International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

Diabetes Prediction Using Optimized Ensemble Machine Learning on Clinical and Lifestyle Indicators

E Pavithra[1], K Bharath Kumar[2], C Venkatesh[3]

[1]Assistant Professor, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur, E-mail: epavithra526@gmail.com, ORC-ID: https://orcid.org/0009-0006-8871-4551

[2]P.G Scholar, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur, E-mail: bharathkumar1418@gmail.com, ORC-ID: https://orcid.org/0009-0002-2009-844X

[3]Assistant Professor, Department of CSE(AI & ML), Sri Venkatesa Perumal College of Engineering & Technology, Puttur, E-mail: chevireddyvenkatesh22@gmail.com, ORC-ID: https://orcid.org/0009-0007-1038-5978

Abstract: Early diagnosis of diabetes is important in reducing the associated health risks and in enabling quick medical interventions. This paper presents a machine learning forecasting model based on an interpretable framework using the Diabetes Health Indicators Dataset on Kaggle. The data set contains clinical and lifestyle health variables that act as input attributes of prediction. The classification of persons at risk was done using different algorithms including Logistic Regression, Random Forest, Decision Tree, CatBoost, and an ensemble Voting Classifier that combines XGBoost and LightGBM. To overcome the issue of data imbalance in the advanced models, Synthetic Minority Oversampling Technique (SMOTE) was applied to increase class representation and model generalization. Explainable Artificial Intelligence (XAI) interpreters, LIME and SHAP were added to provide both local and global interpretability, explaining the most important health variables influencing the prediction of diabetes. The Voting Classifier achieved the best accuracy of 92 with respect to single models such as Logistic Regression and the Random Forest, thus demonstrating the effectiveness of combining sophisticated resampling and ensemble techniques in making transparent and high-performance predictive models in healthcare analytics.

"Index Terms - *Diabetes prediction, Explainable Artificial Intelligence (XAI), Logistic Regression, Random Forest, SMOTE, LIME, SHAP, Healthcare systems".*

1. INTRODUCTION

ML has become an important component of the healthcare business, where the data-based methodology is used to predict diseases accurately and act in time. Diabetes is a major global challenge under the category of chronic diseases due to its increasing rates, and severe chronic outcomes and consequences, particularly in the third world countries [1], [2]. The presence of diabetes, which is characterized by the persistent high levels of glucose in the body because of the impaired functioning of insulin or its effects, substantially elevates the risk of cardiovascular disease, kidney failure and neuropathy. The rising cases of Type 2 Diabetes (T2D) demonstrate the need to develop reliable diagnostic and prognostic tools that enable physicians to promptly make decisions [3], [4].

The traditional statistical models often have limitations when they are used on complex and high-dimensional healthcare data. ML algorithms offer better predicted accuracy and extensiveness to diverse datasets. However, their obscure nature poses clinical implementation challenges, as healthcare professionals desire not accurate predictions but clear explanations of the rationale behind it [5]. Explainable Artificial Intelligence (XAI) has been introduced as a solution to this issue and ensures transparency and trust in ML-based healthcare systems [6], [7]. Elucidation is especially crucial in diabetes prediction, as the understanding of the relative impact of lifestyle, demographic, and clinical variables may result in the increased treatment choice and preventative care [8].

Such techniques as LIME and SHAP can provide an interpretability at both local and global levels, allowing physicians to identify both local and global importance to certain features of specific cases and overall trends among populations [9], [10]. ML models may address the conflict between computational precision and clinical confidence by combining predictive ability and interpretability. The aim is to create powerful diabetes predictive model based on Logistic Regression and RF classifier with SMOTE to address the problem of class imbalance and with the help of LIME and SHAP to have interpretability. The contributions include increased precision of prediction with a balanced impact of features as well as the presentation of interpretable information that

Received: 28-02-2026 | Accepted: 01-04-2026 | Published: 08-04-2026

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

supports the use of ML in the risk assessment of diabetes.

2. RELATED WORK

The latest developments in ML have had a significant effect on the field of healthcare analytics, particularly the early detection and the forecasting of diabetes. Many studies have focused on the use of traditional and ensemble learning to improve the accuracy of the prediction and reduce the shortcomings that are present in classical statistical methodologies. Conventionally, prediction of diabetes has relied on logistic regression and linear models, which are easy to interpret but are often faced with complex nonlinear trends of patient data [11]. In order to solve these problems, improved ML algorithms, such as as DT, SVM, and ensemble methods, have been used to explain complicated correlations between clinical and demographic variables. These models have proven to be very effective in terms of predictive accuracy as compared to the traditional methods [12].

The feature selection would improve the model efficiency and generalizability. Various researches have shown that preprocessing methods, such as, normalization, outlier removal and missing values handling can greatly determine the performance of the model [13]. Mahmud et al. [13] focused on the role of feature selection and preprocessing in improving predicting accuracy in Type 2 Diabetes (T2D) by focusing on the clinically relevant indicators, i.e., fasting blood glucose, BMI, and age. Gradient boosting algorithms and the Ensemble learning techniques, in particular, the RF have gained popularity because of their robustness, as well as, their ability to handle imbalanced data sets. These methods combine a number of weak learners to produce stronger classifiers, which boosts both accuracy and stability of diabetes prediction tasks [14].

The integration of XAI practices to improve interpretability in ML models has been studied by many researchers. Healthcare requires transparency because clinicians must have information about the model logic in order to make informed recommendations. Techniques such as LIME and SHAP have increasingly been used to explain complex models through measuring the contribution of features to the model both at the individual and global levels [15]. Tabaei and Herman [17] presented the efficacy of multivariate logistic regression to identify the risk variable of early diabetes, emphasizing on the importance of the results being in clinically interpretable formats. It has been shown that the use of feature importance measures within ensemble models boosts the process of identifying key predictors, therefore, guiding preventive interventions [16].

The imbalance of data is a major problem of prediction modeling of diabetes. Unbalanced datasets, where there is a large number of non-diabetic cases versus diabetic cases, may lead to biased models which fail to properly identify high-risk patients. Solutions such as oversampling, undersampling and synthetic data generation (e.g. SMOTE) have been successfully used to combat this problem and enhance model sensitivity [18]. Besides, recent research emphasizes the need to use a personalized and context-specific prediction. Individual health profiles, lifestyle factors, comorbidities are more reliable and useful when combined in models that predict risks as compared to generic risk prediction models [19].

DL methods have been explored to predict diabetes, in particular, to learn complex relationships between laboratory and clinical predictors. These approaches, though offering high predictive quality, have been found to have interpretative limitations and this has led to implementation of XAI techniques to ensure clinical relevance is upheld [20]. The combination of model efficacy and interpretability has led researchers to come up with systems that are able not only to make precise predictions concerning the onset of diabetes but also explain the major variables that affect the conclusion made by the model. This approach enhances clinical conviction and leads to the high-quality management of patients.

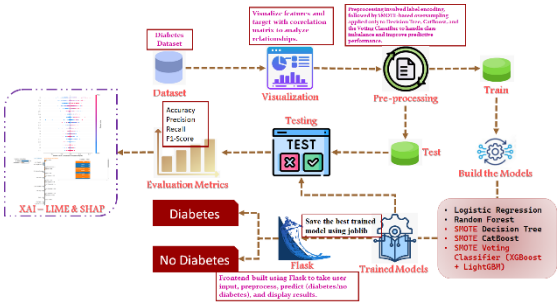
3. MATERIALS AND METHODS

The suggested solution offers a sophisticated machine learning model of the early detection of diabetes based on clinical and lifestyle variables. It employs different algorithms, such as the LR, RF, SMOTE DT, SMOTE CatBoost, and an SMOTE Voting Classifier integrating XGBoost and LightGBM to enhance the accuracy of prediction and model balance. LR is a basic linear model that is improved by the RF and CatBoost based on the ensemble and gradient boosting techniques [21]. The use of SMOTE is an effective way of solving the problem of class imbalance by creating artificial

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research

Paper

samples of underrepresented cases, therefore, encouraging fair model training [22]. The Voting Classifier increases predictive stability by taking the combination of the results of multiple models to make more correct decisions [23]. This hybrid model is more reliable, interpretable, and resilient, offering an entire and advanced predictive model that enables early diagnosis of diabetes and supports the use of data to make clinical judgments to preventive healthcare uses.



"Fig.1 Proposed Architecture"

Figure 1 represents a diabetes prediction system based on ML and explainable AI. The process starts with the collection of raw diabetic data and it is followed by preprocessing process which includes cleaning, integration, transformation, reduction, and discretization. The processed data is further separated into training and test sets to be used to train a ML model, e.g. LR. A model yields a direct output which can be used by a human being in making decisions. An explainable ML component is also part of the system and it has models like LR and RF and explainers like LIME and SHAP, which are used to explain the predictions. This is then explained and the predictive outcome then applied by an individual in order to make a final, informed decision.

i) Dataset Collection:

The Diabetes Health Indicators Dataset was used in this research, and it contains 253,680 records, including 22 attributes of clinical, lifestyle, and demographic data that are relevant to predict diabetes. Necessary variables include blood pressure (HighBP), cholesterol level (HighChol), body mass index (BMI), smoking (yes/no), heart disease history, physical activity, and general health indicators. The socioeconomic and behavioral determinants such as education, income, and access to healthcare are combined to provide a comprehensive description of risk determinants. Attributes are all numerical and there are no missing values thus allowing the direct model training. The data set is also useful in the development of predictive models that may be used to explain a complex association between health markers and risk of diabetes and offers a solid foundation to interpretable ML techniques [24].

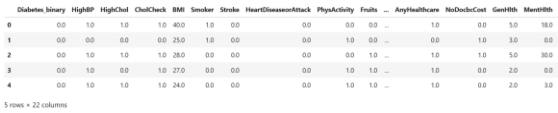


Fig.2 Dataset Collection

ii) Pre-Processing:

Pre-processing is a fundamental stage in the preparation of a dataset to be used in ML to ensure the quality of data, its consistency, and its suitability to be used in training a model. It includes the handling of absent values, encoding of category variables and normalizing numerical features. Good pre-processing reduces noise, mitigates bias, and enhances the effectiveness of the model, creating a platform of accurate and interpretable predictive diabetes.

a) Data Processing: The first step in data processing will be the segregation of input features and the goal variable, which will be denoted by X and y respectively. Label encoding changes categorical variables, such as sex and education, into numerical representation so that they can be compatible with models. The continuous variables e.g. BMI, age and blood pressure are normalized to get a consistent scaling thus ensuring that the features with a wider range do not overwhelm the learning process. The missing values are processed by imputation or drop, which does not compromise the data. This process ensures that the dataset is cleansed, balanced, and suitable to train ML algorithms, thus improving the effectiveness and reliability of prediction in the risk assessment of diabetes.

b) Data Visualization: Data visualization explains the distribution of features, trends and association of the data. The skewness, outliers, and trends of individual feature columns are observed using histograms, boxplots and density plots. The target variable is represented to measure the distribution of the classes with a focus on potential imbalance issues. A correlation matrix is developed to quantify linear correlations among features and the goal that indicates multicollinearity and the most important predictors. Visual exploration enables one to understand the interactions and effects among

Received: 28-02-2026 | Accepted: 01-04-2026 | Published: 08-04-2026

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

numerous health indicators such as BMI, blood pressure and physical activity on diabetes risk. The findings are used to choose features, interpret models, and resampling methods that follow.

c) Data Resampling: The resampling of data helps to reduce the imbalance of classes that can make ML models biased towards majorities. SMOTE is applied in advanced models to generate synthetic samples of the minority group, therefore, reaching a more balanced proportion between diabetes and non-diabetic cases. The SMOTE algorithm works by creating interpolated samples between the already existing minority instances, preserving feature associations and improving representation. Resampling will improve the ability of the model to detect high-risk scenarios, therefore, increasing the sensitivity and general prediction effectiveness. Combining resampling and standardized and encoded features optimizes the data to be used in training robust and interpretable models that provide accurate and reliable predictions of diabetes.

iii) Train & Test:

The dataset is further segmented into training and testing subsets in a ratio of 80: 20 in order to determine the model performance effectively. The 80 percent of the data, known as training set is used to fit and optimize the prediction models so that they are able to identify patterns and correlations between the variables in the input and the diabetic outcome. The remaining 20 percent is the testing set that provides an unbiased evaluation of the ability of the model to extrapolate to new data. This split ensures that the model will not be over-specific to the training data and will have the ability to predict the risk of diabetes in new people. Maintaining this difference enables consistency and application of the same evaluation of performance across experiments.

iv) Algorithms:

Logistic Regression It is an often used supervised method of learning used in classification problems, particularly when the variable of interest is discrete. It approximates the probability of an occurrence using a logistic function on a linear sum of the input characteristics, which enables an easy interpretation of the influence of every component on the outcome. The Logistic Regression is used to predict diabetes by either classifying the patients as being at risk or not, based on clinical and lifestyle variables including BMI, age, and smoking habits [25]. This is because its simplicity and interpretability makes it indispensable in health care, as it allows practitioners to measure the contribution of every health component to the expected risk. Besides, Logistic Regression is computationally efficient, has accurate boundaries of decisions, and enables high-risk patients to be identified earlier.

The formula makes predictions of probabilities of classes using a logistic sigmoidal formula.

$$\hat{y}_i = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}} \qquad (1)$$

Random Forest is an ensemble learning method that produces a large number of decision trees in the training process and combines their predictions to give final classifications. It is able to effectively handle high-dimensional data and explain complex relationships between characteristics, which makes it suitable in diabetes prediction [26]. Random Forest methodology categorizes people based on the health variables of blood pressure, BMI, age, and physical activity, and therefore it is stronger than a single-tree model. The large number of trees helps in reducing the threat of overfitting and also produces consistent and reliable forecasts. The scores of the features of importance obtained in the RF contribute to a better interpretation of the factors that are significant in relation to the risk of diabetes and help to identify these factors. Its strength and adaptability make it a preferred choice of healthcare datasets.

The Gini Equation presented below:

$$Gini = 1 - \sum_{i=1}^{C} (Pi)^2 \qquad (2)$$

The **SMOTE-Decision Tree** model is a hybrid of SMOTE and a DT classifier to address the imbalance in data and enhance the predictive accuracy. SMOTE generates artificial examples of underrepresented classes, therefore, balancing out before training. The DT algorithm then separates data based on feature values in order to form separate, understandable decision paths. This combination will enable fair learning outcomes and improve the generalization of unknown data. It helps to determine the key features of health and correctly estimate the level of risk in diabetes forecasting. The key task of SMOTE-Decision Tree is to solve the problem of class imbalance and maintain the interpretability and consistency of the model in predictive analysis of health.

Received: 28-02-2026 | Accepted: 01-04-2026 | Published: 08-04-2026

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

SMOTE-CatBoost is a combination of synthetic data balancing and a gradient boosting model that can be optimally utilized to work with categorical and numerical variables. CatBoost builds up a series of decision trees iteratively whereby each tree is designed to reduce the residuals of the previous iterations, therefore, improving the predicting accuracy. The combination of SMOTE allows the balance of the diabetic datasets and reduce the bias in favor of the major classes. CatBoost overfitting is alleviated through ordered boosting and advanced regularization, which provide accurate and reliable results in healthcare prediction problems. It is computationally effective and requires minimum preparation. The SMOTE-CatBoost is meant to achieve high-performance diabetes risk categorization through balanced data distribution and advanced boosting algorithms that can make accurate and interpretable clinical predictions.

The **SMOTE-Voting Classifier** is a combination of a range of algorithms, such as XGBoost and LightGBM, that are merged into an ensemble structure to enhance the efficacy of diabetes prediction. Firstly, SMOTE is used to balance the data set to ensure that both diabetic and non-diabetic cases are represented equally. The Voting Classifier then combines the prediction of the two models through hard or soft voting producing more reliable and generalized predictions. XGBoost is a better optimizer of gradient boosting and LightGBM is more efficient through histogram-based learning. Such a combination capitalizes on the strengths of the two algorithms making both less biased and more accurate. The aim is to offer an integrated and strong predictive tool that would deliver uniform results in the different patient health profiles.

4. RESULTS AND DISCUSSIONS

Accuracy: The test accuracy is the ability to make the right diagnosis between a patient and a healthy case. In order to determine the accuracy of a test, the ratio of true positives and true negatives should be calculated, in all the cases that have been assessed. This mathematically can be said to be:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (3)$$

Precision: Precision determines the percentage of correctly classified cases out of the cases that are declared to be positive. Accordingly, the equation of determining the precision is as follows:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (4)$$

Recall: In ML, recall is a measure that determines the ability of a model to identify all relevant examples of a certain class. The proportion of the correct number of positive results that were predicted to the actual number of positive results, which present clues to the effectiveness of a model in detecting events of a particular category.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

F1-Score: The F1 score is a measure that is taken to assess the accuracy of a ML model. It combines the recall and the precision of a model. The accuracy measure will be used to determine how many times a model correctly predicts in the whole dataset.

$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 \quad (6)$$

Table.1 Performance Evaluation

ML Model	Accuracy (CA)	Precision	Recall	F1 Score
Logistic Regression	0.866	0.712	0.571	0.589
Random Forest	0.860	0.681	0.573	0.590
SMOTE-Decision Tree	0.878	0.878	0.878	0.878
SMOTE-CatBoost	0.869	0.869	0.869	0.869
SMOTE-Voting Classifier	**0.920**	**0.925**	**0.920**	**0.920**

The table compares different models, which means that the SMOTE-Voting Classifier has the best balance accuracy, precision, recall, and F1-score when predicting diabetes.

Fig.3 Comparison Graph

Received: 28-02-2026 | Accepted: 01-04-2026 | Published: 08-04-2026

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research

Paper

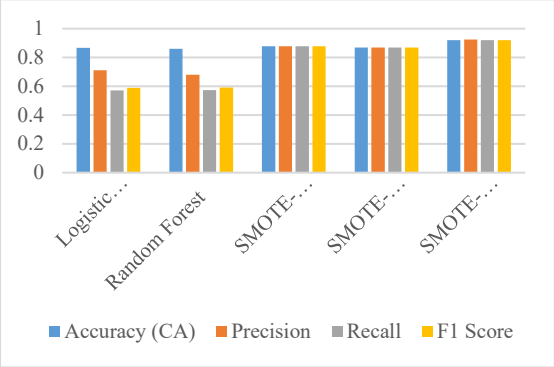


Figure 3 compares the Accuracy, Precision, Recall, F1 Score of five models based on classification, with a high light on improvements in performance between SMOTE and the baseline models.

Fig.4 Upload Input Data

Fig. 4 displays the input information where the user must feed the health indicators such as HighBP, HighChol, CholCheck, BMI, etc., to receive a prognosis.

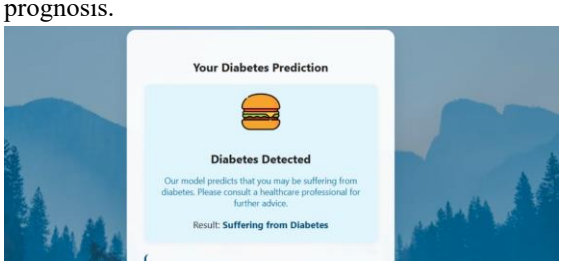


Fig.5 Predict Result

Figure 5 depicts a pop-up with the prediction result, Diabetes Detected, and the suggestion to use the professional consultation.

Fig.6 Upload Another Input Data

In Fig. 6, users are required to key in health indicators in HighBP, HighChol, CholCheck, and BMI in diabetes prediction.

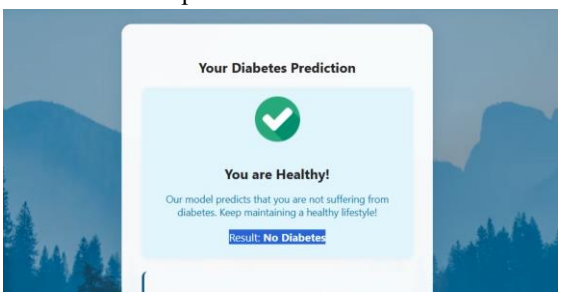


Fig.7 Final Outcome

Fig. 7 shows that in the system, there is a pop-up that shows a prediction outcome of "You are Healthy" and the conclusion of no Diabetes.

5. CONCLUSION

The developed explainable ML system served as a good demonstration of the potential of predictive analytics to determine the risk of diabetes based on health markers. By applying different techniques and using resampling techniques in the sophisticated models, the system achieved significant improvements in prediction accuracy and interpretability in order to deal with class imbalance. Integration of XAI algorithms, LIME and SHAP, provided a detailed explanation of the correlation between features and model decisions, which allows a clear understanding of the key health determinants. The Voting Classifier, a combination of XGBoost and LightGBM, had the highest predictive accuracy of 92, higher than other models such as the Logistic Regression and the RF. This outstanding performance highlights the effectiveness of ensemble learning and evenly distributed data

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

training in improving diagnostic accuracy. This approach will ensure reliable decision support in healthcare analytics that enhances reliability and transparency in prediction outcomes and supports the advancement of interpretable and data-driven medical analysis. The framework provides an example of how explainable ensemble learning can greatly improve medical prediction systems and reconcile model performance with interpretability to support clinical professionals to make informed, evidence-based decisions to detect diabetes early and provide care to patients.

Future studies can focus on exploring a broader spectrum of advanced ML and DL algorithms to enhance the predictive performance and the ability of diabetes detection models to generalize. Combining longitudinal and temporal health data could bring a further knowledge of patterns of progression and early risk identification. Further attention can be paid to the process of feature selection and dimensionality reduction to minimize redundancy and increase computing power. Evaluation of the model on diverse real-world data dealing with diverse demographics would make the model more robust and applicable in the clinical environment. Given that real-time data streams, including wearable measurements, are integrated, it might be possible to improve customized monitoring and enable ongoing assessment of the risk of diabetes in real-life health care systems.

REFERENCES

[1] Khokhar, P. B., Pentangelo, V., Palomba, F., & Gravino, C. (2025). Towards transparent and accurate diabetes prediction using machine learning and explainable artificial intelligence. *arXiv preprint arXiv:2501.18071*.

[2] Allani, U. (2025). Interactive Diabetes Risk Prediction Using Explainable Machine Learning: A Dash-Based Approach with SHAP, LIME, and Comorbidity Insights. *arXiv preprint arXiv:2505.05683*.

[3] Vivek Khanna, V., Chadaga, K., Sampathila, N., Prabhu, S., Chadaga P, R., Bhat, D., & KS, S. (2024). Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers. *Cogent Engineering*, *11*(1), 2330266.

[4] Vinh, T. Q., & Byeon, H. (2024). Towards transparent diabetes prediction: Unveiling the factors with explainable ai. *International Journal of Engineering Trends and Technology*, *72*(5).

[5] Alam, R., Atif, M., Ahmad, N., Kalal, T. V., & Khune, A. (2025, January). Enhancing early-stage diabetes prediction with Explainable Artificial Intelligence (XAI). In *2025 International Conference on Ambient Intelligence in Health Care (ICAIHC)* (pp. 1-6). IEEE.

[6] N. Nipa, M. H. Riyad, S. Satu, Walliullah, K. C. Howlader, and M. A. Moni, ‘‘Clinically adaptable machine learning model to identify early appreciable features of diabetes,’’ Intell. Med., vol. 4, no. 1, pp. 22–32, Feb. 2024.

[7] Y. Zhao, J. K. Chaw, M. C. Ang, M. M. Daud, and L. Liu, ‘‘A diabetes prediction model with visualized explainable artificial intelligence (XAI) technology,’’ in Advances in Visual Informatics (Lecture Notes in Computer Science), vol. 14322, H. B. Zaman et al., Eds., Singapore: Springer, 2024, doi: 10.1007/978-981-99-7339-2_52.

[8] L. P. Joseph, E. A. Joseph, and R. Prasad, ‘‘Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture,’’ Comput. Biol. Med., vol. 151, Dec. 2022, Art. no. 106178.

[9] T. A. Assegie, T. Karpagam, R. Mothukuri, R. L. Tulasi, and M. Fentahun Engidaye, ‘‘Extraction of human understandable insight from machine learning model for diabetes prediction,’’ Bull. Electr. Eng. Informat., vol. 11, no. 2, pp. 1126–1133, Apr. 2022.

[10] N. Gandhi and S. Mishra, ‘‘Explainable AI for healthcare: A study for interpreting diabetes prediction,’’ in Proc. Mach. Learn. Big Data Analytics Int. Conf. Mach. Learn. Big Data Analytics (ICMLBDA). Cham, Switzerland: Springer, 2022, pp. 95–105.

[11] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, ‘‘A decision support system for diabetes prediction using machine learning and deep learning techniques,’’ in Proc. 1st Int. Informat. Softw. Eng. Conf. (UBMYK), 2019, pp. 1–4.

[12] Swapna, G., Sreenivasulu, K., Deepika, M., Baseer, K. K., Neerugatti, V., & Viswanath, G. (2025). Brain tumour detection using MRI images in CNN. Advances in Science, Engineering and Technology, 6.

[13] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, ‘‘Prediction of diabetes using machine learning algorithms in healthcare,’’ in Proc. 24th Int. Conf. Autom. Comput. (ICAC), Sep. 2018, pp. 1–6.

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

[14] M. U. Emon, M. S. Keya, Md. S. Kaiser, Md. A. islam, T. Tanha, and Md. S. Zulfiker, ‘‘Primary stage of diabetes prediction using machine learning approaches,’’ in Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS), Mar. 2021, pp. 364–367.

[15] Md. K. Hasan, Md. 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.

[16] A D Venkatesh, K Bhaskar, G Swapna, & G Viswanath. (2025). Advanced Hybrid Learning Architecture for Precision Cardiovascular Risk Assessment. In International Journal of Health Sciences and Pharmacy (IJHSP) (Vol. 9, Number 1, pp. 50–61). Zenodo. https://doi.org/10.5281/zenodo.15448632

[17] B. P. Tabaei and W. H. Herman, ‘‘A multivariate logistic regression equation to screen for diabetes: Development and validation,’’ Diabetes Care, vol. 25, no. 11, pp. 1999–2003, 2002.

[18] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, ‘‘Accurate diabetes risk stratification using machine learning: Role of missing value and outliers,’’ J. Med. Syst., vol. 42, no. 5, pp. 1–17, May 2018.

[19] A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza-Albarrán, and K. L. Ramaiya, ‘‘Diabetes in developing countries,’’ J. Diabetes, vol. 11, no. 7, pp. 522–539, 2019.

[20] Patel, S., & Patyrykin, K. (2025). Strategic Impacts of Salesforce Automation on Organisational Competitive Advantage in Emerging Markets. Journal of Posthumanism, 5(12), 357–372. https://doi.org/10.63332/joph.v5i12.3782.

[21] H. Naz and S. Ahuja, ‘‘Deep learning approach for diabetes prediction using PIMA Indian dataset,’’ J. Diabetes Metabolic Disorders, vol. 19, no. 1, pp. 391–403, Jun. 2020.

[22] V. C. Bavkar and A. A. Shinde, ‘‘Machine learning algorithms for diabetes prediction and neural network method for blood glucose measurement,’’ Indian J. Sci. Technol., vol. 14, no. 10, pp. 869–880, Mar. 2021.

[23] M. Rout and A. Kaur, ‘‘Prediction of diabetes risk based on machine learning techniques,’’ in Proc. Int. Conf. Intell. Eng. Manage. (ICIEM), Jun. 2020, pp. 246–251.

[24] Kaliappan, J., Saravana Kumar, I. J., Sundaravelan, S., Anesh, T., Rithik, R. R., Singh, Y., ... & Srinivasan, K. (2024). Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. *Frontiers in Artificial Intelligence*, *7*, 1421751.

[25] Hasan, R., Dattana, V., Mahmood, S., & Hussain, S. (2024). Towards transparent diabetes prediction: combining automl and explainable AI for improved clinical insights. *Information*, *16*(1), 7.

[26] Ganesh, B. R., B M, P., Prasad K, K., Swapna, G., & G, Viswanath. (2025). Data Mining-Driven Multi-Feature Selection for Chronic Disease Forecasting. Journal of Neonatal Surgery, 14(5S), 108–124. https://doi.org/10.52783/jns.v14.1993

[27] Poojari, R. Frameworks for Data Management and Lineage in Large-Scale Healthcare Data Systems.

[28] Tanim, S. A., Shrestha, T. E., Emon, M. R. I., Mridha, M. F., & Miah, M. S. U. (2025). Explainable deep learning for diabetes diagnosis with DeepNetX2. *Biomedical Signal Processing and Control*, *99*, 106902.

[29] G Loge, T Sunil Kumar Reddy, G Swapna, & G Viswanath. (2025). Interpretable AI for Precision Brain Tumor Prognosis: A Transparent Machine Learning Approach. In International Journal of Health Sciences and Pharmacy (IJHSP) (Vol. 9, Number 1, pp. 180–195). Zenodo.

[30] Annuzzi, G., Apicella, A., Arpaia, P., Bozzetto, L., Criscuolo, S., De Benedetto, E., ... & Prevete, R. (2023). Exploring nutritional influence on blood glucose forecasting for type 1 diabetes using explainable AI. *IEEE journal of biomedical and health informatics*, *28*(5), 3123-3133.