Implementation of XGBoost for diabetes mellitus risk prediction based on health history

Andi Riansyah[1], **Ghufron**[1], **Lailatul Fitriyah**[1], **Suyanto**[2]
[1]Department of Informatics Engineering, Faculty of Industrial Technology, Universitas Islam Sultan Agung, Semarang, Indonesia
[2]Faculty of Nursing, Universitas Islam Sultan Agung, Semarang, Indonesia

Article Info

Article history:

Received Sep 12, 2024
Revised Oct 30, 2025
Accepted Oct 20, 2025

Keywords:

Decision tree
Diabetes mellitus
Machine learning
Prediction
XGBoost

ABSTRACT

Diabetes mellitus (DM) is a chronic disease with a growing global burden and specific challenges for early management, particularly in regions with limited access to healthcare. This study develops a web-based system to classify diabetes risk from medical history using extreme gradient boosting (XGBoost), an ensemble model of decision trees. The dataset comprised 520 respondents (320 DM, 200 non-DM) and underwent labeling, standardization, and an 80:20 train–test split, followed by hyperparameter selection via grid search and 5-fold cross-validation (CV). On the test set, the model achieved an accuracy of 0.9888, precision of 1.0000, recall of 0.9718, and an F1-score of 0.9857; discriminative performance was also strong with an area under the receiver operating characteristic curve (AUC-ROC) of 0.839. These findings confirm that XGBoost effectively handles complex or imbalanced medical data while providing probabilistic outputs that are clinically meaningful. Deployed as a web application, the system can support early screening, triage, and clinical decision-making, thereby expediting referrals and personalizing interventions in primary care and hospital settings, especially in resource-constrained environments. This work lays the groundwork for further development, including the integration of explainable artificial intelligence (XAI) techniques to enhance clinical transparency.

Corresponding Author:

Andi Riansyah
Department of Informatics Engineering, Faculty of Industrial Technology, Universitas Islam Sultan Agung
Kaligawe Street Km. 4, Semarang 50112, Indonesia
Email: andi@unissula.ac.id.

1. INTRODUCTION

In general, diabetes refers to a group of conditions that affect processes in the body, characterized by a sustained increase in blood sugar levels due to damage to the function of organs such as nerves, kidneys, heart, and blood vessels. This disease can lead to various serious complications that impact many organs in the body, with a range of symptoms [1]. In Indonesia, the prevalence of diagnosed diabetes continues to rise, as reflected in the Riskesdas data since 2018, which provides a detailed picture of the distribution of diabetes across the archipelago. This condition requires a deep understanding to plan effective interventions for the prevention and management of diabetes in the community [2], [3]. Diabetes has evolved into a global health issue with increasing numbers every year. This indicates that diabetes is not just a local health problem but has become a global challenge that requires attention and collective action from various parties who are aware [4], [5].

In recent years, the use of machine learning techniques has become very popular for disease prediction. However, the accuracy of disease predictions remains a significant challenge in the medical field.

With the presence of artificial intelligence, it is hoped that it can predict diseases, especially diabetes mellitus (DM), so that patients can receive prompt and accurate care [6], [7].

The extreme gradient boosting (XGBoost) process can enhance prediction accuracy through multi-threading, which efficiently utilizes CPU cores to significantly accelerate the algorithm's performance. This algorithm applies ensemble techniques and focuses on the gradient of the optimized loss function, resulting in a model that is very robust and accurate [8]–[10]. With XGBoost, this research aims to produce a model that can predict the risk of DM based on an individual's health history to enhance early diagnosis [11]–[13]. The researchers hope to contribute to the development of more accurate predictive tools that can assist in early prevention and management for those at risk, as well as provide additional information about risk factors associated with DM.

This research lies in the implementation of the XGBoost algorithm for predicting DM risk based on patients' health history and its integration into a web-based system [1]. Unlike many previous studies that primarily relied on conventional models such as logistic regression (LR), support vector machine (SVM), or random forest (RF), this study demonstrates the superior ability of XGBoost to handle imbalanced and complex medical datasets, further optimized through hyperparameter tuning with grid search and cross-validation [2]. The main contributions of this work include the development of a high-performance predictive model with the creation of a user-friendly web application to facilitate early detection of diabetes risk, and the provision of probabilistic outputs that offer more nuanced decision support for healthcare professionals. In addition, this research contributes to the scientific community by highlighting the practical advantages of XGBoost in medical informatics and providing a foundation for future exploration of hybrid or ensemble models for disease prediction [3].

2. REVIEW OF LITERATURE

This section reviews prior studies applying the XGBoost algorithm in medical and related fields. Research [14] highlights the importance of feature analysis in XGBoost for cardiovascular disease prediction, showing strong accuracy compared to other models (95% for RF, 90% for SVM, and 86% for LR). Another study applied XGBoost with feature selection and hyperparameter optimization in spam detection, achieving an accuracy of 92.67% and outperforming Chi and principal component analysis (PCA)-based methods [15], [16].

In this study, hyperparameters such as max_depth, learning_rate, and n_estimators were systematically optimized using grid search with 5-fold cross-validation to balance model complexity, learning efficiency, and overfitting risks [4]. Evaluation metrics included accuracy, precision, recall, and F1-score to ensure clinical relevance in diabetes prediction. Precision reduces false diagnoses, recall minimizes undetected cases, and F1-score balances both, especially useful for imbalanced datasets [3].

While XGBoost demonstrates strong performance for diabetes risk prediction, comparisons with other models are essential. RF offers robustness but higher computation costs [17]; LR is simpler yet less effective with non-linear data [18]; SVM performs well with high-dimensional data but struggles with scalability; and ANNs capture complex patterns but require large datasets and careful tuning. Benchmarking against these models, RF, SVM, LR, and artificial neural network (ANN) is needed to confirm XGBoost's superiority in complex and imbalanced medical datasets [19]. Future research should expand to larger and more diverse datasets, including real-time electronic health records (EHRs). Hybrid ensemble methods such as stacking and bagging may enhance robustness, while integrating explainable AI (XAI) is crucial for interpretability and clinical adoption.

2.1. Diabetes mellitus

DM is a common disease characterized by high levels of sugar in urine and blood. The term "diabetes" comes from the Greek diabainen, meaning "to flow continuously," while "mellitus" comes from the Latin mellitus, meaning "honey." This condition occurs when the pancreas does not produce sufficient insulin [5]. Insulin plays a crucial role in transporting blood sugar, derived from carbohydrate breakdown, into cells for energy. A lack of insulin leads to elevated blood sugar levels, which, over time may damage organs and body tissues. Common symptoms include frequent urination, excessive thirst, increased appetite, sudden weight loss, fatigue, obesity, slow-healing wounds, vision problems, and itching [17], [19].

According to the American Diabetes Association (ADA), a new patient is diagnosed with diabetes every 21 seconds, reflecting a high incidence rate [2]. Diabetes is a leading cause of premature death worldwide and contributes to severe complications such as blindness, heart disease, and kidney failure [18].
DM is often called a "silent killer" because many patients remain unaware of their condition until blood vessel damage or other complications occur. Risk factors are not limited to older adults, as anyone from young to elderly can develop the disease [20]. Lifestyle factors play a significant role, with DM frequently associated with chronic conditions such as hypertension and high cholesterol levels [18].

2.2. Prediction

Prediction in the context of machine learning is the process of estimating future outcomes or values based on historical data [21]. This historical information may come from the current dataset or from specific input data [21]. The goal is to develop a model that can learn patterns from existing data and then use it to predict outcomes on unseen data, allowing preventive measures to be implemented earlier [22].

2.3. Decision tree

In general, a decision tree is a decision-making technique that organizes each choice into a branching structure [23]. According to [24], decision trees are a popular and effective method in the fields of classification and prediction that transform data into an easily understandable decision rule tree. This rule can be expressed by searching for or filtering data according to the criteria generated by the classification model. Decision trees reduce a large dataset into a smaller dataset. Where each leaf node represents a class label. The root and internal nodes contain attribute test conditions, usually in the shape of an oval, while the leaf nodes are square-shaped.

A decision tree classifies data by asking questions related to the characteristics of each feature in the data. Each question is represented by an internal node connected to child nodes that represent the answers to the question. In this way, the questions form a hierarchical structure in the shape of a tree. The decision tree is built by gradually adding question nodes, guided by labeled training data. Decision trees naturally divide issues into smaller components in order to solve them, used to divide data into more specific subnets by continuously asking questions until reaching a condition where further division is no longer necessary shown in Figure 1 [15].

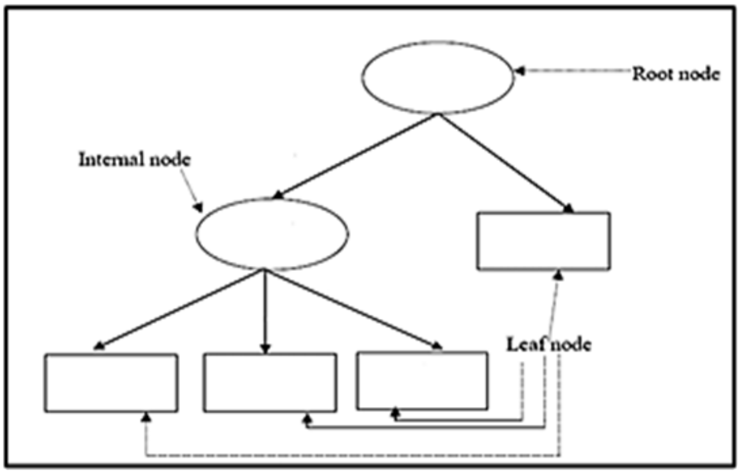


Figure 1. Decision tree structure

2.4. Extreme gradient boosting

XGBoost was introduced by Wang *et al.* [10] as a development of the gradient boosting decision tree (GBDT) previously introduced by [25], [26]. This algorithm builds decision trees iteratively, where each new tree corrects the errors of the previous one. The weights of each tree are also adjusted, resulting in a stronger and more reliable model [27]–[30].

In this study, XGBoost was compared with two basic models commonly used in medical prediction: LR and decision trees. LR works well for simpler cases, but lacks accuracy in handling complex non-linear relationships [6]. Decision trees provide better interpretability but are prone to overfitting when dealing with datasets containing many features or noise. In contrast, XGBoost combines multiple decision trees through ensemble learning, effectively handling feature interactions and delivering more stable and accurate predictions for complex medical datasets.

The main advantage of XGBoost lies in its ability to provide highly accurate predictions of DM risk, achieving 99.79% training accuracy and 98.88% testing accuracy, significantly outperforming LR and decision trees [7]. Moreover, XGBoost is more effective at handling imbalanced data, which is common in disease prediction cases. Although interpretability remains a challenge, XGBoost’s ability to manage data complexity and deliver consistent results makes it a strong candidate for medical applications, particularly in early diabetes risk detection.

Technically, XGBoost works by building regression trees that map data points to leaves with continuous scores. It minimizes a regularized objective function (L1 and L2) and a convex loss function, using gradient descent to reduce error. Each new tree is trained to predict the residuals of the previous ones, and the ensemble of trees ultimately produces the final prediction [25]. Figure 2 illustrates how XGBoost works, where\propto_i is the regularization parameter and r_i is the residue calculated with the first tree, and h_i is a function that has been taught to forecast residuals, r_i using X for the first tree. To calculate\propto_i using the residue calculated r_i it can be computed using (1).

$$\arg_{min_\propto}=\sum_{i-1}^{m} L(Y_i,F_{i-1}(X_i) + \propto h_i(X_i,r_{i-1})) \tag{1}$$

$L\big(Y,F(X)\big)$ is a differentiated loss function.

Information:

- Data collection (*X, Y*): the process begins with a dataset consisting of feature X and target Y.
- Tree *1 (F1(X))*: the first model is trained using a dataset. After the tree is created, r1 is calculated, which is the difference between the predictions of tree 1 and the actual value Y.
- Calculate α1: the value of α1 is calculated, which is a regularization parameter that helps reduce overfitting.
- Tree 2 *(F2(X))*: the second model, trained to predict the residual r1. After tree 2 is trained, the residual r2 is calculated, which is the difference between the predictions of tree 2 and the residual r1
- Calculate: the value of α2 is calculated for tree 2.
- Iteration process: this process is repeated for many trees. (from tree 3 to tree m). Each new tree is trained to predict the residuals from the previous tree, and new residuals are calculated at each step. The value of αi is also calculated each time for each tree i.
- The final model *Fm(X)*: a combination of each tree that has been trained. This combination is written as (2).

  $$F_m\ (X) = F_{m-1}(X) + \alpha_m h_m(X,\ r_{m-1}) \tag{2}$$

  Where hi is a function trained to predict the residual ri at the i-th tree.
- Optimizing α: to calculate the value of α, then minimize the differentiable loss function L, using (3).

  $$arg\ min_\propto \textstyle\sum_{i=1}^{m} L(Y_i,F_{i-1}(X_i)\ + \propto h_i(X_i,r_{i-1})) \tag{3}$$

  Information: L is loss function, Y_i is the actual target value (0 or 1), $F_{i-1}(X_i)$ is prediction of the model in the previous iteration, $\propto$ is the coefficient that determines the extent of the contribution of the new model h_i, $h_i(X_i,r_{i-1})$ is the new model added in the i iteration, and r_{i-1} is residual or error in the previous iteration.

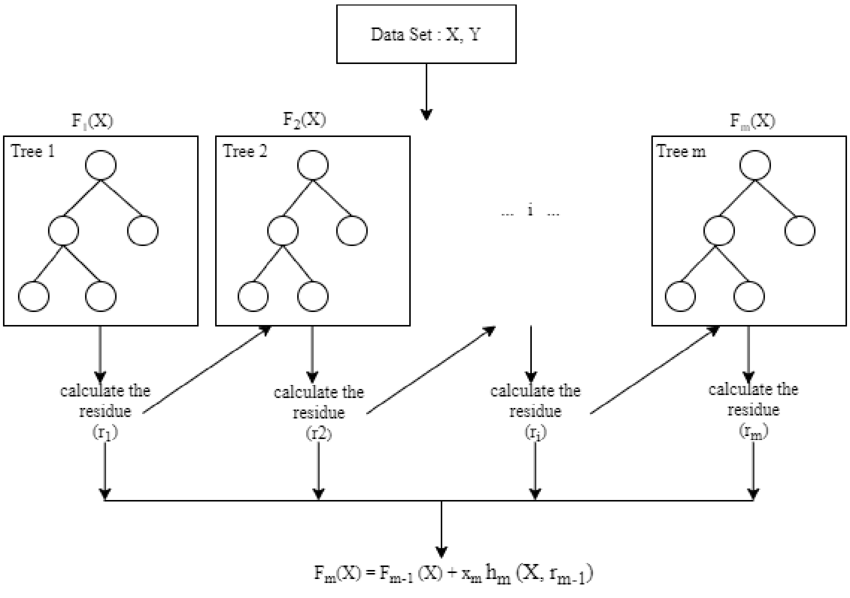

Data Set : X, Y
F1(X)
F2(X)
Fm(X)
Tree 1
Tree 2
Tree m
... i ...
calculate the residue (r1)
calculate the residue (r2)
calculate the residue (ri)
calculate the residue (rm)
Fm(X) = Fm-1 (X) + xm hm (X, rm-1)

Figure 2. How XGBoost works

The use of technology, such as machine learning algorithms in predicting diabetes, is also rapidly advancing, making it important to recognize whether someone falls into the category of patients with diabetes before those with DM experience serious complications, so that they can receive appropriate and timely treatment. The XGBoost algorithm has become popular for predicting diseases. Therefore, in this report, we will utilize the XGBoost algorithm as one of the implementations to generate predictions for the risk of DM. This system is expected to be beneficial for healthcare workers and to help improve hospital services for patients, especially in early detection, particularly of diabetes mellitus, and in personalizing treatment.

3. METHOD

An additional explanation to strengthen the applied science framing and its relevance to Indonesia's healthcare context: XGBoost can be operationalized as an intelligence component in clinical decision support at Puskesmas and hospitals for early diabetes screening, for example, integrated with electronic medical records and routine anamnesis forms so that health workers receive a real-time risk score and its contributing factors. Its efficiency and scalability enable deployment in offline-first mobile applications for areas with limited connectivity, while its interpretability helps clinicians understand the influence of local variables such as body mass index, fasting blood glucose, blood pressure, family history, diet, and physical activity, making follow-up recommendations more actionable. In addition, this approach fits Indonesia's need for resource-efficient solutions: XGBoost handles imbalanced data and many features without costly infrastructure, its accuracy can be monitored across regions, and it can be aligned with local policies so it is ready for use in non-communicable diseases program management dashboards and everyday clinical workflows.

3.1. Dataset

This dataset is called early-stage diabetes risk prediction diabetes and was downloaded from Kaggle. The data used in this research is secondary data obtained from questionnaires filled out by patients at Sylhet Diabetic Hospital in Sylhet, Bangladesh, and has been approved by a doctor. This research involves a total of 520 data points, consisting of 320 data points for the positive DM category and 200 for the non-DM category. This data set covers a variety of symptoms associated with diabetes [29]. The data supporting the findings of this study were obtained from various scientific articles and medical records related to diabetes risk prediction. The data were processed and anonymized prior to use in the development of the prediction model. Due to privacy and ethical considerations, the data are not publicly available but may be obtained from the corresponding author upon reasonable request.

3.2. Preprocessing

This process prepares the data for processing in modeling with the XGBoost algorithm. The initial step involves labeling (label encoding), which is the process of converting data features that are originally text into numeric values [29]. Second, features and targets are separated to ensure that the model is trained only on relevant features and uses the target to make predictions [9]. Next, standardize the data to adjust it so that all features are within a uniform range [25]. Then, the data is divided into two with a distribution ratio of 80:20, resulting in 416 data points for training and 104 data points for testing. Data partitioning is carried out to ensure that the model is trained using specific data and its performance is tested using foreign data that has not been seen during training [29].

3.3. Modelling

After the data splitting process, the format is converted into Dmatrix, which is a special format for XGBoost designed for data storage. The next process is to determine the hyperparameters using grid search with 5-fold cross-validation. The hyperparameters used are as shown in Table 1.

Table 1. Hyperparameters used

Hyperparameter	Value	Information
max_depth	[3, 4, 5]	The maximum depth of each tree.
learning_rate	[0.01, 0.1, 0.2]	Controlling the step size/how quickly the learning from mistakes mode operates in each iteration.
n_estimators	[550, 100, 200]	The total number of trees that will be built by the model in the boosting process.

The best hyperparameter combination based on performance during 5 iterations (cross-validation) will be saved as the model for making predictions on the testing data. This process is carried out using 5-fold cross-validation, where the model is evaluated for each hyperparameter combination to obtain the best parameter combination that yields the most optimal model performance. Then, to maximize the model's performance, the model is trained with training data and evaluated using test data [16]. This process involves repeated iterations over 5 times to minimize prediction errors and reduce overfitting, making the model more reliable in making predictions on data that has not been seen at all during the model training process. After the training and optimization process, the model is saved for future use and can be applied in real-world applications for automatic diabetes risk prediction. With this approach, an accurate and reliable diabetes risk prediction model can be built using the XGBoost algorithm. The following is the system modeling design in the form of a flowchart, as shown in Figure 3.

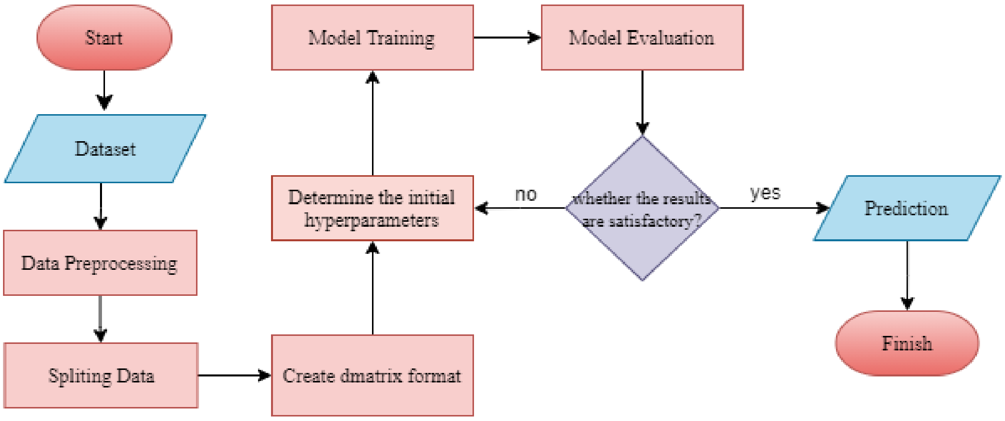


Figure 3. System modeling flowchart

3.4. Model evaluation

The model that will be evaluated in this research uses a confusion matrix, which is one way to measure a model's performance in classification problems [24], where the classification results can belong to two or more classes. Evaluation using the confusion matrix helps in identifying how the model accurately predicts diabetes conditions (positive and negative) [9]. The accuracy value is the comparison between the number of data that has been correctly identified and the total data that has been tested [29]. The precision value is the comparison between the quantity of accurate forecasts as well as the overall number of forecasts [31]. The recall value is a comparison of how many predictions are correct compared to all the actual data [32]. The F1-score value combines precision and recall into a single metric, which is very useful for dealing with imbalanced data [33]–[35].

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision+recall} \tag{7}$$

4. EXPERIMENT ANALYSIS

Modeling using the XGBoost algorithm has been successfully applied to predict the risk of diabetes based on health history. This model was chosen for its advantages in handling complex data and imbalanced data distributions. The results of the modeling show good performance.

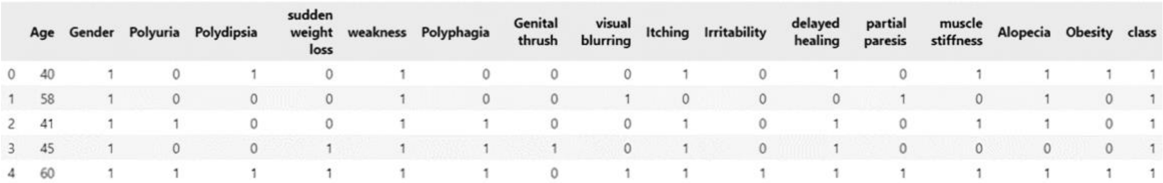
One of the limitations of using XGBoost in this study is its relatively low interpretability, which is a crucial aspect in medical applications. Although XGBoost has demonstrated high predictive accuracy, its ensemble-based and complex nature makes the decision-making process difficult to understand directly by healthcare professionals. In clinical practice, physicians require not only the prediction results but also clear explanations of the contributing factors behind those results in order to ensure ethical use and build trust. Therefore, future research should consider integrating XAI approaches, such as Shapley additive explanations (SHAP) or local interpretable model agnostic explanation (LIME), to provide insights into the dominant features influencing each prediction. By doing so, the model will not only be accurate but also transparent, trustworthy, and better aligned with the needs of medical decision support.

4.1. Preprocessing

This section outlines the preprocessing steps, including label encoding, feature-target separation, and data standardization to prepare the dataset for XGBoost modeling. Figure 4 shows the encoding process converts all categorical features into binary values (0 and 1). Figure 5 shows that the predictor variables (x) consist of patient symptoms and characteristics, while the target variable (y) is the class column representing the diagnosis label.

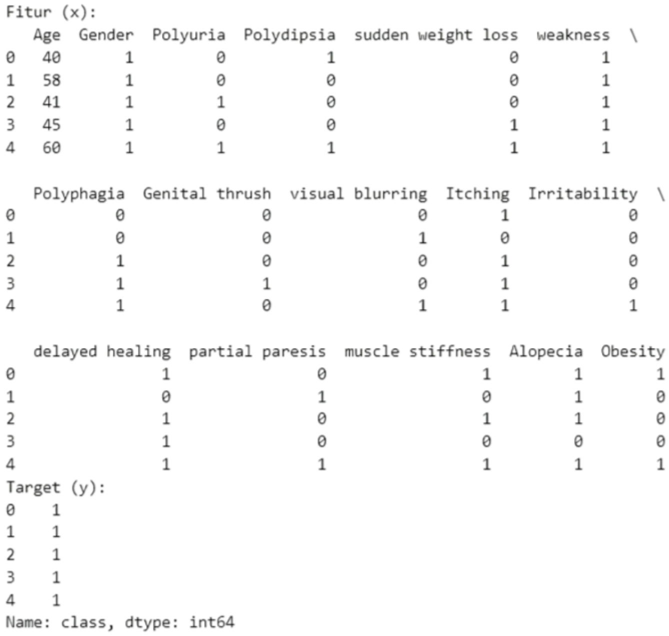
4.2. Determine hyperparameters

This section outlines the determination of hyperparameters for XGBoost modeling. Figure 6 show that the best XGBoost model uses eval_metric=logloss, learning_rate=0.1, max_depth=4, n_estimators=100, and objective=binary:logistic.



	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

Figure 4. Output after encoding data

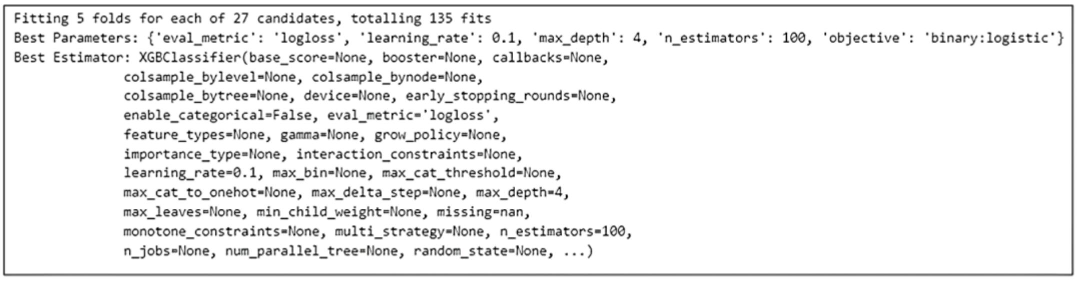


```
Fitur (x):
   Age  Gender  Polyuria  Polydipsia  sudden weight loss  weakness  \
0   40       1         0           1                   0         1
1   58       1         0           0                   0         1
2   41       1         1           0                   0         1
3   45       1         0           0                   1         1
4   60       1         1           1                   1         1

   Polyphagia  Genital thrush  visual blurring  Itching  Irritability  \
0           0               0                0        1             0
1           0               0                1        0             0
2           1               0                0        1             0
3           1               1                0        1             0
4           1               0                1        1             1

   delayed healing  partial paresis  muscle stiffness  Alopecia  Obesity
0                1                0                 1         1        1
1                0                1                 0         1        0
2                1                0                 1         1        0
3                1                0                 0         0        0
4                1                1                 1         1        1
Target (y):
0    1
1    1
2    1
3    1
4    1
Name: class, dtype: int64
```

Figure 5. Output of separating features and targets



```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
Best Parameters: {'eval_metric': 'logloss', 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 100, 'objective': 'binary:logistic'}
Best Estimator: XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, gamma=None, grow_policy=None,
              importance_type=None, interaction_constraints=None,
              learning_rate=0.1, max_bin=None, max_cat_threshold=None,
              max_cat_to_onehot=None, max_delta_step=None, max_depth=4,
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, multi_strategy=None, n_estimators=100,
              n_jobs=None, num_parallel_tree=None, random_state=None, ...)
```

Figure 6. Output defines the hyperparameters

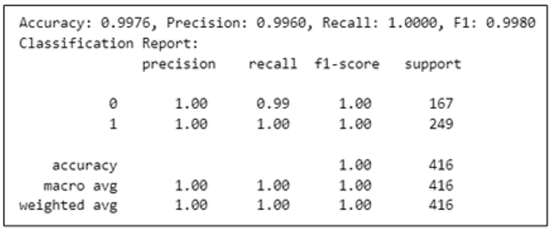
4.3. Result train and test

This section discusses the model's performance, illustrating the training and testing performance in predicting data. Figure 7 shows the result of saving a model that has gone through the grid search stage, a model with the best hyperparameter combination based on performance over 5 iterations. (cross-validation). This model will be used to make predictions on the testing data. The metric results generated during the model training are displayed in a more readable format and displaying four decimal places. Proven by displaying/calculating evaluation metrics on the training data. A confusion matrix is used to see how many correct and incorrect predictions the model made on the training data.

The accuracy produced from the image above is to calculate the model's accuracy in predicting correctly, with an accuracy value of 0.9979. The ratio of real positive forecasts to all positive predictions is known as precision, resulting in 0.9960. The percentage of real positive instances that the model accurately detects is known as recall, yielding a result of 1.000. The F1-score is the harmonization between recall and precision, resulting in 0.9980.

Figure 8 shows that model evaluation is the process of assessing the performance of a model that has been trained and tested using foreign data. The objective is to evaluate the model's performance when applied to unfamiliar data. The following is the result of each score in table form.

Each score displayed in Tables 2 to 4 is the result of the model's performance evaluation against the data used. To measure the model in making accurate predictions (accuracy), to assess the true positive predictions compared to the positive predictions made by the model for each category (precision), then to evaluate how many actual positive cases are correctly detected by the model from the total positive cases (recall), and to calculate the precision and recall harmonic mean, which provides an overall picture of the balance between the two (F1-score). Figure 9 shows that the XGBoost model achieved the best performance with an AUC of 0.839 compared to LR (0.823), linear SVM (0.815), and RF (0.812).

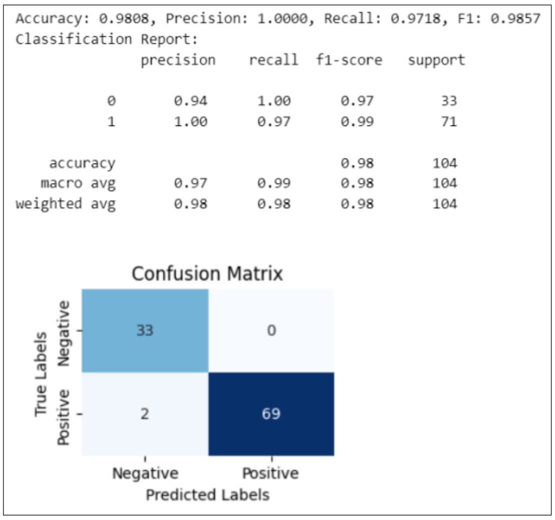


```
Accuracy: 0.9976, Precision: 0.9960, Recall: 1.0000, F1: 0.9980
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       167
           1       1.00      1.00      1.00       249

    accuracy                           1.00       416
   macro avg       1.00      1.00      1.00       416
weighted avg       1.00      1.00      1.00       416
```

Figure 7. Model train



```
Accuracy: 0.9808, Precision: 1.0000, Recall: 0.9718, F1: 0.9857
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        33
           1       1.00      0.97      0.99        71

    accuracy                           0.98       104
   macro avg       0.97      0.99      0.98       104
weighted avg       0.98      0.98      0.98       104
```

Figure 8. Model evaluation

Table 2. Confusion matrix values

True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
33	0	2	69

Table 3. Model performance results

Data	Accuracy	Precision	Recall	F1-score
Training	99%	99%	100%	99%
Testing	98%	100%	97%	98%

Table 4. Category of each class

Category	Precision	Recall	F1-score
Diabetes (1)	94%	100%	97%
Non-diabetes (0)	100%	97%	99%

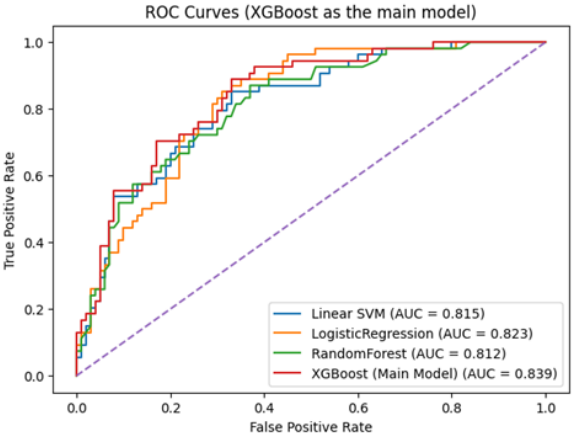
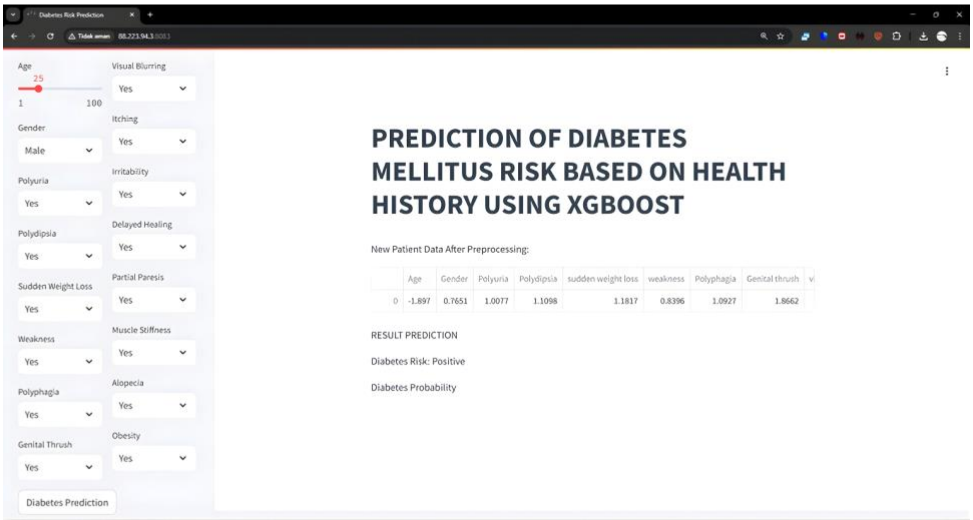


Figure 9. Receiver operating characteristic (ROC) curves

4.4. Website appearance

The developed model has then been implemented in the form of a website to reach many people out there who need predictions of DM risk based on their health history. Figure 10 shows the results of a series of processes that occur after the user inputs their health history data. First, it involves a standardization process for the data that has been entered, so that the newly input values are on a diverse scale based on the parameters that have been processed by the previously trained model, and the results will be displayed on the main page. Then this model will generate predictions for diabetes risk and the probability of diabetes risk based on the input data provided. The model will determine whether the risk of diabetes from the entered data is positive (indicating diabetes) or negative (does not indicate diabetes).

The probability value of diabetes risk is a result with a value between 0 and 1 to indicate how high the model's confidence level is in predicting whether someone has a risk of diabetes based on the given data. A value of 0.5 (50%) is used as the threshold d. If the probability value is above 0.5 (approaching 1), it means the model is very confident that there is a risk of diabetes, and the result will be classified in the positive category (diabetes). Conversely, if the probability value is under 0.5 (approaching 0), it means the model is very confident that there is no risk of diabetes, and the result will be classified in the negative category.



Age
25
1 100

Gender: Male

Polyuria: Yes

Polydipsia: Yes

Sudden Weight Loss: Yes

Weakness: Yes

Polyphagia: Yes

Genital Thrush: Yes

Visual Blurring: Yes

Itching: Yes

Irritability: Yes

Delayed Healing: Yes

Partial Paresis: Yes

Muscle Stiffness: Yes

Alopecia: Yes

Obesity: Yes

Diabetes Prediction

PREDICTION OF DIABETES MELLITUS RISK BASED ON HEALTH HISTORY USING XGBOOST

New Patient Data After Preprocessing:

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	v
0	-1.897	0.7651	1.0077	1.1098	1.1817	0.8396	1.0927	1.8662	

RESULT PREDICTION

Diabetes Risk: Positive

Diabetes Probability

Figure 10. Main display

5. CONCLUSION

This research demonstrates that the XGboost algorithm shows optimal and accurate performance in classifying the risk of DM based on health history to enhance early diagnosis, achieving an accuracy of 0.9808, a precision of 1.000, a recall of 0.9718, and an F1-score of 0.9857. With a high level of accuracy, the model successfully predicts whether an individual is at risk of having diabetes based on their medical history, in order to improve early diagnosis.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by the Universitas Islam Sultan Agung, Semarang, Indonesia in conducting and completing this study. We would like to express our appreciation to the psychologists who verified the dataset we collected; without their contributions, this research would not have been possible. In addition, we are very grateful to our academic advisors for their support and valuable feedback throughout this research process. This research is dedicated to all parties striving to improve the quality of mental health care for adolescents in Indonesia.

FUNDING INFORMATION

This research was funded by Universitas Islam Sultan Agung, Semarang, Indonesia.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Andi Riansyah	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ghufron	✓	✓		✓	✓	✓			✓	✓			✓	
Lailatul Fitriyah	✓	✓	✓		✓	✓	✓		✓	✓	✓			
Suyanto	✓	✓			✓				✓					

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known conflicts of interest, whether financial or non-financial, that could have influenced the conduct or outcomes of this research. The study was carried out independently, and all analyses, interpretations, and conclusions are based solely on the data obtained and the methodologies applied by the authors. The authors alone are responsible for the content and writing of this paper.

DATA AVAILABILITY

The dataset used in this study to develop and evaluate the XGBoost-based diabetes risk prediction model consists of anonymized patient health records derived from historical health data. All personal identifiers were removed prior to analysis to ensure compliance with ethical and privacy standards. Due to institutional and privacy restrictions, the raw data are not publicly available, but they may be obtained from the corresponding author [AR] upon reasonable request and with permission from the relevant data owner.

REFERENCES

[1] Y. Li *et al.*, "Association of age at diagnosis of type 2 diabetes mellitus with the risks of the morbidity of cardiovascular disease, cancer and all-cause mortality: evidence from a real-world study with a large population-based cohort study," *Diabetes Research and Clinical Practice*, vol. 217, Nov. 2024, doi: 10.1016/j.diabres.2024.111870.

[2] J. H. Wersäll, P. Adolfsson, J. Ekelund, G. Forsander, K. Åkesson, and R. Hanas, "Is diabetic ketoacidosis at diagnosis of type 1 diabetes associated with secondary episodes of ketoacidosis? a nationwide longitudinal study of Swedish children from 2012 to 2019," *Diabetes Research and Clinical Practice*, vol. 225, Jul. 2025, doi: 10.1016/j.diabres.2025.112282.

[3] M. A. Pirozzi *et al.*, "In newly diagnosed diabetes, hyperglycemia is associated with increased brain iron deposition, measured by MRI-based quantitative susceptibility mapping (QSM)," *Journal of Advanced Research*, 2025, doi: 10.1016/j.jare.2025.05.047.

[4] R. G. McCoy, L. Faust, H. C. Heien, S. Patel, B. Caffo, and C. Ngufor, "Longitudinal trajectories of glycemic control among U.S. adults with newly diagnosed diabetes," *Diabetes Research and Clinical Practice*, vol. 205, Nov. 2023, doi: 10.1016/j.diabres.2023.110989.

[5] R. M. C.-Larco, W. C. G.-Vilca, X. Xu, and A. B.-Ortiz, "Non-insulin-based markers of insulin resistance at diabetes diagnosis: a pooled analysis of 14 national health surveys," *Primary Care Diabetes*, 2025, doi: 10.1016/j.pcd.2025.06.005.

[6] S. Suyanto *et al.*, "Improving flipped classroom learning for patients with diabetes mellitus: an exploration into the influence of educational factors," *Healthcare in Low-Resource Settings*, vol. 12, no. 3, Oct. 2024, doi: 10.4081/hls.2024.12061.

[7] Z. Cao *et al.*, "Comparative effectiveness of different glycemic criteria for the diagnosis of gestational diabetes mellitus: a target trial emulation," *Diabetes Research and Clinical Practice*, vol. 225, Jul. 2025, doi: 10.1016/j.diabres.2025.112283.

[8] A. Marwanto, A. Riansyah, and M. K. Anwar, "An improvement of apple leaf diseases detection using convolutional neural network methods based on mobile systems," in *International Conference on Electrical Engineering, Computer Science and Informatics*, Sep. 2024, pp. 440–447. doi: 10.1109/EECSI63442.2024.10776345.

[9] S. Khan *et al.*, "XGBoost-enhanced ensemble model using discriminative hybrid features for the prediction of sumoylation sites," *BioData Mining*, vol. 18, no. 1, Dec. 2025, doi: 10.1186/s13040-024-00415-8.

[10] W. Liu, Z. Chen, and Y. Hu, "XGBoost algorithm-based prediction of safety assessment for pipelines," *International Journal of Pressure Vessels and Piping*, vol. 197, Jun. 2022, doi: 10.1016/j.ijpvp.2022.104655.

[11] Q. Wang *et al.*, "XGBoost algorithm assisted multi-component quantitative analysis with Raman spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 323, Dec. 2024, doi: 10.1016/j.saa.2024.124917.

[12] H. Yan, Z. He, C. Gao, M. Xie, H. Sheng, and H. Chen, "Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm," *Advanced Engineering Informatics*, vol. 54, Oct. 2022, doi: 10.1016/j.aei.2022.101789.

[13] J. Kuchyňka and O. Vyšata, "Automated sleep staging using sequential XGBoost and multi-scale temporal fusion," *Discover Artificial Intelligence*, vol. 5, no. 1, Dec. 2025, doi: 10.1007/s44163-025-00356-z.

[14] H. Dong, F. Wang, D. He, and Y. Liu, "The intelligent decision-making of copper flotation backbone process based on CK-XGBoost," *Knowledge-Based Systems*, vol. 243, May 2022, doi: 10.1016/j.knosys.2022.108429.

[15] C. Huang, X. Zhu, M. Lu, Y. Zhang, and S. Yang, "XGBoost algorithm optimized by simulated annealing genetic algorithm for permeability prediction modeling of carbonate reservoirs," *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-99627-z.

[16] B. Bastian, L. G. Smithers, A. Pape, W. Davis, K. Fu, and M. Francois, "Early screening and diagnosis of gestational diabetes mellitus (GDM) and its impact on perinatal outcomes," *Diabetes Research and Clinical Practice*, vol. 217, Nov. 2024, doi: 10.1016/j.diabres.2024.111890.

[17] Z. Lin *et al.*, "Tool wear prediction based on XGBoost feature selection combined with pso-bp network," *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-85694-9.

[18] Z. Zhang *et al.*, "Trajectories of metabolic risk factors prior to the onset of cardiovascular disease in patients with newly diagnosed diabetes," *Public Health*, vol. 243, Jun. 2025, doi: 10.1016/j.puhe.2025.105734.

[19] S. Kaptoge *et al.*, "Life expectancy associated with different ages at diagnosis of type 2 diabetes in high-income countries: 23 million person-years of observation," *The Lancet Diabetes & Endocrinology*, vol. 11, no. 10, pp. 731–742, Oct. 2023, doi: 10.1016/S2213-8587(23)00223-1.

[20] M. Chen, J. Xin, Q. Tang, T. Hu, Y. Zhou, and J. Zhou, "Explainable machine learning model for load-deformation correlation in long-span suspension bridges using XGBoost-SHAP," *Developments in the Built Environment*, vol. 20, Dec. 2024, doi: 10.1016/j.dibe.2024.100569.

[21] A. Fitrianingrum, M. Indriastuti, A. Riansyah, A. Basir, and D. Rusdi, "Business intelligence: alternative decision-making solutions on SMEs in Indonesia," in *Lecture Notes in Computer Science*, vol. 161, Springer, 2023. doi: 10.1007/978-3-031-26281-4_52.

[22] Z. Yan, H. Chen, X. Dong, K. Zhou, and Z. Xu, "Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost," *Expert Systems with Applications*, vol. 207, Nov. 2022, doi: 10.1016/j.eswa.2022.117943.

[23] Y. Qiu, N. Zhang, Z. Yin, Y. Wang, C. Xu, and P. Zhang, "Novel multi-spatial receptive field (MSRF) XGBoost method for predicting geological cross-section based on sparse borehole data," *Engineering Geology*, vol. 338, Aug. 2024, doi: 10.1016/j.enggeo.2024.107604.

[24] H. Sheng, Z. Ren, D. Wang, Q. Li, and P. Li, "Estimation and interpretation of interfacial bond in concrete-filled steel tube by using optimized XGBoost and SHAP," *Structures*, vol. 70, Dec. 2024, doi: 10.1016/j.istruc.2024.107669.

[25] S. Qian, T. Peng, Z. Tao, X. Li, M. S. Nazir, and C. Zhang, "An evolutionary deep learning model based on XGBoost feature selection and Gaussian data augmentation for AQI prediction," *Process Safety and Environmental Protection*, vol. 191, pp. 836–851, Nov. 2024, doi: 10.1016/j.psep.2024.08.119.

[26] Y. C. Chen, H. C. Su, S. M. Huang, C. H. Yu, J. H. Chang, and Y. L. Chiu, "Immune cell profiles and predictive modeling in osteoporotic vertebral fractures using XGBoost machine learning algorithms," *BioData Mining*, vol. 18, no. 1, Dec. 2025, doi: 10.1186/s13040-025-00427-y.

[27] P. M. L. Ching, X. Zou, D. Wu, R. H. Y. So, and G. H. Chen, "Development of a wide-range soft sensor for predicting wastewater BOD5 using an extreme gradient boosting (XGBoost) machine," *Environmental Research*, vol. 210, Jul. 2022, doi: 10.1016/j.envres.2022.112953.

[28] X. Guo, Q. Yang, Q. Wang, Y. Sun, and A. Tan, "Electromagnetic torque modeling and validation for a permanent magnet spherical motor based on XGBoost," *Simulation Modelling Practice and Theory*, vol. 136, Nov. 2024, doi: 10.1016/j.simpat.2024.102989.

[29] S. Xian, K. Chen, and Y. Cheng, "Improved seagull optimization algorithm of partition and XGBoost of prediction for fuzzy time series forecasting of COVID-19 daily confirmed," *Advanced Engineering Software*, vol. 173, Nov. 2022, doi: 10.1016/j.advengsoft.2022.103212.

[30] V. Q. Nguyen, V. L. Tran, D. D. Nguyen, S. Sadiq, and D. Park, "Novel hybrid MFO-XGBoost model for predicting the racking ratio of the rectangular tunnels subjected to seismic loading," *Transportation Geotechnics*, vol. 37, Nov. 2022, doi: 10.1016/j.trgeo.2022.100878.

[31] J. Henson *et al.*, "Sleep disorders in younger and middle-older age adults with newly diagnosed type 2 diabetes mellitus: a retrospective cohort study in >1million individuals," *Diabetes Research and Clinical Practice*, vol. 217, Nov. 2024, doi: 10.1016/j.diabres.2024.111887.

[32] D. Amilo, K. Sadri, E. Hincal, M. Farman, K. S. Nisar, and M. Hafez, "An integrated machine learning and fractional calculus approach to predicting diabetes risk in women," *Healthcare Analytics*, vol. 8, Dec. 2025, doi: 10.1016/j.health.2025.100402.

[33] C. Wang, X. Xu, S. Luo, M. Luo, S. Li, and J. Si, “Interpretable machine learning insights into the association between PFAS exposure and diabetes mellitus,” *Ecotoxicology and Environmental Safety*, vol. 302, Sep. 2025, doi: 10.1016/j.ecoenv.2025.118569.

[34] S. K. Jena, D. K. Behera, A. K. Jena, and J. K. Rout, “A novel ensemble machine learning framework for improved diabetes prediction and complication prevention,” in *Procedia Computer Science*, 2025, pp. 4008–4017. doi: 10.1016/j.procs.2025.04.652.

[35] N. A. J., Z. J. P., and R. M., “Cardiovascular disease (CVD) prediction using machine learning techniques with XGBoost feature importance analysis,” *International Journal of Multidisciplinary Research*, vol. 5, no. 5, pp. 1–17, 2023, doi: 10.36948/ijfmr.2023.v05i05.7715.

BIOGRAPHIES OF AUTHORS

Andi Riansyah is a computer scientist with a bachelor of Computer Science degree from Universitas Islam Sultan Agung and a Master of Information Systems degree from Universitas Diponegoro. He possesses over 7 years of research experience and has acquired a profound understanding of computer science, making noteworthy contributions to scientific advancement. He can be contacted at email: andi@unissula.ac.id.

Ghufron is a computer scientist with a Bachelor of Computer Science degree from Universitas Islam Sultan Agung and a Master of Information Systems degree from Universitas Diponegoro. He has made significant contributions to the development of science and has a deep understanding of his field. He is also active in various activities such as international seminars, training, and research in the fields of artificial intelligence and big data analysis. He can be contacted at email: ghufron@unissula.ac.id.

Lailatul Fitriyah is an undergraduate student at Universitas Islam Sultan Agung, with her main focus on computer science, particularly data engineering and artificial intelligence, after successfully completing numerous research projects and courses. Aside from his academic pursuits. She can be contacted at email: lailafitriyah@std.unissula.ac.id.

Suyanto is a nursing lecturer with a Bachelor of Nursing degree (S.Kep), a Master of Nursing degree (M.Kep), and a Medical-Surgical Nursing Specialist certification. (Sp.Kep.MB). The researcher focuses primarily on diabetes mellitus, with research areas encompassing peripheral neuropathy, diabetic wounds, and peripheral artery disease. The scope of study includes both hospital-based clinical settings and community-based interventions, aiming to improve the quality of life of diabetic patients through promotive, preventive, curative, and rehabilitative approaches. In addition, the researcher is also interested in innovative teaching and learning methods in nursing education to enhance the competencies of students and healthcare professionals in managing diabetes mellitus and its complications. He can be contacted at email: suyanto@unissula.ac.id.