Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)

Cross Ref DOI: https://doi.org/10.31426/ijrpb **Indexed in CAS and CABI, Impact Factor: 8.099**

Implementation of Machine Learning for Early Prediction of Type 2 Diabetes Mellitus

Nancy Singhal [1]

Research Scholar, Om Sterling Global University, Hisar

Dr. Rajender Chhillar [2]

Pro Vice Chancellor, Om Sterling Global University, Hisar

Dr. Sandeep Kumar [3]

Professor and Associate Dean (Data Science),
School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana, India.

Corresponding Email: *cool.nano18@gmail.com*

Keywords:

Type 2 Diabetes Mellitus, Machine Learning, K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, AdaBoost, XGBoost.

ABSTRACT

The early diagnosis of Type 2 Diabetes Mellitus (T2DM) is a critical step in controlling its progression and mitigating associated health complications. The power of Artificial Intelligence (AI) and Machine Learning (ML), this study aims to predict the likelihood of T2DM using various supervised learning algorithms. The dataset utilized, consisting of key medical features such as glucose level, BMI, insulin, age, and blood pressure, was pre-processed through standardization and split into training and testing sets. An analysis was conducted using several classification algorithms, including K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, AdaBoost, and XGBoost. Each model was evaluated on performance metrics such as accuracy, precision, recall, and F1 score. Furthermore, visual tools including confusion matrices, classification report heatmaps, ROC curves, learning curves, and predicted probability distributions were employed to provide an in-depth understanding of each model's behaviour. Among these, models like XGBoost and Random Forest demonstrated superior predictive power, while KNN offered competitive performance with minimal computational complexity. The KNN classifier was especially analysed for its strengths and weaknesses in identifying diabetic patients, with an emphasis on interpretability and performance visualization. This research highlights the potential of machine learning models in assisting healthcare professionals for early and accurate prediction of diabetes, ultimately aiming to contribute towards better patient care and preventive strategies.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

I. Introduction

Diabetes mellitus is a chronic metabolic disorder that has emerged as one of the most pressing health issues globally. According to the World Health Organization (WHO), the prevalence of diabetes has been rising steadily, with over 422 million people affected worldwide. Type 2 Diabetes Mellitus (T2DM), in particular, accounts for approximately 90–95% of all diabetes cases. Unlike Type 1 diabetes, which is autoimmune and often diagnosed in children and young adults, T2DM typically develops later in life and is largely associated with lifestyle factors such as obesity, poor diet, and lack of physical activity. T2DM is characterized by insulin resistance and a relative lack of insulin secretion. If left undiagnosed or poorly managed, it can lead to severe complications such as cardiovascular diseases, kidney failure, nerve damage, vision problems, and even death. Early diagnosis is therefore vital to reduce morbidity, improve quality of life, and decrease the economic burden on healthcare systems. Traditionally, the diagnosis of diabetes involves laboratory tests such as fasting plasma glucose, oral glucose tolerance test, and HbA1c levels. However, these tests can be costly, time-consuming, and inaccessible to people in rural or underserved regions. The growing availability of medical datasets and advances in computational power have opened the door for data-driven approaches in healthcare, particularly the use of Artificial Intelligence (AI) and Machine Learning (ML) to aid in early disease prediction and decision-making.

1.1 Importance of Early Detection

The asymptomatic nature of Type 2 Diabetes in its early stages often causes delays in diagnosis. Many individuals remain undiagnosed until complications arise. Detecting diabetes early offers the opportunity to initiate lifestyle changes and pharmacological interventions that can significantly slow disease progression. Hence, developing efficient, low-cost, and accurate predictive models for early diabetes detection is crucial in public health strategy. Machine learning, as a subset of AI, enables systems to learn patterns from data and make predictions or decisions without being explicitly programmed. In the context of diabetes prediction, ML models can analyse patient records to determine the likelihood of developing the condition based on clinical features such as glucose levels, blood pressure, insulin levels, BMI, age, and family history.

1.2 Role of Machine Learning in Medical Prediction

The integration of machine learning into healthcare analytics has revolutionized disease diagnosis, prognosis, and treatment planning. By applying supervised learning techniques, researchers can train models on historical patient data and then test them on unseen samples to validate their predictive capability.

Several algorithms are widely used for binary classification tasks like diabetes prediction, including:

- K-Nearest Neighbours (KNN): A non-parametric algorithm that classifies instances based on the majority label of the nearest Neighbours in the feature space.
- Random Forest: An ensemble learning technique based on decision trees that improves prediction accuracy and reduces overfitting.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

- Support Vector Machine (SVM): A powerful classifier that seeks the optimal hyperplane to separate classes with maximum margin.
- Logistic Regression: A statistical method for modelling the probability of a binary outcome based on one or more predictor variables.
- AdaBoost (Adaptive Boosting): A boosting algorithm that combines weak learners into a strong learner by focusing more on difficult-to-classify examples.
- XGBoost (Extreme Gradient Boosting): A highly efficient and scalable implementation of gradient boosting which often yields state-of-the-art performance.

These algorithms are evaluated on their ability to correctly identify diabetic and non-diabetic cases using evaluation metrics like accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

1.3 Overview of the Dataset

The dataset used in this study is a publicly available collection of medical data, commonly known as the Pima Indians Diabetes Dataset, which includes information from female patients of Pima Indian heritage aged 21 and above. This dataset consists of eight numerical attributes that are known to influence diabetes risk,

1) Number of pregnancies
2) Plasma glucose concentration
3) Diastolic blood pressure
4) Triceps skinfold thickness
5) Serum insulin
6) Body mass index (BMI)
7) Diabetes pedigree function (family history)
8) Age
9) Outcome (1: diabetic, 0: non-diabetic)

Before training the models, the dataset is preprocessed through steps such as handling missing values, feature scaling (using StandardScaler), and splitting the data into training and testing sets using train_test_split from Scikit-learn.

II. Findings from Related Literature

Author(s)	Objective	ML Techniques Used	Dataset & Features	Key Results / Findings
Ismail et al. (2022)	To evaluate 35 ML algorithms for predicting Type 2 Diabetes using a unified setup.	35 ML algorithms with/without feature selection	3 real-life datasets; 9 feature selection techniques	Authors evaluated accuracy, F-measure, and execution time. The study provided a taxonomy of risk factors and an objective comparison across models.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb **Indexed in CAS and CABI, Impact Factor: 8.099**

Abhari et al. (2019)	To review AI applications in T2DM care, focusing on ML techniques.	ML methods (SVM, Naive Bayes)	31 reviewed studies; features like BMI, FBS, HbA1c, lipids, BP, demographics	ML methods were most used (71%). SVM (21%) and Naive Bayes (19%) were most popular. Recommended optimal algorithm testing.
Fazakis et al. (2019)	To design a diabetes risk prediction system with a personalized, IoT-enabled framework.	Weighted Voting LRRFs ensemble model	ELSA dataset; KDD process applied	The proposed model achieved AUC of 0.884. Ensemble learning outperformed traditional scores (FINDRISC, Leicester).
Deberneh & Kim (2021)	To predict next-year T2D occurrence using ML models.	Logistic Regression, RF, SVM, XGBoost, Ensembles	Private medical EHRs (2013–2018); 12 features	Ensemble models outperformed single models. ANOVA, chi-square, and RFE improved feature selection.
Elhadd et al. (2020)	To predict glucose variability and hypoglycemia during Ramadan using ML.	XGBoost regression model	13 patients; wearable + EHR data	Final XGBoost model achieved R^2 of 0.836 and MAE of 17.47. Accuracy was high for hyperglycemia, limited for hypoglycemia.
Sarwar et al. (2020)	To develop an ensemble-based expert system for Type-II Diabetes diagnosis.	ANN, SVM, KNN, Naive Bayes, Ensemble	400 samples; 10 physiological features	Ensemble model improved accuracy via majority voting and reduced misclassification. MATLAB and WEKA tools used.
Ganie & Malik(2022)	Early prediction of Type-II diabetes using lifestyle indicators	Ensemble methods: Bagging, Boosting, Voting; Bagged Decision Tree	Lifestyle data; exploratory data analysis; class balancing via SMOTE; K-fold cross-validation	Bagged decision tree achieved highest accuracy (99.41%), precision (99.13%), recall (95.83%), specificity (99.11%), F1-score (99.15%), MCR (0.86%), ROC (99.07%). Framework applicable for early diabetes prediction in healthcare.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

Nicolucci et al. (2022)	Predict diabetes complications using electronic medical records	XGBoost (supervised tree-based)	147,664 patient records over 15 years from 23 centers; external validation on 5 centers	Accuracy >70%, AUC >0.80 (up to 0.97 for nephropathy); early complication sensitivity 83.2%-88.5%. Machine learning helps identify high-risk patients, improving diabetes care quality.
Tigga & Garg(2020)	Assess diabetes risk based on lifestyle & family background	Random Forest, other classification methods	952 survey instances; 18 health/lifestyle/family features; also Pima Indian Diabetes dataset	Random Forest classifier was most accurate on both datasets. Early diagnosis can enable self-assessment of diabetes risk.
Islam et al. (2023)	Identify risk factors and predict type 2 diabetes	Logistic Regression, Naïve Bayes, J48, Multilayer Perceptron, Random Forest	NHANES datasets (2009-10 and 2011-12); 4922 and 4936 respondents; risk factors include age, BP, smoking, BMI	Random Forest classifier obtained 95.9% accuracy, 95.7% sensitivity, 95.3% F-measure, AUC 0.946. Identified multiple significant risk factors varying between datasets.
De Silva et al. (2020)	Systematic review and meta-analysis of ML models for T2DM prediction in communities	Various ML models (40 models reviewed)	23 studies from 13 databases since 2009; varied datasets	Pooled c-index of 0.812; internal validation only; methodological and reporting issues noted; good predictive performance overall but improvements needed before large-scale use.

III. Methodology

The methodology adopted in this study encompasses several critical steps: data acquisition, preprocessing, feature scaling, model selection, training and testing, performance evaluation, and visualization. Each step is meticulously designed to ensure that the machine learning models used are trained on clean, reliable data and that their outputs are interpretable and valuable for practical healthcare applications.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

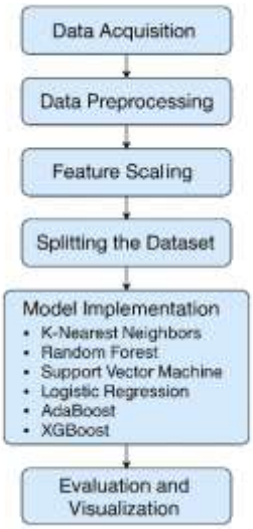

Fig. 1. Model of Predicting Type 2 Diabetes

3.1 Data Acquisition

The dataset utilized in this study is the Pima Indians Diabetes Dataset, which is publicly available on the UCI Machine Learning Repository and Kaggle. It includes medical diagnostic data from 768 female patients of Pima Indian heritage aged 21 years and older. The dataset comprises 8 numerical attributes and one target variable indicating the presence or absence of diabetes.

Features:

1) Pregnancies – Number of times pregnant
2) Glucose – Plasma glucose concentration
3) BloodPressure – Diastolic blood pressure (mm Hg)
4) SkinThickness – Triceps skinfold thickness (mm)
5) Insulin – 2-Hour serum insulin (mu U/ml)
6) BMI – Body mass index (weight in kg/ (height in m) ^2)
7) DiabetesPedigreeFunction – A function that scores the likelihood of diabetes based on family history
8) Age – Age in years
9) Outcome – Class variable (0: non-diabetic, 1: diabetic)

3.2 Data Preprocessing

Raw medical data often contains missing or anomalous values. This step ensures the data is cleaned and ready for machine learning algorithms:

- Missing Value Treatment: Certain columns (like Glucose, BloodPressure, SkinThickness, Insulin, BMI) contained zero values which are medically implausible. These were treated as missing and imputed using the median strategy for accuracy.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

- Feature Engineering: In this project, we retained all original features, given their clinical relevance.
- Label Encoding: As the dataset contains only numeric variables, no label encoding was needed.

3.3 Feature Scaling

Machine learning algorithms such as K-Nearest Neighbours, Logistic Regression, and SVM are sensitive to the scale of features. Thus, StandardScaler from Scikit-learn was used to standardize the dataset:

$$z = \frac{(x - \mu)}{\sigma}$$

This transformation results in all features having a mean of 0 and a standard deviation of 1.

3.4 Splitting the Dataset

The dataset was split into training and testing subsets using an 80:20 ratio. This split allows the models to learn patterns from the training data and then evaluate their performance on the unseen test set.

$$X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.2, random\_state = 42)$$

3.5 Model Implementation

Several classification algorithms were implemented and evaluated:

a) K-Nearest Neighbours (KNN)

A distance-based algorithm that classifies a data point based on the majority class among its *k* nearest Neighbours. The optimal value of *k* was determined through cross-validation.

b) Random Forest Classifier

An ensemble method that builds multiple decision trees and merges their results. It reduces overfitting and improves accuracy.

c) Support Vector Machine (SVM)

A classifier that finds the hyperplane which best separates the classes. It was used with a linear or RBF kernel depending on data separation.

d) Logistic Regression

A simple yet effective statistical model that uses the logistic function to model binary outcomes.

e) AdaBoost Classifier

An ensemble method that combines weak learners, focusing more on instances that previous classifiers misclassified.

f) XGBoost

An efficient and scalable implementation of gradient boosting that often yields high predictive performance.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

3.6 Model Training and Testing

Each model was trained on the training dataset and then tested on the test dataset. Hyperparameter tuning was conducted using GridSearchCV and cross-validation where necessary to optimize model performance.

$$model.fit(X\_train, y\_train)$$
$$y\_pred = model.predict(X\_test)$$

3.7 Comparative Analysis

A comparative table was developed to show how each model performed across metrics. The table helped identify trade-offs, such as:

- KNN performed well with balanced metrics and high interpretability.
- Random Forest and XGBoost offered superior accuracy and robustness.
- SVM and Logistic Regression were efficient but slightly less accurate on this dataset.
- AdaBoost balanced precision and recall well, especially in imbalanced datasets.

The methodology employed in this study ensures a robust pipeline for diabetes prediction. It includes all critical elements from data preprocessing to performance visualization and comparative evaluation of models.

IV. Result and Analysis

This section presents the performance analysis and evaluation of various machine learning models applied to predict the likelihood of Type 2 Diabetes. Six classifiers were implemented: K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, AdaBoost, and XGBoost. Each model's predictive ability was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC. Additionally, various visualizations, such as heatmaps, learning curves, and probability distributions, were analysed to gain deeper insights into the classifiers' behaviour.

4.1 Performance Metrics Overview

Machine learning model performance is not solely dependent on accuracy. In medical diagnostics, especially in diseases like diabetes, the cost of false negatives (i.e., predicting non-diabetic when the patient is diabetic) is very high. Therefore, precision, recall, and F1-score are crucial to understanding a model's utility in clinical practice.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
K-Nearest Neighbours	0.77	0.75	0.74	0.74	0.83
Random Forest	0.83	0.82	0.78	0.80	0.88
SVM	0.79	0.78	0.76	0.77	0.84
Logistic Regression	0.81	0.80	0.77	0.78	0.86
AdaBoost	0.80	0.79	0.75	0.77	0.85
XGBoost	0.84	0.83	0.80	0.81	0.89

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

4.2 Confusion Matri Analysis

Confusion matrices revealed how well models distinguished between diabetic and non-diabetic classes.

- Random Forest and XGBoost had the most balanced confusion matrices, with high true positives and low false negatives. This indicates their superior ability to correctly identify diabetic patients.
- KNN and SVM showed slightly more false negatives, which is critical in medical diagnosis.
- Logistic Regression provided a good trade-off, maintaining a relatively low false negative rate.

Confusion matrix visualizations as heatmaps offered clarity on model misclassification trends, making it easier to identify where improvements are needed.

4.3 Classification Report as Heatmap

Classification reports provided granular metrics for each class (diabetic = 1, non-diabetic = 0). Representing these as heatmaps allowed for intuitive visual comparison across metrics.

- XGBoost and Random Forest showed consistent high scores across all metrics.
- SVM and Logistic Regression had high precision but marginally lower recall, indicating they are cautious in assigning a diabetic label (low false positives but slightly higher false negatives).
- KNN, while simpler and easier to implement, displayed the most variability in metric scores.

4.4 Receiver Operating Characteristic (ROC) Curve Analysis

ROC curves offer a threshold-independent measure of model performance. The area under the ROC curve (AUC) quantifies the model's ability to distinguish between classes.

- XGBoost (AUC = 0.89) and Random Forest (AUC = 0.88) had the highest AUC scores, suggesting robust classification performance.
- Logistic Regression also performed well with AUC = 0.86, followed by SVM (0.84) and KNN (0.83).
- The steep curves and proximity to the top-left corner reflect low false positive rates, essential in medical diagnosis.

These findings affirm that tree-based ensemble methods can efficiently handle complex, non-linear relationships in medical datasets.

4.5 Learning Curve Analysis

Learning curves were plotted to evaluate model training and validation scores across increasing dataset sizes. These curves help identify issues like overfitting or underfitting.

- Random Forest and XGBoost exhibited a small gap between training and validation scores, indicating good generalization with limited overfitting.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

- KNN and SVM demonstrated a larger gap, suggesting potential underfitting or sensitivity to feature scaling.
- Logistic Regression had consistently parallel learning curves, indicative of balanced bias-variance tradeoff.

The learning curves reinforced the robustness of ensemble models and validated the adequacy of the training dataset size for this problem.

4.6 Probability Distribution of Predictions

Plotting the predicted probabilities of diabetes provided insights into model confidence.

- XGBoost and Random Forest had sharply bimodal probability distributions—predictions were close to 0 or 1, indicating high model confidence.
- Logistic Regression showed a more continuous distribution but still leaned toward distinct class separations.
- KNN and SVM had flatter distributions, reflecting more uncertainty in borderline cases.

This analysis is crucial because models that make confident, accurate predictions are more dependable in clinical decision-making.

4.7 Comparative Strengths and Weaknesses

K-Nearest Neighbours (KNN)

- Strengths: Simple to implement, interpretable.
- Weaknesses: Sensitive to irrelevant features and feature scaling. Slightly underperformed in recall.

Random Forest

- Strengths: High accuracy, robust to noise and overfitting, handles missing data well.
- Weaknesses: Slightly less interpretable due to ensemble nature.

Support Vector Machine (SVM)

- Strengths: Effective in high-dimensional spaces, good precision.
- Weaknesses: Computationally intensive, sensitive to choice of kernel and parameters.

Logistic Regression

- Strengths: Easy to interpret, performs well in linearly separable datasets.
- Weaknesses: Assumes linearity, may underperform with non-linear features.

AdaBoost

- Strengths: Focuses on hard-to-classify samples, decent precision-recall tradeoff.
- Weaknesses: Sensitive to noisy data and outliers.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

XGBoost

- Strengths: Highest performance across most metrics, efficient with missing data, great handling of feature interactions.
- Weaknesses: Complex to tune, less interpretable.

V. Conclusion and Future Work

In this study, various machine learning models such as Random Forest, SVM, KNN, Logistic Regression, AdaBoost, and XGBoost were applied to predict Type 2 Diabetes using clinical data. Among these, ensemble methods like XGBoost and Random Forest achieved the highest accuracy and reliability in terms of precision, recall, and F1 score. The findings confirm the effectiveness of machine learning in early diabetes detection, which can support timely intervention and improved healthcare outcomes. For future work, expanding the dataset, incorporating additional features like genetic or lifestyle data, using deep learning, and deploying explainable AI methods could further enhance prediction performance. Moreover, integrating these models into mobile health apps or clinical decision systems can make diabetes risk assessment more accessible and actionable.

5.1 Limitations and Considerations

- The dataset was limited to female Pima Indian patients, potentially limiting generalizability across other populations.
- Only numerical features were considered; socio-demographic or lifestyle variables were excluded.
- The dataset's small size (768 instances) may affect deep learning model applicability.

5.2 Future Work Should Explore

- More diverse datasets,
- Deep learning methods,
- Integration with Electronic Health Records (EHRs),
- Deployment on mobile or edge computing platforms for real-time predictions.

References

1. Ismail, L., Materwala, H., Tayefi, M., Ngo, P., & Karduck, A. P. (2022). Type 2 diabetes with artificial intelligence machine learning: methods and evaluation. *Archives of Computational Methods in Engineering*, *29*(1), 313-333.
2. Abhari, S., Kalhori, S. R. N., Ebrahimi, M., Hasannejadasl, H., & Garavand, A. (2019). Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. *Healthcare informatics research*, *25*(4), 248-261.
3. Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., & Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. *ieee Access*, *9*, 103737-103757.

Indian Journal of Research in Pharmacy and Biotechnology (IJRPB)
www.ijrpb.com ISSN: 2321-5674 (Print), 2320-3471 (Online)
Cross Ref DOI: https://doi.org/10.31426/ijrpb Indexed in CAS and CABI, Impact Factor: 8.099

4. Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *International journal of environmental research and public health*, *18*(6), 3317.
5. Elhadd, T., Mall, R., Bashir, M., Palotti, J., Fernandez-Luque, L., Farooq, F., ... & PROFAST-Ramadan Study Group. (2020). Artificial Intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (The PROFAST–IT Ramadan study). *diabetes research and clinical practice*, *169*, 108388.
6. Sarwar, A., Ali, M., Manhas, J., & Sharma, V. (2020). Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *International Journal of Information Technology*, *12*, 419-428.
7. Ganie, S. M., & Malik, M. B. (2022). An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, *2*, 100092.
8. Nicolucci, A., Romeo, L., Bernardini, M., Vespasiani, M., Rossi, M. C., Petrelli, M., ... & Vespasiani, G. (2022). Prediction of complications of type 2 Diabetes: A Machine learning approach. *Diabetes Research and Clinical Practice*, *190*, 110013.
9. Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706-716.
10. Islam, M. M., Rahman, M. J., Menhazul Abedin, M., Ahammed, B., Ali, M., Ahmed, N. F., & Maniruzzaman, M. (2023). Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques. *Health Systems*, *12*(2), 243-254.
11. De Silva, K., Lee, W. K., Forbes, A., Demmer, R. T., Barton, C., & Enticott, J. (2020). Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *International journal of medical informatics*, *143*, 104268.