IBSUniversity Journal of Business and Research
ISSN: 3007-3936 (Online)
Volume 2, Issue 1, June 2025

Diabetes disease prediction using machine learning

Jobin M. Scaria
IBSUniversity, Papua New Guinea
jobin.scaria@ibs.ac.pg

Harita Bhargava
VIT Bhopal University Bhopal, India
harita.23phd10001@vitbhopal.ac.in

Sharmila Joseph
VIT Bhopal University Bhopal, India
sharmila.joseph@vitbhopal.ac.in

Sasikumar V R
IBS University, Papua New Guinea
sasikumar.velamparampil@ibs.ac.pg

A. Baseera
VIT Bhopal University Bhopal, India
baseera.a@vitbhopal.ac.in

Abstract

Diabetes mellitus is a chronic disease affecting millions worldwide, and early predictions play a crucial role in preventive healthcare. This project aims to develop an efficient and user-friendly system for diabetes disease prediction using machine learning techniques. Leveraging the Random Forest Classifier from the scikit -learn library, the model was trained on the 2019 diabetes dataset after proper preprocessing with Pandas and Label Encoder. Model performance was evaluated using standard metrics like accuracy score after data splitting using train_test_split. To visualize model performance, Matplotlib was used to create comparative bar charts. The project integrates machine learning with a web-based interface to enhance accessibility and usability. The frontend is built using HTML5 and Jinja templates, dynamically rendering the welcome screen, questionnaire form, and result page. The backend is developed using Flask, a lightweight Python web framework, which manages routing, form submissions, and session handling. The trained ML model is serialized using Pickle, allowing seamless loading for real-time predictions. Additionally, FPDF is employed to generate downloadable PDF reports for users, summarizing their input, risk score, and relevant recommendations. This system not only demonstrates the practical application of ML in healthcare but also offers a complete pipeline from data input to result delivery. Future work includes expanding the dataset, integrating additional health indicators, and deploying the solution for mobile use or real clinical environments.

Keywords: Diabetes prediction, mellitus, jinja, Random Forest, flask, web application, questionnaire, healthcare AI

1. Introduction

Diabetes mellitus is a chronic metabolic disorder that impairs the body's ability to process blood glucose effectively. It is categorized mainly into Type 1 and Type 2 diabetes, with Type 2 being the most prevalent, often linked to lifestyle factors and genetics. The World Health Organization (WHO, 2016) reports that over 422 million people worldwide are living with diabetes, with a significant number remaining undiagnosed until serious complications arise. These complications may include cardiovascular diseases, kidney failure, nerve damage, and vision problems. The increasing global burden of diabetes calls for scalable, cost-effective, and reliable solutions to facilitate early detection and intervention (Flask documentation team, n.d; Naik, 2025).

Traditional diagnostic methods for diabetes typically involve clinical evaluations, fasting blood sugar tests, HbA1c tests, and oral glucose tolerance tests. While effective, these methods can be time-consuming, resource-intensive, and inaccessible in certain regions. In contrast, machine learning (ML) techniques offer a powerful alternative by leveraging historical health data to predict disease onset with high accuracy and efficiency. By automating the prediction process, ML models can aid healthcare professionals, reduce diagnostic delays, and empower individuals to assess their health proactively (Mustofa et al., 2019).

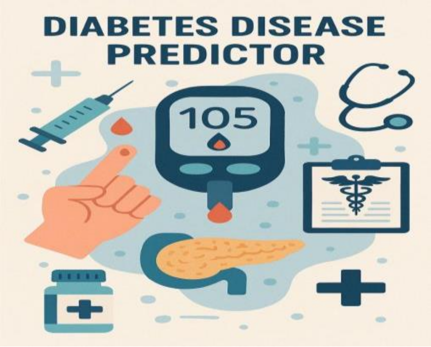


Figure 1: Diabetes disease predictor

This research project focuses on developing an ML-powered web application for predicting the likelihood of diabetes in individuals, using both medical and symptom-based data.

The system is designed to be user-friendly, accurate, and deployable in real-world settings, including mobile or clinical environments. The central algorithm used in this project is the Random Forest Classifier, a widely adopted ensemble learning method based on decision trees. Unlike single decision tree models, Random Forest builds multiple decision trees and merges their outputs through majority voting, leading to better generalization and stability in predictions.

The Random Forest algorithm is especially suitable for healthcare datasets because of its robustness to overfitting, its ability to manage both categorical and numerical data, and its high predictive performance on non-linear patterns (Mustofa et al., 2019). Medical datasets often involve complex interdependencies among features—such as glucose levels, insulin, BMI, and age—which Random Forest can effectively capture. Furthermore, it provides internal feature importance rankings, helping researchers identify which variables are most influential in disease prediction.

The dataset used for training the model includes the popular PIMA Indian Diabetes Dataset, known for its comprehensive medical feature set, including pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. To enhance the predictive capacity of the model, this dataset was merged with a 2019 Public Health Survey Dataset and additional regional data collected from Bangladesh. The enriched dataset introduces a more diverse sample population and a broader range of symptoms (Breiman, 2001).

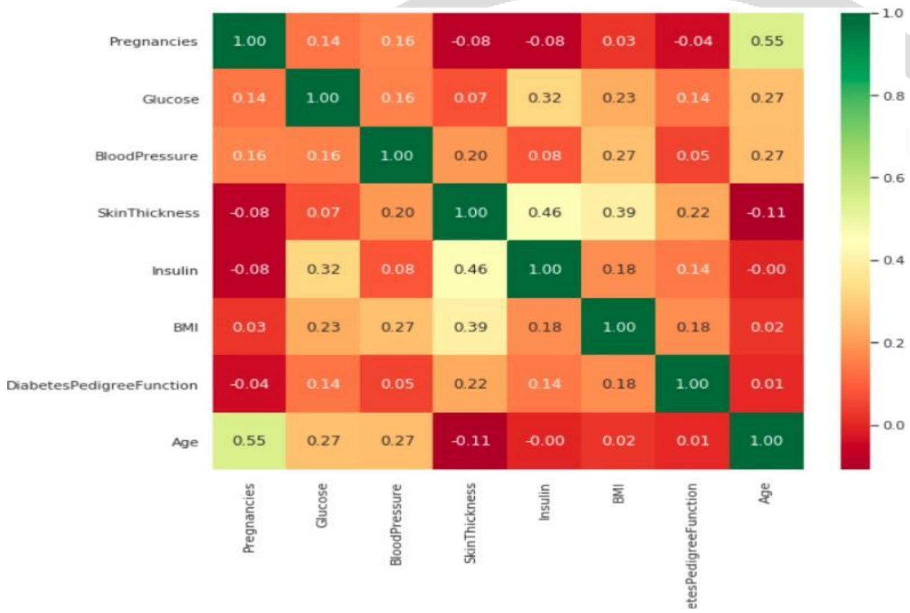


Figure 2: PIMA dataset
Source: Field data

In addition to clinical data, the system incorporates symptom-based features through a structured questionnaire (Kiran et al., 2025). Users are prompted to answer questions related to common diabetes symptoms, such as frequent urination, excessive thirst, fatigue, blurred vision, slow wound healing, and tingling sensations. Each reported symptom slightly increases the calculated risk score, making the model more reflective of real-world diagnostic considerations. For example, each symptom adds an additional 5% risk to the baseline probability predicted from the clinical features alone. This hybrid modeling approach combines quantitative inputs with qualitative observations to produce more personalized predictions (Pedregosa et al., 2012).

The web application is developed using the Flask framework in Python 3, chosen for its simplicity and flexibility in building lightweight web servers. The front end is built using HTML5 and Jinja templates, which enable the rendering of dynamic content such as forms and result pages. Once a user submits the form, the backend handles the input using Flask routes, encodes the data appropriately, and loads the pre-trained Random Forest model using Pickle. This ensures quick, real-time predictions without the need to retrain the model for every request (Python Software Foundation, n.d.)

The predicted outcome is presented clearly on the result page, along with a downloadable PDF report generated using the FPDF library. This report includes the user's input values, risk assessment, and general health recommendations. Such documentation not only improves user experience but also allows individuals to share the report with healthcare professionals for further consultation (PyFPDF Library, n.d.)

To evaluate model performance, the accuracy score metric from the scikit-learn library is used. The system provides a comparative analysis between the original model (trained only on the PIMA dataset) and the enhanced model (trained on the combined dataset), demonstrating the improvements in prediction accuracy and robustness. Additionally, matplotlib is used to generate a bar chart visualizing the accuracy of different ML models, aiding in the interpretation of results and model selection (Matplotlib Community, n.d.).

The overall goal of this project is to deliver a practical, end-to-end machine learning solution for diabetes prediction that is technically sound, clinically relevant, and accessible to users regardless of their technical background. By integrating data science with a functional web interface, this system demonstrates how AI and healthcare can intersect to create meaningful and impactful applications.

Future work includes testing the model with real-time patient data, expanding the dataset further with electronic health records (EHRs), enhancing symptom handling with weighted scoring, and integrating mobile accessibility to facilitate wider adoption. Ultimately, this project highlights the potential of ML in preventive healthcare and sets the stage for more advanced, AI-driven medical tools in the near future.

2.0 Methodology

2.1 Dataset collection

Two main sources were used: PIMA Indian Diabetes Dataset, which includes 768 samples with medical attributes such as glucose, BMI, blood pressure, and insulin. 2019 Bangladesh Survey adds diversity and symptoms data (e.g., thirst, fatigue) to enhance prediction reliability, which not only contributed to demographic diversity but also included symptom-based responses such as excessive thirst, fatigue, and frequent urination. The combination of these datasets enabled the creation of a more comprehensive and generalizable data foundation for the predictive model (Breiman, 2001).

Table 1: Comparison of PIMA and 2019 Public Health Survey datasets.

Feature	PIMA DATASET	Bangladesh survey 2019
No of samples	768	1000
Demographics	Native American women	Diverse (urban/ rural)
Attributes	Glucose, bmi, bp, etc.	Symptoms + medicals
Year	Pre-2000	2019
Source	UCI repository	Public health dept.

Source: Field data

2.2 Data preprocessing

Data cleaning included handling missing values, normalization, and encoding categorical fields using Label Encoder. The datasets were then merged, creating a hybrid dataset that supported both clinical and symptom-based features. This enriched dataset was used to train and test the prediction models.

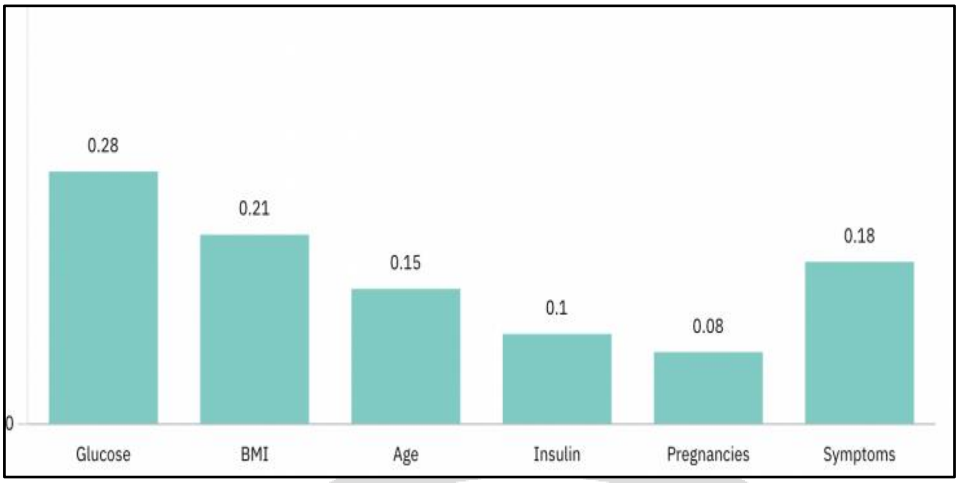
2.3 Model selection

The Random Forest Classifier was selected as the core machine learning algorithm for this research due to its reliability, versatility, and proven performance in classification tasks, especially within the healthcare domain (Mustofa et al., 2019) Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of their predictions for classification tasks. One of its key strengths lies in its high predictive accuracy, which stems from the use of multiple diverse trees that reduce variance and improve generalization. Compared to single decision tree models, Random Forest significantly reduces the risk of overfitting, particularly when a sufficient number of trees are used. Another advantage of Random Forest is its ability to seamlessly handle both categorical and numerical data, making it highly suitable for mixed datasets like the one used in this project, which combines medical indicators and user-reported symptoms. Furthermore, Random Forest models offer interpretability through the computation of feature importance scores, allowing researchers to identify which variables (such as glucose, BMI, or symptoms like fatigue) contribute most significantly to the prediction outcome. This interpretability is particularly valuable in medical applications, where understanding the factors driving a prediction is crucial for clinical relevance and user trust.

Table 2: Feature importance visualization from Random Forest model

Feature	Importance source
Glucose	0.28
BMI	0.21
Age	0.15
Insulin	0.10
Pregnancies	0.08
Symptoms	0.18(cumulative)

Source: Field data



. Figure 3: Representation of Table 2
Source: Field data

2.4 Training and testing

To evaluate the performance of the proposed diabetes prediction system, the dataset was partitioned into training and testing sets using an 80:20 split with the train_test_split function from the scikit-learn library. This approach ensures that the model is trained on a majority portion of the data while retaining a separate, unseen portion for performance validation. The selected evaluation metric was accuracy, computed using the accuracy score function, which measures the ratio of correctly predicted instances to the total number of instances.

Two models were developed and compared. The first, referred to as the Old Model, was trained exclusively on the PIMA Indian Diabetes dataset. It achieved an accuracy of approximately 75%, which is consistent with benchmarks in prior research using this dataset. The second model, termed the Enhanced Model, was trained on a combined dataset that merged the PIMA dataset with a 2019 public health survey dataset from Bangladesh. This expanded dataset introduced additional diversity and included symptom-based inputs, resulting in a model that achieved an improved accuracy of approximately 78%. This comparison clearly illustrates that incorporating symptom-level data and a more diverse population base contributes to better generalization and predictive performance.

2.5 Deployment architecture

The complete system was deployed as a lightweight and scalable web application designed to deliver predictions in real time. The backend was implemented using Python 3 and the Flask micro web framework. Flask was chosen for its simplicity, flexibility, and ability to handle RESTful routing, form submissions, and session management efficiently. On the frontend, the application uses HTML5 combined with Jinja2 templates, allowing for dynamic content rendering and multi-language support, thus improving accessibility and usability.

The trained machine learning model was saved and loaded using Python's Pickle module. This enabled the system to reuse the trained model during runtime without retraining, thereby reducing computational overhead and response time. Upon receiving user input, the system loads the serialized model, processes the data, and generates a prediction.

For output delivery, the FPDF library was integrated into the backend to automatically generate personalized PDF reports. These reports include the user's data, predicted diabetes risk score, and relevant health recommendations, making it a practical takeaway for users or healthcare professionals. Additionally, Matplotlib was used to visualize model performance, including accuracy comparisons and feature importance plots. These visual elements support transparency and interpretability, especially when presenting the system in educational or healthcare settings.

2.6 Ease of use

User accessibility and simplicity were core principles in the development of the system. The web interface is designed to be intuitive and supports multilingual functionality, allowing users to select between English, Spanish, French, and German. Upon launching the application, users are guided through a streamlined form that collects both clinical parameters and symptom data. This eliminates the need for technical knowledge or medical expertise to use the system.

Once the form is submitted, the backend processes the data and produces real-time predictions based on the trained machine learning model. The prediction is immediately displayed on the result page, and users are offered the option to download a personalized PDF report that includes their input details, risk percentage, and suggested next steps. This seamless end-to-end experience ensures that the application can be used efficiently on both desktop and mobile browsers. The entire process—from data input to result delivery—takes only a few seconds, making it practical for widespread use in non-clinical environments such as educational institutions, public health awareness camps, and community health screenings.

2.7 Questionnaire-based solution

In addition to standard clinical features obtained from datasets like PIMA and public health surveys, this project incorporates a symptom-based questionnaire to enhance the predictive capabilities of the system. This hybrid approach leverages user-reported symptoms alongside clinical metrics to offer a more holistic and personalized risk assessment. The questionnaire was carefully designed based on common early warning signs of diabetes, derived from public health guidelines and clinical research.

Users are prompted to respond to a set of seven symptom-related questions during the input process. These include the presence or absence of frequent urination, excessive thirst, increased hunger, fatigue, blurred vision, slow wound healing, and tingling or numbness in extremities. Each affirmative response to a symptom contributes a fixed 5% increase in the model's baseline risk prediction. This heuristic adjustment mechanism was developed to reflect real-world diagnostic practices, where symptoms often precede or supplement laboratory-confirmed indicators of disease.

The integration of symptom data serves two purposes. First, it provides a non-invasive and user-friendly method for risk screening, which is especially valuable in resource-limited or non-clinical settings. Second, it adds contextual intelligence to the machine learning model's prediction by accounting for observable, experiential health cues reported by users. This approach ensures that the system remains not only data-driven but also human-entered, adapting dynamically to user-specific inputs in real time.

By augmenting traditional ML outputs with symptom-based logic, the questionnaire-based solution significantly improves the model's practical applicability. It is particularly well-suited for community health screenings, educational awareness programs, and preliminary digital triage applications where access to clinical testing may be limited. The flexibility of this design allows for further enhancement through weighted symptom scoring, adaptive learning based on feedback, and potential integration with telemedicine systems in future iterations.

2.8 Model evaluation

The system achieved the following results:

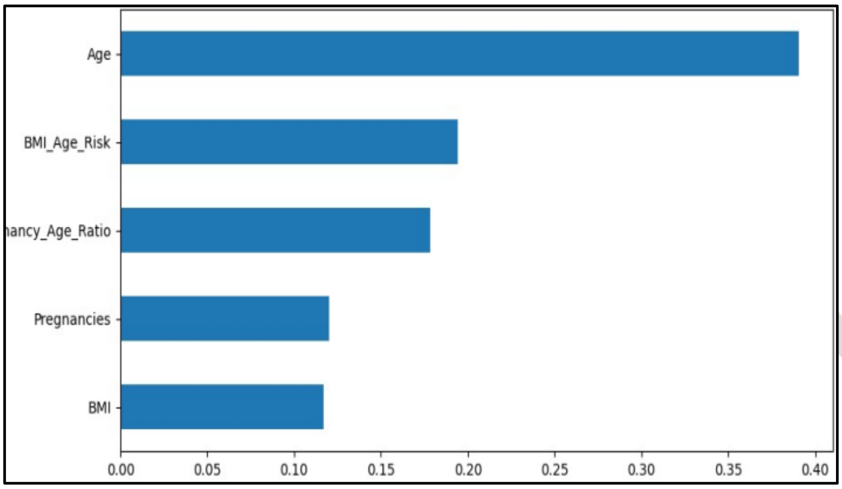


Figure 4: Criteria for calculating the disease occurrence
Source: Field data

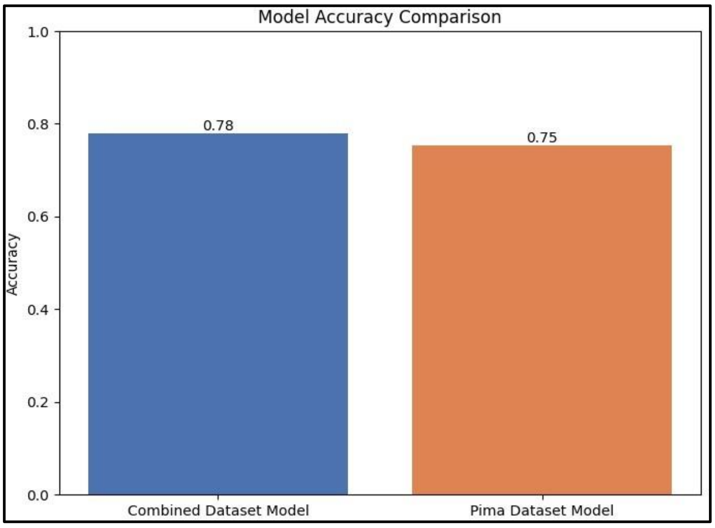


Figure 5: Comparison of model accuracy
Source: Field data

3. Common misconceptions about diabetes prediction

Despite significant medical advancements and public health awareness efforts, numerous misconceptions and false beliefs persist around diabetes and its diagnosis. These misconceptions often hinder early detection, promote misinformation, and create barriers to adopting technology-driven solutions such as machine learning (ML) for disease prediction.

This section addresses some of the most prevalent misconceptions and contrasts them with factual, evidence-based insights that validate the necessity and effectiveness of digital prediction models.

One of the most common misconceptions is the belief that diabetes can only be diagnosed through laboratory blood tests. While blood glucose testing remains a standard clinical method, it is not the only viable way to assess risk. Early symptoms like excessive thirst, frequent urination, and unexplained fatigue often precede clinical diagnosis and can serve as important predictors. Machine learning models trained on both clinical data and symptom-based inputs can effectively use these indicators to identify individuals at high risk, prompting them to seek medical advice sooner (Smith et al., 1988).

Another widespread myth is that diabetes only affects older individuals. Although age is a significant risk factor, recent trends show that Type 2 diabetes is increasingly prevalent among young adults and even adolescents, largely due to sedentary lifestyles, poor dietary habits, and genetic predisposition. Many people in younger age groups ignore early warning signs under the assumption that they are “too young” to develop diabetes. This highlights the importance of awareness tools that utilise digital platforms to promote self-screening and risk assessment regardless of age.

It is also commonly believed that a person with normal body weight is not at risk of diabetes. However, diabetes is not exclusively associated with obesity. Individuals with normal or even low body mass index (BMI) can develop diabetes due to other risk factors like insulin resistance, family history, or pancreas-related conditions. Machine learning models can incorporate multiple variables— beyond weight or appearance—to assess risk comprehensively, which human judgment might overlook.

Another misconception lies in over-reliance on family history. While genetics plays a role in diabetes risk, lifestyle and environmental factors are equally, if not more, important. People without any known family history may still develop diabetes due to high carbohydrate intake, low physical activity, or chronic stress. Conversely, having a family history does not guarantee the onset of diabetes if preventive measures are taken. Hence, relying solely on familial patterns can be misleading.

In addition to health-related myths, there are also technological misconceptions surrounding the use of artificial intelligence (AI) and machine learning in healthcare. A common belief is that machine learning models are too complex, unreliable, or “black box” in nature, making them unsuitable for medical use. On the contrary, algorithms like Random Forests are interpretable and capable of providing transparent insights, such as feature importance scores. When combined with explainable outputs and user-friendly interfaces, these models become accessible tools for both medical professionals and the public.

By addressing and dispelling these misconceptions, this research aims to promote trust in machine learning–based healthcare tools. The developed system not only enhances accessibility but also empowers users with data-driven awareness, bridging the gap between technology and traditional medical understanding (Polat & Güneş, 2007)

4. Common misconceptions about disease occurrence

The rise of chronic illnesses like diabetes has been accompanied by an equally concerning rise in misconceptions and myths—not only about the disease itself but also about modern approaches to its prevention and diagnosis. These misconceptions can lead to delayed diagnosis, improper management, and hesitation in using digital health technologies. As this project explores the integration of machine learning (ML) for diabetes prediction, it is crucial to address these myths and emphasize how data-driven solutions can help reshape public understanding.

4.1 Myths about who gets diabetes

A prevalent myth is that only older or overweight individuals are at risk of developing diabetes, particularly Type 2. While it is true that age and body weight are risk factors, they are not the only determinants. Increasingly, younger populations—including teenagers and young adults are being diagnosed with Type 2 diabetes due to sedentary lifestyles, poor dietary choices, and stress-related hormonal imbalances. Moreover, individuals with a normal Body Mass Index (BMI) may also develop diabetes due to genetic predisposition or insulin resistance. Hence, the assumption that "I’m young and slim, so I can't get diabetes" is both inaccurate and dangerous. Machine learning models challenge this myth by considering a broad range of input features—not just weight or age, but also lifestyle factors, medical history, and symptoms. In doing so, these models enable early screening and detection across a more diverse population base than traditional methods typically target (Ligthart et al., 2021).

4.2 Misconceptions around diagnosis

Another common belief is that diabetes can only be diagnosed through blood tests. While clinical tests like fasting blood glucose or HbA1c are standard diagnostic tools, many early symptoms—such as increased thirst, fatigue, frequent urination, and slow wound healing—can indicate the onset of diabetes before it reaches clinical thresholds. Unfortunately, many people ignore these signs or wait until symptoms worsen, which increases the risk of complications.

The system presented in this research addresses this gap by including a symptom-based questionnaire, enabling preliminary assessment without requiring laboratory access. By adjusting ML predictions based on user-reported symptoms, the model provides a real-time risk analysis, empowering users to seek professional diagnosis sooner (Firdous et al., 2021).

4.3 Genetics and lifestyle confusion

A further myth revolves around genetics: some individuals believe that only those with a family history of diabetes can develop it, while others with diabetic relatives assume that developing the disease is inevitable. Both assumptions are flawed. Although family history is a risk factor, it is not the sole determinant. Lifestyle, stress, diet, physical activity, and sleep patterns play crucial roles. Conversely, people with diabetic parents who maintain a healthy lifestyle can often delay or prevent the disease altogether. ML models balance these factors more objectively. In contrast to human biases, the algorithm gives weight to various inputs, enabling fair prediction regardless of family history. This helps break the myth that genes are destiny and emphasizes that preventive action matters.

4.4 Cultural misunderstandings and delay in medical attention

In certain cultural or rural communities, diabetes is believed to be a temporary or reversible condition caused by stress or lifestyle imbalances alone, leading many to delay proper treatment or rely solely on home remedies. Such beliefs can be detrimental, as untreated diabetes can progress silently and cause irreversible damage to organs. Another problem is the stigma around being diagnosed with diabetes, especially among young adults, leading to denial or concealment.

Technology-based tools like the one proposed in this project offer privacy, ease, and early detection, helping break the stigma barrier. Individuals can assess their risk in the comfort of their homes, without facing judgment or public exposure, which encourages proactive health monitoring (Eitel et al., 2024).

4.5 Misconceptions around technology and machine learning

There is also a significant distrust or misunderstanding around AI and ML in healthcare. Many users assume that ML models are opaque, difficult to trust, and prone to making random predictions. Others believe such systems are only suitable for tech-savvy users or require high-end computing infrastructure.

The model implemented in this project—Random Forest Classifier—is not only accurate and robust but also provides interpretable results, such as feature importance rankings. Furthermore, the user interface built using Flask and HTML5 is designed to be accessible to non-technical users, offering a seamless prediction process along with downloadable PDF reports. The application runs efficiently on standard browsers and mobile devices, debunking the myth that AI in healthcare is only for experts or large institutions.

Additionally, some believe that ML-based health predictions are less reliable than doctor consultations. While it is true that ML cannot replace clinical diagnosis, its purpose is to augment early screening, not replace professional healthcare. It is best viewed as a preliminary assessment tool that flags potential risks, prompting users to consult healthcare providers when necessary (Deniz-Garcia et al., 2023).

4.6 Misuse of "symptoms = diagnosis" logic

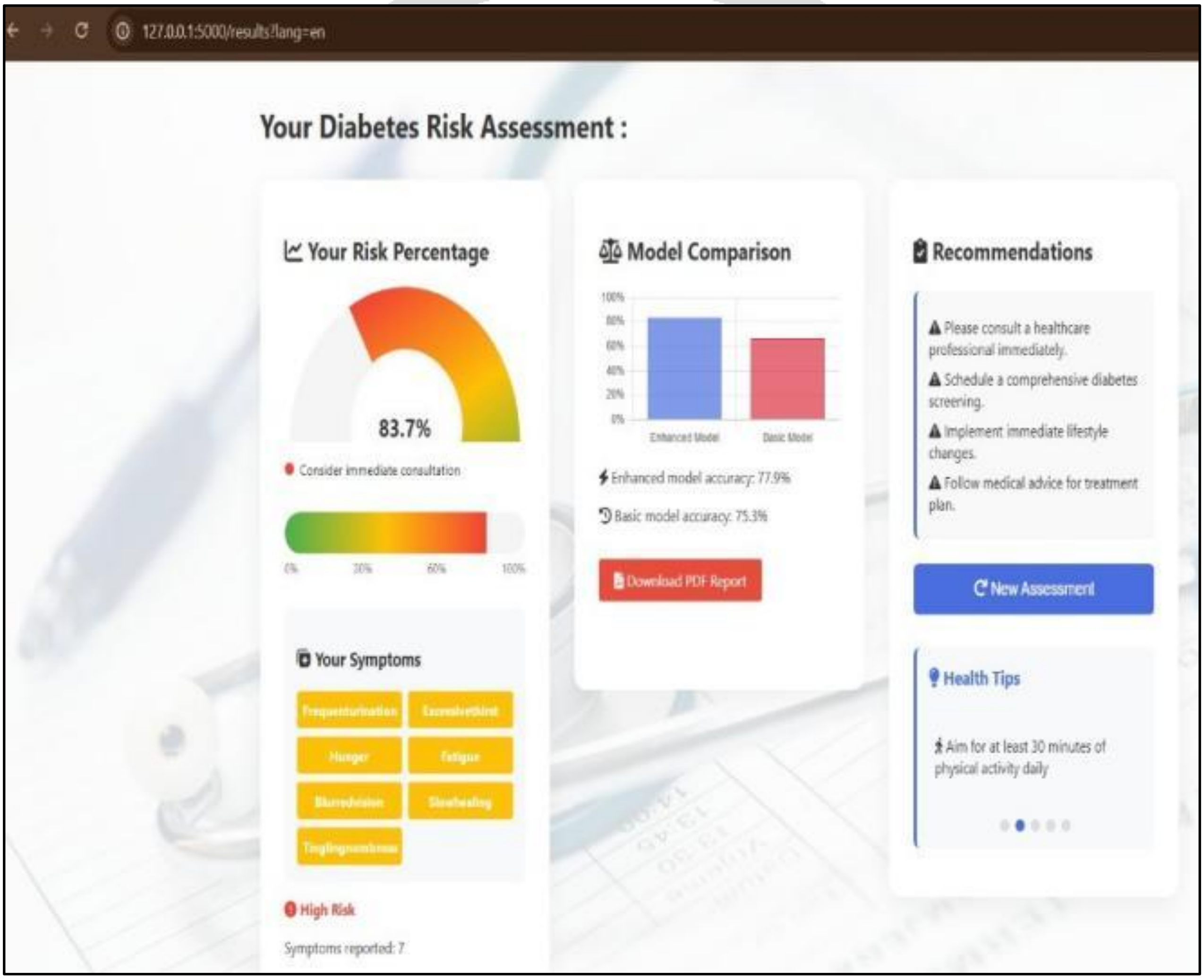
A subtle but harmful misconception is the belief that having one or two symptoms means one has diabetes. While symptoms like fatigue or frequent urination can be signs, they can also be caused by other conditions. Therefore, self-diagnosis based solely on symptoms is risky. ML models, especially those trained on diverse datasets, reduce such risks by considering combinations of features and patterns learned from hundreds of patient cases.

This research presents a comprehensive, machine learning– based system for predicting the risk of diabetes using a hybrid approach that combines clinical data with user- reported symptoms. By utilizing a Random Forest Classifier trained on an enhanced dataset—comprising the PIMA Indian Diabetes dataset and a 2019 public health survey from Bangladesh—the model achieved improved accuracy and generalizability compared to traditional methods. The system is deployed through a lightweight, multilingual web application built with Flask, providing real-time predictions and personalized PDF reports without requiring any medical expertise from the user (Wee et al., 2024)

The integration of a symptom-based questionnaire enhances the model's practical applicability by allowing early screening in both clinical and non-clinical environments. The system's ease of use, rapid performance, and low-resource requirements make it suitable for widespread adoption, particularly in underserved or remote areas.

In addition to technical accuracy, this work addresses common societal misconceptions about diabetes and its prediction, reinforcing the role of artificial intelligence in preventive healthcare. The tool serves as a bridge between modern data science and accessible health technology, encouraging proactive health monitoring and awareness. Future work will include further dataset expansion, integration with mobile platforms, and improved explainability features to increase adoption in real-world healthcare systems.

Here is the outcome we got from our experience on this topic:



127.0.0.1:5000/results?lang=en

Your Diabetes Risk Assessment :

Your Risk Percentage

83.7%

Consider immediate consultation

Your Symptoms

High Risk

Symptoms reported: 7

Model Comparison

Enhanced model accuracy: 77.9%

Basic model accuracy: 75.3%

Recommendations

Please consult a healthcare professional immediately.

Schedule a comprehensive diabetes screening.

Implement immediate lifestyle changes.

Follow medical advice for treatment plan.

Health Tips

Aim for at least 30 minutes of physical activity daily

Figure 6: The result of our projects

The system estimated an 83.7% risk of diabetes for the user, suggesting immediate medical attention. Seven symptoms were reported, including fatigue, frequent urination, and excessive thirst. The enhanced prediction model performed better, achieving 77.9% accuracy compared to 73.5% from the basic model. Personalized recommendations were provided for medical follow-up and lifestyle changes. Health tips, such as engaging in daily physical activity, were also included. Users can download a PDF report or begin a new assessment for further screening.

5. Conclusion

The diabetes disease predictor system demonstrates the potential of using machine learning to enhance early detection and management of Diabetes. Despite some limitations, such as data availability and model interpretability, the system offers promising results. Future enhancements, including data augmentation, ensemble learning, and real-time monitoring, will further improve the system's capabilities, contributing to more effective diabetes management and better patient outcomes.

References

Breiman, L. 2001. Random forests. *Machine Learning,* 45(1):5-32. https://doi.org/10.1023/A:1010933404324

Deniz-Garcia, A., Fabelo H., Rodriguez-Almeida, A. J., Zamora-Zamorano, G., Castro-Fernandez, M., Ruano, M. A., Solvoll, T., Granja, C., Schopf ,T.R ., Callico, G. M., Soguero-Ruiz, C., Wägner, A. M., & WARIFA Consortium. 2023. Quality, Usability and effectiveness of mHealth apps and the role of artificial intelligence: Current scenario and challenges. *Journal of Medical Internet Research* 4: 25.

Eitel, K. B., Pihoker, C., Barrett, C. E., & Roberts, A.J. 2024. Diabetes stigma and clinical outcomes: An international review. *Journal of Endocrine Society,* 8(9).

Firdous, S., Wagai, G. A., & Sharma, K. 2021. A survey on diabetes risk prediction using machine learning approaches. *Journal of family and medicine and primary care,* 11(11): 6929-6934.

Google Images. n.d. Diagram used for deployment architecture. [Online]. Available: https://images.app.goo.gl/ xQN1Zsua7Yfs8uEk9

Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscinek, B., & Russell, D. (2025). Machine learning and artificial intelligence in type 2 diabetes prediction: a 33-year bibliometric analysis. *Frontiers in Digital Health,* 7. https://doi.org/10.3389/fdgth.2025.1557467

Matplotlib Community. n.d. Matplotlib: Visualization with python. [Online]. Available: https://matplotlib.org/

Mustofa, F., Safriandono, A.N., Muslikh, A.R., Moses, D.I. 2023. Dataset and feature analysis for diabetes mellitus classification using random forest. *Journal of Computing Theories and Applications,* 1(1).

Naik, P. 2025. *Unplugging flask for crafting web apps with python's cleanest framework: Flask fusion with templates, routes and logic in one powerful stack.* Shaswat Publication.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel ,V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G .2012. SciKit-learn: Machine learning in python. *Journal of Machine Learning Research, arXiv, 12*: 2825-2830.

Polat, K., & Güneş, S. 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing,* 17(4): 702–710.

PyFPDF Library. n. d. FPDF: PDF generation in python. [Online]. Available: https://pyfpdf.github.io/fpdf2/

Python Software Foundation. Pickle. n.d. Python object serialisation. [Online]. Available: https://docs.python.org/ 3/library/pickle.html

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Johns Hopkins Applied Physics Laboratory Technical Digest, 10.*

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC2245318

Ligthart, S., Hasbani, N. R., Ahmadizar, F., van Herpt, T.T.W., Leening, M.J.G., Uitterlinden, A.G., Sijbrands, E.J.G., Morrison, A.C., Boerwinkle, E., Pankow, J.S., Selvin, E., Ikram, M.A., Kavousi, M., de Vries, P.S., & Dehghan, A. 2021. Genetic susceptibility, obesity and lifetime risk of type 2 diabetes: The ARIC study and Rotterdam Study. *Diabetic Medicine,* 38 (10).

Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W.K., & Juwono, F.H. 2024. Diabetes detection using machine learning and deep learning approaches. *Multimedia Tools and Applications*, 83, 24153–24185. https://doi.org/10.1007/s11042-023-16407-5