International Journal of Web of Multidisciplinary Studies

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

website: https://ijwos.com OPEN ACCESS E-ISSN : 3049-2424

Volume 3 Issue 8 , August 2026 DOI: https://doi.org/10.71366/ijwos

IJWOS International Journal of Web of Multidisciplinary Studies

HealthAI: An Intelligent Machine Learning Approach to Early Prediction of Diabetes Risk

Rajashekhara G C[1], Kadli Sagar[2]

[1] Assistant Professor and Director, Faculty of Computing & IT, GM University, Davanagere, Karnataka, India

[2] 2nd Year MCA Student, Faculty of Computing & IT, GM University, Davanagere, Karnataka, India.

Article Info

Article History:

Published: 27 August 2026

Publication Issue:
Volume 3, Issue 8
August-2026

Page Number:
649-663

Corresponding Author:
Kadli Sagar

Abstract:

Diabetes is a persistent disease that needs to be diagnosed early and effectively managed to prevent the emergence of any further complications. In most instances, individuals may not be aware of their risk factors of diabetes until their symptoms become evident.This creates a need for systems that can examine commonly available health information and provide an early indication of possible diabetes risk. Machine learning offers a practical approach for analysing multiple patient-related factors and identifying patterns that may be difficult to assess manually.

The current paper describes HealthAI - an intelligent machine learning approach that can be used for predicting the risk of developing diabetes at the early stages and for decision support in healthcare. Clinical and demographic features, including patient's age, body mass index, glucose level, related to blood pressure, level of HbA1c, heart diseases, smoking habits, and other features of the health condition are taken into account. The proposed model structure includes data pre-processing, feature selection, modeling, prediction, and evaluation steps. Different classification algorithms like Logistic Regression, Decision Tree, Random Forest, SVM, K-Nearest Neighbor, and Naïve Bayes can be used. The performance of the model could be evaluated based on accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.HealthAI model could provide early information regarding the risk that will allow preventing diabetes. It is important to mention that the suggested solution cannot be considered as a diagnostic tool, but rather it serves as an assistance for making decisions.

Keywords: Diabetes Mellitus, Machine Learning, Diabetes Risk Prediction, Predictive Analytics, Healthcare Decision Support, Feature Selection, HealthAI

1. Introduction

Diabetes is an illness caused by complications related to blood sugar levels in the body. In case this problem is not diagnosed or treated, then different diseases may arise related to the heart, kidneys, eyes, and nerves. Hence, early diagnosis of diabetes can help one take steps to prevent it.

Machine learning could assist in analysing multiple health parameters and finding correlations related to diabetes. Parameters such as glucose level, BMI, age, HbA1c, blood pressure, heart disease, and smoking could help in making such a prediction.

This paper offers a HealthAI framework for machine learning aimed at the early detection of diabetes risks and decision-making in health care sector. The proposed framework would include several stages: data preprocessing, feature selection, machine learning algorithm training, and performance evaluation.

Some of the algorithms that can be compared include Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and Naïve Bayes. Accuracy, Precision, Recall, F1-Score, and ROC-AUC can be some of the measures used to evaluate these models.

HealthAI framework is supposed to serve as preliminary risk prediction tool and help make decisions in health care area. It is not expected to replace doctors and make medical diagnosis.

2. Problem Statement

Diabetes risk can be difficult to identify at an early stage because different health factors can contribute to the condition. Manually considering several factors together can also make risk assessment more difficult when large amounts of patient data are involved.

Many existing approaches focus on a limited number of health parameters or use a single machine learning model. There is therefore a need for a structured approach that can preprocess healthcare data, identify relevant features, compare different machine learning algorithms, and provide a useful early risk indication.

HealthAI addresses this problem by using clinical and demographic information to build and evaluate machine learning models for early diabetes risk prediction. The framework is intended to support preventive healthcare and decision-making while keeping professional medical evaluation as the final authority.

3. Objectives

The main objective of **HealthAI** is to develop a machine learning framework that can provide an early indication of diabetes risk using relevant health information.

The specific objectives are:

1. To collect and prepare healthcare data for diabetes risk prediction.
2. To preprocess the dataset by handling missing, inconsistent, and categorical data.
3. To identify relevant health features for diabetes prediction.
4. To train different machine learning models for diabetes risk classification.
5. Comparing Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and Naïve Bayes.
6. Evaluating the models on the basis of Accuracy, Precision, Recall, F1-Score, and ROC-AUC.
7. To identify a suitable model based on the actual experimental results.
8. To provide an early diabetes risk indication through the HealthAI framework.
9. To support preventive healthcare and healthcare decision-making.
10. To provide a foundation for future improvements and clinical validation.

4. Research Gap

Several studies have proved that it is possible to use machine learning approaches in order to help predict the onset of diabetes through the analysis of patient's health data. The use of several types of supervised learning approaches such as logistic regression, decision tree, random forest, SVM, KNN, and naïve Bayes has shown promising results when used to increase the accuracy of prediction.

Although these studies have reported encouraging results, several research gaps still exist. Many existing models are developed and evaluated using a single dataset, making it difficult to determine whether the prediction model can generalize well across different healthcare environments. In addition, some studies focus primarily on improving classification accuracy without providing a comprehensive framework that integrates data preprocessing, feature engineering, model comparison, and healthcare decision support into a single system.

Another limitation observed in the literature is that several prediction models rely mainly on performance metrics while offering limited support for practical clinical use. Most systems generate only a prediction result and do not provide structured risk assessment or decision-support information that could assist healthcare professionals during the initial evaluation of patients. Consequently, the practical usefulness of many prediction models remains limited despite their high reported accuracy.

The reviewed literature also indicates that different machine learning algorithms perform differently depending on the characteristics of the dataset and the selected clinical features. Therefore, relying on a single algorithm may not always produce the most reliable prediction model. Comparative evaluation of multiple supervised learning algorithms using common performance metrics can provide a more balanced understanding of their effectiveness for diabetes prediction.

To address these gaps, this research proposes **HealthAI**, an intelligent machine learning framework for early diabetes risk prediction and healthcare decision support. The proposed framework combines data preprocessing, feature selection, comparative evaluation of multiple supervised machine learning algorithms, diabetes risk prediction, and healthcare decision support within a unified architecture. Unlike approaches that emphasize prediction accuracy alone, the proposed framework aims to support both reliable risk assessment and informed healthcare decision-making, thereby contributing to preventive healthcare and improving the practical applicability of machine learning in diabetes prediction.

5. Literature Review

Machine learning (ML) techniques that have been developed recently have helped to enhance the accuracy and efficiency of diabetic prediction models. The researchers have looked into several machine learning algorithms and techniques for feature and data selection to predict those people who are prone to develop diabetes.

Study	Methodology	Dataset / Features	Key Findings	Limitations
Diabetes Detection Based on Machine Learning and Deep Learning Approaches (2024)	Review of ML and DL techniques for diabetes prediction	Multiple public datasets and non-invasive features	ML and DL significantly improve prediction accuracy and support early diagnosis.	No unified framework or standard data collection procedure is proposed.
Diabetes Prediction Using Machine Learning Model (2022)	Logistic Regression, Feature Selection and Ensemble Learning	PIMA Indian Diabetes Dataset and Vanderbilt Dataset	Ensemble methods and feature selection improve prediction performance compared to a single model.	Primarily focuses on algorithm performance rather than clinical decision support.
Survey on Diabetes Risk Prediction Using Machine Learning Approaches (2022)	Comparative survey of supervised ML algorithms	Multiple diabetes prediction studies	Different algorithms perform differently depending on datasets and evaluation metrics.	Does not propose a practical prediction framework.
Diabetes Prediction Using Machine Learning (2020)	Logistic Regression, Decision Tree, Random Forest, KNN and SVM	Diabetes dataset with clinical health parameters	Comparative evaluation identifies suitable algorithms for diabetes prediction.	Focuses mainly on prediction accuracy without healthcare decision support.
Diabetes Prediction Using Machine Learning Algorithms (2019)	Data Preprocessing, Clustering, Multiple ML Algorithms and Pipeline Model	Clinical parameters including Glucose, BMI, Age and Insulin	Data preprocessing and pipeline-based learning improve prediction accuracy.	Limited integration of personalized healthcare recommendations and unified framework.

6. Machine Learning Models and Mathematical Formulation

HealthAI considers multiple supervised machine learning algorithms for diabetes-risk classification. Each algorithm approaches the prediction problem differently. Comparing these models helps determine which approach is more suitable based on the actual experimental results.

Logistic Regression

Logistic regression can be applied to binary classification problems like determining if someone belongs to a diabetes-predisposing group.where,

$P(Y = 1 \mid X) = 1 / (1 + e^{(-z)})$

where,

$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$

Here, represent the input health features and values represent the model coefficients.

Decision Tree

In a Decision Tree, the predictions are made by splitting the data into smaller sets based on some criteria at each step. The nodes denote the conditions of features, and the leaf denotes the prediction class.

A common splitting measure is **Gini Impurity**:

Gini=1−i=1∑Cpi2

where is the proportion of samples belonging to class .

Random Forest

Random Forest algorithm employs several decision trees in order to get the output through the combined prediction of these decision trees.

Final classification can be shown as:

y^=mode{T1(x),T2(x),...,TB(x)}

where represent the individual decision trees.

Support Vector Machine

Support Vector Machine (SVM) tries to obtain a separation line that segregates classes.

The basic decision rule is:

f(x)=wTx+b

where stands for the weight vector and stands for the bias term.

K-Nearest Neighbor

KNN classifies the class of the new observation based on the classes of its k nearest training observations.

One common way to calculate distance is to use Euclidean distance:

d(x,y)=i=1∑n(xi−yi)2

where k is the number of neighbors considered.

Naïve Bayes

Naïve Bayes classifies the observation based on its likelihood of belonging to a class using probability.

P(C|X) = P(X|C)P(C)

P(X)

where represents a class and represents the observed feature values.

Performance Evaluation Metrics

The models are evaluated using standard classification measures.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics provide different views of model performance. In diabetes-risk prediction, **Recall is particularly useful because missing a person who may be at risk can be more concerning than generating an additional false positive.**

7. Algorithm for HealthAI

Algorithm HealthAI_Diabetes_Prediction

Input: Dataset D

Output: Risk Prediction

1. Load D

2. Clean D

3. Handle missing values

4. Encode categorical features

5. Scale numerical features when required

6. Select relevant features

7. Split D into training and testing data

8. Train:

    Logistic Regression

    Decision Tree

    Random Forest

    SVM

    KNN

    Naïve Bayes

9. Predict using test data

10. Calculate:

    Accuracy

    Precision

    Recall

    F1-Score

    ROC-AUC

11. Compare model performance

12. Select suitable model based on actual results

13. Predict diabetes risk for new input

14. Display preliminary risk assessment

15. End

8. Dataset and Data Preprocessing

Dataset Description

HealthAI is a framework using data related to patients in terms of healthcare for early prediction of risk of diabetes. Data has various attributes relating to diabetes. Features which are considered by the framework include:

Feature	**Description**
Age	Age of the individual
Gender	Gender information
BMI	Body Mass Index
Glucose Level	Blood glucose measurement
HbA1c Level	Indicator related to average blood glucose
Hypertension	Indicates the presence of hypertension
Heart Disease	Indicates the presence of heart disease
Smoking History	Information about smoking status
Diabetes Outcome	Target variable for prediction

The **Diabetes Outcome** is used as the target variable, while the remaining attributes are used as input features.

9. Proposed Methodology

The proposed **HealthAI** framework is designed to provide an intelligent approach for early diabetes risk prediction by integrating healthcare data preprocessing, feature engineering, machine learning model development, and healthcare decision support into a unified framework. The methodology aims to improve prediction reliability while supporting healthcare professionals in identifying individuals at high risk of developing diabetes.

The proposed framework consists of six major phases, as illustrated in the methodology workflow.

Phase 1: Data Collection

The first phase involves collecting healthcare data containing clinical and demographic information related to diabetes. The dataset is expected to include important health parameters such as:

- Age
- Gender
- Body Mass Index (BMI)
- Blood Glucose Level
- HbA1c Level
- Hypertension
- Heart Disease
- Smoking History
- Diabetes Outcome

These attributes provide essential information for developing an effective diabetes prediction model. The collected dataset will serve as the foundation for training and evaluating the machine learning models. The selected health parameters are consistent with commonly used diabetes prediction datasets discussed in the reviewed studies.

Phase 2: Data Preprocessing

Incomplete, inconsistent, or duplicate data is common in healthcare data sets; hence, data preprocessing will take place prior to model construction.

The preprocessing stage will include:

- Handling missing values
- Removing duplicate records
- Data normalization
- Feature encoding (where applicable)
- Dataset validation

These preprocessing steps are expected to improve data quality and enhance the overall performance of the prediction models. Similar preprocessing strategies have been emphasized in previous diabetes prediction studies.

Phase 3: Feature Selection

Not every attribute contributes equally to diabetes prediction. Therefore, feature selection techniques will be applied to identify the most significant clinical parameters.

The selected features will help:

- Reduce unnecessary data
- Improve prediction efficiency
- Minimize computational complexity
- Reduce overfitting
- Increase model interpretability

Important predictive features may include blood glucose level, HbA1c level, BMI, age, hypertension, and family history-related indicators, depending on the available dataset.

Phase 4: Machine Learning Model Development

After preprocessing and feature selection, multiple supervised machine learning algorithms will be trained using the prepared dataset.

The proposed framework will evaluate algorithms such as:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- K-Nearest Neighbour (KNN)
- Naïve Bayes (NB)

Instead of relying on a single classifier, the proposed framework will compare the performance of multiple algorithms to determine the most suitable model for early diabetes prediction.

Phase 5: Model Evaluation and Prediction

The trained models will be evaluated using standard classification metrics to measure prediction performance.

The evaluation metrics include:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Score

- Confusion Matrix

The algorithm achieving the best overall performance will be considered the most suitable model for diabetes risk prediction.

Phase 6: Healthcare Decision Support

The final phase focuses on supporting healthcare decision-making based on the prediction results.

The proposed framework will provide:

- Diabetes risk prediction
- Risk classification (Low, Moderate, or High)
- Prediction summary
- Health recommendations
- Decision-support information for healthcare professionals

Rather than replacing clinical diagnosis, the proposed framework is intended to function as a supportive tool that assists healthcare professionals in identifying high-risk individuals and encourages early preventive healthcare.

10. HealthAI Framework

HealthAI is structured into a number of well-defined steps, namely healthcare dataset, data cleaning, missing value handling, categorical variables transformation, feature scaling, feature selection, splitting into train/test sets, candidate classifiers, model evaluation, model selection, risk prediction, healthcare decision support. Every step is kept separately as an artefact/procedure allowing to review data pre-processing and model behavior. Selection depends on actual validation results and is not predetermined

Fig. 1. Overall Workflow of the HealthAI Framework.

11. System Architecture

The architecture segregates the analysis pipeline from the decision-support interface. A verified feature vector is fed into candidate models in the training phase, but only the selected and well-documented model would be considered for deployment post evaluation. The output component needs to provide a risk assessment along with its relevant limitations and the identity of the model in an auditable manner. It should not provide any directives or diagnosis.

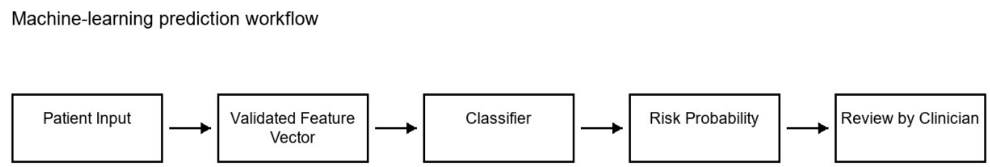

Fig. 2. Machine Learning Prediction Workflow.

12. Performance Evaluation

Performance evaluation is used to measure how well the machine learning models in HealthAI classify diabetes-risk cases. Different evaluation metrics are considered because a single metric may not provide a complete picture of model performance.

Accuracy

Accuracy measures the proportion of correctly classified cases among all tested cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Precision shows how many of the cases predicted as positive are actually positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall measures how many of the actual positive cases are correctly identified by the model.

$$Recall = \frac{TP}{TP + FN}$$

For diabetes-risk screening, recall is particularly important because failing to identify a potentially high-risk case may delay further medical evaluation.

F1-Score

F1-Score combines Precision and Recall into a single measure.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROC-AUC

ROC-AUC is used to evaluate how well a classification model separates the two outcome classes across different decision thresholds. A higher AUC generally indicates better class-separation ability.

Confusion Matrix

A confusion matrix represents the classification results using:

- **TP (True Positive):** Correctly predicted positive cases.
- **TN (True Negative):** Correctly predicted negative cases.
- **FP (False Positive):** Negative cases incorrectly predicted as positive.
- **FN (False Negative):** Positive cases incorrectly predicted as negative.

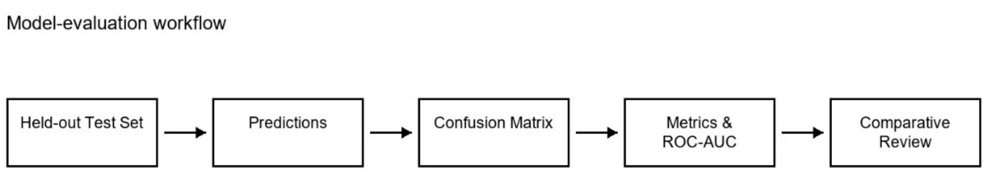

Fig. 3. Model Evaluation Workflow.

13. Result and Discussion

Experimental performance figures are currently unavailable in this experiment's stage of execution. Therefore, accuracy, precision, recall, F1 score, ROC-AUC, confusion matrix entries, number of samples, and classifier ranking are not presented. No actual experimental performance measures are currently available.

The experiments will compare performance of all six classifiers using the same held-out sample and preprocessing strategy. Confusion matrices will present true positives, false positives, true negatives, and false negatives, making it possible to analyze errors instead of hiding them behind a single score. ROC-AUC will serve to measure classification ranking ability at different thresholds. At the same time, calibration and threshold-dependent recall metrics will also help in discussing the model's deployment. A final model choice must be made solely based on the results above.

14. Advantages

- Enables early risk assessment based on various clinical and demographic features.
- Allows for comparison of different conventional supervised approaches, instead of presuming any of them being superior.
- Demarcates between data pre-processing, feature selection, model training, testing, and communication of risk.
- Fosters data-driven approach with proven evaluation process.
- May facilitate preventive care discussions following proper validation.
- Employs modular architecture that could be further developed following governance and clinical approval.

15. Limitations

- Relies on the quality, representativeness, completeness and outcome definition of the dataset.
- Class imbalance problem needs to be addressed explicitly and not ignored.
- Model predictions might reflect biases and artefacts contained within the training set.
- Lacks clinical validation and experimental results of performance in the current study.
- Research prototype cannot substitute professional diagnostics, physical examination and guidelines-based practice.
- Lack of external validity, calibration and safe workflow integration.

16. Future Scope

The future scope of work should include working on large datasets, external validation, calibration, use of explainable AI techniques, and balanced performance in relevant populations. Some other areas of future work could include integrating with electronic health records, mobile/web interface development, real-time monitoring, input from wearable devices, personalized risk communication, translation into multiple languages, and prospective clinical validation. These are areas of future research and not current features.

17. Conclusion

HealthAI is an elaborate research methodology for predicting the risk of diabetes in early stages based on clinical and demographic information in a patient. The methodology provides the data preprocessing techniques, feature selection evaluation techniques, and a comparative study of the performance of Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and Naïve Bayes. HealthAI does not confuse risk assessment with medical diagnosis as it simply helps to organize information for further expert analysis. There is no performance evaluation conducted in actual experiments at present; hence there is no performance claim and no preferred model. The first step after that should be verification of the dataset and reproducible evaluation.

References

[1] World Health Organization, “Diabetes,” Fact sheet, Nov. 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] American Diabetes Association Professional Practice Committee, “2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025,” Diabetes Care, vol. 48, suppl. 1, pp. S27–S49, 2025, doi: 10.2337/dc25-S002.

[3] F. Mohsen, H. R. H. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, “A scoping review of artificial intelligence-based methods for diabetes risk prediction,” npj Digital Medicine, vol. 6, art. 197, 2023, doi: 10.1038/s41746-023-00933-5.

[4] U.S. Food and Drug Administration, Health Canada, and Medicines and Healthcare products Regulatory Agency, “Good Machine Learning Practice for Medical Device Development: Guiding Principles,” Oct. 2021. [Online]. Available: https

[5] L. Lama, O. Wilhelmsson, E. Norlander, and L. Gustafsson, "Machine learning for prediction of diabetes risk in middle-aged Swedish people," Heliyon, vol.

[6] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. Xu, "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators," Healthcare Analytics, vol. 2, p. 100118, 2022, doi: 10.1016/j.health.2022.100118.

[7] Y. Mao, Z. Zhu, S. Pan, W. Lin, J. Liang, H. Huang, L. Li, J. Wen, and G. Chen, "Value of machine learning algorithms for predicting diabetes risk: A subset analysis from a real-world retrospective cohort study," Journal of Diabetes Investigation, vol. 14, no. 2, pp. 309–320

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[9] E. Christodoulou et al., “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” Journal of Clinical Epidemiology, vol. 110, pp. 12–22, 2019, doi: 10.1016/j.jclinepi.2019

[10] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[11] J. D. Hunter, “Matplotlib: A 2D graphics environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007, doi: 10.1109/MCSE.2007.55.

[12] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30, 20