TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 22, No. 4, August 2024, pp. 941~948
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v22i4.25656

Optimized decision tree classification method for diabetes prediction

Elly Muningsih[1], Fabriyan Fandi Dwi Imaniawan[1], Aprih Widayanto[2], Eva Argarini Pratama[1], Sutrisno[2], Sri Kiswati[1]

[1]Department of Information System, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia
[2]Department of Computer Technology, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia

Article Info

Article history:

Received Aug 27, 2023
Revised Mar 6, 2024
Accepted Mar 29, 2024

Keywords:

Decision tree
Diabetes prediction
Highest accuracy
Particle swarm optimization
Parameter optimization

ABSTRACT

Diabetes is one of the most deadly chronic diseases because most sufferers do not realize they have it. A more accurate prediction of diabetes disease must be made to reduce the risk of bad things happening to sufferers. This research will optimize the decision tree (DT) classification method for diabetes prediction. Optimization is done by splitting criteria, splitting data, particle swarm optimization (PSO), and parameter optimization to find the highest and most accurate forecast of diabetes. Splitting criteria is done by comparing the results of three criteria, namely gain ratio (GR), information gain (IG), and gini index (GI). Splitting data is done by dividing training data and testing data into three comparison groups, namely 70:30, 80:20, and 90:10. The application of PSO and parameter optimization is carried out to increase the accuracy value. The processed data is taken from the UCI machine learning repository with 520 records and 17 attributes (1 class/label attribute). From the experiments, the GI criterion with splitting data 90:10 obtained the greatest accuracy of 98.08%, and the combination with PSO resulted in an accuracy of 97.66%. Meanwhile, parameter optimization with splitting data 90:10 combined with GR criteria resulted in the highest accuracy of 97.90%.

Corresponding Author:

Elly Muningsih
Department of Information System, Faculty of Engineering and Informatics
Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, Kwitang, Senen, Jakarta City, Indonesia
Email: elly.emh@bsi.ac.id

1. INTRODUCTION

Diabetes mellitus (DM) or diabetes is a chronic metabolic disorder that affects the body’s ability to utilize energy sources contained in food and is one of the diseases that causes a healthcare crisis throughout the world regardless of geographical, racial, or ethnic context [1]-[3]. Nearly 50% of diabetic patients have genetic characteristics considered essential features of DM, and diabetes occurs due to the failure of the pancreas to produce insulin and the inability to use insulin [4]. We generally know of three types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes [2], [4], [5]. The cause of diabetes is lack of insulin which can result in serious damage such as heart attack, weight loss, cardiovascular dysfunction, blindness, ulcers, damage of the nervous system, and coma [6], [7]. In 2014, it was known that an estimated global prevalence of 9% was among adults aged 18+ and that 1.5 million deaths were directly caused by diabetes, which was later corroborated by the WHO, which stated that diabetes would be the seventh leading cause of death in 2030 [8].

In today’s modern technological era, computer technology can help us detect diseases accurately and save time and money, one of which is data mining, a field of computer science for making predictions [2], [9],

[10]. One of the methods in data mining is the classification method, widely used in the medical field to classify data into different classes according to several constraints, which are comparatively individual classifiers [11]. Several studies that have discussed diabetes prediction were carried out by [12]-[14]. Research on splitting criteria has been carried out [15], where this paper tries to find splitting criteria derived from gini index (GI) criteria that improve the performance of GI criteria. Splitting criteria are used in the data mining decision tree (DT) training process with the GI criteria (G) and two derivative criteria. Jain *et al.* [16] investigated the criteria of separation together using the two most frequently used criteria, namely information gain (IG) and the GI. This paper proposes to divide data points when the IG is maximum and the GI is minimum. The proposed approach was rigorously tested and compared to building a DT-based random forest. From the experimental results, it is known that the proposed splitting criteria work satisfactorily.

Meanwhile, another study [5] compared several classification methods, namely logistic regression (LR), k-nearest neighbor (kNN), support vector machine (SVM), gradient boost (GB), DT, multilayer perceptron (MLP), radio frequency (RF), and gaussian naïve bayesian (GNB), and then evaluated their performance with mean-absolute-error (MAE), root mean squared error (RMSE), receiver operating characteristic (ROC), test accuracy, precision, and recall. The study results concluded that the LR and GB methods achieved a higher testing accuracy of 79% than the other methods. Khanam and Foo [17] designed a system that can predict diabetes with high precision using the waikato environment for knowledge analysis (WEKA) tool. The study compared seven algorithms, namely DT, KNN, RF, NB, adaboost (AB), LR, and SVM, on Pima Indian diabetes dataset (PIDD) to predict diabetes and evaluate performance on various measures. All models provide over 70% accuracy. LR and SVM give an accuracy of around 77%–78% for the train/test split and k-fold cross-validation methods. The neural network (NN) with two hidden layers is considered the most efficient and promising for analyzing diabetes, with an accuracy rate of about 86%.

Optimization of the classification method with a combination of particle swarm optimization (PSO) and artificial neural network (ANN) that can categorize between ripeness stages and orange's immaturity [18]. This quantified data is used to train and test the optimized ANN model. Accuracy is based on ROC performance. Experiments in the study showed that the accuracy achieved for optimization was 70.5%, with a sensitivity and specificity of 60.1% and 80.0%, respectively. Meanwhile, research has been carried out for parameter optimization by, who applied a combination of the k-means method with Davies Bouldin index optimization and the DT method with parameter optimization. The parameter optimization that was carried out was the number of folds parameter in cross-validation and the criteria, maximal depth, apply to prune and apply preponing parameters in the DT. From the experiments, it was found that parameter optimization can increase the accuracy value to >90% for the DT method. PSO is a flock-based stochastic algorithm first proposed by Kennedy and Eberhart (1995) that exploits the concept of social behavior in animals such as flocks of birds and schools of fish [19], [20].

This research will optimize the DT classification method for diabetes prediction. Optimization is done by splitting criteria, splitting data, PSO, and parameter optimization to find the highest and most accurate forecast of diabetes. Splitting criteria is done by comparing the results of three criteria, namely gain ratio (GR), IG, and GI. Splitting data is done by dividing training data and testing data into three comparison groups, namely 70:30, 80:20, and 90:10. The application of PSO and parameter optimization is carried out to increase the accuracy value. The processed data is taken from the UCI machine learning repository with 520 records and 17 attributes (1 class/label attribute). Evaluation is carried out using the confusion matrix to obtain the accuracy value and the area under curve (AUC) value based on the ROC curve. The tools used in this research experiment were RapidMiner.

2. METHOD

The research method used in this experiment uses the cross industry standard process for data mining (CRISP-DM) model, which has six research stages [21], [22]. The six stages of the research are described as a research framework, as in Figure 1. The stages of the CRIS-DM method are:

2.1. Business understanding

This stage identifies problems, namely diabetes predictions that are not yet accurate or whose accuracy is still low. This stage aims to identify problems related to the inaccuracy of diabetes predictions or their low accuracy. Low accuracy will result in a suboptimal model, leading to inaccurate diabetes predictions. This can have a negative impact on potential diabetes patients. Based on the analysis and comparison of previous research, this study aims to optimize the Decision Tree classification method by separating criteria, splitting data, applying PSO, and optimizing parameters to achieve higher accuracy values.

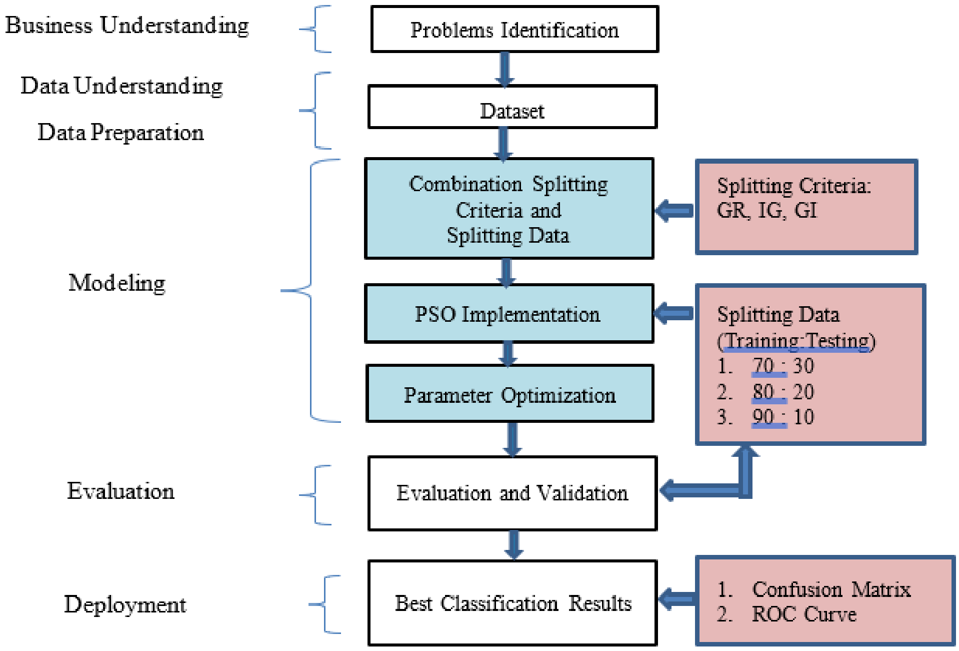

Figure 1. Research framework

2.2. Data understanding and data preparation

The dataset processed in the experiment was data taken from the UCI machine learning repository, with 520 records and 17 attributes (1 class/label attribute). This dataset was collected from patients using a direct questionnaire from Sylhet Diabetes Hospital in Bangladesh. This data has 16 attributes, namely age, sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class. A class attribute is an attribute that indicates whether a person is at risk for diabetes or not. The class attribute has a positive value for those with diabetes risk and a negative value for those who do not have diabetes risk. As for the value of each attribute, age has a value of 1. 20–35, 2. 36–45, 3. 46–55, 4. 56–65, and above 65. The attribute sex has a value of 1 for men and a value of 2 for women. For other attributes containing the values yes and no.

2.3. Modeling

In the processing carried out in this experiment, there are three main stages or processes related to optimizing the DT classification method for diabetes prediction. The optimization in question is:

a. Optimization with splitting data combination splitting criteria.
b. Data splitting is done by comparing the amount of training data and testing data. The ones used in the experiment were splitting data 70:30, 80:20, and 90:10. This comparison can be explained by the fact that the first number represents the amount of training data, and the second number represents testing data. In contrast, the criteria used in the experiment were GR, IG, and GI.
c. Optimization with the implementation of PSO.
d. PSO is one of the many techniques widely used in solving optimization problems of unstructured, constrained, or unconstrained continuous or discrete functions. PSO is a popular choice for prediction problems to a certain degree because its representation is intuitively simple, and the number of parameters that can be adjusted is relatively low [19]. PSO can improve classification by maximizing attribute weighting or attribute selection [23].
e. Parameter optimization.
f. At this stage, the parameters that are optimized to produce the highest accuracy are iteration and DT. Criteria, DT.maximal_depth, DT.apply_pruning, DT.apply_prepruning, and CV.number_of_folds, where DT, and cross validation (CV).

2.4. Evaluation

The processed model will be evaluated using the confusion matrix to determine its accuracy and the AUC value based on the ROC curve.

a. Confusion matrix

The confusion matrix is a table widely used to evaluate the performance of an algorithm on a group of test data and allows a more comprehensive analysis than just accuracy [24]. An n×n confusion matrix associated with classifiers shows predicted and actual classifications, where n is the number of different classes [25]. Table 1 a confusion matrix with $n = 2$, which means:

- a: represents the quantity of accurate negative predictions
- b: represents the quantity of inaccurate positive predictions
- c: represents the quantity of inaccurate negative predictions
- d: represents the quantity of accurate positive predictions

Table 1. The confusion matrix for two-class classification problem (n=2)

	Predicted negative	Predicted positive
Actual negative	a	b
Actual positive	c	d

And for the value of the accuracy of the (1),

$$Accuracy = \frac{a+d}{a+b+c+d} \tag{1}$$

b. ROC and AUC

A ROC curve is a graph that describes the performance of an algorithm at the limit of classification, where the curve plots the two parameters with maximum sensitivity. While AUC is an area under the ROC curve that gives the cumulative performance value at all classification limits, AUC can also be explained as a probability in which the algorithm places positive random samples at a higher rank than negative random samples [22]. The AUC ranges from 0 to 1. Models whose predictions are 100% wrong have an AUC of 0.0, while those with 100% correct have an AUC of 1.0 [24]. A clear guide to classifying the accuracy of diagnostic tests using the AUC in the traditional system [26] is presented as follows:

- 0.90–1.00: classification of excellent.
- 0.80–0.80: classification of good.
- 0.70–0.80: classification of fair.
- 0.60–0.70: classification of poor.
- 0.50–0.60: failure.

2.5. Deployments

At this stage, the implementation of the DT classification method is carried out with optimization, which produces the highest accuracy value. Optimization is carried out by separating criteria and data, followed by the application of PSO for optimization, and finally, parameter optimization. The use of classification methods with the highest accuracy values is expected to provide better results and contributions to diabetes prediction.

3. RESULTS AND DISCUSSION

3.1. Optimization of splitting criteria and splitting data

Optimization of the first method is to do a combination of splitting criteria, namely GR, IG, and GI, with splitting data and a comparison of training data with testing data in 3 categories, namely 70:30, 80:20, and 90:10. A data ratio of 70:30 means that the experiment uses 70% of the dataset as training data and the remaining 30% is used for data testing. The optimization results with a combination of splitting criteria and splitting data are shown in Table 2, which shows the accuracy and AUC values.

The table shows that at splitting data 70:30 and 80:20, the GR criterion produces the highest accuracy values, namely 96.79% and 97.12%. Then followed GI and GI criteria with the same accuracy values, namely 96.15% (70:30) and 96.15% (80:20). Whereas in splitting data 90:10, GI criteria produced the highest accuracy, namely 98.08%, followed by GI and GI criteria of 96.15%. Meanwhile, for AUC values, still, from Table 2, it is known that the GR criteria in splitting data 70:30 and 80:20 produce the highest AUC values, namely 0.968 and 0.982. They were followed by IG and GI criteria with the same AUC value of 0.951 for splitting data 70:30 and 0.966 for splitting data 80:20. Figure 2 shows that the GR criteria produce the highest accuracy and AUC values at splitting data at 70:30 and 80:20. GI has the highest precision and AUC values on splitting data at 90:10.

Table 2. Optimization splitting criteria and splitting data

Criteria	70:30		80:20		90:10	
	Accuracy (%)	AUC	Accuracy (%)	AUC	Accuracy (%)	AUC
GR	**96.79**	**0.968**	**97.12**	**0.982**	96.15	0.965
IG	96.15	0.951	96.15	0.966	96.15	0.965
GI	96.15	0.951	96.15	0.966	**98.08**	**0.969**

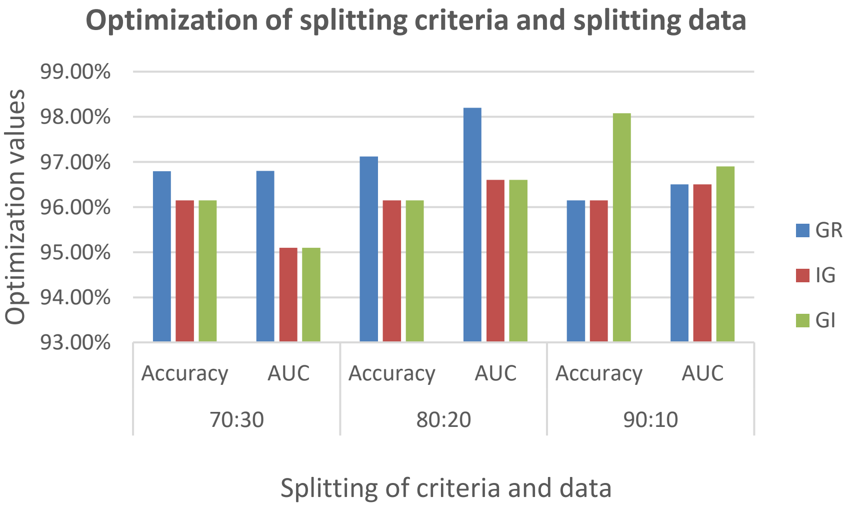

Figure 2. Optimation splitting data and splitting criteria

3.2. Optimizing the implementation of PSO

The second optimization carried out in this study is a combination of the first optimization, namely splitting criteria and splitting data, with the application of PSO. The results of the accuracy and AUC values are shown in Table 3. And it can be seen that with the application of PSO, the average accuracy and AUC values increase even though there are slight differences in the criteria compared to the results of the first previous optimization.

Table 3. PSO implementation

Criteria	70:30		80:20		90:10	
	Accuracy (%)	AUC	Accuracy (%)	AUC	Accuracy (%)	AUC
GR	**97.00**	**0.968**	97.13	**0982**	97.22	0.965
IG	96.43	0.951	97.35	0.966	97.65	0.965
GI	96.97	0.951	**97.36**	0966	**97.66**	**0.969**

From Table 3, it is known that in splitting data 70:30, GR got the highest accuracy result, namely 97.00%, followed by GI criteria of 96.97% and GI criteria with an accuracy of 96.43%. In splitting data 80:20 and 90:10, the GI criteria produced the highest accuracy, namely 97.36% and 97.66%. In splitting data 80:20, the second order is GI with an accuracy value of 97.36%, and the last is the GI criteria of 97.35%. In splitting data 90:10, the second order of highest accuracy is the GI criterion of 97.65%, and the last sequence is the GR of 97.22%. In the AUC value of the PSO application, the GI criterion produced the highest AUC value in all splitting data categories, namely a value of 0.970 for a 70:30 ratio, followed by the second highest, the GR criterion of 0.969, and the last, the IG criterion of 0.962. In splitting data 80:20, the highest AUC value is 0.970. The GI criterion produces the largest AUC value of 0.974 for splitting data 90:10, followed by the GR criterion of 0.973, and finally, the GI of 0.968. In general, the AUC value increases with the application of PSO. Figure 3 shows that with the application of PSO, the accuracy value of most criteria increases except for the GI criteria with splitting data of 90:10. For the AUC value, in general, the value increases with the application of PSO.

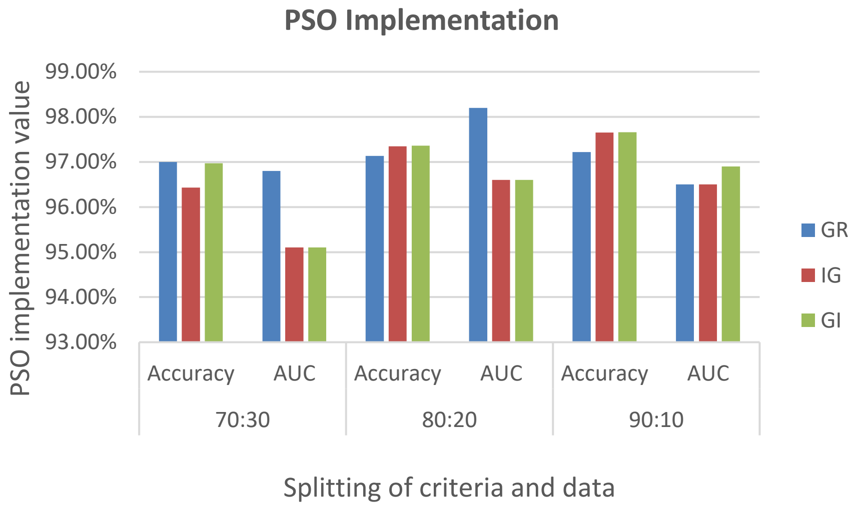

Figure 3. PSO implementation

3.3. Parameter optimization

Parameter optimization is applied to three categories of Splitting data: 70:30, 80:20, and 90:10. The parameters in question are Iteration, DT.criteria, DT.maximal_depth, DT.apply_pruning, DT.apply_prepruning, and CV.number_of_folds. The evaluation is based on the resulting accuracy value in optimizing this parameter. The results of parameter optimization are shown in Table 4.

Table 4. Parameter optimization

Parameter	70:30	80:20	90:10
Iteration	360	336	391
DT.criteria	GI	GI	GR
DT.maximal_depth	90	90	90
DT.apply_pruning	True	True	False
DT.apply_prepruning	False	False	False
CV.number_of_folds	7	7	7
Accuracy	97.80%	97.40%	97.90%

Table 4 that the highest accuracy is 97.90% on splitting data 90:10, Iteration: 391, DT. Criteria: GR, DT. maximal_depth: 90, DT. Apply_pruning: False, DT. Apply_prepruning: false, and number_of_folds (Cross Validation): 7. The second highest accuracy value is 97.80% for splitting data 70:30 and criteria GI. And the lowest accuracy value is 97.40% with splitting data 80:20 and criteria: GI.

4. CONCLUSION

This paper has presented optimizing method for diabetes prediction. The process of implementing the DT classification method has been executed with optimization, following a systematic approach combination of splitting data and splitting criteria can produce the highest accuracy value of 98% for GI criteria with splitting data of 90:10. Second, the application of PSO has proven to be able to increase the accuracy and AUC values in almost all categories for the combination of splitting data and splitting criteria. Third, with parameter optimization, accuracy values can increase to >97% for all types of splitting data. Of the three optimizations that have been carried out, all of them fall into the category of excellent classification. Optimization with the highest accuracy value can be used as a reference basis for applying the best DT classification method. In the future, research could involve more data and other methods.

ACKNOWLEDGMENS

Special thank you to all academia Universitas Bina Sarana Infomatika for its support so this research can be carried out properly.

REFERENCES

[1] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model," *International Journal of Information Technology*, vol. 12, no. 2, pp. 419–428, 2020, doi: 10.1007/s41870-018-0270-5.

[2] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," *Computer Vision and Machine Intelligence in Medical Image Analysis*, vol. 992, pp 113–125, February 2022, pp. 113–125, 2020, doi: 10.1007/978-981-13-8798-2_12.

[3] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728–1737, 2023, doi: 10.11591/eei.v12i3.4412.

[4] R. T. Selvi and I. Muthulakshmi, "Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 1717–1730, 2021, doi: 10.1007/s12652-020-02242-1.

[5] S. A. D. Alalwan, "Diabetic analytics: Proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 85–95, 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95.

[6] I. N. Mahmood and H. S. Abdullah, "Analyzing the behavior of different classification algorithms in diabetes prediction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, pp. 201–206, 2024, doi: 10.11591/ijai.v13.i1.pp201-206.

[7] R. N. Patil, S. Rawandale, N. Rawandale, U. Rawandale, and S. Patil, "An efficient stacking based NSGA-II approach for predicting type 2 diabetes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1015–1023, 2023, doi: 10.11591/ijece.v13i1.pp1015-1023.

[8] S. Selvakumar, A. Sheik Abdullah, and R. Suganya, "Decision support system for type II diabetes and its risk factor prediction using bee-based harmony search and decision tree algorithm," *Int. J. Biomed. Eng. Technol.*, vol. 29, no. 1, pp. 46–67, 2019, doi: 10.1504/IJBET.2019.096880.

[9] E. Muningsih, C. Kesuma, Sunanto, Suripah, and A. Widayanto, "Combination of K-Means method with Davies Bouldin index and decision tree method with parameter optimization for best performance," *AIP Conference Proceedings*, vol. 2714, 2023, doi: 10.1063/5.0129119.

[10] Z. Saringat, A. Mustapha, R. D. R. Saedudin, and N. A. Samsudin, "Comparative analysis of classification algorithms for chronic kidney disease diagnosis," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1496–1501, 2019, doi: 10.11591/eei.v8i4.1621.

[11] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[12] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and prediction of diabetes disease using decision tree method," *CEUR Workshop Proceedings*, vol. 2824, pp. 163–172, 2021.

[13] L. Shrinivasan, R. Verma, and M. D. Nandeesh, "Early prediction of diabetes diagnosis using hybrid classification techniques," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 3, pp. 1139–1148, 2023, doi: 10.11591/ijai.v12.i3.pp1139-1148.

[14] G. Alfian, Y. M. Saputra, L. Subekti, A. D. Rahmawati, F. T. D. Atmaji, and J. Rhee, "Utilizing deep neural network for web-based blood glucose level prediction system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, pp. 1829–1837, 2023, doi: 10.11591/ijeecs.v30.i3.pp1829-1837.

[15] L. A. Badulescu, "Experiments for a better Gini index splitting criterion for Data Mining Decision Trees algorithms," *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)*, Sinaia, Romania, 2020, pp. 208-212, 2020, doi: 10.1109/ICSTCC50638.2020.9259691.

[16] V. Jain, A. Phophalia, and J. S. Bhatt, "Investigation of a Joint Splitting Criteria for Decision Tree Classifier Use of Information Gain and Gini Index," *TENCON 2018 - 2018 IEEE Region 10 Conference*, Jeju, Korea (South), 2018, pp. 2187-2192, doi: 10.1109/TENCON.2018.8650485.

[17] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[18] A. D. Rosli, N. S. Adenan, H. Hashim, N. E. Abdullah, S. Sulaiman, and R. Baharudin, "Application of Particle Swarm Optimization Algorithm for Optimizing ANN Model in Recognizing Ripeness of Citrus," *IOP Conference Series: Materials Science and Engineering*, vol. 340, no. 1, 2018, doi: 10.1088/1757-899X/340/1/012015.

[19] S. Sengupta, S. Basak, and R. Peters, "Particle Swarm Optimization: A Survey of Historical and Recent Developments with Hybridization Perspectives," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 157–191, 2018, doi: 10.3390/make1010010.

[20] A. G. Gad, "Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review," *Archives of Computational Methods in Engineering*, vol. 29, no. 5, 2022, doi: 10.1007/s11831-021-09694-4.

[21] S. Peker and Ö. Kart, "Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review," *Journal of Data, Information and Management*, vol. 5, no. 1–2, pp. 1–21, 2023, doi: 10.1007/s42488-023-00085-x.

[22] B. Ziv and Y. Parmet, "Improving nonconformity responsibility decisions: a semi-automated model based on CRISP-DM," *International Journal of System Assurance Engineering and Management*, vol. 13, no. 2, pp. 657–667, 2022, doi: 10.1007/s13198-021-01318-1.

[23] S. I. Novichasari and I. S. Wibisono, "Particle Swarm Optimization For Improved Accuracy of Disease Diagnosis," *Journal of Applied Intelligent System*, vol. 5, no. 2, pp. 57–68, 2020, doi: 10.33633/jais.v5i2.4242.

[24] A. A. Abdullah, S. A. Hafidz, and W. Khairunizam, "Performance Comparison of Machine Learning Algorithms for Classification of Chronic Kidney Disease (CKD)," *Journal of Physics: Conference Series*, vol. 1529, no. 5, 2020, doi: 10.1088/1742-6596/1529/5/052077.

[25] N. Nofriani, "Comparations of Supervised Machine Learning Techniques in Predicting the Classification of the Household's Welfare Status," *Journal Pekommas*, vol. 4, no. 1, p. 43, 2019, doi: 10.30818/jpkm.2019.2040105.

[26] F. Gorunescu, "Data mining: Concepts, models and techniques," *Intelligent Systems Reference Library*, vol. 12. 2011, doi: 10.1007/978-3-642-19721-5.

BIOGRAPHIES OF AUTHORS

Elly Muningsih currently is pursuing her doctoral program in Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada Yogyakarta, Indonesia. She took her undergraduate (S.Kom) at STMIK AMIKOM Yogyakarta in 2003, and Master (M.Kom) in STMIK Nusa Mandiri Jakarta in 2012. Her research areas of interest are data mining, machine learning, big data and information system. She can be contacted at email: elly.emh@bsi.ac.id.

Fabriyan Fandi Dwi Imaniawan is a lecturer at Bina Sarana Informatika University in the Information Systems Study Program, Banyumas Regency Campus. He has studied S1 Information Systems (2014) and S2 Computer Science (2016) at STMIK Nusa Mandiri. He is currently actively writing papers in nationally accredited scientific journals, writing books and being a resource person in several seminars and trainings on digital marketing. He can be contacted by email: fabriyan.fbf@bsi.ac.id.

Aprih Widayanto is a lecturer at Bina Sarana Informatika University, Computer Technology since 2021. Bachelor degree at STMIK Nusa Mandiri majoring in Informatics Engineering (2014) and Masters degree majoring in Computer Science (2016). Computer networks and computer technology are the main topics of her study. He can be contacted at email: aprih.apz@bsi.ac.id.

Eva Argarini Pratama completed her undergraduate program at the Faculty of Computer Science, Dian Nuswantoro University, Semarang and Master of Computer Science at STMIK Nusa Mandiri Jakarta. She is one of the lecturers at Bina Sarana Informatika University in the Department of Information Systems who teaches several courses related to system analysis and design, besides that several research publications have also been carried out and published in various journals with informatics and computer scope. She can be contacted at email: eva.eap@bsi.ac.id.

Sutrisno is a lecturer at Bina Sarana Informatika University in the Computer Tecnology Study Program, Banyumas Regency Campus. He has studied S1 Arabic Education (2013) and S2 Computer Science (2016) at Dian Nuswantoro University. He is currently actively writing papers in nationally accredited scientific journals and freelance web programmer. He can be contacted by email: sutrisno.stz@bsi.ac.id.

Sri Kiswati is a lecturer at Bina Sarana Informatika University in the Information Systems Study Program, Yogyakarta City Campus. She has studied S1 Industrial Engineering at Islamic Indonesian Univercity (1999) and S2 Magister Management at Diponegoro Univercity (2010). She is currently actively writing papers in nationally accredited scientific journals and writing books. She can be contacted by email: sri.srk@bsi.ac.id.