

**ADVANCED DIABETES PREDICTION USING SUPERVISED MACHINE LEARNING  
TECHNIQUE: RANDOM FOREST**

**Ugboaja Samuel Gregory<sup>1</sup>, Edeh Michael Onyema<sup>2\*</sup>, Madubuezi Christian Okoronkwo<sup>3</sup>,  
Anichebe Gregory Emeka<sup>4</sup>, Udeh Chukwuma Callistus<sup>5</sup>, Ogbuoka Oby Modest<sup>6</sup>**

<sup>1</sup>Department of Computer Science, Michael Okpara University of Agriculture Abia State, Nigeria  
[ugboaja.samuel@mouau.edu.ng](mailto:ugboaja.samuel@mouau.edu.ng),

<sup>2</sup>Department of Mathematics and Computer Science, Coal City University, Enugu, Nigeria.  
[mikedreamcometrue@gmail.com](mailto:mikedreamcometrue@gmail.com)

<sup>3</sup>Department of Computer Science, Michael Okpara University of Agriculture Abia State, Nigeria  
[Okmaduchris@gmail.com](mailto:Okmaduchris@gmail.com)

<sup>4</sup>Department of Computer Science, University of Nigeria, Nsukka. Nigeria. [gregory.anichebe@unn.edu.ng](mailto:gregory.anichebe@unn.edu.ng)

<sup>5</sup>Dept of Computer Science, Enugu state university of Science and Technology, Enugu, Nigeria [Chukwuma.udeh@esut.edu.ng](mailto:Chukwuma.udeh@esut.edu.ng)

<sup>6</sup>Department of Education Foundation, Faculty of Education, Coal City University, Enugu, Nigeria  
[obymessages@gmail.com](mailto:obymessages@gmail.com)

**Corresponding Author:** Edeh Michael Onyema ([mikedreamcometrue@gmail.com](mailto:mikedreamcometrue@gmail.com))

## Abstract

Diabetes remains one of the major causes of untimely death globally. Over 11% of the global population is diabetic, possibly due to late disease detection, inadequate interventions, and lifestyle choices etc. The growing severity of diabetes is driving scientific interest in leveraging Digital Health Technologies (DHTs) for improved management and treatment. Early diagnosis of diabetes is essential for effective interventions, reducing complications, and lowering the mortality rate associated with the disease. Thus, this study focuses on prediction of diabetes using supervised machine learning technique, specifically Random Forest Algorithm (RFA) for timely detection and prevention of the disease. The model was trained using Pima Indian dataset (diabetes), which is freely available on Kaggle database. Trial result indicate that the model was promising, with an accuracy of 92%, 89% precision, 88% recall, and a 90% F1-score. The study shows that applying the Random Forest algorithm significantly improves the accuracy and efficiency of early diabetes detection and diagnosis. However, in spite of the prospects of ML models in diabetes management, there are still concerns about its drawbacks including algorithmic bias, legal and ethical issues, and implementation challenges in clinical environment. Thus, we recommend that legal framework should be put in place to guide the use ML algorithms, and other digital health technologies in clinical diabetes care delivery.

**Keywords:** Diabetes, Machine learning, Random Forest algorithm, Digital diagnosis, Healthcare.

## INTRODUCTION

Technology has revolutionized every segment of human society especially the health sector (Muhammad *et al.*, 2024). Innovative technologies such as mobile technologies are helping healthcare professionals to improve their skills and professionalism (Onyema *et al.*, 2022). One area that researchers are trying to leverage the potentials of Digital Health technologies is in Diabetes care delivery. Diabetes is a major health problem that has caused many untimely deaths across the world (World Health Organization, 2023). Diabetes is categorized into four types: type 1, type 2, gestational, and pre-diabetes (February *et al.*, 2021). The danger associated with diabetes can be worse if the disease is not detected and prevented on time (Olisah *et al.*, 2022). Ignorance and poor interventions increases the prevalence of the disease especially in third world countries (Tahir and Farhan, 2023). Statistics from International Diabetes Federation (IDF) indicates over 425 million persons spread across the globe are diabetic (International Diabetes Federation, 2015).

Technological advances offer opportunities to improve Diabetes education (Onyema *et al.*, 2020), data security (Onyema *et al.*, 2021) and also early diagnosis and therapy. Timely detection of diabetes is critical to the proper management and inventions against the disease. Research has shown that late diagnosis or treatment of diabetes can increase its severity and lower the survivability of patients (Herman *et al.*, 2015). The use of cutting edge technologies especially Artificial Intelligence (AI) algorithms such as deep learning and machine learning have shown prospects for digital diagnosis, customized diabetes management and enhanced treatment outcome. Several studies including (Saxena *et al.*, 2012 ; Hossain *et al.*, 2014; Choubey *et al.*, 2020; Edeh *et al.*, 2022, and Khandakar *et al.*, 2022; Victor *et al.*, 2022) indicated that technology are being deployed at different levels of healthcare to manage diabetic conditions.

Considering the growing concerns about diabetes, and the need for more robust technique for its management, this study attempted to leverage the potentials of supervised machine learning algorithm (Random Forest) for diabetes prediction with a view to contribute to the ongoing efforts towards diabetes care and reduction of mortalities due to the disease. The

researchers believe that the outcome of this research would be critical in creating more awareness about the disease and also assisting physicians in improving their knowledge and approach to support patients.

## **REVIEW OF CLOSELY RELATED LITERATURE**

Diabetes constitutes a grave danger to humanity. Diabetes-related insulin resistance impairs the ability of immune cells to operate, which leads to immunological suppression. Deformations in phagocytic cells, which are essential mediators for managing and curing infections caused by bacteria, are brought on by faulty insulin signaling leading to diabetic condition. An important health consequence linked to diabetic is a heightened susceptibility to bacteria-related infections, characterized by elevated incidence and intensity in comparison to those without the disease. Research into early diabetes diagnosis is crucial due to the global prevalence of the disease. Report from the World Health Organization indicates a stable increase in the number of diabetes related cases. For instance, In 1980, there were 108 million persons with diabetes; by 2014, there were 422 million (WHO, 2023). Diabetes condition has been projected to rise in coming years as can be seen in figure 1 unless more urgent steps are taken globally to halt the continuous spread of the disease.

The rise in cases is largely attributed to lifestyle factors and poor dieting. Automated diagnosis can help detect serious complications on time (Noviyanti and Alamsyah, 2024). According to Zou *et al.* (2018), AI-driven algorithms could be used to improve diabetes mellitus prediction for hospitalized patients, and enhance their chances of survival. Human errors often associated in diabetes management can be bridged by digital health technologies (Guan *et al.* (2023), such as AI to better understand risk factors of diabetes at early stage and apply necessary precautionary treatments to treat the disease. An intelligent mobile diabetes management system was created by Alotaibi *et al.* (2016), and an initial trial showed that it could considerably facilitate users' comprehension of information relating to their health. Table 1 highlights summary of previous efforts made by different scholars towards the use of technology in diabetes healthcare delivery. The study updates the understanding of computational algorithms' application in diabetes care delivery. The outcome will contribute to the efforts towards combating the disease by leveraging the potentials of technology.

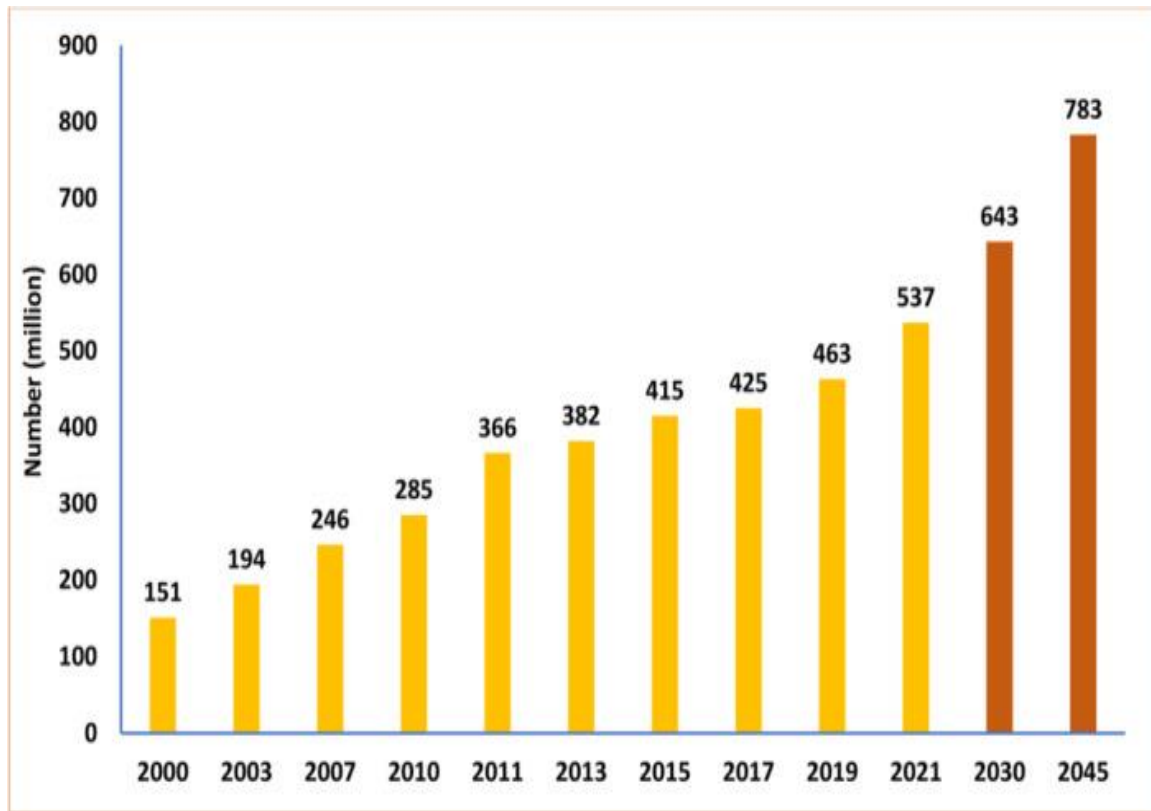


Figure 1: Global Adult Diabetes Statistics, with Projected Numbers for 2030 and 2045  
 ( Source: International Diabetes Federation diabetes atlas. 10th edition, 2021; Hossain et al, 2024)

**TABLE 1: SUMMARY OF RELATED WORKS**

| Authors/Year                  | Technique   | Performance /Outcome   |
|-------------------------------|---|--|
| Choubey <i>et al.</i> , 2020. | SVM, KNN, and NB were used to predict diabetes occurrence | The result showed that SVM performed better than Naive Bayes, And KNN                    |
| Noviyanti and Alamsyah, 2024  | Machine Learning approach using The Pima Indian Dataset   | The finding showed accuracy rate of 87% indicating its potential for diabetes diagnosis. |

|                                |   |   |
|--------------------------------|---|---|
| Edeh <i>et al.</i> , 2022      | Different ML algorithms, Random Forest, SVM, Naïve Bayes, and Decision Tree, were utilized and their performances were compared for diabetes prediction | The output indicated that the random forest algorithm surpassed other modeling methods with a 97.6% accuracy. The study demonstrated that computational algorithms are capable of identifying diabetes. |
| Saxena <i>et al.</i> , 2012    | Different machine learning algorithms were used and performance compared.   | KNN performed better than other algorithms that were compared including SVM, Decision tree and KNN  |
| Khandakar <i>et al.</i> , 2022 | The CNN and k-means clustering techniques were employed to classify diabetic foot complications.  | The widely used VGG 19 CNN model demonstrated notable performance in stratifying severity diabetic foots.   |
| Zou <i>et al.</i> , 2018       | RF was used to evaluate patients diabetes proneness in in Luzhou,China  | The findings revealed that the model was promising and achieved Accuracy of 0.8084  |
| Deepa <i>et al.</i> , 2021     | ML approach   | Outcome showed 92% precision by SVM   |
| Butt <i>et al.</i> , 2021      | Logistic regression and IoT   | The study affirmed the suitability of IoT-enabled devices in diabetes tracking and prevention.  |
| Kopitar <i>et al.</i> , 2020   | RF approach using a dataset of 3,723 participants   | The result achieved AUC 0.84–0.85 which indicates its reliability and potential for assistive diabetic treatment.   |

## CONCEPT OF MACHINE LEARNING

Machine learning (ML) is a technique that deals with training automated machines or algorithms to acquire specific behaviors or qualities and then thinking and making judgments based on that knowledge (Edeh *et al.*, 2020b). An intelligent models or algorithms, whether supervised or unsupervised, are able to grow its capabilities, such as preciseness, without being explicitly designed. Machine learning algorithms learn from their experiences or

insights before becoming extremely active in order to perform actions similar to humans (Flach, 2012). Machine learning advancements have made possible automated forecasting of diabetes compliance risks and individualized therapy outcomes, as shown in a variety of studies such as (Mackenzie *et al.*, 2024). Most modern artificial intelligence now relies on machine learning, which can be unsupervised (seeking a pattern in the data presented to it) or supervised (learning from information fed into it by a human who has labeled it). ML remains one of the most potential fields of research for solving many issues in society.

### **RANDOM FOREST ALGORITHM (RFA)**

Random Forest is a strategy introduced by Leo Breiman (Breiman, 2001), and popularly used for classification (Benbelkacem and Atmani, 2019). It is a method of ensemble learning that generates a large number of trees of decisions during training and returns the mode (classification) or mean prediction (regression) of each tree. RFA is notable for durability, precision, capacity for managing big, complex datasets and less prone to overfitting challenges. It can be deployed in solving classification problems such as Disease prediction, image classification, and also for regression including demand forecasting and market trends. Despite its potentials, Random forest has its limitations such as high memory demand and also algorithmic bias.

### **METHODOLOGY**

The study utilized a method based on the approach developed by Noviyanti and Alamsyah (2024), as illustrated in Figure 2. This method involves several steps: data collection, data pre-processing, data splitting, modeling, and evaluation.

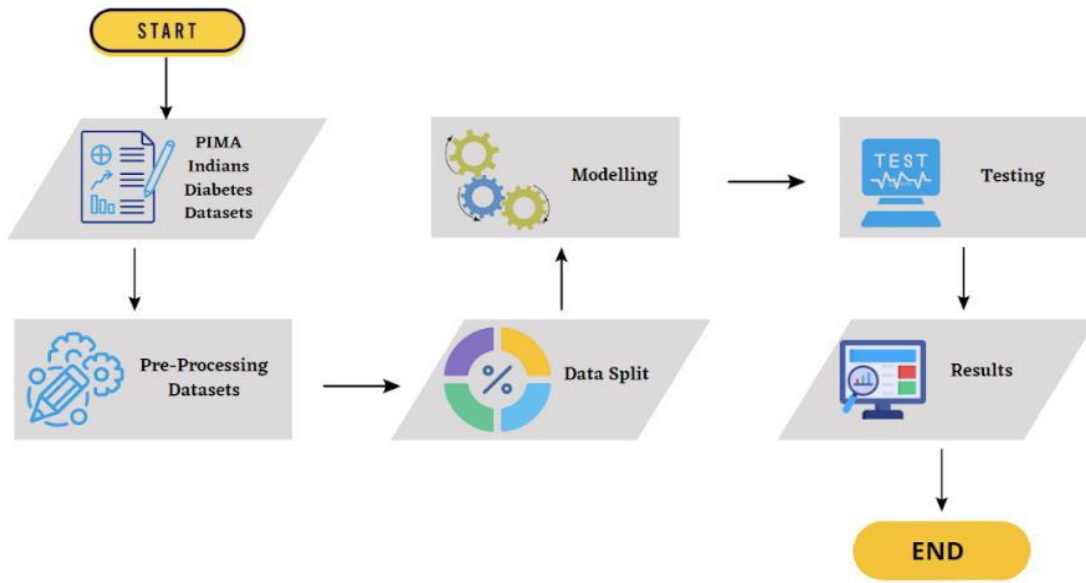


Fig. 2: Dataflow of the proposed model (Noviyanti and Alamsyah, 2024).

### Data Collection

This study utilizes the Pima Indian dataset, which is freely available on the Kaggle database (<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>) to diagnose diabetes in individuals. Because of its huge sample size and a variety of factors such as medical information and demographics, the Pima Indian Dataset is one of the most commonly used datasets in machine learning to diagnose diabetes (Rousyati *et al.*, 2021). The dataset contains 768 unique data points, ranging in age from 21 to 81 years. A total of 500 data records are in the negative class, which includes people that have not been identified as having hyperglycemia. The remaining quantity of data, 268 in total, comes from people who have diabetes.

### Preprocessing Data

Data is preprocessed to ensure that the simulation procedure runs smoothly. Data preparation can be accomplished by filtering data, adjusting data formats, and so on. During the initial processing stage, NaN values or empty rows in the dataset were discovered. Conversely, following this identification, it is discovered that the dataset is full, with no rows involving NaN. Then, each feature was checked for a value of zero. However, several properties, such as

sugar levels, cardiac output, skin dimension, insulin, and BMI, are unattainable if the number is zero. Default value is then inserted into the data, which has zeros in each of these variables.

### Split Data

At this stage, the processed patient data is divided into three sets: training data, validation data, and testing data. The training data is used to train the Random Forest model, while the validation data is employed to assess the model's performance during training. The testing data was reserved for evaluating model performance after training.

### Model Performances Metrics

By utilizing the confusion matrix, we can assess the model's performance through key metrics such as accuracy, precision, recall, and F1-score. A detailed breakdown of these metrics is provided in equation 1, 2, 3 and 4 below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy represents how well the model predicts the correct outcomes across the entire dataset. It is calculated using the stated formula above.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Precision indicates how accurately the model identifies instances predicted as fake that are actually fake. It can be calculated using the mentioned formula:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall measures how effectively the model identifies all instances of fake jobs. It is calculated using the formula shown above.

$$F1 = 2 \frac{Precision * recall}{Precision + recall} \quad (4)$$

The F1-Score combines both precision and recall into a single metric that reflects model overall performance. It was calculated using the formula in equation 4.

## RESULTS AND DISCUSSIONS

The study leverages a machine learning model to predict diabetes risk based on individual medical history and demographic information. By analyzing factors such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, and blood glucose level, this application assists healthcare professionals in identifying patients at risk of developing diabetes. The results indicate that the model was promising, with an accuracy of 92%, 89% precision, 88% recall, and a 90% F1-score as can be seen in figure 3. The study shows that Random Forest algorithm can significantly facilitate early diabetes detection and diagnosis. However, in spite of the prospects of ML models in diabetes management, there are still concerns about its drawbacks including algorithmic bias, legal and ethical issues, and implementation challenges in clinical environment.

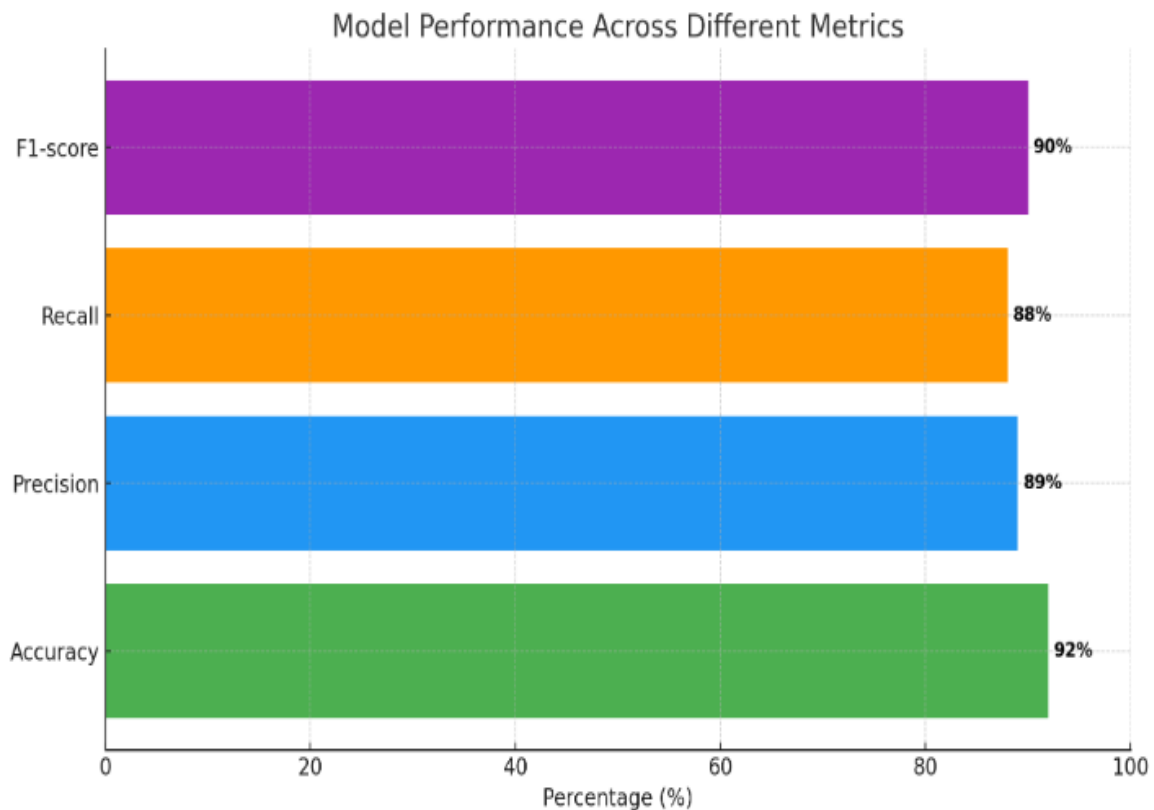


Figure 3: Model Performance Metrics

The graph (Figure 3) displays the model's performance across metrics such as accuracy (92%), precision (89%), recall (88%), and F1-score (90%), highlighting their relative performance. Similarly, figure 4 depicts the form for Collecting Data on Diabetes Risk Prediction, while figure 5 shows the codes blocks in Python environment.

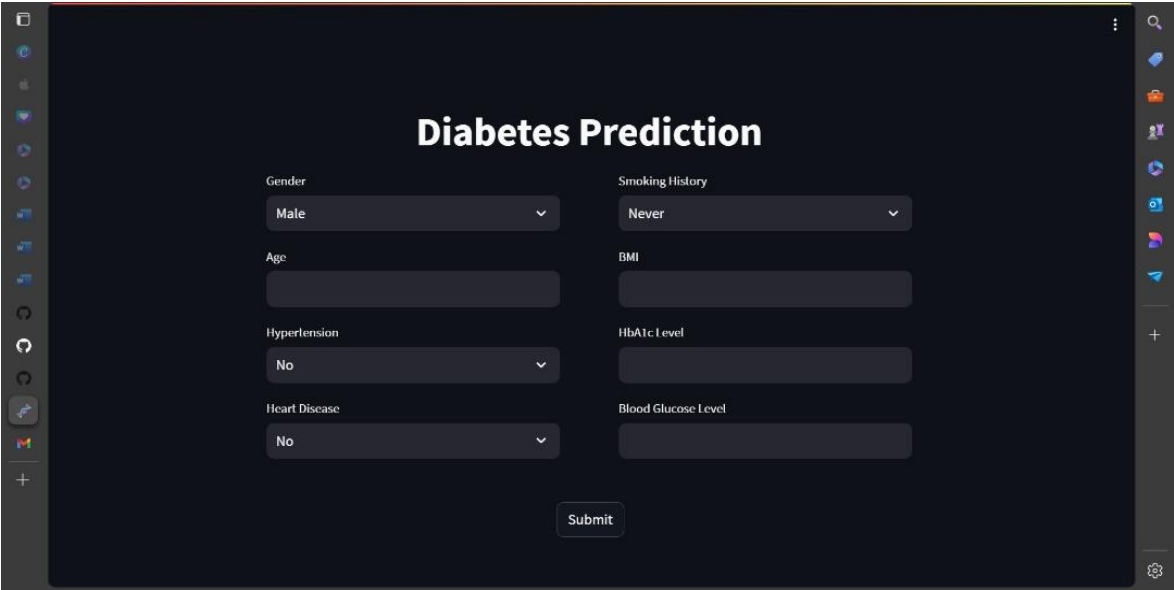
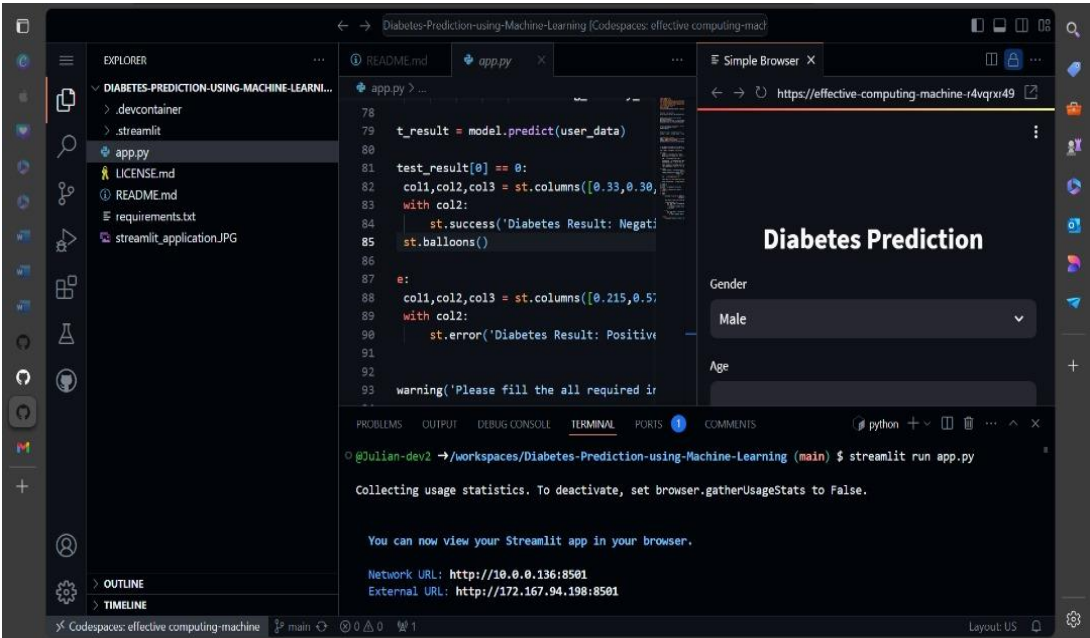


Fig 4 Form for Collecting Data on Diabetes Risk Prediction



```
78
79 t_result = model.predict(user_data)
80
81 test_result[0] == 0:
82     col1,col2,col3 = st.columns([0.33,0.30,
83     with col2:
84         st.success('Diabetes Result: Negati
85     st.balloons()
86
87 e:
88     col1,col2,col3 = st.columns([0.215,0.5;
89     with col2:
90         st.error('Diabetes Result: Positiv
91
92
93 warning('Please fill the all required ir
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS COMMENTS

@Julian-dev2 → /workspaces/Diabetes-Prediction-using-Machine-Learning (main) \$ streamlit run app.py

Collecting usage statistics. To deactivate, set browser.gatherUsageStats to False.

You can now view your Streamlit app in your browser.

Network URL: http://10.0.0.136:8501  
External URL: http://172.167.94.198:8501

### Fig 5: Code Blocks in Python

Table 2: Performance of the Proposed model with and without balancing data.

| Performance Metric | Without data balancing | With data balancing |
|--------------------|------------------------|---------------------|
| Accuracy (%)       | 83                     | 92                  |
| Precision (%)      | 80                     | 89                  |
| Recall (%)         | 79                     | 88                  |
| F1-score           | 80                     | 90                  |

Table 2 shows that data balancing can increase predictive model performance, particularly when dealing with uneven datasets. By ensuring that each class is represented evenly, the model can learn more effectively and generate better predictions. In the example of forecasting diabetes, employing balanced data allows the model to better distinguish between distinct classes and decreases biases caused by an unequal distribution of cases.

## CONCLUSION

This research investigated the efficacy of Random forest approach, in forecasting diabetes. The experimental trial revealed that Random Forest, with its ensemble learning technique, provides a reliable and accurate solution for diabetes prediction. The model's high accuracy, precision, and recall scores indicate that Random Forest could be very supportive for early diabetes detection and treatment. Using this strategy, healthcare practitioners can improve their diagnostic abilities and adopt more effective preventive measures. However, while the Random Forest algorithm produces promising results, it is critical to note its limits, such as the possibility of overfitting with very big datasets and the necessity for adequate processing resources. The further refining and validation of the model will lead to more accurate, timely, and tailored diabetic therapy especially in clinical setting.

## FUTURE WORK

Future work could compare the Random Forest approach to various supervised machine learning methods to assess performance gains. More Investigations can also be conducted to investigate the effect of new features on the efficiency of models such as lifestyle, hereditary, or environmental components.

## REFERENCES

- Alotaibi M.M., Istepanian R., Philip N. (2016). A mobile diabetes management and educational system for type-2 diabetics in Saudi Arabia (SAED) *mHealth*. **2**:33.
- Benbelkacem S. and B. Atmani, (2019) “Random forests for diabetes diagnosis,” *2019 International Conference on Computer and Information Sciences, ICCIS* pp. 1–4.
- Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR.(2021). Machine learning based diabetes classification and prediction for healthcare applications. *J Healthcare Eng.* 2021:9930985.
- Breiman L. (2001). “Random Forests,” *Machine Learning*, **45**(1): 5–32.
- Choubey DK, Kumar M, Shukla V, Tripathi S, Dhandhanika VK.(2020). Comparative analysis of classification methods with PCA and LDA for diabetes. *Curr Diabetes Rev.* **16**:833–50.
- Deepa N, Prabadevi B, Maddikunta PK, Gadekallu TR, Baker T, Khan MA, et al.(2021). An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. *J Super Comput.* **77**:4–16.
- Edeh MO, Khalaf OI, Tavera CA, Tayeb S, Ghouali S, Abdulsahib GM, Richard-Nnabu NE and Louni A (2022) A Classification Algorithm-Based Hybrid Diabetes Prediction Model. *Frontiers Public Health*, **10**:829519. 1-7.
- Edeh MO, Almuzaini, KK; Onu, FU; Verma, D; Gregory, US; Puttaramaiah, M and Afriyie, RK. (2022b). Prospects and Challenges of Using Machine Learning for Academic Forecasting (2022). *Hindawi Computational Intelligence and Neuroscience*, Article ID 5624475: 1-7.
- February J., O. S., Abe, O. O., Obe, O. K., Boyinbode, and O. N. Biodun, (2021) “Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review,” *International Journal of Advanced Trends in Computer Science and Engineering*, **10**:430–439.

- Flach, P. (2012). *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.
- Guan Z, Li H, Liu R, Cai C, Liu Y, et al (2023). Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. *Cell Rep Med*. **4**(10):101213.
- Herman WH, Ye W, Griffin SJ, Simmons RK, et al (2015). Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity and Mortality: A Simulation of the Results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care. *Diabetes Care*. **38**(8):1449-55.
- Hossain MJ, Al-Mamun M, Islam MR.(2004). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Sci Rep*. **7**(3):e2004.
- International Diabetes Federation (IDF, 2015). IDF diabetes atlas, 7th edition. Brussels, Belgium: *International Diabetes Federation*, 2015. International Diabetes Federation (IDF, 2021) *diabetes atlas*. 10th edition. *International Diabetes Federation*; 2021. <https://diabetesatlas.org/atlas/tenthedition/Kaggledataset:https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.
- Khandakar, A., Chowdhury, M.E.H., Reaz, M.B.I., Ali, S.H.M., Kiranyaz, S., Rahman, T., Chowdhury, M.H., Ayari, M.A., Alfkey, R., Bakar, A.A.A., et al. (2022). A Novel Machine Learning Approach for Severity Classification of Diabetic Foot Complications Using Thermogram Images . *Sensors* **22** :4249.
- Kopitar L, Kocbek P, Cilar L, et al.(2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep*. **10**(1):11981.
- Muhammad H, Venkataramaiah G, Onyema EM, Fahad M, Wajid U.K, Muhammad I, Nwosu, O. F. (2024). Cloud-Enhanced Machine Learning for Handwritten Character Recognition in Dementia Patients. In Book: In Book: Driving Transformative Technology Trends With Cloud Computing. Chap 17. 1-14.
- Mackenzie, S.C., Sainsbury, C.A.R. & Wake, D.J.(2024). Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges. *Diabetologia* **67**: 223–235 .
- Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. *Journal of Information System Exploration and Research*, **2**(1). <https://doi.org/10.52465/joiser.v2i1.245>.
- Olisah CC, Smith L, Smith M.(2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput Methods Programs Biomed*. **220**:106773.

- Onyema EM, Akindutire OR, Emelisana CE, Ani NC and Osijirin A (2022). Awareness and Perception of COVID-19 Vaccine among Computer Science Students in Higher Education in Southeast Nigeria. *Journal of Computer Science and its Application* **29** (1): 34-40.
- Onyema, E.M; Quadri, N.N; Alhuseen, O.A; Nwafor,C.E; Abdullahi, I. and Faluyi S.G. (2020). Development of a Mobile-Learning Platform for Entrepreneurship Education in Nigeria. *British Journal of Science* (BSJ), **18** (2):123-141.
- Onyema, E.M., Edeh, C. D., Gregory, U.S., Edmond, V.U., Charles, A.C. and Richard-Nnabu, N.E. (2021). Cybersecurity Awareness Among Undergraduate Students in Enugu Nigeria. *International Journal of Information Security, Privacy and Digital Forensic* , **5** (1): 34 -42.
- Rousyati R, A. N. Rais, E. Rahmawati, and R. F. Amir(2021). “Prediksi Pima Indians Diabetes Database Dengan Ensemble Adaboost Dan Bagging,” *EVOLUSI : Jurnal Sains dan Manajemen*, **9**(2): 36–42.
- Tahir F and Farhan M (2023) Exploring the progress of artificial intelligence in managing type 2 diabetes mellitus: a comprehensive review of present innovations and anticipated challenges ahead. *Front. Clin. Diabetes Healthc.* **4**:1316111.
- UCI Machine Learning. Pima Indians Diabetes Database.” [Online]. Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.
- Victor, U.E; Onyema EM, Osijirin, A.N. and Obasi, O (2022). Application of Innovative Technologies in Computer Science Education during Covid-19 School Closure in Enugu. *International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)*, **12** (43): 5129-5139.
- Saxena S., D. Mohapatra, S. Padhee, and G. K. Sahoo,(2021). “Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms,” *Evolutionary Intelligence*, no. 0123456789, 2021, doi: 10.1007/s12065-021-00685-9.
- WHO (2023) Diabetes. <https://www.who.int/news-room/fact-sheets/detail/diabetes>.
- Zou Q, Qu K, Luo Y, et al.(2018). Predicting diabetes mellitus with machine learning techniques. *Front Genet.* **9**:515.