SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Early Detection of Diabetes Using Random Forest Algorithm

Cindy Nabila Noviyanti[1*], **Alamsyah**[2]
[1,2] Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang, Indonesia

DOI: https://doi.org/10.52465/joiser.v2i1.245
Received 10 November 2023; Accepted 12 January 2024; Available online 29 January 2024

Article Info

Keywords:
Diabetes;
Machine learning;
Random forest;
Early detection

Abstract

Diabetes is one of the most chronic and deadly diseases. According to data from WHO in 2021, there were approximately 422 million adults living with diabetes worldwide, and this number is expected to continue to increase in the future due to various factors. Many studies have been conducted for early detection of diabetes by focusing on improving accuracy. However, a big problem in diabetes prediction is the selection of the right classification algorithm. This study aims to improve the accuracy of early detection of diabetes by implementing the Random Forest algorithm model. This research was conducted with the stages of data collection, data preprocessing, split data, modeling, and evaluation. This research uses the Pima Indian Diabetes data set. The results showed that the diabetes early detection model using the Random Forest algorithm produced an accuracy of 87%. This research shows that by using the Random Forest algorithm model, the performance of early detection of diabetes can be improved. However, there is still room for optimization of this performance, which is recommended for further research to carry out feature selection, data balancing, more complex model building, and exploring larger data.

1. Introduction

Diabetes mellitus is a chronic and deadly disease that has become a global health problem due to its continuously increasing prevalence from year to year [1]. According to the World Health Organization (WHO), diabetes mellitus is a chronic condition that occurs when the pancreas is no longer able to produce enough insulin or when the body is not able to use insulin effectively, which leads to the occurrence of increased levels of glucose (sugar) in the blood, which can cause various health complications [2]. Diabetes can be classified into four types: type 1, type 2, gestational, and pre-diabetes [3]. The long-term effects of the disease include blindness, kidney failure, amputation, and even death. According to WHO data, by 2021, there will be approximately 422 million adults living with

[*] ***Corresponding Author:***

Cindy Nabila Noviyanti,
Department of Computer Science,
Faculty of Mathematics and Natural Science, Universitas Negeri Semarang,
Semarang, Indonesia
Email: cindynabila1403@gmail.com

diabetes worldwide. This number is expected to continue to rise in the future to 625 million is due to factors such as an aging population, lifestyle changes, and increasing rates of obesity [4].

Research into early diagnosis of diabetes is a general need because of the number of people who have diabetes around the world. The increase in cases is due to lifestyle training and unhealthy diet. Automatic diagnosis can identify fatal complications such as heart disease, kidney problems, nerve damage, and eye problems. Nevertheless, automatic screening for diabetes can increase a huge financial burden for people and the health system in general. Based on the many cases of people with diabetes nowadays, it is necessary to take early action to address problems in the future, namely by making early predictions about diabetes.

This prediction of diabetes can be done by utilizing some of the data of patients with diabetes that has been stored in a database to create a pattern for determining diabetes [5]. Machine learning (ML) technology has been widely used in a variety of fields [6], especially in the early detection of diabetes. Because of this over the years, Machine learning has solved many sophisticated and complex problems in a variety of fields such as marketing, business and retail, natural language processing, health, robotics, imaging, sound, gaming, etc [7].

The previous researchers [8], created a model for diabetes prediction using two datasets: Pima Indian Diabetes data and early-stage diabetes risk prediction to detect diabetes using several models such as Naïve Bayes, K-nearest neighbor, Decision Tree, Logistic Regression, Random Forest, SVM, AdaBoost classifier, Gradient classifier, and extra tree classifier. Among the several proposed methods, the Super Learner Classifier model has the highest accuracy of 86% for the Pima Indian dataset. For the data set of early-stage diabetes risk prediction, the KNN model has the highest accuracy of 97%. The study [9] used the DLPD (Deep Learning for Predicting Diabetes) model to predict diabetes and produced an accuracy of 94,02,174% for the Diabetes-type data sets and the Indian Pima-diabetics data sets are accurate at 99.4112%. The experimental results show improvements in the recommended formats compared to the target method. In addition, the model of prediction of diabetes disease also has been developed [10], use of short-term memory (LSTM), convoluted nerve network (CNN), and combinations to extract complex time dynamic characteristics from HRV input data. These characteristics are transferred to a supported vector machine (SVM) for classification. The proposed classification system can help doctors diagnose diabetes with an ECG signal with a very high accuracy of 95.7 percent. Further studies conducted by [11], used Gradient Boosting, Logistic Regression, and Naive Bayes for diabetes diagnosis to obtain 86% accuracy for gradient boosting, 79% for logistical regression, and 77% for naive bayes. Further, [12] uses logistical regression, support vector machines, nearest K neighbors, random forests, naive Bayes, boot gradient algorithms, and predictive machine learning models built and monitored in this study. With predictive capacities of 86.28% and 86.29%, respectively, learning-based models from random forest predictions and booting gradients proved to be the best prediction models. Besides, research conducted by [13], has detected miletus diabetes early using predictive analysis. The results showed the decision tree algorithm and the random forest had the highest specifications of 98.20% and 98.00%, each holding the best for the analysis of diabetes data. Naive Bayesian results state the best accuracy of 82.30%. Further, [14] developed a new super-learning model and managed to obtain the best accuracy results in the detection of diabetes mellitus compared to the base-learning for the prediction of risk of early-stage diabetes (99.6%), PIMA (92%), and diabetes 130-US hospitals (98%) datasets, respectively.

2. Method

The method used in this research is proposed in the form of a flowchart in Figure 1 which starts with data collection, data pre-processing, split data, modeling, and evaluation.

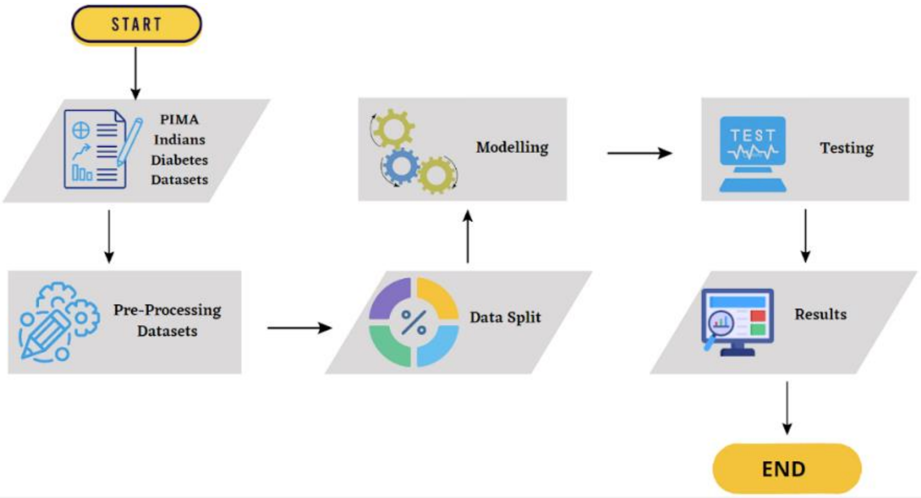

Figure 1. Flowchart of the proposed method

Each step in this study was conducted sequentially. A detailed explanation of each step taken in this research can be seen as follows.

2. 1. Data Collection

To detect diabetes in a person, this research uses the Pima Indian dataset which is publicly accessible on the Kaggle platform [15]. The Pima Indian Dataset has been one of the most frequently used datasets in machine learning and deep learning research to detect diabetes because it has a large sample size and a variety of features that include health history and demographic characteristics [16]. The dataset consists of 768 individual data with an age range of 21 to 81 years. A total of 500 data records are in the form of a negative class, namely individuals who are not diagnosed with diabetes. While the rest, namely 268 is the amount of data from individuals diagnosed with diabetes. The dataset consists of 8 variables, namely number of pregnancies, glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour insulin, body time index, hereditary disease history, and age.

2. 2. Preprocessing Data

Preprocessing data is done to prepare the data to run well in the modeling process. Data preprocessing can be done by cleaning data, changing data types, and others. In the preprocessing stage of this research, NaN values or empty rows in the dataset were identified. However, after this identification, it is found that the dataset is complete, there are no rows containing NaN. Then checking the value of 0 in each feature is done. It is possible because as in the example of the pregnancies feature, it makes sense if someone has never been pregnant, and is indicated by the pregnancies data 0. However, there are features such as glucose, blood pressure, skin thickness, insulin, and BMI, which are impossible if the value is 0. Then the mode value is entered into the data that contains 0 in each of these variables.

2. 3. Split Data

At this stage, dataset division is carried out where patient data that has been processed in the previous stage is divided into 3 sets, namely training data, validation data, and testing data. The training data here is used to train the Random forest model. Validation data is used to test the performance of the model during training. Testing data is used to test the performance of the model after training. The percentage of data division is 60% as training data, 25% as validation data, and 20% as testing data.

2. 4. Random Forest

The Random Forest algorithm is proposed to be able to increase the accuracy of early diabetes prediction. This algorithm is a machine learning classifier that works by building a decision tree. These algorithms can also be used for regression and classification [17]. Random Forest is part of the Supervised Learning group developed by Leo Breiman [18]. This method is one of the most accurate classification methods used in making predictions, can handle enormous amounts of variable inputs without overfitting, and helps eliminate correlations between decision trees such as the characteristic ensemble methods [19], [20]. The random forests can be used in biomedical research [21], especially in the detection of diabetes. In addition, Random Forest has a lower error rate on diabetes data than its other classification algorithms and has good performance in diabetes classification [22].

2.5 Evaluation

Confusion metric is one of the assessment methods used to analyze the performance of the ensemble learning model which is also used in this study. An incorrect matrix can give an overall picture of the predictions and real circumstances given by the algorithmic model. The mixed matrix has four important elements. Among them is true positive (TP), which is the real amount of data contained in a positive class and predicted by the model as a positive. Then True Negative (TN), which reflects the real portion of data that is in the negative class, and the prediction of the model is in a negative class. False positive (FP) is the amount of information that is actually in the negativity class but predicts it as a positive class. Using the confusion matrix, we can calculate and evaluate the performance of the model created using indicators such as accuracy, precision, recall, and F1 scores. Detailed estimates are listed below.

1. Accuracy

Accuracy is a value that indicates how accurate the model is in predicting the entire data. Measured by the formula:

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

2. Precision

Precision measures the degree to which work predicted to be fake is fake. We can calculate it with the following formula:

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

3. Recall (Sensitivity)

Recall measures the extent to which the model successfully detects fake jobs overall. Recall calculated by the following formula:

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

4. F1-Score

The F1-Score is a combination of precision and recall into a single metric that yields the overall model performance. This can be calculated by the following formula:

$$F1 - Score = \frac{2 * (Presisi * Recall)}{(Presisi + Recall)} \tag{4}$$

3. Results and Discussion

In this study, random forest model algorithms were used to recognize patterns of diabetes. Random forest models are designed using Google Collab. In the early stages, Pima Indian was divided into three categories, data training, data validation, and data testing. After that, we input the median of the features glucose, blood pressure, skin thickness, insulin, and BMI, which is not possible if the value is 0. The next step is Exploratory Data Analysis to identify patterns, relationships, trends, or anomalies in the data that can provide valuable insights into decision-making. Where in this step we find a comparison between patients with diabetes and patients without diabetes, as shown in Figure 2.

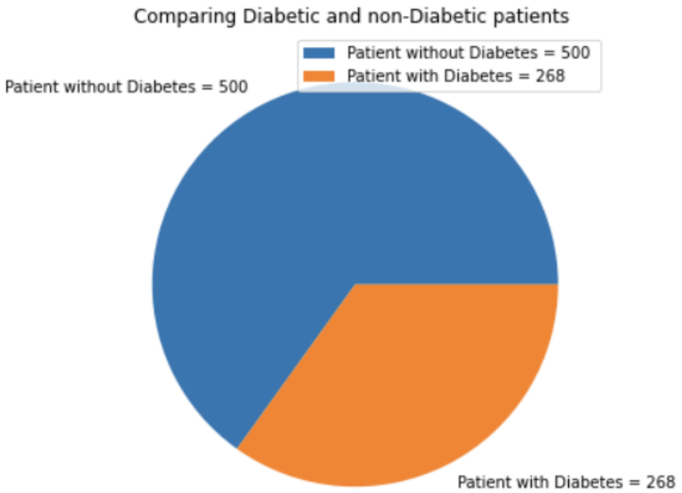

Figure 2. Comparison between patients with diabetes and patients without diabetes

The relationship pattern between the features in the dataset and its class, namely 'outcome' which consists of classes 1 (diabetics) and 0 (healthy individuals) can be seen in Figure 3 below.

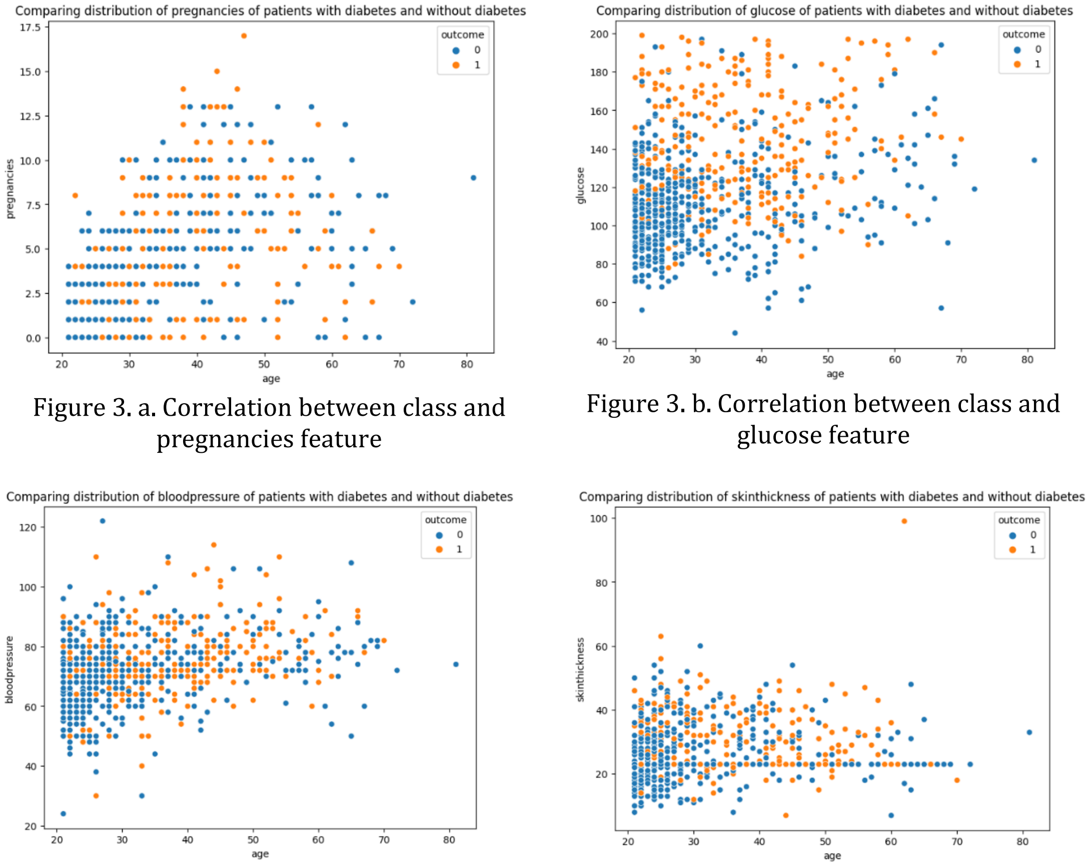

Figure 3. a. Correlation between class and pregnancies feature

Figure 3. b. Correlation between class and glucose feature

Figure 3. c. Correlation between class and blood pressure feature

Figure 3. d. Correlation between class and skinthickness feature

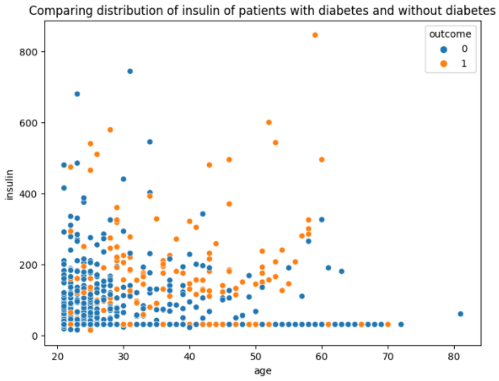

Figure 3. e. Correlation between class and age feature

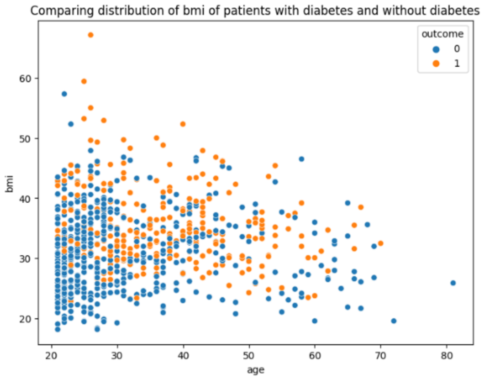

Figure 3. f. Correlation between class and BMI feature

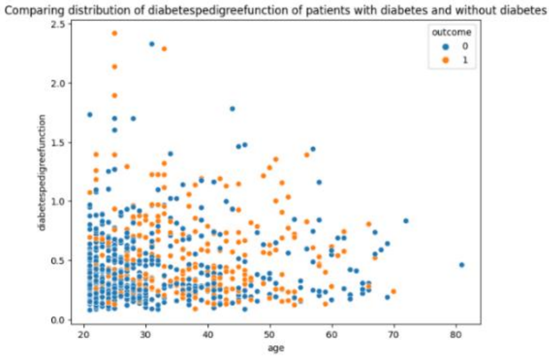

Figure 3. g. Correlation between class and diabetes pedigree function feature

Figure 3. Correlation between data and class

Based on the correlation between the pregnancy variable and the age feature, it can be seen that age and pregnancies are spread across healthy and diabetic individuals. However, there is a slight tendency that healthy individuals (non-diabetics) are those who are around 30 years old and below and do not often get pregnant. For the correlation of glucose and age, we can see a pattern where healthy individuals (non-diabetics) are those who are mostly less than 40 years old and their glucose is less than 140. For the correlation of blood pressure and age, we can see a pattern where healthy individuals (non-diabetics) are those who are around 30 years old and less than 100. For the correlation between skinthickness and age, we can see a pattern where healthy individuals (not diabetics) are those whose ages are around 30 years and below and whose skinthickness is around 40 and below. For the correlation of insulin and age, we can see a pattern where healthy individuals (non-diabetics) are those whose age is around 30 years and below and their insulin is less than 400. However, there are also some individuals aged 50 to 60 years who do not suffer from diabetes because their insulin is very minimal. For the correlation of BMI and age, we can see a pattern where healthy individuals (not diabetics) are those whose ages are around 40 years and below and their BMI is less than 40. For the correlation of diabetes pedigree function and age, we can see a pattern where healthy individuals (not diabetics) are those whose ages are around 40 years and below and their diabetes pedigree function is less than 0.8.

The model evaluation stage is carried out to obtain measurement performance, where the metric used in this research is accuracy. From the process carried out and the model that has been created, this research obtained a training accuracy of 78.18% and an accuracy on testing of 87%. The following is a comparison of the findings of the research with previous research in Table 1.

Table 1. Comparison of accuracy

Algorithm	Accuracy result
Super Learner Classifier [8]	86%
Random Forest (Proposed Method)	87%

Research conducted by applying the Random Forest algorithm has higher accuracy results compared to previous research using the super learner classifier model algorithm on the same dataset. In the data preprocessing stage before modeling, the percentage of data division into training, validation, and testing data can be said to influence and help in obtaining these results. So this research shows that the model built, namely Random Forest together with the preprocessing stage carried out, is superior in classifying individuals suffering from diabetes with healthy individuals.

4. Conclusion

In this study, early detection of diabetes was carried out using the Random Forest algorithm. This research shows the superiority of the Random Forest algorithm over the use of other algorithms in the early detection of diabetes using the same data. In the model built, this study achieved an accuracy of 87%. This result shows an increase in performance from previous research. However, there is room for further development. Future research is expected to consider other factors to improve detection accuracy such as the implementation of data balancing methods, application of feature selection, building more complex models, and exploring larger data.

References

[1] S. Kaul and Y. Kumar, "Artificial Intelligence-based Learning Techniques for Diabetes Prediction: Challenges and Systematic Review," *SN Computer Science*, vol. 1, no. 6. Springer, Nov. 01, 2020. doi: 10.1007/s42979-020-00337-2.

[2] Institute of Electrical and Electronics Engineers, *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS).*

[3] J. February, O. S. Abe, O. O. Obe, O. K. Boyinbode, and O. N. Biodun, "Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 1, pp. 430–439, 2021, doi: 10.30534/ijatcse/2021/641012021.

[4] G. Li, S. Peng, C. Wang, J. Niu, and Y. Yuan, "An energy-efficient data collection scheme using denoising autoencoder in wireless sensor networks," *Tsinghua Science and Technology*, vol. 24, no. 1, pp. 86–96, 2019, doi: 10.26599/TST.2018.9010002.

[5] P. Arsi and O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 3, no. 3, pp. 290–294, 2018, doi: 10.30591/jpit.v3i3.1008.

[6] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

[7] F. A. Jaber and J. W. James, "Early Prediction of Diabetic Using Data Mining," *SN Computer Science*, vol. 4, no. 2, pp. 1–7, 2023, doi: 10.1007/s42979-022-01594-z.

[8] S. Saxena, D. Mohapatra, S. Padhee, and G. K. Sahoo, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms," *Evolutionary Intelligence*, no. 0123456789, 2021, doi: 10.1007/s12065-021-00685-9.

[9] H. Zhou, R. Myrzashova, and R. Zheng, "Diabetes prediction model based on an enhanced deep neural network," *Eurasip Journal on Wireless Communications and Networking*, vol. 2020, no. 1, Dec. 2020, doi: 10.1186/s13638-020-01765-7.

[10] A. S. Mahajan, "Medical Diagnosis of Diabetes Using Deep Learning Techniques and Big data Analytics," *Journal of Emerging Technologies and Innovative Research*, vol. 7, no. 4, pp. 1490–1497, 2020.

[11] R. Birjais, A. K. Mourya, R. Chauhan, and H. Kaur, "Prediction and diagnosis of future diabetes risk: a machine learning approach," *SN Applied Sciences*, vol. 1, no. 9, pp. 1–8, 2019, doi: 10.1007/s42452-019-1117-9.

[12] L. J. Muhammad, E. A. Algehyne, and S. S. Usman, "Predictive Supervised Machine Learning Models for Diabetes Mellitus," *SN Computer Science*, vol. 1, no. 5, pp. 1–10, 2020, doi: 10.1007/s42979-020-00250-8.

[13] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.

[14] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk," *Medical and Biological Engineering and Computing*, vol. 61, no. 3, pp. 785–797, 2023, doi: 10.1007/s11517-022-02749-z.

[15] "UCI Machine Learning. Pima Indians Diabetes Database." [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[16] R. Rousyati, A. N. Rais, E. Rahmawati, and R. F. Amir, "Prediksi Pima Indians Diabetes Database Dengan Ensemble Adaboost Dan Bagging," *EVOLUSI : Jurnal Sains dan Manajemen*, vol. 9, no. 2, pp. 36–42, 2021, doi: 10.31294/evolusi.v9i2.11159.

[17] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, pp. 1–14, 2020, doi: 10.1007/s13755-019-0095-z.

[18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[19] T. N. Nuklianggraita, A. Adiwijaya, and A. Aditsania, "On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier," *Jurnal Infotel*, vol. 12, no. 3, pp. 89–96, 2020, doi: 10.20895/infotel.v12i3.485.

[20] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest," *Journal of Information System Exploration and Research*, vol. 1, no. 1, pp. 49–70, 2023.

[21] S. Shah, X. Luo, S. Kanakasabai, R. Tuason, and G. Klopper, "Neural networks for mining the associations between diseases and symptoms in clinical notes," *Health Information Science and Systems*, vol. 7, no. 1, pp. 1–9, 2019, doi: 10.1007/s13755-018-0062-0.

[22] S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," *2019 International Conference on Computer and Information Sciences, ICCIS 2019*, pp. 1–4, 2019, doi: 10.1109/ICCISci.2019.8716405.