

Implementation of the Random Forest Algorithm for Website-Based Diabetes Mellitus Risk Prediction

Muhammad Rizky Ardiansyah¹ and Supriyadi^{2*}

^{1,2*} Universitas Nusa Mandiri, Indonesia

MEDINFTEch is licensed under a Creative Commons 4.0 International License.



ARTICLE HISTORY

Received: 31 July 26
Final Revision: 14 September 26
Accepted: 16 September 26
Online Publication: 30 September 26

KEYWORDS

Diabetes Mellitus, DiAnalyze, Machine Learning, Random Forest, Website

CORRESPONDING AUTHOR

supriyadi.spy@nusamandiri.ac.id

DOI

10.37034/medinftech.v4i3.168

ABSTRACT

Diabetes mellitus is a chronic health condition that continues to increase worldwide and may lead to various severe complications when early risk identification is not performed. This study addresses the limited implementation of machine learning models in practical web-based applications that can assist users in analyzing diabetes risk based on health-related parameters. Therefore, this research aims to develop DiAnalyze, a web-based diabetes risk prediction platform utilizing the Random Forest algorithm. The study used the Diabetes Prediction Dataset obtained from Kaggle, initially containing 100,000 records with eight input features and one target variable. After data preprocessing and class balancing using undersampling, 18,482 records were retained and divided into training, validation, and testing sets using a 64:16:20 ratio. The Random Forest model achieved an accuracy of 90.70%, precision of 89.09%, recall of 90.97%, F1-score of 90.02%, and ROC-AUC of 97.61%. The trained model was successfully integrated into the DiAnalyze web-based platform, enabling users to obtain diabetes risk classifications and corresponding risk probabilities based on the provided health parameters. DiAnalyze is intended as a supporting tool for diabetes risk assessment and does not replace professional medical diagnosis.

1. Introduction

Diabetes mellitus has become one of the major global health concerns due to the continuous increase in the number of affected individuals [1]. This disease is associated with impaired glucose regulation in the human body and may contribute to several long-term complications, including cardiovascular disorders, nerve damage, eye-related disorders, and kidney problems. Previous studies indicate that diabetes occurrence is influenced by various factors, such as unhealthy lifestyle patterns, obesity conditions, and insufficient early health monitoring [2]. Therefore, identifying diabetes risk at an earlier stage is an important step to reduce the potential impact caused by this disease.

The advancement of information technology, particularly in artificial intelligence and machine learning, provides new opportunities for processing and analyzing healthcare data. Machine learning approaches have been widely applied to develop

predictive models by discovering relationships between medical attributes and disease conditions. Several algorithms, including Random Forest, Logistic Regression, and Decision Tree, have been utilized for diabetes classification based on clinical parameters [3]. Among these approaches, Random Forest is considered an effective classification method because it combines multiple decision trees to improve prediction stability and reduce the possibility of overfitting.

Despite the increasing application of machine learning in healthcare prediction, several challenges remain in implementing diabetes prediction systems. Existing studies often focus primarily on developing classification models without providing an accessible system that can be directly utilized by users [4]. In addition, healthcare datasets frequently contain complex characteristics, such as imbalanced class distributions and various clinical attributes, which may influence model performance. Furthermore, several important parameters, including HbA1c levels and blood glucose measurements, require medical

examination and are not always easily accessible by general users.

Previous research has demonstrated that machine learning algorithms can provide promising results for diabetes prediction. Safitri et al. [2] implemented machine learning techniques for diabetes prediction and showed that data-driven approaches can assist classification processes. Syahri et al. [3] conducted a comparative evaluation between Logistic Regression, Random Forest, and AdaBoost algorithms, demonstrating differences in classification performance among these methods. Furthermore, Anggreini et al. [5] improved Random Forest performance through hyperparameter tuning techniques, showing that optimization of model parameters can enhance prediction capability.

Based on these considerations, this research proposes the development of a web-based diabetes risk prediction system named DiAnalyze. The system integrates a trained Random Forest model into a website application to provide users with diabetes risk analysis based on health parameters. Unlike studies that only focus on algorithm evaluation, this research emphasizes the implementation of machine learning models into an accessible application platform.

The objectives of this research are to develop a diabetes risk prediction model using the Random Forest algorithm, evaluate model performance using accuracy, precision, recall, F1-score, and ROC-AUC metrics, and implement the developed model into a web-based application. The proposed system is intended to function as a decision support tool that assists users in understanding potential diabetes risk based on available health information rather than replacing professional medical diagnosis.

2. Research Method

The dataset utilized in this study is the Diabetes Prediction Dataset obtained from the Kaggle platform. This dataset contains health-related attributes that represent several factors associated with diabetes risk and is used as the primary data source for developing the classification model.

The dataset contains 100,000 records with nine attributes. Eight attributes are used as input features, while one attribute is assigned as the prediction target. The input variables consist of gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level. Meanwhile, the target attribute is diabetes, which represents the classification label where value 0 indicates non-diabetes and value 1 indicates diabetes.

The utilization of healthcare datasets in machine learning enables the identification of patterns between patient characteristics and disease risks. Previous studies have shown that machine learning approaches

can support diabetes prediction by analyzing relationships between medical features and classification outcomes [6], [7].

2.2. Data Preprocessing

Data preprocessing is performed to prepare the dataset before the training process. This stage aims to improve data quality, transform data into a suitable format, and minimize potential problems that may affect model performance. Data preparation is an important stage in machine learning because raw datasets commonly require cleaning, transformation, and normalization processes before being used for model training [8].

The preprocessing stages performed in this research include:

2.2.1. Missing Value Checking

The dataset was examined to identify missing values in each attribute. Based on the analysis results, all attributes contained complete data without missing values. Therefore, no missing value handling technique was required.

2.2.2. Duplicate Data Removal

Duplicate data checking was performed to eliminate repeated records that could affect the learning process. The initial dataset contained 3,854 duplicated records, which were removed. After the removal process, the dataset consisted of 96,146 records.

2.2.3. Categorical Feature Encoding

Since the Random Forest algorithm requires numerical input values, categorical attributes were converted into numerical representations using the Label Encoding technique. The categorical attributes processed in this stage were Gender and Smoking History. The encoding process allows categorical information to be processed by machine learning algorithms while maintaining the representation of each category [9].

2.2.4. Dataset Balancing

The original dataset experienced an imbalance problem because the number of non-diabetes samples was significantly higher than diabetes samples. This condition may cause classification models to become biased toward the majority class.

To overcome this issue, an undersampling technique was applied to reduce the number of majority-class samples. After the balancing process, the dataset consisted of 10,000 non-diabetes records and 8,482 diabetes records.

Handling class imbalance is important in classification problems because an imbalanced dataset may cause the model to favor the majority class and reduce prediction performance on minority classes [10].

2.2.5. Dataset Splitting

After preprocessing and balancing, the dataset was divided into three subsets: 11,828 records (64%) for training, 2,957 records (16%) for validation, and 3,697 records (20%) for testing. The training dataset was used to build the classification model, the validation dataset was utilized to observe model performance during development, and the testing dataset was used for final evaluation using unseen data.

2.3. Training Model

The classification model developed in this research uses the Random Forest Classifier algorithm. Random Forest is an ensemble learning method that combines multiple decision trees to generate predictions through majority voting.

Random Forest has been widely applied in classification problems because it can handle complex datasets, manage multiple features, and provide stable prediction performance [11]. Previous research has demonstrated that Random Forest can effectively classify diabetes risk based on health-related attributes and achieve good predictive performance [12].

The model training process was conducted using the preprocessed dataset. The Random Forest model was configured with 100 estimators, a maximum depth of 10, a minimum of 10 samples required to split an internal node, a minimum of 5 samples per leaf node, the Gini criterion, and the square root (sqrt) method for determining the maximum number of features considered at each split. These parameters were selected to control model complexity and reduce the possibility of overfitting during the learning process.

After training, the generated model was saved and integrated into the DiAnalyze web-based application. The implementation of machine learning models into web applications allows prediction systems to become more accessible and easier to use by end users [13].

2.4. Evaluation

The evaluation stage aims to measure the performance of the Random Forest model in predicting diabetes risk. The evaluation was performed using the testing dataset, which contains data that was not previously used during the training process. Several evaluation metrics were used, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

Accuracy measures the proportion of correct predictions compared with the total testing samples, while precision measures the proportion of correctly predicted positive cases compared with all predicted positive cases. Recall measures the ability of the model to identify actual positive diabetes cases, whereas the F1-score represents the harmonic mean between precision and recall and provides a balanced measurement of classification performance. The

confusion matrix was used to analyze classification results based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

In addition, ROC-AUC was used to evaluate the model's ability to distinguish between diabetes and non-diabetes classes based on prediction probability. The evaluation approach follows previous diabetes prediction studies that assess machine learning model performance using classification metrics such as accuracy, precision, recall, and F1-score [7], [14].

3. Result and Discussion

3.1. Model Training and Evaluation Results

The developed diabetes risk prediction model was implemented using the Random Forest Classifier algorithm [15]. The training process was performed after completing the preprocessing stage, including duplicate removal, categorical encoding, and dataset balancing. The final dataset consisted of 18,482 records, which were divided into training, validation, and testing datasets.

The dataset distribution used during model development is presented in Table 1.

Table 1. Dataset Distribution

Dataset	Number of Data	Percentage
Training	11.828	64%
Validation	2.957	16%
Testing	3.697	20%

The Random Forest model was trained using several parameter configurations to control model complexity. The parameters used include 100 decision trees, Gini criterion, maximum depth of 10, minimum samples split of 10, minimum samples leaf of 5, and sqrt maximum features.

After the training process was completed, the generated model was exported into a pickle format [16]. The saved model components consist of:

1. random_forest_model.pkl as the main prediction model.
2. gender_encoder.pkl for converting gender categorical values.
3. smoking_history_encoder.pkl for converting smoking history categorical values.

These model components were integrated into the backend service using FastAPI to perform prediction inference based on user input parameters.

3.2. Random Forest Performance Analysis

The performance evaluation was conducted using the testing dataset containing 3,697 records that were not involved during the training process. The evaluation aims to measure the capability of the Random Forest model in classifying diabetes and non-diabetes

categories [17]. The evaluation results are shown in Table 2.

Table 2. Random Forest Model Evaluation Results

Evaluation Metric	Result
Accuracy	90.70%
Precision	89.09%
Recall	90.97%
F1-score	90.02%
ROC-AUC	97.61%

Based on the evaluation results, the Random Forest model achieved an accuracy of 90.70%, indicating that most testing samples were correctly classified. The precision value of 89.09% shows that the model was able to minimize incorrect positive predictions, while the recall value of 90.97% indicates that the model successfully identified most samples belonging to the diabetes risk category.

The F1-score value of 90.02% demonstrates a balanced performance between precision and recall. Furthermore, the ROC-AUC value of 97.61% indicates that the model has strong capability in distinguishing between diabetes and non-diabetes classes.

The testing classification results are presented in Table 3.

Table 3. Classification Report on Testing Dataset

Class	Precision	Recall	F1-score
Non-Diabetes (0)	0.92	0.90	0.91
Diabetes (1)	0.89	0.91	0.90

The classification results show that the model provides relatively balanced performance for both classes. This indicates that the dataset balancing process using undersampling successfully reduced the impact of class imbalance during model training [18].

The confusion matrix analysis is presented in Figure 1.

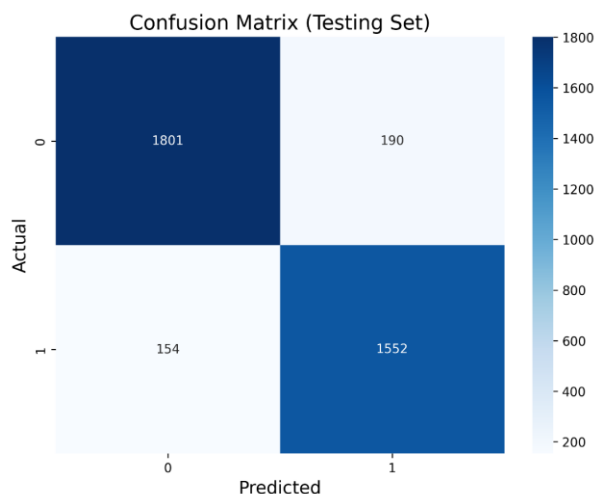


Figure 1. Confusion Matrix of Random Forest Model

The confusion matrix results consisted of 1,801 True Negative (TN), 1,552 True Positive (TP), 190 False Positive (FP), and 154 False Negative (FN) cases. The

model successfully identified 1,552 diabetes risk cases and 1,801 non-diabetes cases correctly. Meanwhile, 154 diabetes cases were incorrectly classified as non-diabetes. The relatively small number of false negatives indicates that the model has good capability in identifying potential diabetes risk cases.

3.3. DiAnalyze Web-Based System Implementation

The trained Random Forest model was implemented into a web-based application called DiAnalyze [19]. The application was developed to provide users with an accessible platform for analyzing diabetes risk based on health parameters [20].

The system architecture consists of three main components:

3.3.1. Frontend

The user interface was developed using React.js with Tailwind CSS to provide a responsive and interactive interface.

3.3.2. Backend

The backend service was developed using FastAPI [21]. The backend handles user authentication, receives prediction requests, performs preprocessing, and executes inference using the Random Forest model.

3.3.3. Database

MySQL was used to store user account information, including username, email, password hash, and account creation timestamp [22].

The application was deployed using Docker Compose on a Virtual Private Server (VPS) environment. The server specifications used are shown in Table 4.

Table 4. Server Hosting Specification

Component	Specification
Hosting Service	Virtual Private Server
Processor	2vCPU
Memory	2 GB RAM
Storage	40 GB SSD
OS	Ubuntu 24.04
Server Region	Jakarta

The deployed application can be accessed through the domain: <https://dianalyze.web.id>.

The implementation of Docker allows each application component to run independently within separate containers, making the deployment process easier to manage and maintain [23].

The developed DiAnalyze system provides three main pages: the Homepage, Authentication Page, and Diabetes Prediction Page.

3.3.4. Homepage

The homepage provides general information about the application, diabetes education, and information regarding the technology used. It consists of three main

sections: the landing section, diabetes information section, and platform information section, as illustrated in Figures 2–4.



Figure 2. Homepage Landing Section Interface

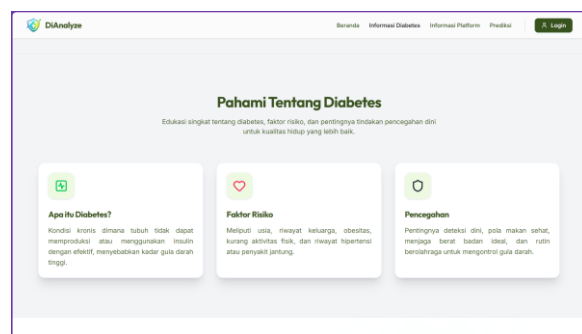


Figure 3. Homepage Diabetes Information Section Interface

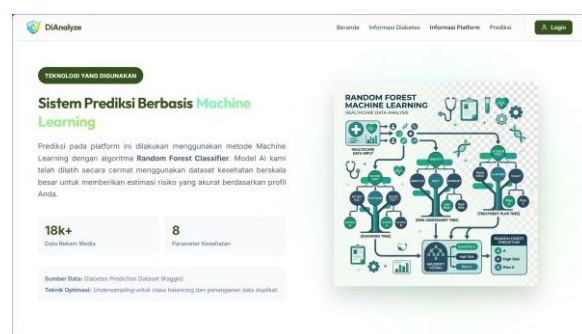


Figure 4. Homepage Platform Information Section Interface

3.3.5. Authentication Page

The authentication page provides user registration and login functionalities to control access to the diabetes prediction feature. As shown in Figure 6, registered users can access the system by entering their username or email address and password. Users who do not yet have an account can create a new account through the registration page.

As illustrated in Figure 7, the registration page requires users to provide a username, email address, password, and password confirmation. After successful registration, the account information is stored in the MySQL database, while the password is stored in hashed form for security purposes. Registered users

can then log in to access the diabetes prediction feature.

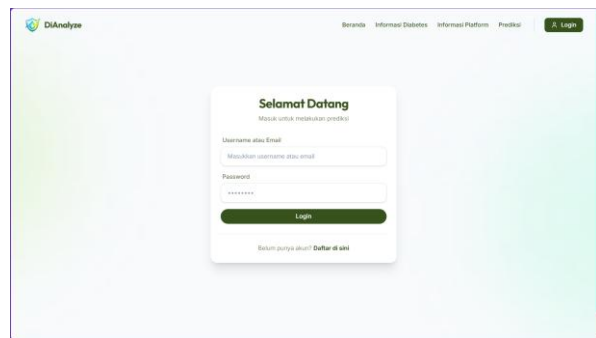


Figure 5. User Login Interface

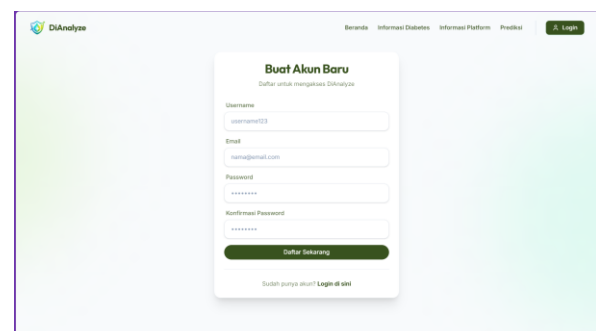


Figure 6. User Registration Interface

3.3.6. Diabetes Prediction Page

The diabetes prediction page allows authenticated users to perform diabetes risk analysis by entering the required health parameters, including gender, age, hypertension history, heart disease, smoking history, BMI, HbA1c level, and blood glucose level.

As shown in Figure 7, users are required to complete the input form before initiating the prediction process.

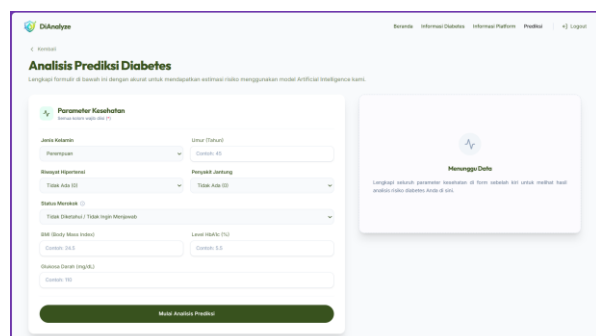


Figure 7. Diabetes Prediction Page Interface

After the user submits the health parameters, the system processes the input data through the prediction API and applies the trained Random Forest model to generate the prediction result. The system displays the predicted diabetes risk classification and the corresponding risk probability.

As illustrated in Figure 8, the prediction result is presented together with the probability percentage and an informational message emphasizing that the model output is not a medical diagnosis.

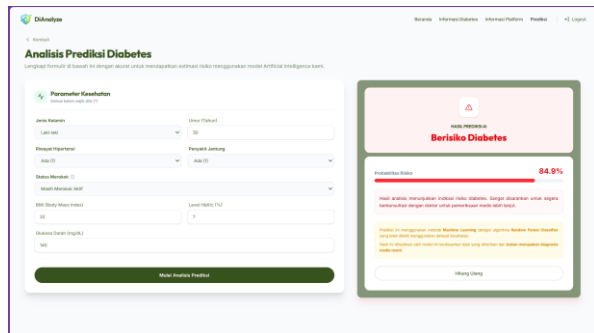


Figure 8. Diabetes Prediction Result Interface

4. Conclusion

This study successfully developed a web-based diabetes risk prediction system named DiAnalyze, incorporating the Random Forest algorithm for classification. The model development progressed through several structured phases, including data preprocessing, categorical feature encoding, dataset balancing via undersampling, model training, and performance evaluation on testing data.

Evaluation metrics demonstrated the Random Forest model's strong performance in predicting diabetes risk, achieving an accuracy of 90.70%, precision of 89.09%, recall of 90.97%, and an F1-score of 90.02%. These results validate the model's ability to categorize diabetes risk using health parameters such as gender, age, hypertension, heart disease history, smoking status, BMI, HbA1c level, and blood glucose level.

Furthermore, the trained model was seamlessly deployed into the DiAnalyze web application using FastAPI for the backend and React.js for the frontend. The platform offers educational resources, user authentication, and an interactive prediction interface.

However, DiAnalyze serves as a decision support tool for early risk assessment rather than a substitute for professional medical diagnosis. Future research could enhance system capability by leveraging larger clinical datasets, evaluating alternative machine learning algorithms, expanding health parameters, and implementing continuous model updating mechanisms to improve predictive accuracy.

References

- [1] F. Teufel *et al.*, "Diabetes in Adults : Findings From the 2025 IDF Diabetes Atlas Global , Regional , and National Estimates of Undiagnosed Diabetes in Adults : Findings From the 2025 IDF Diabetes Atlas," *Diabetes Care*, vol. 49, no. 3, pp. 490–496, 2026, doi: 10.2337/dc25-2583.
- [2] E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, and S. Kamila, "Prediksi Penyakit Diabetes Melitus Menggunakan Algoritma Machine Learning Diabetes Mellitus Disease Prediction using Machine Learning Algorithms," *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 12, no. 4, pp. 760–766, 2024, doi: 10.26418/justin.v12i4.84620.
- [3] A. Syahri, U. Fariha, R. Afandi, and I. Nurliyana, "Comparison of Logistic Regression , Random Forest and Adaboost Algorithms for Diabetes Mellitus Classification," *IJATIS Indones. J. Appl. Technol. Innov. Sci.*, vol. 1, no. February, pp. 41–46, 2024, doi: 10.57152/ijatiss.v1i1.11116.
- [4] M. C. Ibrahim, "Comparison of Diabetes Prediction Data Using Machine Learning Perbandingan Data Prediksi Diabetes Menggunakan Machine Learning," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 4, pp. 1423–1436, 2025, doi: 10.57152/malcom.v5i4.2301.
- [5] N. L. Anggreini, A. Yuliana, D. S. Ramdan, and W. Al-dayyeni, "Improving Diabetes Prediction Performance Using Random Forest Classifier with Hyperparameter Tuning," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1847–1860, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.4.4755.
- [6] O. P. Handayani, I. A. Ashari, and R. Ardianto, "Systematic Literature Review: Penerapan Machine Learning dalam Diagnosis dan Prediksi Penyakit Diabetes," *KOMPUTA J. Ilm. Komput. dan Inform.*, vol. 14, no. 2, pp. 3–7, 2025, doi: 10.34010/komputa.v14i2.
- [7] C. C. Olisah, L. Smith, and M. Smith, "Computer Methods and Programs in Biomedicine Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, 2022, doi: 10.1016/j.cmpb.2022.106773.
- [8] Rianto and P. I. Santosa, "Data Preparation untuk Machine Learning & Deep Learning," Edisi Pert., Yogyakarta: ANDI, 2025.
- [9] F. D. A. and G. F. Fitriana, *PENGENALAN BAHASA PEMROGRAMAN PYTHON DAN IMPLEMENTASI MACHINE LEARNING*. Yogyakarta: TEKNOSAIN, 2022.
- [10] H. Aulia and A. Wibowo, "Integration of Random Forest , ADASYN , and SHAP for Diabetes Prediction and Interpretation," *Sci. J. Informatics*, vol. 12, no. 2, pp. 211–222, 2025, doi: 10.15294/sji.v12i2.24314.
- [11] K. and A. H. Pratama, *Belajar Mudah dan Singkat MACHINE LEARNING*. Yogyakarta: ANDI, 2024.
- [12] C. N. Noviyanti, "Journal of Information System Early Detection of Diabetes Using Random Forest Algorithm," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, pp. 41–48, 2024, doi: 10.52465/joiser.v2i1.245.
- [13] A. P. Suryotomo, B. M. Akbar, and R. Husaini, "Performance Analysis of FastAPI Framework on Lost Circulation Handling Management Application in Oil Well Drilling," *J. Inform. dan Teknol. Inf.*, vol. 21, no. 1, pp. 110–121, 2024, doi: 10.31515/telematika.v21i1.13259.
- [14] B. S. Ahamed, M. S. Arya, and A. O. N. V, "Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques," *Front. Comput. Sci.*, vol. 4, no. May, pp. 1–5, 2022, doi: 10.3389/fcomp.2022.835242.
- [15] L. Li *et al.*, "Machine learning for predicting diabetes risk in western China adults," *Diabetol. Metab. Syndr.*, vol. 15, no. 165, pp. 1–12, 2023, doi: 10.1186/s13098-023-01112-y.
- [16] L. Jiang *et al.*, "Diabetes risk prediction model based on community follow-up data using machine learning," *Prev. Med. Reports*, vol. 35, no. July, p. 102358, 2023, doi: 10.1016/j.pmedr.2023.102358.
- [17] Q. Ou, W. Jin, L. Lin, D. Lin, and K. Chen, "LASSO-based machine learning algorithm to predict the incidence of diabetes in different stages," *Aging Male*, vol. 26, no. 1, p., 2023, doi: 10.1080/13685538.2023.2205510.

- [18] H. Yousef, S. F. Feng, and H. F. Jelinek, "Exploratory risk prediction of type II diabetes with isolation forests and novel biomarkers," *Sci. Rep.*, pp. 1–11, 2024, doi: 10.1038/s41598-024-65044-x.
- [19] D. Nafisah, A. Ahli, S. R. Nudin, and D. A. Dermawan, "Enhancing Web-Based Diabetes Prediction Using Random Forest Optimization and SMOTE," *Proceeding Int. Jt. Conf. UNESA*, vol. 3, no. 2, pp. 1–10, 2023.
- [20] H. El-sofany, S. A. El-seoud, O. H. Karam, Y. M. A. El-latif, and I. A. T. F. Taj-eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App," *Int. J. Intell. Syst.*, vol. 2024, 2024, doi: 10.1155/2024/6688934.
- [21] S. Effati, A. Kamarzardi-torghabe, F. Azizi-froutaghe, and I. Atighi, "Web application using machine learning to predict cardiovascular disease and hypertension in mine workers," *Sci. Rep.*, pp. 1–14, 2024.
- [22] B. Rawat, S. Purnama, and M. Mulyati, "MySQL Database Management System (DBMS) On FTP Site LAPAN Bandung," *Int. J. Cyber IT Serv. Manag.*, vol. 1, no. 2, pp. 173–179, 2021, doi: 10.34306/ijcitsm.v1i2.47.
- [23] R. Malhotra, A. Bansal, and M. Kessentini, "Assessing security vulnerabilities in a docker-enabled federated learning framework with hyperparameter tuning for software bug prediction," *Int. J. Inf. Secur.*, vol. 25, no. 2, pp. 1–18, 2026, doi: 10.1007/s10207-026-01213-5.