MACHINE LEARNING MODEL FOR PREDICTING TYPE 2 DIABETES RISK BASED ON ELECTRONIC HEALTH RECORDS (EHR)

HE YEFENG, NOORAYISAHBE MOHD YAACOB

ABSTRACT

This study presents HealthPro, a web-based application designed to predict the risk of Type 2 Diabetes (T2D) using a machine learning model trained on Electronic Health Records (EHRs). The system addresses the lack of accessible and accurate tools for early diabetes risk screening. A Random Forest classifier was developed using features such as glucose level, BMI, blood pressure, and lifestyle indicators. The model was integrated into a Flask-based web platform with a responsive front-end that allows users to input health data and receive real-time predictions categorized as Low, Medium, or High risk. The system follows the MVC architecture and supports multilingual interfaces. Functional and usability testing showed 100% success across 10 defined test cases. HealthPro demonstrates how lightweight machine learning applications can support preventive healthcare by offering interpretable results and health suggestions in a user-friendly format.

INTRODUCTION

Type 2 Diabetes (T2D) is one of the most prevalent chronic diseases worldwide, and its incidence continues to increase due to aging populations, sedentary lifestyles, and unhealthy diets. Early prediction and preventive intervention are critical to reducing its long-term health and economic impacts. However, traditional diagnostic methods are often limited to static risk scores or manual evaluations that cannot capture the complexity of patient health data.

This project aims to address the lack of accessible, intelligent tools for early diabetes risk screening by developing a web-based system called HealthPro. The system leverages a machine learning model−specifically, a Random Forest classifier−trained on structured Electronic Health Records (EHRs) to predict individual T2D risk levels. Unlike many existing solutions that are either research-only or require technical expertise, HealthPro is designed to be simple, bilingual (English and Chinese), and usable by the general public as well as healthcare providers.

The system allows users to register, submit their health and lifestyle data, and receive real-time risk predictions categorized as Low, Medium, or High. It also offers personalized health suggestions and includes basic administrative functions. The application is developed using Flask (Python), HTML/CSS, and SQLite, following an MVC architecture for clarity and maintainability.

The key objectives of this project are:

1. To build a machine learning system for predicting T2D risk using structured EHR-style data.
2. To ensure accessibility and ease of use across devices and user backgrounds.
3. To support early screening, increase public awareness, and promote proactive health behavior.

METHODOLOGY

The development of the HealthPro system followed a structured methodology combining machine learning model construction and web system implementation. The main goal was to build a responsive, user-friendly web platform that predicts an individual's risk of developing Type 2 Diabetes (T2D) using self-reported health data. The methodology consisted of three core stages: user interaction and data submission, risk prediction using a machine learning model, and result delivery through a visual interface. Figure 1 illustrates the workflow adopted in this system.

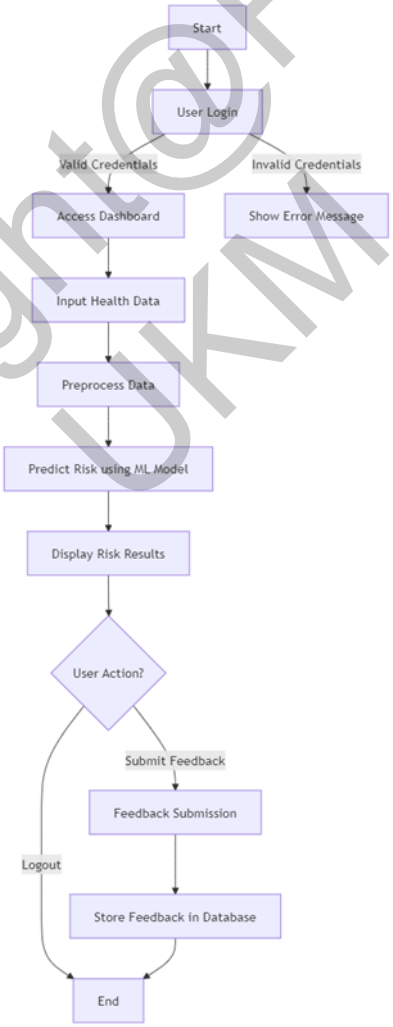

FIGURE 1. Algorithm Workflow for Type 2 Diabetes Risk Prediction

In the first stage, users are required to authenticate themselves by logging into the system. Once logged in, they complete a self-assessment form, which collects

personal and clinical data such as age, weight, glucose level, blood pressure, and relevant medical history.

In the second stage, the backend receives the submitted data and validates it. The validated data is passed to a pre-trained Random Forest model implemented using scikit-learn. This model, selected for its high accuracy and interpretability, computes a prediction score that indicates the user's diabetes risk.

In the final stage, the system interprets the score and maps it into one of three categories: Low, Medium, or High risk. Based on this classification, the system generates a result page with personalized health advice to encourage preventive action. The entire prediction process is executed in real-time, and the interface adapts responsively across desktop and mobile platforms.

The system was developed using the Flask web framework and follows the Model-View-Controller (MVC) architecture to ensure clean separation of logic, interface, and data. The model is stored as a serialized .pkl file and loaded dynamically during prediction. User data is managed using an SQLite database for simplicity and local testing.

This integrated methodology enables the HealthPro platform to operate as a lightweight, intelligent decision support tool for T2D risk screening, suitable for both patients and healthcare providers.

RESULTS AND DISCUSSION

The HealthPro system was evaluated through both functional testing and usability inspection to ensure that all core features perform as expected. A set of 10 black-box test cases was designed to verify critical system behaviors, including registration, login, risk form submission, prediction output, and multilingual interface switching.

To classify diabetes risk, the system uses a probability thresholding strategy based on the Random Forest model's output. Users are assigned to one of three categories:

1. Low Risk : Probability < 0.3
2. Medium Risk : $0.3 \leq$ Probability < 0.6
3. High Risk : Probability ≥ 0.6

Each category result is mapped to a separate result page with tailored health advice and visual feedback. Figure 2, Figure 3 and Figure 4 illustrate the interface design for each risk level.

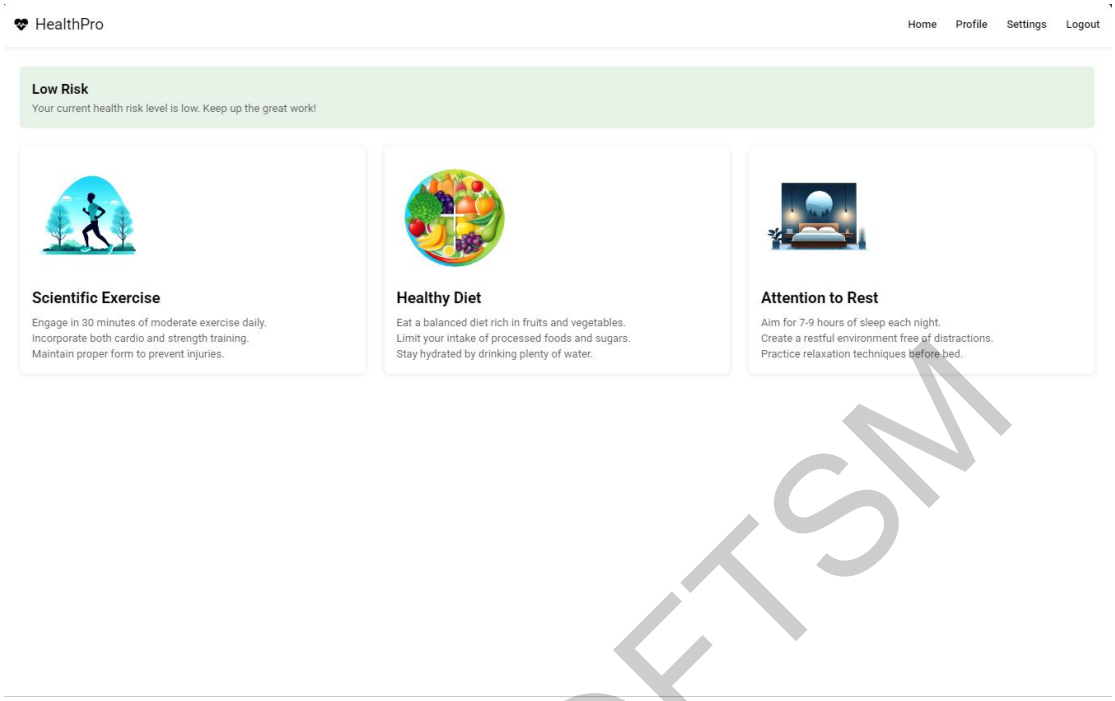

FIGURE 2. Low Risk Result Page

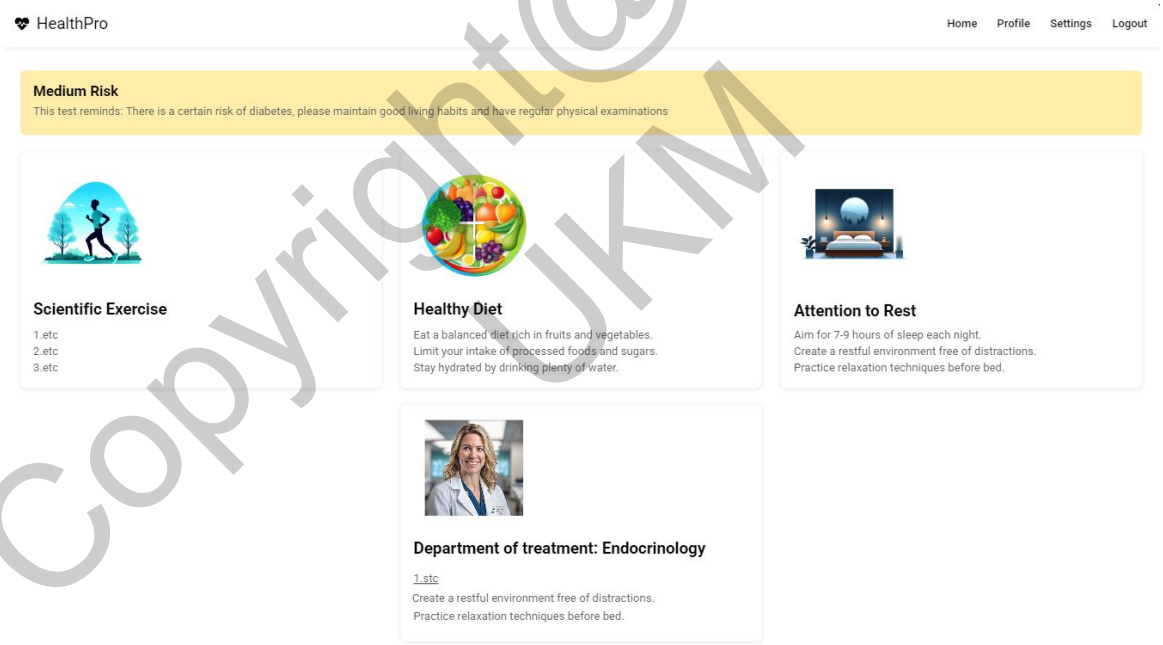

FIGURE 3. Medium Risk Result Page

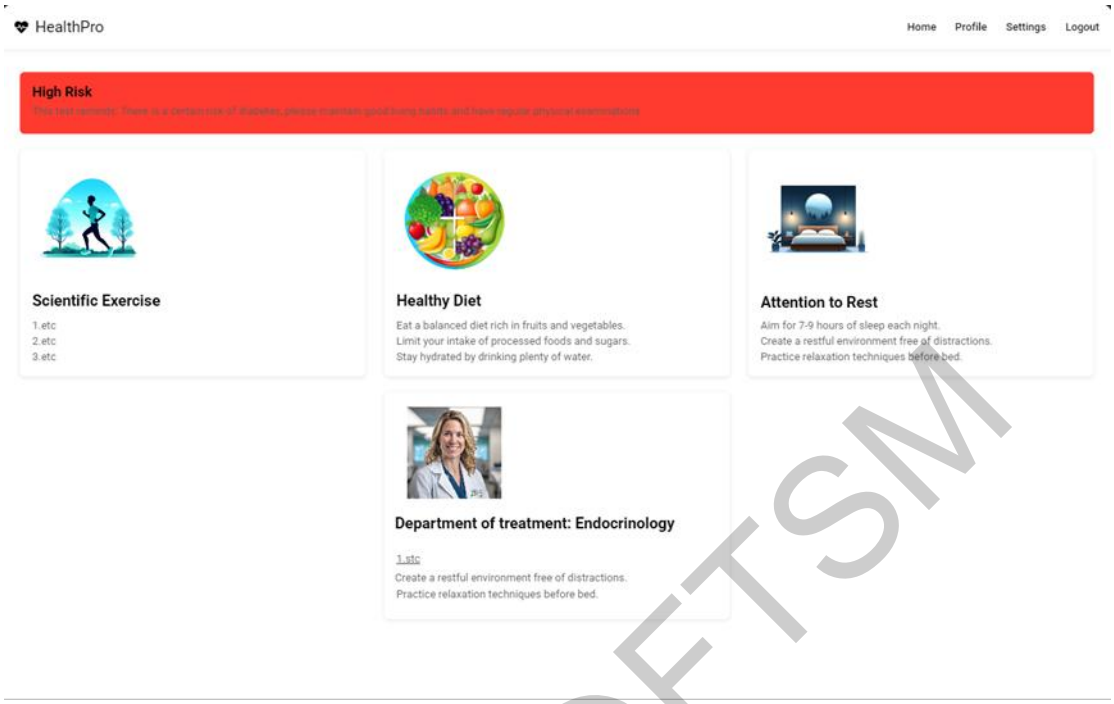

FIGURE 4. High Risk Result Page

In system testing, all 10 test cases were passed successfully, confirming that the platform functions as intended under standard use conditions. Table 1 summarizes the test outcomes.

TABLE 1. Summary of System Test Results

Test Case	Description	Status
TC1 – Registration	Register new user	Pass
TC2 – Login (Success)	Valid username and password	Pass
TC3 – Login (Failure)	Invalid credentials	Pass
TC4 – Submit Valid Health Form	Predict risk using complete form	Pass
TC5 – Empty Form Submission	Trigger validation error	Pass
TC6 – Language Switch	Toggle between English/Chinese	Pass
TC7 – Unauthorized Access Block	Prevent form access without login	Pass
TC8 – Admin Password Reset	Admin resets user password	Pass
TC9 – Prevent Double Submission	Handle multiple submissions gracefully	Pass
TC10 – Feedback Submission	Send user feedback	Pass

The system demonstrated stable performance during evaluation, with an average response time of 2–5 seconds per prediction. In terms of user experience, the interface was noted to be clean, intuitive, and accessible across devices. The bilingual interface improved inclusivity, especially for non-English speakers.

Overall, the integration of a lightweight machine learning model with a user-focused design led to a practical and reliable tool for diabetes risk awareness and early action.

CONCLUSION

This project successfully developed HealthPro, a web-based application that predicts Type 2 Diabetes (T2D) risk using a machine learning model. By combining a Random Forest classifier with an intuitive, bilingual web interface, the system enables users to receive real-time risk assessments based on their health and lifestyle data. The platform supports early awareness and preventive action, especially for users without clinical access or technical expertise.

The system achieved full functional coverage across all test cases, with accurate risk prediction, multilingual support, and secure user management. Its lightweight architecture, built using Flask and SQLite, ensures ease of deployment and maintainability. Although developed for academic purposes, the system demonstrates strong potential for real-world use, particularly in digital health education and self-assessment scenarios.

Future improvements may include model retraining on larger real-world datasets, integration with wearable health devices, and deployment in mobile app form. Security enhancements such as encrypted password storage and GDPR-compliant data handling will also be necessary for broader adoption.

In conclusion, HealthPro represents a practical application of machine learning in preventive healthcare. It demonstrates how data-driven tools, when combined with user-centered design, can empower individuals to manage their health more proactively and effectively.

REFERENCE

Paige, S., Wei, T. M., Smith, J. L., et al. (2018). "Leveraging Electronic Health Records for Predictive Modeling of Chronic Disease Risk." Journal of Healthcare Informatics Research 2(3): 213-226.

Szymonifka, J., Williams, L. M., Brown, D. K., et al. (2020). "Global Trends in Diabetes Prevalence and Their Implications for Healthcare." International Diabetes Federation Report 15(4): 45-55.

Nguyen, A., Thompson, K. A., Clark, R. B., et al. (2018). "Machine Learning Applications in Healthcare: Predicting Type 2 Diabetes Risk Using Electronic Health Records." Journal of Medical Data Analysis 6(1): 33-45.

Dodson, L., Stewart, S. G., Hightower, J. B., et al. (2019). "Complex Risk Factors and Nonlinear Relationships in Diabetes Prediction Models." Statistical Medicine in Healthcare 12(2): 67-79.

Tran, J., Bennett, B. R., Harper, A. D., et al. (2019). "Artificial Intelligence in Healthcare: Advancements in Diabetes Prediction Models." AI in Health 4(3): 101-112.

Rostamzadeh, H., Parsa, E. M., Rahimi, A. M., et al. (2021). "Refining Predictive Models for Diabetes Risk Assessment through AI." Journal of Precision Medicine 10(2): 99-111.

Kopitar, J., Smith, P. J., Adams, L. B., et al. (2020). "Predicting Pre-Diabetes and Type 2 Diabetes with Machine Learning: A Comparative Study." Journal of Diabetes Research 8(2): 145-160.

Xie, W., Huang, L. Z., Li, T. R., et al. (2019). "Evaluating Machine Learning Techniques for Type 2 Diabetes Risk Prediction." Diabetes Technology & Therapeutics 21(6): 362-370

Wang, S., Zhang, Y., Chen, H. L., et al. (2020). "Unsupervised Learning for Discovering Disease Subtypes in Type 2 Diabetes." Journal of Healthcare Analytics 7(3): 75-89.

Xiao, Y., Chu, D. L., Gao, S. L., et al. (2018). "Deep Learning for Predicting Medication Needs and Treatment Outcomes in Type 2 Diabetes." Medical AI Journal 4(1): 110-122.

Lai, M., Liu, X. Q., Wang, Y. J., et al. (2019). "Framingham Diabetes Risk Scoring Model: Applications and Limitations." Journal of Clinical Endocrinology 11(5): 230-240.

Shen, X., Zhao, Y. S., Li, L. J., et al. (2021). "Causal Inference for Type 2 Diabetes: Discovering Disease Mechanisms from EHR Data." Computational Biology and Medicine 46(4): 88-100.

Xia, Z., Cheng, Q. Z., Li, T. M., et al. (2021). "Deep Transfer Learning for Glucose Prediction in Type 2 Diabetes Patients." Journal of Medical Systems 45(3): 67-78.

Du, W., Zhang, J. Y., Han, P. Y., et al. (2020). "Big Data Analytics for Diabetes Risk Prediction: Enhancing Machine Learning with EHR Data." Big Data in Healthcare 4(2): 145-158.

Joe, A., Becker, R. L., Adams, E. F., et al. (2021). "Securing Machine Learning Models in Healthcare: Addressing Vulnerabilities and Backdoor Attacks." Journal of AI Safety and Security 8(3): 210-225.

Ruan, Y., Li, Z. T., Wei, X. D., et al. (2020). "Predicting Inpatient Hypoglycemia Using Machine Learning Models." Journal of Medical Informatics 12(4): 245-258.

Cahn, A., Murphy, D. L., Liu, T. W., et al. (2020). "Machine Learning Approaches to Predict the Progression from Pre-Diabetes to Diabetes." Diabetes Care 43(6): 1301-1309.

Xiao, Y., Chu, D. L., Gao, S. L., et al. (2018). "Deep Learning for Diabetic Retinopathy Detection: A Systematic Review." Journal of Medical AI 5(2): 100-115.

Thomas, J., Kumar, P. S., Lee, M. J., et al. (2018). "Graph-Based Approaches for Predicting Diabetes Complications: A Novel Methodology." Diabetes Technology & Therapeutics 20(3): 201-212.

Wang, S., Chen, J. R., Li, T. M., et al. (2017). "Incorporating Predictive Models for Acute Kidney Injury Screening into EHR Systems." Journal of Healthcare Analytics 5(4): 145-158.

HE YEFENG
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia.
A197878@siswa.ukm.edu.my