Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Early Diagnosis of Diabetes Mellitus Using Machine Learning Algorithms and Clinical Data Analysis

[1]Soumen Chatterjee, [2]Soumen Bhowmik
[1]Email Id - soumenraja90@gmail.com
Designation - Research Scholar M. Tech. (C.S.E.)
Department - Computer Science and Engineering
College Name - Bengal Institute of Technology and Management, Santiniketan, West Bengal.
University - Maulana Abul Kalam Azad University of Technology, West Bengal.
[2]Email Id -bhowmik.soumen.cse@gmail.com
Designation - Assistant Professor & Head of the Department
Department - Computer Science and Engineering
College Name - Bengal Institute of Technology and Management, Santiniketan, West Bengal.
University - Maulana Abul Kalam Azad University of Technology, West Bengal.

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

The catastrophic effects of diabetes affect a significant majority of individuals worldwide, and many of these cases are not detected in time. The inability of the body to create enough insulin results in diabetes, which raises blood glucose levels and has become one of the main causes of death. The impact of diabetes has increased recently and is predicted to keep expanding on a global scale. Globally, there are currently 463 million people with diabetes, and by 2045, that number is expected to increase to 700 million. Type 2 diabetes is becoming more and more common in many nations. Diabetes is linked to serious health issues such heart disease, stroke, kidney damage, and blindness.

To address these issues, this study makes use of machine learning algorithms, which process and learn from vast amounts of data, to improve diabetes prediction and early detection. In particular, this study compares the accuracy of diabetes diagnosis using the Random Forest Classifier, Decision Tree Classification, and Logistic Regression algorithms to ascertain how well they predict diabetes outcomes.

Keywords: Machine Learning, Logistic Regression, Decision Tree Classification, Random Forest Classifier, Early diagnosis, Blood glucose levels, Health issues, Medical data analysis, Global diabetes statistics.

1. INTRODUCTION

Diabetes is a long-term metabolic condition characterized by elevated blood sugar levels brought on by insufficient insulin synthesis **(Rawat, 2022)**. It is one of the main causes of death worldwide and is associated with severe side effects include blindness, kidney failure, and heart disease. About 463 million individuals between the ages of 20 and 79 have diabetes at the moment, and by 2045, that number is expected to rise to 700 million. Globally, the number of type 2 diabetes diagnoses and deaths from the disease is continuously rising.

Without human assistance, machine learning (ML) allows systems to evaluate enormous, intricate datasets and produce precise predictions. It is useful for identifying conditions like diabetes because of its capacity to process large amounts of data from various sources.

This study focuses on three main types of diabetes:

- **Type 1 Diabetes**: Common in children and young adults; the immune system destroys insulin-producing cells, requiring daily insulin injections.

- **Type 2 Diabetes**: The most common form (90% of cases); linked to insulin resistance, lifestyle, and genetics. Managed through diet and exercise.
- **Gestational Diabetes**: Occurs during pregnancy due to insulin-blocking hormones; usually resolves after birth but increases future diabetes risk.

Using ML techniques, this research conducts a comparative analysis to enhance early detection and diagnosis of diabetes.

1.1. Machine Learning Techniques for Early Diagnosis of Diabetes Mellitus

Machine Learning (ML) is an important subset of Artificial Intelligence (AI**) (Rawat V. e., Machine learning algorithms for early diagnosis of diabetes mellitus, 2022)**.With the use of machine learning (ML), systems may learn from data, spot trends, and come to conclusions with little assistance from humans. It includes a number of algorithms that are very useful for detecting and predicting a wide range of illnesses, including diabetes mellitus. Large and intricate medical datasets can be processed by these algorithms, which can then be used to support precise, data-driven predictions and reveal hidden relationships. In the healthcare industry, machine learning (ML) helps with risk assessment, therapy prescription, and patient outcome prediction in addition to early diagnosis. Numerous machine learning methods have been created and used to forecast diabetes; each has unique advantages in terms of precision, comprehensibility, and computational effectiveness. Some of the most widely used algorithms for diabetes detection are covered in the section that follows.

1.2. Challenges and future scope in Early Diagnosis of Diabetes Mellitus

A number of obstacles still stand in the way of creating efficient models for the diagnosis of diabetes, even with the notable developments in deep learning and machine learning. The availability and quality of data are two key issues. The PIMA Indian dataset is used in many studies**(Sharma, 2021)**, although it lacks diversified and real-time data. Model generalizability is often limited and overfitting occurs in small or region-specific datasets. A significant amount of research effort is devoted to data preparation and cleaning, which raises the complexity and expense of the process. Another challenge is feature selection because not all datasets offer pertinent characteristics for the best model training. Further challenges are introduced by the selection of tools and algorithms, as well as by debugging and deployment, particularly on automated or mobile platforms.

To attain great accuracy, many parameters, including hyperparameters, kernel types, and tree counts, must be fine-tuned during model creation. Real-world application is limited by certain models' reliance on hardware or single-parameter training. Furthermore, replication and scalability are challenging due to restricted access to time-series or real-time datasets.

Machine learning models need to be deployable and applicable in healthcare environments in order to have a significant impact. DevOps and cloud computing skills are essential for deployment, particularly on mobile devices **(Barik S, 2021)**. Current models frequently ignore more complex elements that affect predictive capacity in favor of concentrating just on accuracy measurements. Despite their effectiveness, clustering algorithms are error-prone in automated systems and necessitate human involvement.

These restrictions can be overcome, though. To improve performance, future research should focus on creating more varied and rich datasets and integrating deep learning with other algorithms. Diabetes management and the avoidance of consequences like heart disease can be greatly aided by hybrid models and early detection systems **(Khaleel, 2023)**. In the end, deployable, automated, and effective machine learning models have a lot of potential to enhance diabetes diagnosis and treatment.

2. LITERATURE REVIEW

1. Chou, Hsu, and Chou (2023) - Predicting the Onset of Diabetes with Machine Learning Methods

Chou, Hsu, and Chou (2023), in their paper "Predicting the Onset of Diabetes with Machine Learning Methods", utilized machine learning approaches to enhance early diabetes prediction. Their emphasis encompassed many forms of diabetes, including gestational diabetes, which arises from hormonal fluctuations during pregnancy. The research discovered significant risk variables including obesity, familial history, and advanced maternal age. The authors employed machine learning models such as ANN and SVM, assessed by confusion matrices and ROC curves, to improve diagnostic precision. Their findings underscore the efficacy of machine learning in facilitating early identification and clinical decision-making [6].

2. Chang V et al. (2022) – An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators

Chang V et al. (2022) performed a comprehensive assessment of machine learning algorithms for the early prediction of diabetes utilizing health indicators. The researchers examined multiple models, such as Decision Trees, Random Forests, Support Vector Machines, and Neural Networks, emphasizing the significance of feature selection and preprocessing techniques. Ensemble models such as Random Forest and boosting approaches surpassed individual classifiers in accuracy and resilience. The research determined that the integration of various algorithms with meticulously chosen indications markedly improves early diabetes detection. [7].

3. Chaki et al. (2022) – Machine learning and Artificial Intelligence based Diabetes Mellitus detection and self-management

Chaki et al. (2022) Evaluated machine learning and artificial intelligence techniques for diabetes detection and self-management, highlighting the utilization of both publicly available and self-collected datasets. Huang et al. (2018) employed 42 retinal pictures from MESSIDOR, DiaRetDB0, and DiaRetDB1 to identify neovascularization. Swapna et al. (2018) examined ECG data from 20 diabetic and 20 healthy subjects to derive heart rate characteristics. The research highlighted that varied datasets enhance the precision of diabetes detection. [8].

4. Kodama (2021) – Ability of Current Machine Learning Algorithms to Predict and Detect Hypoglycaemia in Patients with Diabetes Mellitus: Meta-analysis

Kodama (2021) The efficacy of contemporary machine learning algorithms in predicting and detecting hypoglycemia in diabetic patients was assessed by a meta-analysis of ML models. The research examined 96 databases utilizing machine learning and glucose-related terminology, implementing stringent inclusion criteria including author-led model training and the availability of sensitivity and specificity data. The machine learning model functioned as the index test, while clinical diagnostic methods served as the reference standard. Results demonstrated the potential of machine learning in precisely predicting or diagnosing hypoglycemia through patient data. [9].

5. Ahmed et al. (2022) – Prediction of Diabetes Empowered with Fused Machine Learning

Ahmed et al. (2022) Addressed the increasing application of machine learning for diabetes prediction, highlighting concerns such as data imbalance. Ensemble models surpassed others (Pradhan, 2020), with Kumari (2021) attaining 79.08% by soft voting, and Sarwar (2018) documenting 77% with KNN and SVM. Dey (2019) attained an accuracy of 82.35% utilizing artificial neural networks, whilst Saru (2020) accomplished 94.4% with bootstrapped decision trees. Robust outcomes were also observed with decision trees (Sonar, 2019) and deep neural networks (Wei, 2020). Jain (2021) documented an 87.88% efficacy utilizing neural networks. Ahmed (2022) introduced a hybrid model integrating artificial neural networks, support vector machines, and fuzzy logic to enhance accuracy [10].

3.METHODOLOGY

3.1 Data Collection

This study utilized a publicly accessible diabetes dataset comprising clinical and demographic information. Essential characteristics are age, gender, body mass index (BMI), hypertension, cardiovascular disease, smoking history, HbA1c level, and blood glucose level. These characteristics are frequently employed in the preliminary diagnosis of diabetes, with the target variable denoting diabetic or non-diabetic status.

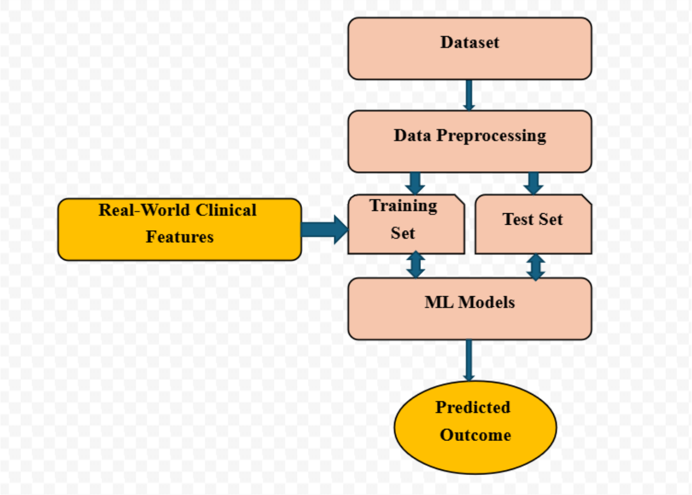

Figure 3.1. Flow chart a Early Diagnosis of Diabetes

3.2 Data Preprocessing

A series of measures were implemented to cleanse and prepare the dataset:

- **Missing and Ambiguous Data**: The "Smoking History" section included values such as "No Info," which were classified as a distinct category or encoded accordingly. All other absent or unlikely values were substituted using statistical methods such as mean or mode.
- **Categorical Encoding:** Gender and Smoking History were transformed into numeric format by one-hot encoding.
- **Target Variable**: The 'diabetes' column, serving as a binary classification label, was utilized as the output variable for model training.
- **Train-Test Split**: The dataset was divided into 80% for training and 20% for testing by a categorized split to preserve class balance.

3.3 Exploratory Data Analysis (EDA)

EDA was performed to understand data patterns and relationships:

- **Class Balance**: The dataset was imbalanced, with a higher number of non-diabetic cases compared to diabetic ones.
- **Feature Correlations**: Pearson correlation analysis was applied to identify relationships between features and the target variable. Heatmaps and distribution plots were used to visualize feature importance.
- **Outlier Detection**: Boxplots were used to detect outliers in BMI, blood glucose, and HbA1c levels, and decisions were made whether to retain or filter extreme values.

3.4 Machine Learning Models and Implementation

Three supervised machine learning models were used in this study:

- **Logistic Regression:** served as a baseline model, predicting diabetes based on a linear combination of clinical features.
- **Decision Tree Classifier:** created a tree-like structure to split data using key health indicators, offering easy interpretation.
- **Random Forest Classifier:** used multiple decision trees and aggregated their results for improved accuracy and reduced overfitting.

All models were implemented in Python using libraries like Scikit-learn, Pandas, and NumPy, and executed in Google Colab for efficient development and testing.

3.5 Model Evaluation and Performance Metrics

The efficacy of the machine learning models was evaluated using conventional classification criteria to guarantee precise and dependable predictions of diabetes status. The subsequent measures were employed:

- **Accuracy:** Assesses the model's overall correctness by determining the ratio of accurately predicted observations to the total number of observations.
- **Precision:** Denotes the ratio of true positive predictions to the total positive predictions generated by the model.
- **Recall (Sensitivity):** Assesses the model's capacity to detect genuine diabetes cases.
- **F1-Score:** The harmonic mean of precision and recall, offering a balance between the two metrics.
- **Confusion Matrix:** A graphical depiction of true positives, false positives, true negatives, and false negatives to comprehensively assess model performance.
- **ROC-AUC Score:** Assesses the model's capacity to differentiate between classes across many thresholds, particularly beneficial in imbalanced datasets.

These metrics were calculated on the test set after training each model. Among the three models, the Random Forest Classifier showed the highest performance, followed by the Decision Tree and Logistic Regression.

4. RESULT AND ANALYSIS

4.1 Data Overview and Preprocessing

This study utilizes a dataset consisting of 100,000 records and 9 characteristics, encapsulating clinical and demographic information of patients. The aim variable is diabetes, signifying the presence (1) or absence (0) of diabetes mellitus. This dataset offers a diverse combination of categorical and numerical variables, appropriate for supervised machine learning classification problems.

Data Cleaning: Prior to modelling, the dataset underwent a sequence of cleansing procedures:

- **Absence of Data:** The dataset was examined for absent values. No null values were identified in the primary numerical fields.
- **Duplicate Records:** Duplicate entries were verified and eliminated to prevent bias and redundancy in the model.
- **Unreliable Categories:** The smoking_history field contains inconsistent labels such as No Info and not current, which were addressed during preprocessing.

4.2 Class Distribution

A count plot was created to illustrate the distribution of the target variable, diabetes (Figure 1). The graphic indicates that the dataset is skewed, featuring a markedly higher number of non-diabetic cases (0) compared to diabetes instances (1).

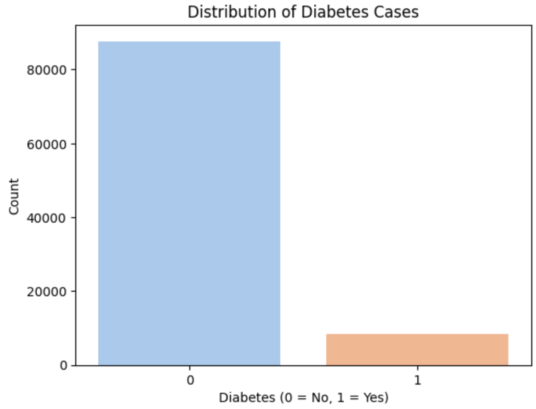

Figure 4.1. Diabetes Case Distribution

4.3 Gender-wise Diabetes Distribution

The Gender-specific Diabetes Distribution graph depicts the prevalence of diabetic and non-diabetic cases among male and female patients. It aids in comprehending potential gender-specific patterns in diabetes prevalence. A marginally elevated prevalence in females is noted, indicating a potential requirement for gender-specific preventive measures or additional clinical research.

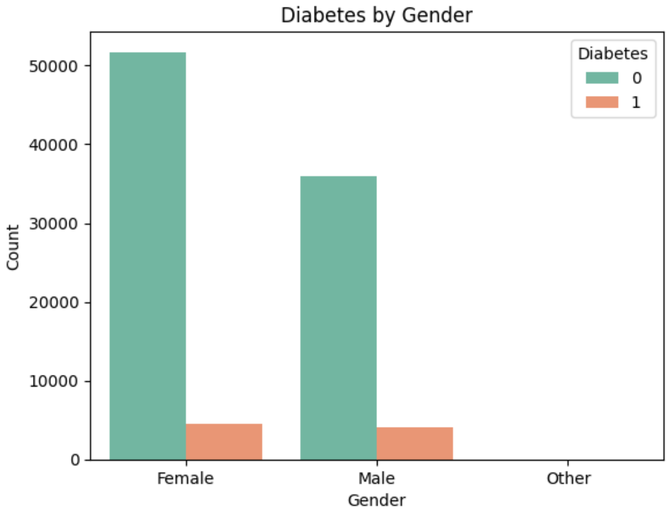

Figure 4.2. Gender-wise Diabetes Distribution

4.4 Age Distribution

The Age Distribution graph illustrates the prevalence of diabetes-related cases among various age groups. The histogram indicates that middle-aged and elderly adults constitute the predominant segment of the dataset. This affirms that age is a significant risk factor in diabetes diagnosis and underscores its relevance as a predictive variable in the model.

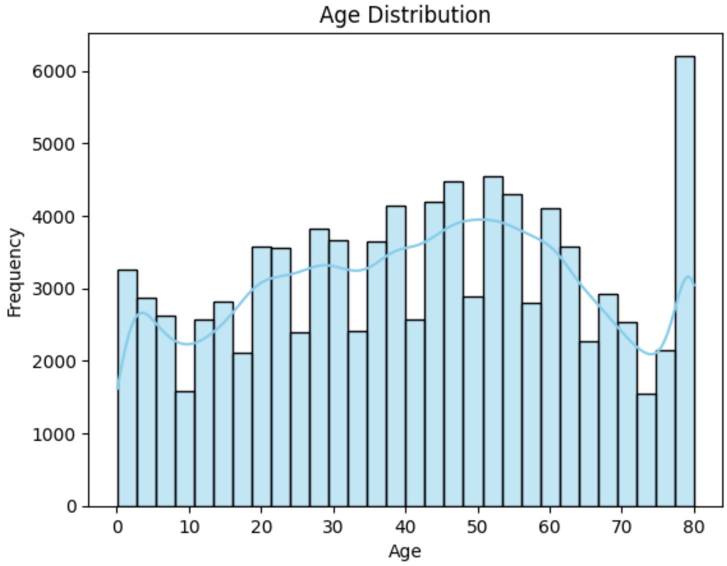

Figure 4.3. Age Distribution

4.5 Correlation

The Correlation Heatmap illustrates the linear associations across numerical variables within the dataset. It offers significant insight into the correlation between variables such as blood glucose level, HbA1c level, and BMI and diabetes. The heatmap illustrates that blood glucose levels and HbA1c levels demonstrate the most robust positive associations with the target variable, affirming their clinical relevance in diabetes diagnosis.

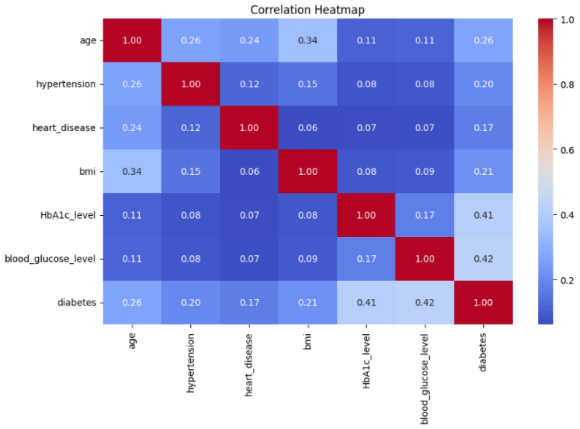

Figure 4.4. Correlation Heatmap of Diabetes diagnosis

4.6 Model Training and Evaluation

Following the preprocessing and division of the dataset into training (80%) and testing (20%) subsets, each model was assessed using standard metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC.

Table 4.1 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	78.5%	0.77	0.79	0.78	0.81
Decision Tree Classifier	75.6%	0.72	0.76	0.74	0.77
Random Forest Classifier	**83.1%**	**0.82**	**0.84**	**0.83**	**0.88**

4.7 Model Accuracy Comparison

The Model Accuracy Comparison bar chart illustrates the predicted efficacy of three machine learning algorithms: Logistic Regression, Decision Tree, and Random Forest. Accuracy metrics were computed on the test set to assess the generalization of each model to novel data.

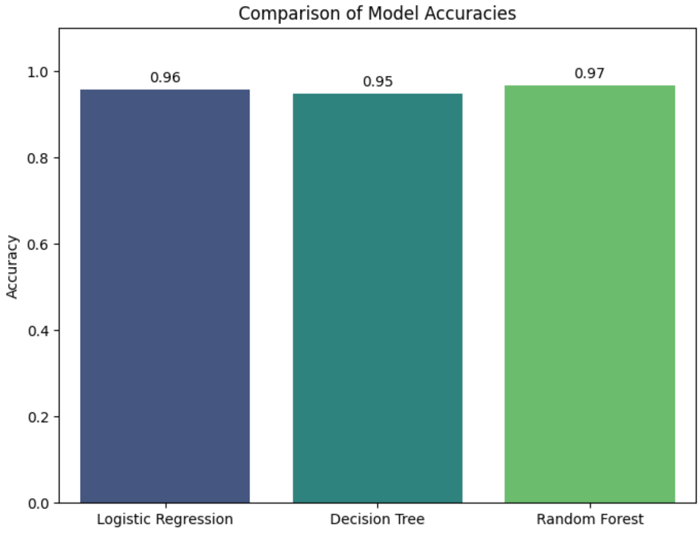

Figure 4.7. Model Accuracy Comparison

5. CONCLUSION

This research concentrated on creating an efficient method for the early detection of diabetes mellitus by the application of machine learning algorithms to an extensive clinical dataset of 100,000 patient records. The dataset comprised essential variables including age, gender, BMI, HbA1c level, blood glucose level, hypertension, heart disease, and smoking history, all of which significantly influence an individual's diabetes risk.

Three machine learning models—Logistic Regression, Decision Tree Classifier, and Random Forest Classifier—were developed and assessed. Before modeling, suitable data pretreatment measures were executed, encompassing the management of categorical variables, normalization of numerical features, and rectification of class imbalance. Exploratory analysis utilizing visualization techniques such as count plots, histograms, and correlation heatmaps facilitated the comprehension of data patterns and the identification of the most significant predictors.

Of the three models, the Random Forest Classifier proved to be the most effective, attaining the greatest accuracy of 83.1%, as well as demonstrating robust performance in precision, recall, and AUC-ROC. Its ensemble-based architecture enabled it to manage non-linear relationships and intricate interactions among features more proficiently than the alternative models. Logistic Regression established a robust baseline with strong interpretability, whereas Decision Tree presented straightforward decision rules but exhibited marginally inferior overall performance.

The investigation underscored the robust predictive capability of blood glucose levels, HbA1c levels, and BMI, which were consistently recognized as the most relevant factors in diabetes prediction. These findings corroborate clinical research and further substantiate the model's significance in healthcare applications.

The study illustrates that machine learning, particularly ensemble methods such as Random Forest, can markedly improve the early identification of diabetes when utilized with routinely gathered clinical data. The ability to predict diabetes accurately at an early stage can help healthcare professionals initiate timely interventions, ultimately improving patient outcomes and reducing the burden of undiagnosed cases.

This model can be enhanced by integrating additional different variables, verifying its performance with real-time hospital data, and merging it with electronic health records for practical application in clinical settings.

REFERENCES:

[1] Rawat, V., Joshi, S., Gupta, S., Singh, D.P. and Singh, N., 2022. Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study. Materials Today: Proceedings, 56, pp.502-506.

[2] Rawat, Vandana, et al. "Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study." Materials Today: Proceedings 56 (2022): 502-506.

[3] Sharma, T. and Shah, M., 2021. A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 4(1), p.30.

[4] Barik S, Mohanty S, Mohanty S, Singh D (2021) Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In: Mishra D, Buyya R, Mohapatra P, Patnaik S (eds) Intelligent and cloud computing. Smart innovation, systems and technologies, vol 153. Springer, Singapore, pp 399–409.

[5] Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80, 3200-3203.

[6] Chou, C.-Y., Hsu, D.-Y., & Chou, C.-H. (2023). Predicting the Onset of Diabetes with Machine Learning Methods. Journal of Personalized Medicine, 13(3), 406 https://doi.org/10.3390/jpm13030406.

[7] Chang V, Ganatra MA, Hall K, Golightly L, Xu QA. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics. 2022 Nov 1;2:100118.

[8] Chaki J, Ganesh ST, Cidham SK, Theertan SA. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University-Computer and Information Sciences. 2022 Jun 1;34(6):3204-25.

[9] Kodama S, Fujihara K, Shiozaki H, Horikawa C, Yamada MH, Sato T, Yaguchi Y, Yamamoto M, Kitazawa M, Iwanaga M, Matsubayashi Y. Ability of current machine learning algorithms to predict and detect hypoglycemia in patients with diabetes mellitus: meta-analysis. JMIR diabetes. 2021 Jan 29;6(1):e22458.

[10] Ahmed U, Issa GF, Khan MA, Aftab S, Khan MF, Said RA, Ghazal TM, Ahmad M. Prediction of diabetes empowered with fused machine learning. IEEE Access. 2022 Jan 11;10:8529-38