Content lists available at Pubmedia

International Journal of Health and Information System (IJHIS)

journal homepage : https://ijhis.pubmedia.id/index.php/ijhis

Article

Diagnosis of Stroke and Diabetes Mellitus With Classification Techniques Using Decision Tree Method

Mudafiq Riyan Pratama [1*], Arinda Lironika Suryana [2], Gamasiano Alfiansyah [3], Zora Olivia [4], Ida Nurmawati [5], Prawidya Destarianto [6]

[1] Health Information Management Study Program, Politeknik Negeri Jember; mudafiq.riyan@polije.ac.id
[2] Clinical Nutrition Program, Politeknik Negeri Jember; arinda@polije.ac.id
[3] Health Information Management Study Program, Politeknik Negeri Jember; gamasiano.alfiansyah@polije.ac.id
[4] Clinical Nutrition Study Program, Politeknik Negeri Jember; zora@polije.ac.id
[5] Health Information Management Study Program, Politeknik Negeri Jember; ida@polije.ac.id
[6] Informatics Engineering Study Program, Politeknik Negeri Jember; prawidya@polije.ac.id

* Correspondence: mudafiq.riyan@polije.ac.id

Abstract: Stroke is a cerebral vascular disease characterized by the death of brain tissue that occurs due to reduced blood and oxygen flow to the brain. Ischemic stroke is associated with diabetes mellitus, therefore it is important to identify the risk factors that cause stroke and DM by diagnostic cause of the disease. This study aimed to classify and compare accuracy tests on medical record data sets for stroke and DM. This study analyzed the diagnosis of stroke and DM using Decision Tree. The risk factors consisted of gender, age, blood pressure, nutritional status, smoking, history of DM, and history of hypertension. The results of the analysis using the Decision Tree method showed that the accuracy rate was 86.67%, which means that the modeling has a good level of correctness of the prediction results. We conclude that the Decision Tree method was an accurate method for detecting stroke and DM.

Keywords: Classification; Decision Tree; Diabetes Mellitus; Stroke

Citation: M. R. Pratama, A. L. Suryana, G. Alfiansyah, Z. Olivia, I. Nurmawati, and P. Destarianto, "Diagnosis of Stroke and Diabetes Mellitus with Classification Techniques using Decision Tree Method at dr. Soebandi Hospital", *IJHIS*, vol. 2, no. 1, pp. 1–8.

Received: 11-01-2024
Accepted: 20-02-2024
Published: 09-03-2024

1. Introduction

The development of information technology can be utilized in processing health data in hospitals. The health industry has a large amount of health data, but most of this data has not been processed properly so that the resulting information is less effective for use, such as in making decisions regarding predictions of patient disease and patient health [1]. One of the hospitals located in Jember Regency is Dr Soebandi Regional Hospital. Health services at Dr Soebandi Regional Hospital include health services for stroke patients and diabetes mellitus patients.

Stroke and Diabetes mellitus (DM) are non-communicable diseases whose prevalence increases every year. DM is a chronic disease that affects many people throughout the world [2]. DM is characterized by an increase in blood glucose levels due to the inability of the pancreas to produce insulin effectively [3]. Meanwhile, stroke is the second biggest disease that causes death [4], [5]. Stroke is a multicausal disease caused by many factors [6]. DM is the second most common risk factor after hypertension with an increase in the relative risk of ischemic stroke of 1,6 to 8 times [7]–[9].

Early diagnosis for DM and stroke sufferers is important. Uncontrolled increases in blood glucose levels can cause damage to the body's organs and increase the risk of diseases including heart disease, kidney disease, stroke and other complications [10], [11]. Meanwhile, stroke is the main cause of physical disability in adults and the second leading cause of death in upper middle income countries [12]. The long-term impacts of stroke

can be depression, functional dependency, and separation from society [13]. Therefore, the development of accurate classification technology to diagnose DM and stroke is important.

The method that can be used to diagnose DM and stroke is the Decision Tree method. Decision Tree is a classification method that uses a tree representation, each node represents an attribute, tree branches represent the value of the attribute, and leaves represent the class [14]–[16]. Decision tree has advantages in data processing, (1) decision tree is easy for users to understand; (2) decision tree has a high level of performance with a minimum of large amounts of data in a short time; and (3) decision trees can be used in various data processing applications on various platforms or software [17], [18].

The aim of this study was to classify and compare accuracy tests on medical record data sets for stroke and DM. By using the decision tree algorithm, it was expected that the classification method developed in this research can provide accurate results in diagnosing stroke and DM. This will enable doctors to identify stroke and DM patients early so that patients can be treated appropriately. Therefore, this research has the potential to contribute to improving the quality of life of stroke and DM patients, as well as preventing disease complications that threaten the patient's health. So this research can be used as decision support in determining stroke and DM as early prevention.

2. Materials and Methods

2.1 Data collection

This research was carried out at the Dr Soebandi Regional Hospital, Jember, Indonesia. Data collection was carried out using secondary data from patient medical records. The inclusion criteria used in selecting medical records were medical records that had complete data regarding gender, age, blood pressure, nutritional status, smoking history, diabetes mellitus, blood sugar level, and history of hypertension. Meanwhile, the exclusion criteria were outpatient medical records and data in medical records that were illegible. This research had received approval from the Jember State Polytechnic Health Research Ethics Commission with ethical approval letter number 1060/PL17.4/PG/2023.

2.2 Data Preprocessing

Data obtained from patient medical records would be preprocessed before being processed by eliminating data that was not appropriate based on predetermined inclusion criteria and exclusion criteria, so that 98 patient data were obtained.

3. Results and Discussion

The data processed from this research amounted to 98 patient medical record data at the Neurology Polyclinic of Dr. Soebandi Regional Hospital that consisting of 8 variables: gender, age, blood pressure, nutritional status, smoking history, Diabetes Mellitus, blood sugar level, and history of hypertension. Of these 8 variables, the diagnosis label consists of 2 diagnoses, namely Stroke and Diabetes Mellitus (DM).

3.1 Modeling Decision Tree With C4.5 Algorithm

The Decision Tree method modeling was carried out using the RapidMiner Studio tools. Before modeling, the 98 data were divided into two parts, namely training data and testing data. Training data is used as a knowledge base to create data modeling whose output becomes a decision tree. Meanwhile, data testing is used to test data and calculate the accuracy of the model that has been created.

The C4.5 algorithm is an algorithm used in data mining to build decision tree models. This algorithm was developed by Ross Quinlan in 1993 which was a development of the previous algorithm known as ID3. The C4.5 algorithm uses machine learning concepts to generate a prediction model based on the given data. Essentially, these algorithms focus on understanding the structure of the data and classifying it into appropriate groups. This process allows users to explore relationships between variables and gain a deep understanding of the existing dataset.

The percentage distribution of training and testing data was 70%:30% with 68 training data and 30 testing data. The 68 training data consisted of 34 data with a DM diagnosis label and 34 data with a Stroke diagnosis label. The training data processed can be seen in the table below.

Table 1. Training Data

No.	Sex	Age	Blood Pressure	Nutrition Status	Smoking	DM	Blood Sugar Level	Hypertension	Diagnosis
1.	Female	Adult	Pre-Hypertension	Normal	No	No	Normal	No	Stroke
2.	Female	Elderly	Pre-Hypertension	Normal	No	No	Prediabetes	Yes	Stroke
3.	Female	Young Old	Pre-Hypertension	Underweight	No	No	Prediabetes	Yes	Stroke
4.	Male	Old	Pre-Hypertension	Normal	No	No	Prediabetes	Yes	Stroke
5.	Female	Middle Age	Pre-Hypertension	Normal	No	Yes	Diabetes	Yes	Stroke
6.	Female	Elderly	Hypertension Stage 1	Obesity	No	Yes	Diabetes	Yes	DM
7.	Female	Adult	Pre-Hypertension	Normal	No	Yes	Diabetes	Yes	DM
8.	Female	Elderly	Pre-Hypertension	Normal	No	Yes	Diabetes	No	DM
9.	Male	Middle Age	Pre-Hypertension	Normal	Yes	Yes	Prediabetes	No	DM
…	…	…	…	…	…	…	…	…	…
68.	Female	Elderly	Level 1 Hypertension	Overweight	No	Yes	Diabetes	Yes	DM

The training data is used as a knowledge base for the c4.5 algorithm in forming a rule base in the form of a decision tree. This training data amounts to 68 data, which is 70% of the total available data. From the training data, a modeling framework was created using RapidMiner Studio according to the image below.

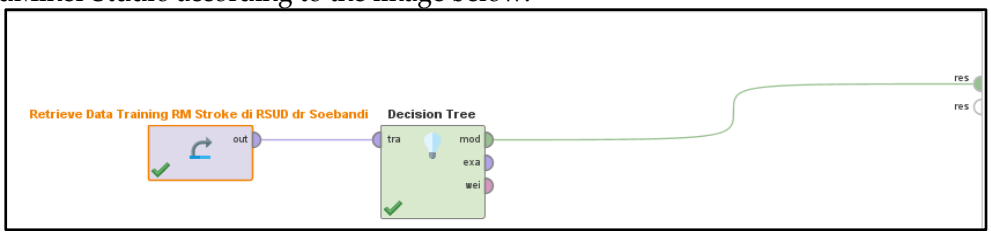

Figure 1. Decision Tree Model in RapidMiner Studio

From the modeling created, the Decision Tree structure is as follows.

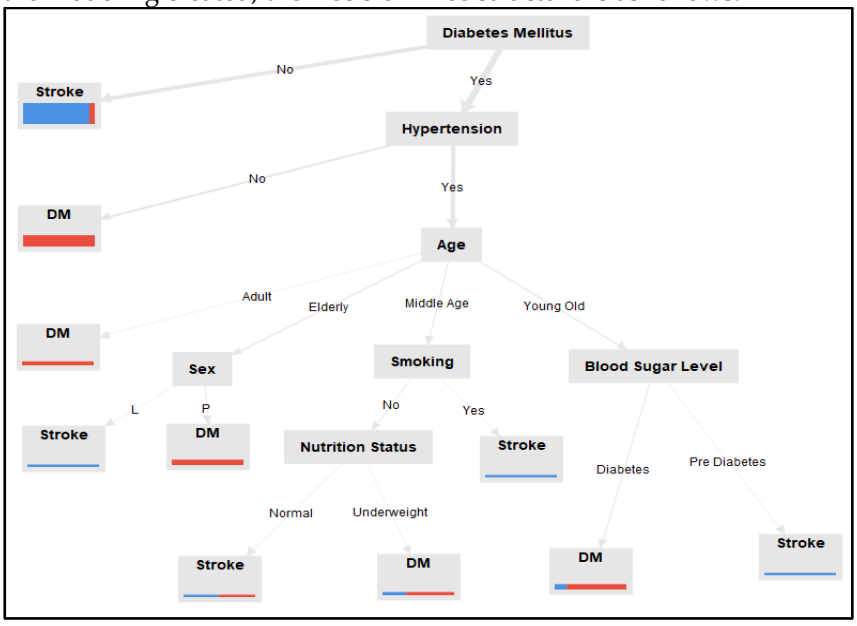

Figure 2. Decision Tree Model for Classification of Stroke and DM Using the C4.5 Algorithm

Figure 2 is a role model obtained from processing training data which forms a decision tree. The root of the decision tree is Diabetes Mellitus (DM), which is the main factor determining DM and stroke.

1. If the patient has a history of DM, they must be checked for hypertension.
2. If the patient does not have hypertension, then he is identified as having DM, whereas if he has hypertension then his age must be checked.
3. If the person is an adult, the algorithm identifies it as DM. If the person is elderly, the gender must be checked.
4. If the gender is male then the algorithm identifies it as Storek's disease.
5. And other roles according to figure 2.

3.2 Model Performance Testing

Performance testing was carried out to test the level of accuracy, precision and recall of the model that has been created at the modeling stage. The testing was carried out by adding testing data, Apply Model, and Performance. The testing data used was 30 data.

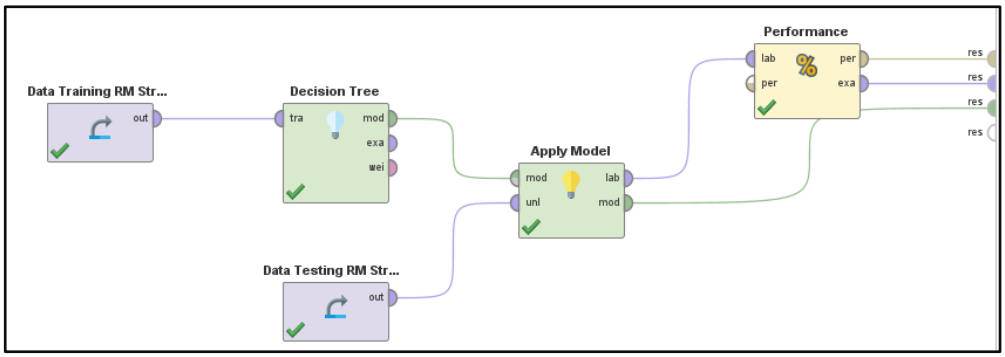

Figure 3. Decision Tree Model Performance Test Settings

The test results can be seen in the following table. The Factual Diagnosis column is real diagnosis data that occurs in patients, while the Predictive Diagnosis is the classification result predicted by the Decision Tree model.

Table 2. Testing Results

No.	Sex	Age	Blood Pressure	Nutrition Status	Smoking	DM	Bloos Sugar Levels	Hypertension	Diagnosis Actual	Diagnosis Prediction
1	Male	Adult	Pre Hypertension	Underweight	No	Yes	Diabetes	No	DM	DM
2	Male	Adult	Normal	Underweight	No	Yes	Diabetes	No	DM	DM
3	Female	Adult	Pre Hypertension	Normal	No	Yes	Diabetes	No	DM	DM
4	Female	Adult	Pre Hypertension	Normal	No	Yes	Diabetes	No	DM	DM
5	Female	Adult	Pre Hypertension	Normal	No	Yes	Pre Diabetes	No	DM	DM
6	Male	Elderly	Pre Hypertension	Normal	No	No	Pre Diabetes	No	Stroke	Stroke
7	Male	Elderly	Pre Hypertension	Normal	Yes	Yes	Diabetes	No	DM	DM
8	Female	Young Old	Normal	Obesity	No	Yes	Pre Diabetes	No	DM	DM
9	Male	Young Old	Normal	Normal	No	Yes	Pre Diabetes	No	Stroke	DM
10	Male	Young Old	Pre Hypertension	Normal	No	No	Pre Diabetes	No	Stroke	Stroke
11	Female	Middle Age	Pre Hypertension	Underweight	No	Yes	Diabetes	No	DM	DM
12	Female	Middle Age	Pre Hypertension	Normal	No	Yes	Diabetes	No	DM	DM
13	Female	Middle Age	Pre Hypertension	Normal	No	Yes	Diabetes	No	DM	DM
14	Female	Middle Age	Hypertension Stage 2	Normal	No	No	Pre Diabetes	No	Stroke	Stroke

No.	Sex	Age	Blood Pressure	Nutrition Status	Smoking	DM	Bloos Sugar Levels	Hypertension	Diagnosis Actual	Diagnosis Prediction
15	Male	Elderly	Hypertension Stage 1	Normal	Yes	No	Normal	Yes	DM	Stroke
16	Male	Elderly	Isolated Systolic Hypertension	Normal	Yes	Yes	Pre Diabetes	Yes	DM	Stroke
17	Male	Elderly	Pre Hypertension	Normal	Yes	Yes	Pre Diabetes	Yes	DM	Stroke
18	Male	Elderly	Hypertension Stage 2	Normal	Yes	No	Diabetes	Yes	Stroke	Stroke
19	Male	Elderly	Pre Hypertension	Overweight	Yes	Yes	Diabetes	Yes	Stroke	Stroke
20	Male	Elderly	Pre Hypertension	Normal	Yes	No	Pre Diabetes	Yes	Stroke	Stroke
21	Male	Elderly	Pre Hypertension	Normal	No	Yes	Pre Diabetes	Yes	Stroke	Stroke
22	Female	Elderly	Hypertension Stage 2	Normal	No	Yes	Diabetes	Yes	DM	DM
23	Female	Elderly	Hypertension Stage 2	Normal	No	Yes	Diabetes	Yes	DM	DM
24	Female	Elderly	Hypertension Stage 1	Overweight	No	Yes	Diabetes	Yes	DM	DM
25	Male	Old	Pre Hypertension	Underweight	Yes	No	Pre Diabetes	Yes	Stroke	Stroke
26	Male	Middle Age	Hypertension Stage 2	Overweight	No	No	Pre Diabetes	Yes	Stroke	Stroke
27	Female	Middle Age	Hypertension Stage 1	Overweight	No	No	Pre Diabetes	Yes	Stroke	Stroke
28	Male	Middle Age	Hypertension Stage 2	Normal	Yes	No	Pre Diabetes	Yes	Stroke	Stroke
29	Male	Middle Age	Pre Hypertension	Normal	No	No	Normal	Yes	Stroke	Stroke
30	Female	Middle Age	Pre Hypertension	Normal	No	No	Pre Diabetes	Yes	Stroke	Stroke

One of the method used to measure data mining performance using classification techniques is the Confusion Matrix. Evaluation with the Confusion Matrix is carried out by predicting the level of truth of the data. Confusion Matrix can measure accuracy, precision and recall levels. The Confusion Matrix is depicted in the following form.

Table 3. Confusion Matrix Form

		Predictive Value	
		+	-
Actual Value	+	TP	FN
	-	FP	TN

Information:

1. TP (True Positive): data that is actually positive and predicted to also be positive. In this case, it means that the actual diagnosis data is worth a Stroke and the predictive diagnosis is also worth a Stroke
2. FN (False Negative): data that is actually positive, but is predicted to be negative. In this case, it means that the actual diagnosis data is worth Stroke, but the predictive diagnosis is worth DM
3. FP (False Positive): data that is actually negative, but is predicted to be positive. In this case, it means that the actual diagnosis data is worth DM, but the predictive diagnosis is worth Stroke

4. TN (True Negative): data that actually has a negative value and is predicted to also have a negative value. In this case, it means that the actual diagnosis data is worth DM, and also the predictive diagnosis is worth DM.

Therefore, based on the 30 test data, the Confusion Matrix is obtained in the form below.

Table 4. Confusion Matrix from Test Data

TP = 13	FN = 1
FP = 3	TN = 13

Then, performance measurements are calculated based on the following levels of Accuracy, Precision, and Recall.

3.2.1 Accuracy

Accuracy is the ratio of correct predictions (positive and negative) to the entire data. So the output from this calculation is the percentage of Stroke and DM data that was predicted correctly. The formula for this Accuracy calculation is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{13 + 13}{13 + 13 + 3 + 1} = \frac{26}{30} = 0.8667 = 86.67\,\%$$

So the accuracy level of the correctness results is 86.67%, that means that the modeling has a good level of correctness of the prediction results.

3.2.2 Precision

Precision is the ratio of true positive predictions compared to the total positive prediction results. This means that data that is worth a Stroke diagnosis is compared with all predictions of a Stroke diagnosis. Precision measures the level of truth in predicting Stroke. The formula for calculating Precision is as follows.

$$Precision = \frac{TP}{TP + FP} = \frac{13}{13 + 3} = \frac{13}{16} = 0.8125 = 81.25\,\%$$

Based on the Precision results, it means that the level of truth in predicting a stroke diagnosis is 81,25%, so the error in predicting stroke is only 18,75%. So it can be stated that this modeling has a good level of truth in stroke prediction results.

3.2.3 Recall

Recall is the ratio of true positive predictions compared to all actual positive data. Actual positive data was obtained from TP and FN. So in this sudy, recall was used to calculate the percentage of Stroke data that was predicted correctly compared to the total actual Stroke data. The Recall calculation formula is:

$$Recall = \frac{TP}{TP + FN} = \frac{13}{13 + 1} = \frac{13}{14} = 0.9286 = 92.86\,\%$$

Based on the Recall results, it means that Decision Tree modeling can predict stroke with an accuracy of 92,86%. So only 7,14% of Stroke data were not successfully predicted as Stroke.

3.3 Analysis

Based on the modeled decision tree structure, the results obtained show that the root of the Decision Tree is Diabetes Mellitus (DM). If the patient does not have DM, the decision tree leads to the decision that the patient has a stroke. This is because in the training data there were 25 patients who did not have DM but were diagnosed with stroke. Meanwhile, only 9 patients have DM and were diagnosed with stroke. This means that the data on patients who do not suffer from DM but suffer from stroke is 73%.

4. Conclusions

This study tested the model using the C4.5 decision tree algorithm using medical record data from patients suffering from stroke and DM. The resulting model has an accuracy of 86.67%, precision of 81.25%, and recall of 92.86%. Thus, it can be concluded that the research results can provide accurate problem solving for stroke and DM.

5. Acknowledgments

The researcher would like to thank the Directorate of Research and Community Service of the Ministry of Education, Culture, Research and Technology for funding so that this research could be completed. Thanks were also conveyed to RSD Dr Soebandi for the research permission granted so that it could run smoothly.

References

[1] A. Rohman, "Application of the Adaboost-Based C4.5 Algorithm for Heart Disease Prediction," *Pandanaran Univ. Sci. Mag.*, vol. 11, no. 26, pp. 40–49, 2013.

[2] N. Camerlingo, M. Vettoretti, S. Del Favero, A. Facchinetti, P. Choudhary, and G. Sparacino, "Generation of post-meal insulin correction boluses in type 1 diabetes simulation models for in-silico clinical trials: More realistic scenarios obtained using a decision tree approach," *Comput. Methods Programs Biomed.*, vol. 221, p. 106862, 2022, doi: 10.1016/j.cmpb.2022.106862.

[3] D. R. Ente, S. A. Thamrin, S. Arifin, H. Kuswanto, and A. Andreza, "Classification Of Factors Causing Diabetes Mellitus At Unhas Hospital Using The C4.5 Algorithm," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 80–88, 2020, doi: 10.29244/ijsa.v4i1.330.

[4] V. L. Feigin, "Global, regional, and national burden of stroke and its risk factors, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019," *Lancet Neurol.*, vol. 20, no. 10, pp. 1–26, 2021, doi: 10.1016/S1474-4422(21)00252-0.

[5] M. Katan and A. Luft, "Global Burden of Stroke," in *Seminars in Neurology*, 2018, vol. 38, pp. 208–211.

[6] V. Azzahra and S. Ronoatmodjo, "Factors Associated with Stroke in Population Aged >15 Years in Special Region of Yogyakarta (Analysis of Basic Health Research 2018)," *J. Epidemiol. Kesehat. Indones.*, vol. 6, no. 2, pp. 91–96, 2023, doi: 10.7454/epidkes.v6i2.6508.

[7] N. Antonios and S. Silliman, "Diabetes mellitus and stroke," *Northeast Florida Med.*, pp. 17–22, 2005.

[8] B. Daneshfard, S. Izadi, A. Shariat, M. A. Toudaji, Z. Beyzavi, and L. Niknam, "Epidemiology of stroke in Shiraz, Iran.," *Iran. J. Neurol.*, vol. 14, no. 3, pp. 158–63, 2015.

[9] J. Shou, L. Zhou, S. Zhu, and X. Zhang, "Diabetes is an Independent Risk Factor for Stroke Recurrence in Stroke Patients: A Meta-analysis," *J. Stroke Cerebrovasc. Dis.*, vol. 24, no. 9, pp. 1961–1968, 2015, doi: 10.1016/j.jstrokecerebrovasdis.2015.04.004.

[10] R. Chen, B. Ovbiagele, and W. Feng, "Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes," *Am J Med Sci.*, vol. 351, no. 4, pp. 380–386, 2016, doi: 10.1016/j.amjms.2016.01.011.Diabetes.

[11] S. U. Putri, E. Irawan, and F. Rizky, "Implementation of Data Mining to Predict Diabetes Using the C4.5 Algorithm," *J. Appl. Inf. Syst. (Computers Manag.*, vol. 2, no. 1, pp. 39–46, 2021.

[12] S. J. Murphy and D. J. Werring, "Stroke: causes and clinical features," *Medicine (Baltimore).*, vol. 48, no. 9, pp. 561–566, Sep. 2020, doi: 10.1016/j.mpmed.2020.06.002.

[13] S. Parikh, S. Parekh, and N. Vaghela, "Impact of stroke on quality of life and functional independence," *Natl. J. Physiol. Pharm. Pharmacol.*, vol. 8, no. 12, pp. 1595–1598, 2018, doi: 10.5455/njppp.2018.8.0723807092018.

[14] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and Prediction of Diabetes Disease using Decision Tree Method," in *IT&AS 2021: Symposium on Information Technologies & Applied Sciences*, 2021, pp. 1–10. doi: 10.1109/ELIT53502.2021.9501151.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2012.

[16] A. Iyer, J. S, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, pp. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.

[17] A. Bisri and R. S. Wahono, "Application Of Adaboost To Resolve Class Imbalances in Determining Student Graduation Using The Decision Tree Method," *J. Intell. Syst.*, vol. 1, no. 1, pp. 27–32, 2015.

[18] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 2, pp. 26–31, 2018, doi: 10.9781/ijimai.2018.02.004.