Jurnal Mantik, 7 (2) (2023), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Analysis of the random forest and grid search algorithms in early detection of diabetes mellitus disease

Andi[1], Thamrin[2], Agus Susanto[3], Elyzabeth Wijaya[4], Deva Djohan[5]
[1,2,3,4,5]Institut Bisnis Informasi Teknologi dan Bisnis, Indonesia

A R T I C L E I N F O

Article history:

Accepted Jul 22, 2023
Revised Jul 26, 2023
Accepted Aug 07, 2023

Keywords:

Diabetes mellitus;
Grid search algorithm;
Prediction system;
Random forest algorithm.

ABSTRACT

This research focuses on implementing the Random Forest and Grid Search algorithms for the early detection of diabetes mellitus, aiming to modernize and enhance medical practices using technology. The proposed model achieved an accuracy of 77.06%, a precision of 71.43%, a recall of 47.30%, and a misclassification error of 22.94%. Comparative analysis with other data mining algorithms, including Decision Tree, Random Forest without Grid Search, and Cat Boost, demonstrated that the Random Forest with Grid Search algorithm outperformed the others. By utilizing Grid Search, the accuracy of the Random Forest algorithm increased by 2.03%. These findings indicate the potential effectiveness of machine learning in early diabetes detection. While the research offers promising results, there are limitations in terms of the dataset size and the number of detection variables used. Future studies should explore larger datasets and alternative algorithms to further enhance accuracy and aid in the early detection of diabetes mellitus.

Corresponding Author:

Andi,
Information System Study Program,
Institut Bisnis Informasi Teknologi dan Bisnis,
Jl. Mahoni No.16, Gaharu, Medan City, North Sumatra, 20235, Indonesia
Email: andi.lecture1995@gmail.com

1. INTRODUCTION

In the fast-paced advancement of the era, humanity encounters a multitude of diseases, some of which can be highly dangerous and even fatal. Among the most well-known illnesses is diabetes mellitus (Siregar et al., 2023). Diabetes mellitus is a persistent and non-communicable health condition characterized by the body's impaired ability to utilize glucose, a type of sugar crucial for providing energy. In Indonesia, diabetes mellitus stands as the sixth leading cause of mortality, trailing conditions associated with childbirth. As of 2021, the country accommodates approximately 19.5 million individuals living with diabetes mellitus, ranking Indonesia fifth globally in terms of the highest number of diabetes patients (Perdana et al., 2023). Untreated and unidentified diabetes can lead to serious complications (Apriliah et al., 2021).

In general, diabetes mellitus disease has symptoms that are almost similar to ordinary illness conditions so many people do not realize that they have the disease and

even have complications. Therefore, it is very important to do early detection of diabetes mellitus disease because if the disease is left too long without treatment it can result in dangerous complications such as kidney failure, damage to other organ function to heart attack (Anissa et al., 2023). Many ways can be done in early detection of a person experiencing diabetes mellitus or not. From a medical point of view, the diagnosis of diabetes mellitus can be done through blood tests. However, sometimes from a medical point of view, the process of checking blood only looks at one parameter, namely the patient's blood sugar level. In cases where a diabetes mellitus patient has low blood sugar levels, considering other parameters is crucial for achieving more accurate early detection results (Yusnaeni & Widiarina, 2022).

From a computer science perspective, the early detection process of diabetes mellitus in humans can be accomplished through an information system by applying data mining techniques. Data mining involves a series of actions or processes to discover meaningful relationships through patterns and trends within large datasets using various methods and algorithms (Airi et al., 2023). The advantages of utilizing information systems and data mining techniques are the ability to perform early detection of diabetes mellitus in humans quickly (Andi et al., 2023). Additionally, with data mining, predictions can be made involving multiple parameters, not just relying on a single parameter (Robet et al., 2022). Parameters such as age, weight, blood pressure, and other relevant factors can be included, resulting in more accurate detection outcomes (Ramayu et al., 2022).

Research that discusses the analysis of early detection of diabetes mellitus through the implementation of data mining has been done before by implementing the KNN algorithm (Siregar et al., 2023). Then, the next research combines the KNN algorithm with Naïve Bayes in order to increase the accuracy of the results of early detection of diabetes mellitus through the proposed model (Ikhromr et al., 2023). Subsequent research implements the C4.5 algorithm in the classification and prediction of diabetes mellitus in humans (Wahyu et al., 2023). All of the previous studies conducted were quite good and obtained a fairly high accuracy. Therefore, this study will also take the same approach as previous studies, namely implementing data mining in the early detection of diabetes mellitus in humans.

In this study, the Random Forest algorithm is implemented, which is one of the types of machine learning algorithms based on ensemble learning techniques. Ensemble learning combines predictions from multiple models (decision trees) to improve performance and prediction accuracy (Sari et al., 2023). This study chose the Random Forest algorithm because based on previous research that made comparisons between the Random Forest, Naïve Bayes, and Decision Tree algorithms, it was found that the Random Forest algorithm was superior to the two algorithms in making predictions. (Napiah et al., 2023)(Pamuji & Ramadhan, 2021). Additionally, to improve the accuracy of the Random Forest algorithm, it will be combined with the Grid Search algorithm, which is a technique used to find the best combination of parameters in the model (Kohsasih et al., 2022). By using Grid Search, we can search for the optimal parameters in the Random Forest algorithm, such as the number of decision trees, the maximum depth of the trees, and the criteria for selecting the best features, thereby obtaining a more optimal model and enhancing the prediction accuracy in the early detection of diabetes mellitus disease (Ramadhan et al., 2017).

This study contributes to analyzing the performance of the Random Forest and Grid Search algorithms in conducting early detection of diabetes mellitus so that it is hoped that this model can be implemented in the world of health, especially for the community so that they can quickly and accurately detect diabetes mellitus.

2. RESEARCH METHOD

The stages of the research method to be conducted in this study are illustrated in Figure 1.

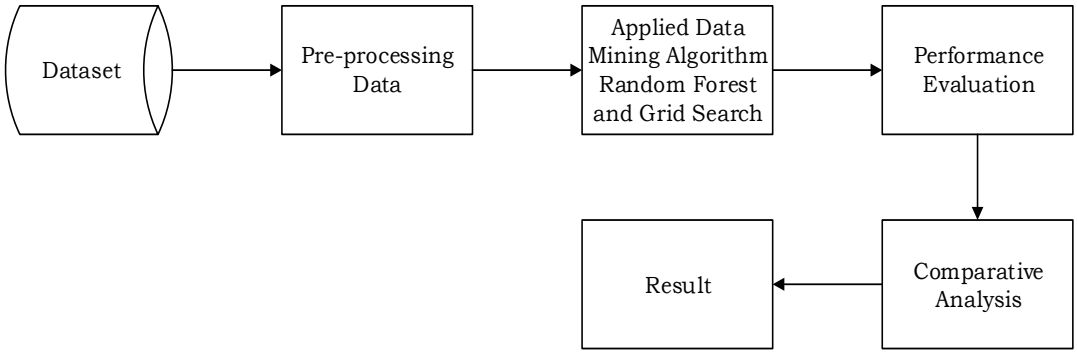

Figure 1. Research Method Diagram

From the proposed research framework depicted in Figure 1, it can be explained as follows:

a. Dataset

The dataset used in this study was taken from the Kaggle website, namely the Pima Indians Diabetes Database, with a total of 768 records. The data was split into 70% for training and 30% for testing. Table 1 below shows the dataset used in this study.

Table 1. Research Datasets

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
…	…	…	…	…	…	…	…	…	…
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

b. Pre-processing data

The data pre-processing steps in this study are divided into two parts: (a) preprocessing with mean imputation is a method to handle missing data in the dataset (Karrar, 2022). In this process, the missing values in a feature (column) will be replaced with the mean value of that feature. This method helps to maintain consistency and data integrity in the dataset (Prasetya & Priyatno, 2023), (b) preprocessing with feature standard scaling is a process in data processing before inputting it into a machine learning model. The goal of standard scaling is to transform the values of each feature (column) in the dataset to have the same scale so that no feature dominates or has a too significant impact on the model (Ambarwari et al., 2020).

c. Applied data mining algorithm Random Forest and Grid Search: (a) the Random Forest algorithm, introduced by Leo Breiman in 2001, is a statistical method and a machine learning expert from the University of California. The concept of Random Forest

is to combine multiple decision trees (Alhabib et al., 2022). This algorithm combines the results from several decision trees built randomly, resulting in a more accurate model. The way Random Forest works is by building multiple decision trees in parallel and making predictions based on the majority vote of the trees. This ensemble approach helps to improve the overall prediction accuracy and reduce the risk of overfitting (Sriyanto & Supriyatna, 2023), (b) Grid Search is a method for determining the combination of models and hyperparameters by testing each combination one by one and performing validation on each combination (Wirasasmita & Anisa, 2023). Hyperparameters are variables that can determine the results of a model in data mining. Grid Search is one of the techniques used for hyperparameter tuning in machine learning models. When using machine learning algorithms, several hyperparameters need to be determined before the model can be trained, such as the learning rate and the number of trees in the Random Forest. Grid Search works by testing all combinations of predetermined hyperparameter values. For example, if there are two hyperparameters (learning rate and number of trees), and we want to test 3 learning rate values and 4 number of tree values, then Grid Search will test a total of 3 x 4 = 12 combinations of these hyperparameters (Putri et al., 2023). Each combination will be trained on the training data and evaluated using cross-validation techniques to measure the model's performance on data that was not used in training (Nugroho & Amrullah, 2023). Evaluation metrics such as accuracy or mean squared error will be used to select the best combination of hyperparameters that provide the most optimal performance for the model (Fatmawati & Rifai, 2023).

d. Performance evaluation

The performance evaluation of the proposed model in this research is conducted using a Confusion Matrix, which is a cross-tabulation of positive and negative class data that are classified into predicted and actual classes (Andi et al., 2021). The Confusion Matrix consists of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) (Dikka et al., 2023). The confusion matrix table is presented in Table 2.

Table 2. Confusion Matrix

Actual Label	Actual Label	
	1	2
1	TP	FP
2	FN	TN

In this study, the positive class represents individuals with diabetes, and the negative class represents individuals without diabetes. Based on the Confusion Matrix, calculations are performed to determine accuracy, recall, precision, and misclassification error values.

e. Comparative analysis

At this stage, a comparative analysis is conducted between the research model using the Random Forest algorithm before combining it with Grid Search and the Decision Tree algorithm. The results of the comparative analysis show the accuracy, recall, precision, and misclassification error of each algorithm.

f. Result

The results of the research consist of a comprehensive discussion of the conducted analysis, elaborated in detail and linked to previous studies.

3. RESULTS AND DISCUSSIONS

This research utilized the Python programming language and Google Colab to implement the Random Forest and Grid Search algorithms for early detection of diabetes mellitus disease. The Grid Search algorithm was applied to find the best parameters for the

decision tree, resulting in the optimal settings: max depth of 5, minimum leaf samples of 4, minimum split samples of 10, and n_estimators of 50. The implementation of both algorithms produced a Confusion Matrix, as depicted in Figure 2.

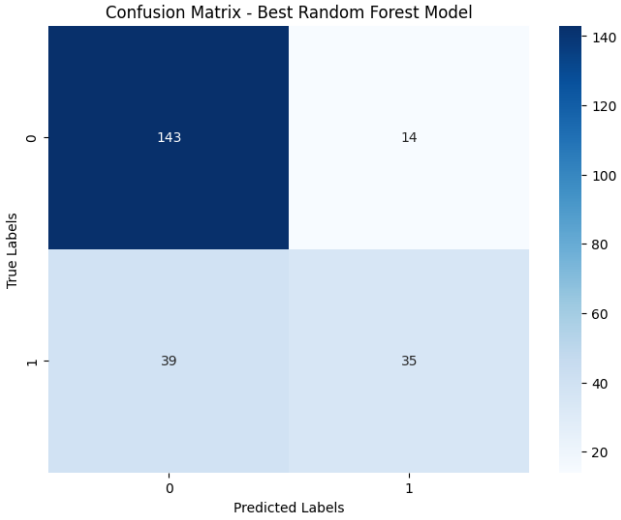

Figure 2. Confusion Matrix Plot

Based on the Confusion Matrix plot, the performance of the proposed research model yielded an accuracy of 77.06%, precision of 71.43%, recall of 47.30%, and a misclassification error of 22.94%. The results of this study demonstrate that the Random Forest model optimized with Grid Search is capable of providing reasonably accurate predictions in the early detection of Diabetes Mellitus. Although the accuracy and precision are quite good, the recall value still needs improvement to detect more diabetes cases effectively.

Next, a comparative analysis was conducted with the Decision Tree algorithm and the Random Forest algorithm before implementing Grid Search, as shown in Table 3.

Table 3. Comparative Analysis

Algorithm	Accuracy	Precision	Recall	Misclassification Error
Decision Tree	71.86	55.84	58.11	28.14
Random Forest Without Grid Search	75.03	69.23	46.44	24.97
Cat Boost	76.19	64.62	56.76	23.91
Random Forest With Grid Search	77.06	71.43	47.30	22.94

Based on Table 3, a comparative analysis of four different machine learning algorithms has been tested and evaluated in the early detection of diabetes mellitus disease. The four compared algorithms are Decision Tree, Random Forest without Grid Search, Cat Boost, and Random Forest with Grid Search. From the evaluation results, it can be observed that accuracy is a high evaluation metric for all four algorithms, with accuracy ranging from 71.86% to 77.06%. This indicates that all four algorithms have good overall classification capabilities.

The Random Forest algorithm with Grid Search showed the highest accuracy of 77.06%. This indicates that the model is capable of correctly classifying data in most cases. Next, the second-highest accuracy is achieved by the Cat Boost algorithm, showing good performance with an accuracy of 75.03%. The third-highest accuracy is obtained by the Random Forest algorithm without Grid Search, with an accuracy of 75.03%. Although

the Decision Tree algorithm has a lower accuracy compared to other algorithms, which is 71.86%, it still demonstrates its ability to perform correct classification.

However, when analyzing precision and recall values, there is a significant variation among the algorithms. The Decision Tree algorithm has a low precision value of 55.84%, but it has a better recall value compared to the Random Forest without Grid Search. On the other hand, Cat Boost has better precision and recall values than the Decision Tree, but slightly lower than the Random Forest without Grid Search. In this context, the Random Forest with Grid Search demonstrates the best performance among the four algorithms. This algorithm has a high precision rate of 71.43%, indicating that most of the cases predicted as positive are indeed cases of diabetes mellitus. Additionally, it also has a better recall value than the Cat Boost and Decision Tree algorithms, with a value of 47.30%, indicating that this algorithm is quite effective in detecting most cases of diabetes.

The high accuracy results from these four algorithms indicate that early detection of diabetes mellitus using machine learning approaches has the potential to be an effective method. By employing data mining and appropriate preprocessing techniques, these algorithms can provide better outcomes in detecting and identifying patients at risk of developing diabetes early on. It is crucial to continue developing and improving the quality of these models to support more effective prevention and treatment efforts for diabetes in the future (Iparraguirre-villanueva et al., 2023).

The research findings also indicate that the addition of the Grid Search algorithm can increase the accuracy of the Random Forest algorithm by 2.03%. This is consistent with the research conducted by (Anggoro & Afdallah, 2022) where the findings concluded that Grid Search can increase the accuracy of the Random Forest algorithm by 0.1-0.2%.

In addition, the researchers also conducted a survey directly to the hospital to test the model proposed in this study. From 10 patients suffering from diabetes mellitus, samples were taken and then tested on the proposed model. The test results showed that the proposed model was able to detect early in 9 out of 10 patients tested in this study.

4. CONCLUSION

This research implemented the Random Forest and Grid Search algorithms in the early detection of diabetes mellitus patients to assist medical professionals to become more modern and effective in utilizing technology. Based on the results of the conducted research, an accuracy of 77.06%, precision of 71.43%, recall of 47.30%, and a misclassification error of 22.94% were obtained from the proposed research model. Furthermore, to provide a more measured evaluation of the proposed research model, a comparison was made with several other data mining algorithms such as Decision Tree, Random Forest without Grid Search, and Cat Boost. The research findings showed that the Random Forest algorithm outperformed the other three data mining algorithms. The findings also revealed that the addition of the Grid Search algorithm increased the accuracy of the Random Forest algorithm by 2.03%. The implications of this research are significant for the early detection of diabetes mellitus patients and the advancement of medical technology. The utilization of these algorithms can empower medical professionals to make more informed decisions in diagnosing diabetes, potentially leading to early interventions and improved patient outcomes. The research highlights the importance of incorporating modern data mining techniques in the healthcare domain, as it can enhance the accuracy and efficiency of medical diagnoses. However, the study also acknowledges the limitations of a relatively small dataset and a limited number of detection variables. Therefore, future research endeavors should focus on utilizing larger datasets and exploring other advanced algorithms to further enhance the accuracy and reliability of early diabetes detection models. This research makes a valuable contribution to the field by demonstrating the effectiveness of Random Forest with Grid Search and

encouraging the adoption of data-driven approaches in medical diagnosis, paving the way for more sophisticated and accurate diabetes detection systems in the future.

REFERENCES

Airi, F. A. H., Suprapti, T., & Bahtiar, A. (2023). Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke. *E-Link: Jurnal Teknik Elektro Dan Informatika*, *18*(1), 73. https://doi.org/10.30587/e-link.v18i1.5271

Alhabib, I., Faqih, A., & Dikananda, F. (2022). Komparasi Metode Deep Learning, Naïve Bayes Dan Random Forest Untuk Prediksi Penyakit Jantung. *INFORMATICS FOR EDUCATORS AND PROFESSIONAL : Journal of Informatics*, *6*(2), 176–185. https://doi.org/10.51211/itbi.v6i2.1881

Ambarwari, A., Jafar Adrian, Q., & Herdiyeni, Y. (2020). Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(1), 117–122. https://doi.org/10.29207/resti.v4i1.1517

Andi, A., Juliandy, C., & David, D. (2023). Clustering Analysis of Tweets About COVID-19 Using the K-Means Algorithm. *Sinkron*, *8*(1), 543–533. https://doi.org/10.33395/sinkron.v8i1.12145

Andi, Juliandy, C., Robet, R., Pribadi, O., & Wijaya, R. (2021). Image Authentication Application with Blockchain to Prevent and Detect Image Plagiarism. *2021 6th International Conference on Informatics and Computing, ICIC 2021*, *December*. https://doi.org/10.1109/ICIC54025.2021.9632966

Anggoro, D. A., & Afdallah, N. A. (2022). Grid Search CV Implementation in Random Forest Algorithm to Improve Accuracy of Breast Cancer Data. *International Journal on Advanced Science, Engineering and Information Technology*, *12*(2), 515–520. https://doi.org/10.18517/ijaseit.12.2.15487

Anissa, K., Rumahorbo, H., & Wahyuni, S. (2023). Development of Instruments Test to Detect Diabetes Mellitus in Pregnancy. *Jurnal Kebidanan*, *12*(1), 27–36. https://doi.org/10.26714/jk.12.1.2023.27-36

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Jurnal Sistem Informasi*, *10*(1), 163–171. http://sistemasi.ftik.unisi.ac.id

Dikka, G., Prana, W., & Gede, L. (2023). Analisis Performa Algoritma K-Nearest Neighbor dalam Klasifikasi Tingkat Kerontokan Rambut. *Jurnal Nasional Teknologi Informasi Dan Aplikasinya*, *1*(3), 941–950.

Fatmawati, & Rifai, N. A. K. (2023). Klasifikasi Penyakit Diabetes Retinopati Menggunakan Support Vector Machine dengan Algoritma Grid Search Cross-Validation. *Jurnal Riset Statistika (JRS)*, *3*(1), 79–86.

Ikhromr, F. N., Sugiyarto, I., Faddillah, U., & Sudarsono, B. (2023). Implementasi Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Algoritma Naives Bayes dan K-Nearest Neighbor. *INTECOMS: Journal of Information Technology and Computer Science*, *6*(1), 416–428.

Iparraguirre-villanueva, O., Espinola-linares, K., Ornella, R., Castañeda, F., & Cabanillas-carbonell, M. (2023). *Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes*.

Karrar, A. E. (2022). The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values. *Indonesian Journal of Electrical Engineering and Informatics*, *10*(2), 375–384. https://doi.org/10.52549/ijeei.v10i2.3730

Kohsasih, K. L., Hayadi, B. H., Robet, Juliandy, C., Pribadi, O., & Andi. (2022). Sentiment Analysis for Financial News Using RNN-LSTM Network. *2022 4th International Conference on Cybernetics and Intelligent System, ICORIS 2022*. https://doi.org/10.1109/ICORIS56080.2022.10031595

Napiah, M., Astuti, R. D., & Pratama, E. K. (2023). Komparasi Algoritma Machine Learning untuk Klasifikasi Gejala Coronavirus Disease 19 (Covid-19). *Computer Science (CO-SCIENCE)*, *3*(2), 78–83.

Nugroho, A., & Amrullah, A. (2023). EVALUASI KINERJA ALGORITMA K-NN MENGGUNAKAN K-FOLD CROSS VALIDATION PADA DATA DEBITUR KSP GALIH MANUNGGAL. *Jurnal Informatika Teknologi Dan Sains (JINTEKS)*, *5*(2), 294–300.

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 46–50. https://doi.org/10.26905/jtmi.v7i1.5982

Perdana, A., Hermawan, A., & Avianto, D. (2023). Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *12*(1), 70–75. https://doi.org/10.32736/sisfokom.v12i1.1598

Prasetya, M. R. A., & Priyatno, A. M. (2023). Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining. *Jurnal Informasi Dan Teknologi*, *5*(2), 56–62. https://doi.org/10.37034/jidt.v5i1.324

Putri, T. A. E., Widiharih, T., & Santoso, R. (2023). Penerapan Tuning Hyperparameter Randomsearchcv Pada Adaptive Boosting Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung. *Jurnal Gaussian*, *11*(3), 397–406. https://doi.org/10.14710/j.gauss.11.3.397-406

Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., & Ghifari, A. (2017). Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency. *DEStech Transactions on Computer Science and Engineering*, *cece*. https://doi.org/10.12783/dtcse/cece2017/14611

Ramayu, I. M. S., Susanto, F., & Mahendra, G. S. (2022). Penerapan Data Mining Dengan Algoritma C4.5 Dalam Pemesanan Obat Guna Meningkatkan Keuntungan Apotek. *Prosiding Seminar Nasional Manajemen, Desain & Aplikasi Bisnis Teknologi (SENADA)*, *5*, 237–245. http://senada.idbbali.ac.id

Robet, Juliandy, C., Andi, Hendri, Hendrik, J., & Tarigan, F. A. (2022). Image Road Surface Classification Based on GLCM Feature Using LGBM Classifier. *IOP Conference Series: Earth and Environmental Science*, *1083*(1). https://doi.org/10.1088/1755-1315/1083/1/012006

Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, *14*(1), 155–162. https://doi.org/10.35970/infotekmesin.v14i1.1751

Siregar, S. D., Uli, Y. R. G., Sintami, N., Butar-butar, H. S., & Simanjuntak, R. M. (2023). Implementation of KNN algorithm in classifying diabetic ulcers in patients with diabetes mellitus. *Jurnal Mantik*, *7*(2), 691–701.

Sriyanto, & Supriyatna, A. R. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. *Teknika*, *17*(1), 163–172.

Wahyu, B. A. R. P., Fayi, F. A., Mahendra, C. P., & Hapsari, R. K. (2023). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. *Journal of Information Technology*, *8*(1), 80–89. https://doi.org/10.47970/siskom-kb.v4i1.173

Wirasasmita, D., & Anisa, E. (2023). Analisis Sentiment Twitter Berbasis Grid Search Algorithm (GSA) dengan Metode Support Vector Machine (SVM). *Asiimetrik*, *5*(1), 35–42.

Yusnaeni, W., & Widiarina. (2022). Penerapan Algoritma C4.5 Dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes). *Jurnal Teknik Komputer AMIK BSI*, *8*(1), 56–60. https://doi.org/10.31294/jtk.v4i2