SmartHealth: A Web-Integrated Machine Learning Framework for Real-Time Type 2 Diabetes Prediction

Balogun Temitayo E.
Department of Information Systems
Federal University of Technology
Akure, Nigeria

Ogundumila Olanrewaju A.
Department of Information Systems
Federal University of Technology
Akure, Nigeria

Aderiye Daniel A.
Department of Information Systems
Federal University of Technology
Akure, Nigeria

Ekundayo M. Omotehinse
Department of Computer Science
Louisiana Tech University
USA

ABSTRACT
Diabetes mellitus, especially Type 2 diabetes, is a major global health challenge with increasing prevalence and serious complications if not detected early. Early prediction of individuals at risk is therefore essential for effective management and prevention. This project was undertaken to design and implement a web-based diabetes prediction system using machine learning techniques. The study employed the Pima Indians Diabetes Dataset obtained from Kaggle, which contains 768 patient records with eight medical attributes, including glucose level, blood pressure, body mass index (BMI), age, and family history. The dataset was preprocessed to handle missing values, outliers, and feature scaling before being divided into training and testing sets. A Logistic Regression model was developed to perform the prediction task. The choice of this algorithm was based on its simplicity, efficiency, and interpretability in binary classification problems. The model was trained and evaluated using standard metrics such as accuracy, precision, recall, and F1-score, and the results confirmed its reliability in predicting diabetes outcomes. For practical implementation, the trained model was integrated into a Flask-based web application with a user-friendly HTML/CSS interface and deployed on Vercel. The system enables users to input clinical details and receives instant predictions of diabetes risk. While not intended to replace professional medical diagnosis, the application provides a cost-effective and accessible tool for early screening and awareness. This project demonstrates the potential of machine learning in healthcare and establishes a foundation for future improvements, including the use of larger datasets, and additional predictive features.

General Terms
Machine Learning, Predictive Algorithms, Werb Application

Keywords
Type 2 Diabetes, Logistic Regression, Preventive Healthcare, Pima Indians Diabetes Dataset, Random Forest, Smart Health.

1. INTRODUCTION
Diabetes mellitus is a chronic metabolic disease characterized by elevated blood glucose levels resulting from defects in insulin production or utilization [1]. Type 2 diabetes is the most common type of diabetes. Type 2 diabetes accounts for over 90% of all diabetes cases globally and has become a leading cause of morbidity and mortality. According to the International Diabetes Federation [2], approximately 537 million adults live with diabetes, and this number is projected to rise to 783 million by 2045. Unlike Type 1 diabetes, which normally becomes apparent at early ages and is associated with the inability of the body to produce insulin, Type 2 diabetes occurs mostly in adults and is associated with insulin resistance. This implies that the body continues to produce insulin, but the cells fail to respond to insulin. The Centers for Disease Control and Prevention [3] reports that Type 2 diabetes affects millions of people around the globe, and many of them do not even know that they have it since the symptoms are not always severe and may go unnoticed before complications develop. Early detection is critical for preventing severe complications such as kidney failure, blindness, and heart disease [4].

Traditional diagnostic methods such as Fasting Plasma Glucose (FPG), Oral Glucose Tolerance Test (OGTT), and HbA1c tests are accurate but costly and time-consuming, limiting accessibility in low-resource settings [5]. Machine learning (ML) offers a promising alternative by analyzing clinical data to identify individuals at risk of diabetes early.

Several studies have explored ML applications for diabetes prediction. Wilson et al. [6] used logistic regression to develop a diabetes risk score from clinical and demographic data. Kopitar et al. [7] compared logistic regression and random forest models and found that ensemble methods improved prediction accuracy. Kwon et al. [8] utilized deep learning on electronic health records for predicting diabetes onset. Nguyen et al. [9] proposed a hybrid 'wide and deep' model achieving an AUC of 0.84. Zou et al. [10] applied decision trees and neural networks in China and highlighted the importance of dimensionality reduction for improved computation. Gundogdu [11] used XGBoost with feature selection, achieving high accuracy and interpretability using SHAP values. Collectively, these studies show that ML techniques enhance early detection, but challenges remain in deployment for practical use. This study contributes by comparing Logistic Regression and Random Forest for early diabetes detection, emphasizing model transparency and practical deployment through a user-friendly web-based platform.

2. LITERATURE REVIEW
Several studies have contributed to the growing field of

diabetes prediction using statistical models and machine learning techniques.

One of the earliest influential works was by Wilson et al. [6], who employed logistic regression on the Framingham Offspring Study to develop a diabetes risk score. Their dataset included demographic variables such as age and sex, as well as clinical indicators like BMI, fasting glucose, and family history. The study revealed that age, BMI, and glucose levels were strong predictors of diabetes onset. However, logistic regression assumes a linear relationship between predictors and outcomes, which limits the model's ability to capture complex interactions. Furthermore, the model was developed using data from a specific population in the United States, making its applicability to other populations questionable without recalibration. Building on this, Kopitar et al. (2020) conducted a study using Slovenian electronic health records to compare logistic regression with random forest algorithms. They applied extensive data preprocessing, including outlier detection, normalization, and multiple imputations for missing data. By creating time-sliced datasets spanning six to thirty months, they assessed not only predictive accuracy but also the stability of feature importance over time. Their results indicated that random forest outperformed logistic regression, particularly in handling nonlinear relationships. However, the authors noted that their work remained confined to algorithmic comparison and did not provide a deployable screening tool for use in hospitals or clinics. This highlights the persistent gap between research experiments and practical clinical applications.

In a Korean context, Kwon et al. [8] utilized electronic health records from over 8,000 patients to build models for predicting the onset of Type 2 diabetes within five years. They tested several models, including logistic regression, linear discriminant analysis, quadratic discriminant analysis, and k-nearest neighbors. The dataset incorporated demographic factors, clinical laboratory results, and lifestyle indicators. Using ten-fold cross-validation, logistic regression emerged as the most reliable model, with an AUC of 0.78. However, the study also revealed that model performance was highly sensitive to the quality of data used, suggesting that feature engineering and data preprocessing are just as critical as the choice of algorithm. More advanced models have also been tested.

Nguyen et al. [9] proposed a "wide and deep" learning architecture on the Practice Fusion EHR dataset, which contained nearly 10,000 patients. Their hybrid model combined a linear "wide" component to capture memorized associations and a "deep" neural network component to learn higher-order interactions. They used the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, improving sensitivity from 31% to over 70%, though with a small drop in specificity. The model achieved an AUC of 0.84, outperforming traditional methods. The main drawback, however, was its lack of interpretability, as the internal mechanisms of deep learning are often opaque to medical practitioners.

Similarly, Zou et al. [10] employed decision trees, random forests, and neural networks on hospital datasets in China, coupled with feature selection methods like Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR). They found that random forest achieved the highest accuracy of about 81%, outperforming neural networks in this case. The study's methodological contribution lay in its use of dimensionality reduction to improve computational efficiency. However, the dataset was region-specific, and the authors acknowledged the risk of overfitting their local population. In recent years, researchers have turned to ensemble learning techniques such as XGBoost and LightGBM.

Gundogdu [11] applied random forest-based feature selection followed by XGBoost classification to predict early-stage diabetes. This two-step strategy enhanced performance by first identifying the most relevant clinical features and then using a gradient boosting algorithm to achieve high predictive accuracy. The study also employed explainability tools such as SHAP (Shapley Additive Explanations) to interpret model outputs, which increased the model's transparency and potential clinical acceptance. The emphasis on interpretability reflects a growing recognition that black-box models, while accurate, may not gain trust in healthcare settings unless they can explain their predictions.

Further contributions include works such as Fazakis et al. [12], who explored the use of support vector machines (SVMs) for diabetes classification. Their model was effective in handling high-dimensional datasets and demonstrated robustness against overfitting. However, SVMs remain less interpretable than regression-based approaches, and their computational cost increases with larger datasets.

3. METHODOLOGY.

The system architecture, as seen in Figure 2, is structured framework designed to deliver a seamless and Engaging Diabetes Prediction Process. This architecture consists of several layers that work together to ensure efficient prediction, data management, and user interaction

3.1 Data Collection

The dataset was sourced from Kaggle, a well-known platform for machine learning and data science competitions. Kaggle provides thousands of datasets for free and allows researchers to benchmark their models against established studies.

The Pima Indians Diabetes Dataset was originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data records the medical history of adult female patients of Pima Indian heritage, aged 21 years and above. This specific population was chosen because the Pima Indians were reported to have one of the highest rates of Type 2 diabetes worldwide at the time of data collection. The dataset is composed of 768 patient records and 9 attributes. Among these, 8 are input features (independent variables) and 1 is the output feature (dependent variable). Figure 1 shows the count of the dataset where 1 indicates the presence of diabetes and 0 shows non-diabetic.

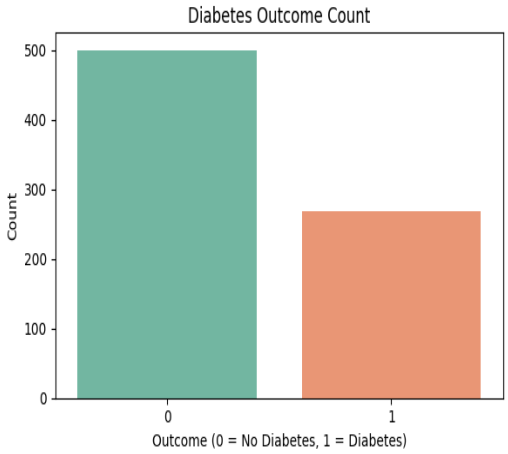

Fig 1: DIABETES OUTCOME COUNT

3.2 Data Preprocessing

Before the dataset could be used for model training and evaluation, it required several preprocessing steps to address inconsistencies, missing values, and scaling issues. Preprocessing ensures that the dataset is clean, balanced, and suitable for machine learning algorithms. The main preprocessing tasks carried out includes handling missing and invalid values such as Blood Pressure values with "0" mmHg and Skin Thickness and Insulin with zero entries were corrected calculating the median of each column and used to replace these unrealistic zeros. Median imputation was chosen because it is less affected by extreme outliers compared to the mean.

3.3 Feature Extraction and Data Splitting

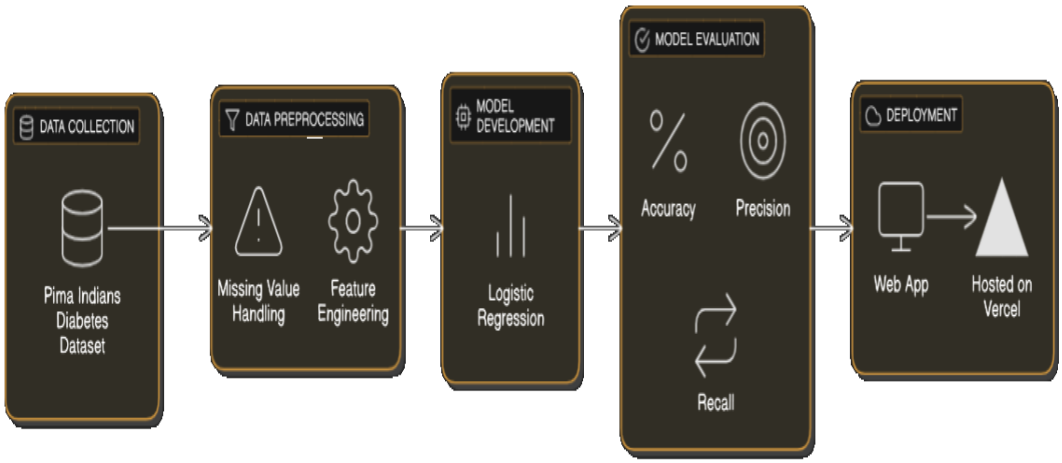

Fig 2: Architecture of the system

Since the dataset contains attributes measured in different ranges, for example, glucose values can go above 150, while Diabetes Pedigree Function is below 3, scaling was necessary. Standardization was applied to transform the features into a common scale with a mean of 0 and a standard deviation of 1. This ensured that no single variable dominated the learning process, especially in models like Logistic Regression, which are sensitive to feature magnitudes

To evaluate the models effectively, the dataset was divided into:

Training Set (80%) – used to train the machine learning models.

Testing Set (20%) – used to evaluate how well the models generalize to unseen data.

The train_test_split() function from Scikit-learn was used to perform this split randomly, ensuring that both diabetic and non-diabetic cases were fairly represented in each set.

3.4 Model Selection and Training

The ready data is then subjected to different machine learning algorithms at this point to get trained. The system investigates various classification models, including Logistic Regression, and Random Forest to establish the most accurate and interpretable algorithm to predict diabetes. The models were each trained on the training set and tested on the test set, and their performance is measured using their accuracy, precision, recall, F1-score, and AUC-ROC.

Both models were trained using the Scikit-learn library in Python. A 5-fold cross-validation strategy was applied to reduce bias and variance. Hyperparameter tuning was conducted using Grid Search:

Logistic Regression: tuning the regularization parameter C.

Random Forest: tuning the number of trees (n_estimators), maximum depth, and minimum samples per split

Following extensive comparison, logistic regression was chosen as the major predictive model. Logistic regression has several strengths, such as interpretation, computational performance, and good performance on small to medium-sized datasets. In addition, it yields probabilistic results, and these are specifically applicable in medical contexts where estimations of risk are more informative than binary classification.

3.5 Model Deployment

This is the final layer, and it involves deploying the machine learning model as a web-based application for end-user interaction. It integrates the trained model with a web framework (Flask) and provides a user-friendly interface that is accessible from any device with an internet connection. It allows users to input clinical parameters, trigger predictions, and view results in real time.

Additionally, this layer connects with a backend relational database to store patient data, historical predictions, and system logs. These records can be used for audit purposes, clinical decision support, and future model retraining, thereby ensuring continuous system improvement.

4. RESULT AND DISCUSSION

Feature importance analysis was conducted as shown in Figure 3 to determine which clinical variables had the greatest influence on the prediction of Type 2 diabetes. This provides valuable insights into both technical and medical interpretation of the results. The importance of features is indicated by the magnitude of their coefficients. Features with larger absolute coefficient values contribute more strongly to the prediction outcome. Positive coefficients increase the likelihood of being classified as diabetic, while negative coefficients reduce it.

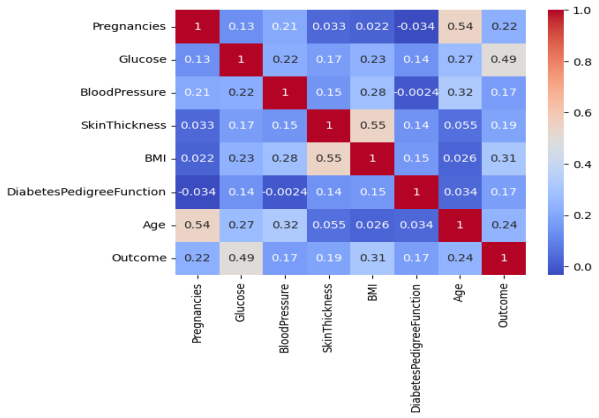

Fig 3: Feature Correlation Matrix

From the correlation matrix, the most influential features included:

Glucose Level – the strongest predictor, as higher glucose readings directly increase diabetes risk.

BMI (Body Mass Index) is an important risk factor, as obesity is a major contributor to diabetes.

Age – older patients showed a higher risk of developing diabetes.

Other features, such as Pregnancies and Diabetes Pedigree Function, also contributed but had relatively smaller coefficients compared to the primary predictors.

The Logistic Regression and Random Forest models were evaluated on the test dataset. Logistic Regression achieved an accuracy of 77%, as shown in figure 4 while Random Forest achieved an accuracy of 76% as shown in figure 5.

```
Accuracy: 0.7686567164179104
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.90      0.83        86
           1       0.74      0.54      0.63        48

    accuracy                           0.77       134
   macro avg       0.76      0.72      0.73       134
weighted avg       0.77      0.77      0.76       134


print("Training Score:",model.score(X_train,y_train)*100,'%')

print("Testing Score:",model.score(X_test,y_test)*100,'%')

Training Score: 77.98507462686567 %
Testing Score: 76.86567164179104 %
```

Fig 4: Performance result of the Logistic Regression

```
Accuracy: 0.7611940298507462
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.86      0.82        86
           1       0.70      0.58      0.64        48

    accuracy                           0.76       134
   macro avg       0.74      0.72      0.73       134
weighted avg       0.76      0.76      0.76       134


print("Training Score:",model.score(X_train,y_train)*100,'%')

print("Testing Score:",model.score(X_test,y_test)*100,'%')

Training Score: 100.0 %
Testing Score: 76.11940298507463 %
```

Fig 5: Performance result of the Random Forest

The Logistic Regression was deployed as a web application on Vercel, making it accessible via the internet and usable on different devices. This ensured that the predictive system could reach a wider audience beyond the local development environment.

Deployment aimed to transform the trained Logistic Regression model into an interactive web-based tool where users can input their clinical details and receive an instant prediction on their diabetes risk. Hosting the system on Vercel provides scalability, reliability, and global accessibility.

The deployment process followed these steps:

1. Model Preparation: The Logistic Regression model was trained and saved using joblib. A preprocessing pipeline was included to ensure user inputs were standardized before prediction.

2. Application Development: A Flask backend was developed to handle requests and generate predictions. A frontend interface (HTML, CSS, Bootstrap) was created to collect user input and display the prediction results.

3. Code Integration for Deployment: The project files (Flask app, model file, templates, static assets) were organized in a structured format. A requirements.txt file was created to specify dependencies (Flask, Scikit-learn, Pandas, NumPy, Joblib). A configuration file (vercel.json) was added to guide deployment settings on Vercel.

4. Hosting on Vercel: A GitHub repository was created and connected to Vercel. The repository was imported into Vercel's dashboard. Deployment was triggered automatically by Vercel, which built and hosted the application.

5. Testing on Vercel: After successful deployment, the web app was tested on different browsers (Chrome, Firefox, Edge) to confirm that the input form, preprocessing, and prediction output worked as intended. The app was also tested on both desktop and mobile devices to ensure responsiveness.

Once deployed, the web application allowed users to enter clinical details such as glucose, BMI, blood pressure, and age into a form, submit the data for real-time analysis by the trained model and receive a prediction result (Diabetic or Non-Diabetic). The interface of the application is shown in Figure 6 while Figures 7 and 8 shows a prediction for a diabetic and non-diabetic patient.

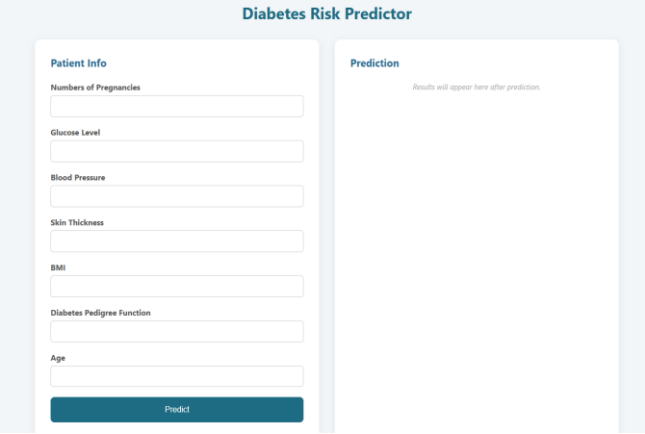

Fig 6. User Interface of the application

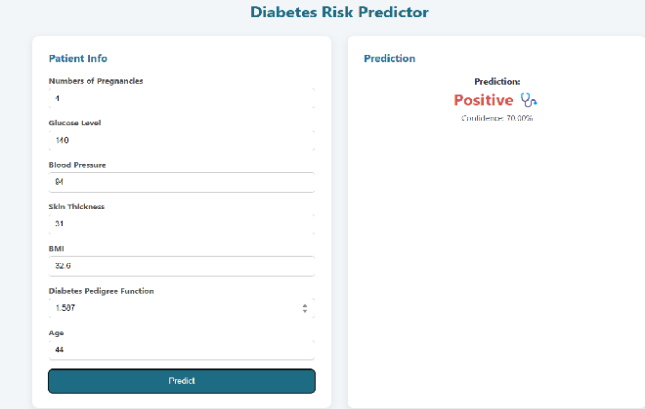

Fig 7. Prediction for a diabetic patient

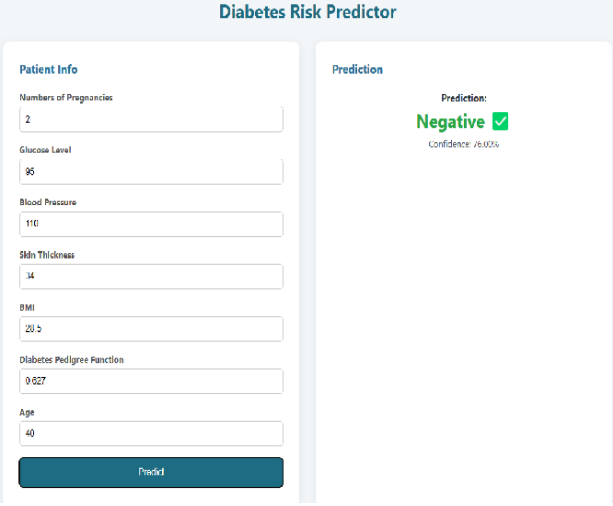

Fig 8. Prediction for a non-diabetic patient

The system was also designed to provide a clean and user-friendly interface so that users without technical expertise could interact with it easily

5. CONCLUSION

The prevalence of diabetes continues to rise globally, particularly in developing countries where limited access to healthcare facilities makes early screening difficult. This study set out to address the challenge of early detection of Type 2 diabetes through the application of machine learning techniques to clinical data. The research utilized the Pima Indians Diabetes dataset, which provided relevant clinical variables such as glucose level, BMI, age, and family history of diabetes. Logistic Regression and Random Forest were trained, validated, and evaluated using multiple performance metrics. The evaluation results showed that while Random Forest had a slight advantage Beyond evaluation, the project went a step further by deploying the adopted Logistic Regression model on Vercel as a web application. To increase accessibility, the system could be extended to mobile platforms (Android/iOS). This would allow users to easily check their diabetes risk on smartphones, making it more practical for daily use.

6. ACKNOWLEDGMENTS

Our thanks to everyone who contributed to this research, most especially the Department of Information Systems, Federal university of technology, Akure for their lab in carrying out this research.

7. REFERENCES

[1] American Diabetes Association, 2013. Diagnosis and classification of diabetes mellitus. Diabetes care, 36(Supplement_1), pp.S67-S74.

[2] Magliano, D.J., Boyko, E.J. and Atlas, I.D., 2021. COVID-19 and diabetes. In IDF DIABETES ATLAS [Internet]. 10th edition. International Diabetes Federation.

[3] Centers for Disease Control and Prevention. (2020). National diabetes statistics report, 2020. [online] U.S. Department of Health and Human Services. https://www.cdc.gov/diabetes/data/statistics-report/index.html

[4] Balogun, T., Saliu, R., Faluyi, S. and Fapohunda, K., 2022, November. Comparative analysis of deep learning models for the detection and classification of Diabetes Retinopathy. In 2022 5th Information Technology for Education and Development (ITED) (pp. 1-6). IEEE.

[5] Ayesha, K., Ajmal, A., Akram, A., Sadaf, J., Usman, F. and Hafeez, S., 2025. Diagnostic Accuracy of Fasting Blood Sugar and Oral Glucose Challenge Test for Gestational Diabetes Mellitus: Fasting Blood Sugar and Oral Glucose Challenge Test for GDM. Pakistan Journal of Health Sciences, pp.139-143.

[6] Wilson, P.W., D'Agostino, R.B., Fox, C.S., Sullivan, L.M. and Meigs, J.B., 2011. Type 2 diabetes risk in persons with dysglycemia: the Framingham Offspring Study. diabetes research and clinical practice, 92(1), pp.124-127..

[7] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. and Stiglic, G., 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 10(1), p.11981.

[8] Kwon, O., Na, W., Kang, H., Jun, T.J., Kweon, J., Park, G.M., Cho, Y., Hur, C., Chae, J., Kang, D.Y. and Lee, P.H., 2022. Electronic Medical Record–Based Machine Learning Approach to Predict the Risk of 30-Day Adverse Cardiac Events After Invasive Coronary Treatment: Machine Learning Model Development and Validation. JMIR Medical Informatics, 10(5), p.e26801.

[9] Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T., ... & Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer methods and programs in biomedicine, 182, 105055.

[10] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H., 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, p.515.

[11] Gündoğdu, S., 2023. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. Multimedia Tools and Applications, 82(22), pp.34163-34181.

[12] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N. and Moustakas, K., 2021. Machine learning tools for long-term type 2 diabetes risk prediction. ieee Access, 9, pp.103737-103757.