level explanations. LIME (Local Interpretable Model-Agnostic Explanations) provides local explanations by approximating a complex model around an individual observation. Studies such as [2] and [15] demonstrate the application of SHAP and LIME in diabetes prediction. However, XAI is generally treated as an interpretation component rather than being integrated with risk assessment and personalized healthcare guidance.

## VIII. LITERATURE REVIEW

The 22 reviewed studies can be broadly grouped into four research directions.

**Conventional ML:** Studies [1], [5]–[7], [9], [14], [16], and [21] investigate conventional ML or deep learning approaches for diabetes prediction using PIMA and other clinical datasets. These studies establish the effectiveness of supervised learning for structured diabetes data.

**Ensemble and Hybrid Learning:** Studies [3], [4], [8], [11], [18], [19], [20], and [22] investigate ensemble, hybrid, or deep learning approaches. Their findings indicate that combining multiple models, feature-selection techniques, or resampling strategies can provide competitive prediction performance.

**XAI and Interpretability:** Studies [2] and [15] incorporate explainability techniques, particularly SHAP and LIME, to improve the transparency of diabetes prediction models.

These studies demonstrate the importance of understanding feature contributions in healthcare-oriented ML systems.

**Risk Assessment and Data Optimization:** Studies [10], [12], [13], [17], and [20] investigate class

balancing, risk assessment, feature importance, and feature-selection approaches. These works demonstrate the importance of appropriate data handling and risk-oriented prediction.

Overall, the reviewed literature shows that prediction, optimization, ensemble learning, explainability, and risk assessment have each received considerable attention. However, these capabilities are rarely combined with multi-level risk classification and personalized preventive recommendations within a unified framework.

## IX. DIABETES RISK ASSESSMENT AND PERSONALIZED HEALTHCARE

Conventional diabetes prediction generally produces a binary diabetic/non-diabetic outcome. Studies [12], [13], and [21] demonstrate that risk-oriented diabetes prediction has also been investigated, although risk definitions and assessment strategies vary.

A multi-level approach can provide more informative results by combining prediction probability with relevant clinical characteristics. Possible categories include Non-Diabetic, Low Risk, Medium Risk, and High Risk, with thresholds requiring systematic validation.

Personalized healthcare recommendations can further extend prediction systems by converting risk information and influential clinical characteristics into general preventive guidance related to diet, physical activity, monitoring, and lifestyle improvement. Such recommendations should support decision-making rather than replace professional medical diagnosis or treatment.

## X. COMPARATIVE DISCUSSION AND RESEARCH GAPS

The reviewed studies demonstrate that machine learning has been extensively investigated for diabetes prediction using different datasets, pre-processing techniques, feature-selection methods, classification algorithms, and ensemble approaches. The PIMA Indians Diabetes Dataset remains one of the most commonly used benchmark datasets, while some studies have also investigated larger population-based and clinical datasets.

The literature indicates that prediction performance is influenced not only by the choice of classifier but also by data pre-processing, class balancing, feature optimization, and hyperparameter tuning. Ensemble and hybrid approaches have frequently shown competitive performance, while recent studies have incorporated Explainable Artificial Intelligence techniques such as SHAP and LIME to improve model interpretability.

The review also shows that diabetes risk assessment has already been investigated in some studies. Therefore, risk assessment and explainability should not be claimed as completely unexplored areas. Instead, an important research opportunity lies in integrating these capabilities with personalized preventive healthcare recommendations.

Based on the reviewed studies, the major research gaps are:

1. Limited integration of multiple components: Most studies focus on individual aspects such as prediction, optimization, explainability, or risk assessment rather than integrating them into one framework.

2. Limited integration of XAI and risk assessment: Patient-level explanations are generally not directly connected with the generated diabetes risk category.

3. Limited personalized recommendations: Few studies convert prediction results and influential clinical characteristics into personalized preventive healthcare guidance.

4. Dataset generalization challenges: The frequent use of benchmark datasets such as PIMA limits the ability to assess model generalization across different populations.

5. Inconsistent evaluation strategies: Differences in pre-processing, validation methods, datasets, and evaluation metrics make direct comparison of reported model performance difficult.

6. Limited end-to-end decision support: There is scope for frameworks that combine optimized prediction, explanation, risk classification, and preventive recommendations into a unified decision-support system.

The comparison of the reviewed studies is presented in Table I.

Table I. Comparative Analysis of the Reviewed Diabetes Prediction Studies

| Ref. | Dataset / Data                                    | Main Approach                                                        | Pre-processing / Optimization                                                                               | XAI                                   | Risk / Recommendation                                            |
|------|---------------------------------------------------|----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|---------------------------------------|------------------------------------------------------------------|
| [1]  | PIMA                                              | ML classifiers                                                       | Pre-processing, outlier handling, class balancing                                                           | No                                    | Binary diabetes classification                                   |
| [2]  | BREDS, Diabetes 2019                              | ML classification, Extra Trees                                       | Random oversampling, quantile transform, GridSearchCV                                                       | SHAP, PDP, LIME                       | No integrated personalized recommendation                        |
| [3]  | PIMA Indians Diabetes Database                    | XGBoost, Gradient Boosting, Random Forest; DT and SVM for comparison | SMOTE, Borderline-SMOTE, ADASYN, K-Means SMOTE, Gaussian SMOTE; Information Gain, Chi-square, LASSO, Boruta | No dedicated XAI                      | Early diabetes prediction; no integrated recommendation module   |
| [4]  | PIMA, Frankfurt Hospital, Iraqi diabetes datasets | Ensemble deep learning (ANN, LSTM, CNN with stacking)                | Data pre-processing, Extra Trees feature extraction, stacking ensemble                                      | No                                    | Clinical decision support; no personalized recommendation module |
| [5]  | PIMA                                              | Naïve Bayes, Random Forest, J48                                      | PCA, K-means clustering, importance ranking / feature selection                                             | Interpretable ML models; no SHAP/LIME | Binary diabetes classification; no personalized recommendation   |
| [6]  | PIMA                                              | ANN, NB, DT, Deep Learning                                           | Data pre-processing, feature analysis, model parameter optimization                                         | No dedicated XAI                      | Early detection / prognostic support                             |

| Ref. | Dataset / Data                                   | Main Approach                                             | Pre-processing / Optimization                                                                    | XAI              | Risk / Recommendation                                          |
|------|--------------------------------------------------|-----------------------------------------------------------|--------------------------------------------------------------------------------------------------|------------------|----------------------------------------------------------------|
| [7]  | PIMA + BRFS                                      | Decision Tree, AdaBoost, Gradient Boosting, Extra Trees   | Outlier removal, up-sampling                                                                     | No dedicated XAI | Binary diabetes classification; no personalized recommendation |
| [8]  | BRFS Health Indicators                           | Ensemble ML: XGBoost, RF, CatBoost, LightGBM, KNN         | SMOTE-ENN, XGBoost-based feature selection                                                       | No dedicated XAI | Risk-related feature analysis; no personalized recommendations |
| [9]  | Pima Indians Diabetes Dataset                    | Machine Learning based classifiers                        | Pre-processing, model comparison                                                                 | No dedicated XAI | Early diabetes prediction; no personalized recommendations     |
| [10] | PIMA, BRFS                                       | RF, GB, LightGBM, DT                                      | Mean imputation, IQR, random oversampling, PCA/Boruta                                            | No dedicated XAI | No personalized risk/recommendation                            |
| [11] | PIMA                                             | RF, R-SVM, KNN voting ensemble                            | Ensemble learning                                                                                | No dedicated XAI | Diabetes risk prediction                                       |
| [12] | PIMA Indians Diabetes Dataset                    | AdaBoost                                                  | Standardization, feature selection, data pre-processing/handling                                 | No dedicated XAI | Diabetes Risk assessment                                       |
| [13] | PIMA                                             | Logistic Regression, Gradient Boosting, Decision Tree     | Feature engineering, pre-processing, SMOTE, standardization                                      | No dedicated XAI | Risk assessment and prediction                                 |
| [14] | PIMA Indian Diabetes Dataset                     | KNN, Linear SVM, Random Forest                            | Pre-processing, model comparison                                                                 | No dedicated XAI | No risk/recommendation module                                  |
| [15] | PIMA + private RTML dataset                      | Multiple ML classifiers; XGBoost + ADASYN best            | Mutual Information, SMOTE, ADASYN, Min-Max normalization, GridSearchCV                           | SHAP, LIME       | Web/Android prediction application                             |
| [16] | PIMA Indians Diabetes Dataset                    | Support Vector Machine (SVM)                              | Data cleaning, standardization, train/test split                                                 | No dedicated XAI | No risk/recommendation module                                  |
| [17] | Fasa Adult Cohort Study (FACS), 5-year follow-up | RF, Gradient Boosting, DT, MLP                            | SMOTE, ADASYN, SMOTEENN, Random Over Sampling, KMeansSMOTE; feature-importance analysis          | No dedicated XAI | 5-year diabetes incidence prediction                           |
| [18] | PIMA + second diabetes dataset                   | KNN, NB, SVM, DT, RF, LR + TabNet, XGBoost, MLP ensemble  | Duplicate removal, zero-value imputation, standardization, GridSearchCV                          | No dedicated XAI | No risk/recommendation module                                  |
| [19] | Indian Pima Diabetes Dataset                     | Stacking ensemble: RF, XGBoost, MLP + LR meta-model       | LOF outlier analysis, feature engineering, normalization, GridSearchCV                           | No dedicated XAI | No dedicated risk/recommendation module                        |
| [20] | PIMA + Early Diabetes Risk Prediction dataset    | Boruta + K-Means++ + stacking ensemble (NB, KNN, DT, SVM) | Missing-value handling, outlier removal, Z-score standardization, Boruta, K-Means++, Grid Search | No dedicated XAI | Early diabetes prediction; no recommendation module            |
| [21] | PIMA Indians Diabetes Database                   | KNN, DT, RF, GB, XGBoost, SVM, NN, LR; NN & LR            | EDA, correlation heatmap, stratified split, 10-fold CV                                           | No dedicated XAI | Early risk estimation, no personalized recommendations         |

| Ref. | Dataset / Data                 | Main Approach | Pre-processing / Optimization                                                | XAI              | Risk / Recommendation    |
|------|--------------------------------|---------------|------------------------------------------------------------------------------|------------------|--------------------------|
| [22] | PIMA Indians Diabetes Database | VAE, SAE, CNN | VAE-based data augmentation, SAE-based feature augmentation, class balancing | No dedicated XAI | Early diabetes detection |

## XI. FUTURE RESEARCH DIRECTION

The findings of this review indicate the need for diabetes prediction frameworks that extend beyond conventional binary classification. Future research should investigate the combined use of data pre-processing, feature optimization, comparative machine learning, and hyperparameter tuning to develop reliable prediction models.

The optimized model can then be integrated with SHAP-based patient-level explanations to identify the clinical factors contributing to individual predictions. Prediction probability and relevant clinical characteristics can subsequently be investigated for developing and validating multi-level diabetes risk categories.

A recommendation module can further use the predicted risk level and influential clinical features to provide general preventive healthcare guidance. The complete framework should be evaluated using multiple performance measures and compared with existing approaches under consistent validation conditions.

Such an integrated approach may improve the interpretability, informativeness, and practical usefulness of machine-learning-based diabetes prediction systems while providing a foundation for future clinical decision-support research.

## XII. CONCLUSION

Machine learning has become an important research approach for diabetes prediction, with numerous studies investigating conventional classifiers, ensemble methods, deep learning, pre-processing techniques, feature optimization, and model tuning. The reviewed literature demonstrates that these approaches can achieve promising predictive performance on both benchmark and clinical datasets. The review also shows increasing interest in Explainable Artificial Intelligence, particularly SHAP

and LIME, as well as diabetes risk assessment. However, these capabilities are generally investigated independently, and relatively limited attention has been given to their integration with personalized preventive healthcare recommendations.

The major research opportunity identified in this review is therefore the development of an integrated and interpretable diabetes decision-support framework that combines optimized prediction, patient-level explanation, multi-level risk assessment, and personalized preventive guidance.

Future work should experimentally investigate these components using appropriate datasets, validation strategies, and comprehensive evaluation metrics. The resulting framework can provide a basis for more informative and transparent diabetes prediction systems and support further research toward practical healthcare decision-support applications.

## REFERENCES

- [1] O. Mecili, H. Barkat, F. Nouioua, A. Attia, and S. Akhrouf, "Efficient Machine Learning Approach for Diabetes Mellitus Disease Prediction," in Proceedings of the 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), El Oued, Algeria, Apr. 24–25, 2024, pp. 1–8, doi: 10.1109/PAIS62114.2024.10541247.
- [2] M. M. Islam, H. R. Rifat, M. S. B. Shahid, A. Akhter, M. A. Uddin, and K. M. M. Uddin, "Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, SHAP Analysis, Partial Dependency, and LIME," Engineering Reports, vol. 7, no. 1, Art. no. e13080, Jan. 2025, doi: 10.1002/eng2.13080.
- [3] D. Tripathi, S. K. Biswas, S. Reshmi, A. N. Boruah, and B. Purkayastha, "Diabetes Prediction Using Machine Learning Analytics: Ensemble Learning Techniques," in Proceedings of the 2nd Asian Conference on Innovation in

- Technology (ASIANCON), Ravet, Pune, India, Aug. 26–28, 2022, pp. 1–6, doi: 10.1109/ASIANCON55314.2022.9908975.
- [4] M. S. Al Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, A. Shaikh, and M. A. Elmagzoub, "An Innovative Ensemble Deep Learning Clinical Decision Support System for Diabetes Prediction," *IEEE Access*, vol. 12, pp. 123232–123253, 2024, doi: 10.1109/ACCESS.2024.3436641.
- [5] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.
- [6] H. Naz and S. Ahuja, "Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Jun. 2020, doi: 10.1007/s40200-020-00520-5.
- [7] M. A. H. Saeed, "Diabetes Type 2 Classification Using Machine Learning Algorithms with Up-Sampling Technique," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Art. no. 8, 2023, doi: 10.1186/s43067-023-00074-5.
- [8] D. Das, Aayushman, S. Kumar, M. A. Hussain, and B. R. Reddy, "Diabetes Prediction Using Ensemble Learning Techniques," *Procedia Computer Science*, vol. 258, pp. 3155–3164, 2025, doi: 10.1016/j.procs.2025.04.573.
- [9] R. Sinha, B. S. Vennela, and S. Babu, "Early Diabetes Prediction Using Machine Learning Algorithms," in *Proc. 2024 5th IEEE International Conference on Advances in Information, Communication and Computing Technology (ICAAIC)*, 2024, doi: 10.1109/ICAAIC60222.2024.10575581.
- [10] G. R. Ashisha, X. Anitha Mary, E. G. M. Kanaga, J. Andrew, and R. Jennifer Eunice, "Random Oversampling-Based Diabetes Classification via Machine Learning Algorithms," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Art. no. 270, 2024, doi: 10.1007/s44196-024-00678-3.
- [11] B. Amma N. G., "En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus," *Egyptian Informatics Journal*, vol. 25, Art. no. 100441, 2024, doi: 10.1016/j.eij.2024.100441.
- [12] A. A. Ibrahim and A. Alhialy, "Enhancing Diabetes Risk Assessment in PIMA Indians: A Machine Learning Approach Using AdaBoost," *International Journal of Computer Engineering in Research Trends (IJCERT)*, vol. 10, no. 6, pp. 1–8, 2023, doi: 10.22362/ijcert.v10i6.857.
- [13] S. S. Bhat, M. Banu, G. A. Ansari, and V. Selvam, "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms," *Healthcare Analytics*, vol. 4, Art. no. 100273, 2023, doi: 10.1016/j.health.2023.100273.
- [14] M. Pal, S. Parija, and G. Panda, "Improved Prediction of Diabetes Mellitus using Machine Learning Based Approach," in *Proc. 2021 2nd International Conference on Range Technology (ICORT)*, Chandipur, Balasore, India, 2021, pp. 1–6, doi: 10.1109/ICORT52730.2021.9581774.
- [15] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.
- [16] J. Nedeljković, "Diabetes Prediction Using Machine Learning," in *Proc. 17th Int. Scientific Conf. on Computer Science and Education (IEEESTEC)*, 2024, doi: 10.46793/IEEESTEC17.063N.
- [17] M. T. Moghaddam, Y. Jahani, Z. Arefzadeh, A. Dehghan, M. Khaleghi, M. Sharafi, and G. Nikfar, "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," *BMC Medical Research Methodology*, vol. 24, Art. no. 220, 2024, doi: 10.1186/s12874-024-02341-z.
- [18] P. Chowdhury, P. Barua, and M. N. Uddin, "Diabetes Prediction Using Machine Learning and Hybrid Deep Learning Ensemble Technique," in *Proc. 2024 IEEE International Conference on Computer, Power and Communications Systems (COMPAS)*, 2024, doi: 10.1109/COMPAS60761.2024.10796486.
- [19] A. Jain and S. Jain, "A Novel Hybrid Model for Efficient Prediction of Diabetes Using Machine Learning," in *Proc. International Conference on*

Intelligent Data Communication Technologies and Internet of Things (IDICAIEI), 2024, doi: 10.1109/IDICAIEI61867.2024.10842728.

- [20] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, Art. no. 224, 2023, doi: 10.1186/s12859-023-05300-5.
- [21] N. Tripathy, S. Hota, P. Satapathy, S. K. Nayak, D. S. K. Nayak, and P. P. Nayak, "Diabetes Prediction in Healthcare Utilizing Machine Learning Algorithms," in *Proc. IEEE Int. Conf. on Artificial Intelligence and Healthcare Computing (ICAHC)*, 2025, doi: 10.1109/ICAHC64101.2025.10957296.
- [22] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Computer Methods and Programs in Biomedicine*, vol. 202, Art. no. 105968, 2021, doi: 10.1016/j.cmpb.2021.105968.