Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 1 Iss. 1 July 2023, pp: 41-46
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Random Forest Optimization Using Particle Swarm Optimization for Diabetes Classification

Pangeran Fadillah Pratama[1*], Desvita Rahmadani[2], Rahma Sani Nahampun[3], Della Harmutika[4], Akhas Rahmadeyan[5], Muhammad Fikri Evizal[6]

[1,2,3,4,5]Department of Information Systems, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[5]Puzzle Research Data Technology, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[6]Computer Science and Information Engineering Majority, College of Science and Engineering,
National Dong Hwa University, Hualien, Taiwan

E-Mail: [1]12050310337@students.uin-suska.ac.id, [2]12050323639@students.uin-suska.ac.id,
[3]12050321674@students.uin-suska.ac.id, [4]12050320481@students.uin-suska.ac.id,
[5]11950314479@students.uin-suska.ac.id, [6]410921346@gms.ndhu.edu.tw

Received March 19th 2023; Revised Jun 13th 2023; Accepted Jul 20th 2023
Corresponding Author: Pangeran Fadillah Pratama

Abstract

Diabetes mellitus is a chronic degenerative disease caused by a lack of insulin production in the pancreas or the body's ability to use insulin less effectively. According to a report by the World Health Organization (WHO), 4% of the total deaths in the world are caused by diabetes. The International Diabetes Federation (IDF) notes that in 2013 there has been an increase in diabetes sufferers. Indonesia is the seventh place with the largest number of cases of diabetes mellitus. In this study, the method used to classify diabetes is using a random forest algorithm with Particle Swarm Optimization (PSO) optimization. This study resulted in an accuracy of the random forest classification algorithm of 78.2% and 82.1 using PSO optimization with an increase in value of 3.9%. It can be concluded that PSO optimization can provide a better increase in classification accuracy values when compared to the random forest algorithm without PSO optimization.

Keywords**:** Classification, Diabetes, International Diabetes Federation, Particle Swarm Optimization, Random Forest

1. INTRODUCTION

Diabetes mellitus is a non-communicable disease that causes a slow decline in the function of the patient's body cells, characterized by increased blood sugar levels in the urine due to disturbed metabolism when the function of the hormone insulin cannot normally run [1]. The metabolic disease known as diabetes mellitus is caused by a malfunction of insulin secretion and action [2]. High blood sugar levels can result in damaged body cell functions such as kidney failure, heart arteries, stroke, blindness, and death [3].

The World Health Organization (WHO) reports diabetes causes 4% of deaths worldwide [4]. This is recorded by the International Diabetes Federation (IDF) that there was an increase in diabetes sufferers in 201 3 [5]. Deaths due to diabetes occur in countries with low and middle incomes, one of which is Indonesia [6]. Indonesia is listed as the country with the seventh-largest number of cases of Diabetes Mellitus (DM) in the world. Diabetes Mellitus is classified into type 1, type 2, and type 2 DM which often occurs in pregnancy [7]. Clinical symptoms that often occur are caused by factors of polyphagia (eating a lot), polyurea (urinating a lot), and drinking a lot [8].

WHO defines Diabetes Mellitus as a chronic degenerative disease that occurs due to insufficient production of insulin in the pancreas, so the insulin produced by the body cannot be used effectively [9]. Some people with diabetes find it difficult to recognize that they have been affected by the disease and are even close to complications. This is because the symptoms they feel are similar to common illnesses [10]. Nearly half of diabetics are caused by hereditary factors [11].

Seeing the problems that occur, early detection of diabetes is necessary. Early detection is expected to reduce the risk of complications in diabetes patients in the future. To analyze diabetes patients from an early age, many records of this disease are carried out so that prevention can be done. One of the records that can be done is by utilizing technology, namely by using data processing methods, namely data mining.

Data mining is a method for acquiring knowledge. According to Daniel T. Larose (2014), data mining is a procedure for creating bonds that have meanings, patterns, and inclinations by observing large data groups that are in storage using pattern identification techniques. The methods that are usually operated in data mining include description or depiction, prediction or prediction, clustering or grouping, classification and association, and estimation [12]. Classification is a process for creating functions or models explaining classes in data or concepts in order to predict the class of an object whose label has not been obtained [13]. This study used classification techniques to predict which people had diabetes and who did not. Several algorithms can be used to calculate the classification process. One of the classification algorithms is the random forest with Particle Swarm Optimization (PSO).

Research conducted by Nurlaelatul Maulidah et al. (2020), using PSO and Naïve Bayes for diabetes classification, concluded that the accuracy of the classification algorithm with a PSO is 77.34% with an increased accuracy value of 2.73%. It can be concluded that applying the PSO optimization method can produce a higher level of accuracy than using only individual methods [14]. Another study conducted by Fely Dany Prasetya et al. (2022) in this study using the DecisionTree C4.5 algorithm for predicting hepatitis with PSO resulted in the conclusion that optimization with the PSO Algorithm can increase the accuracy of the DecisionTree C4.5 algorithm higher than using only the DecisionTree algorithm C4.5 only, with an accuracy ratio of 99.35% and 99.67% [15].

Based on the above references, researchers are enthusiastic about researching to optimize the random forest algorithm using PSO to classify diabetes with the hope that this optimization can improve the accuracy of the random forest algorithm. This research aims to classify diabetes and optimize the random forest algorithm using PSO to classify diabetes.

2. MATERIALS AND METHODS

This research uses the random forest algorithm to optimize the Particle Swarm Optimization (PSO) algorithm. This study aimed to prove the performance of PSO accuracy in the random forest algorithm in classifying diabetes. Figure 1 shows the methodology of this study.

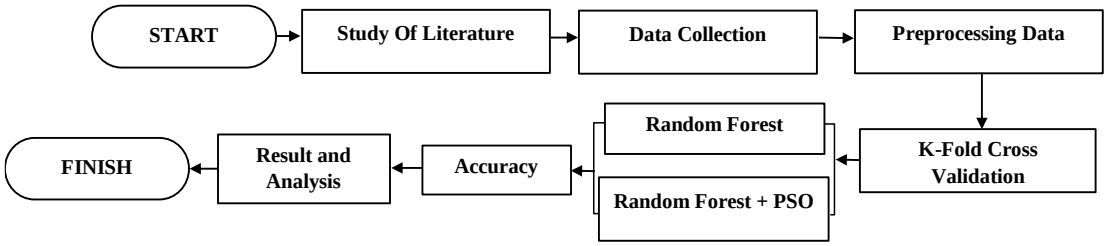

Figure 1. Research Methodology

2.1 Study of Literature

This research conducted a literature study to obtain information and collect the required data through several sources, including scientific publications such as journals, theses, books, and other validated sources [16].

2.2 Data collection

Determining the data to be processed, searching for available data, obtaining additional data needed, and integrating everything will be done at this stage. For the research objectives to be achieved, some important information will be collected to support the research [17]. Data in this study using a diabetes dataset collected through the Kaggle website. The diabetes dataset obtained will be classified using the random forest algorithm and optimized using the PSO algorithm. The attributes used in this research are (1) Pregnancies, (2) Glucose, (3) BloodPressure, (4) SkinThickness, (5) Insulin, (6) BMI, (7) DiabetesPedigree Function, (8) Age, (9) Outcomes.

2.3 Preprocessing Data

Data preprocessing is a machine learning technique that transforms actual data into a logical or intelligent structure. Preprocessing data is a standard method for minimizing sound problems. Data cleaning, data transition, data discretion, and data reduction are some tasks that fall under the category of preprocessing [18]. Data preprocessing is carried out to prepare datasets processed using data mining algorithms. Data preprocessing is done by handling missing or incomplete values [19].

2.4 K-Fold Cross Validation

The data distribution in this study uses the Cross Validation technique, which will assess the results of the best statistical model with the fold value used, which is 10 [20]. The application of Cross Validation is considered capable of obtaining maximum accuracy results. This method has an advantage over repeated random subsampling because training and validation are performed for each of them at least once [21].

2.5. Random Forest Algorithm Implementation

One of the methods used for classification and regression is random forest. This method uses a decision tree as a classifier built and combined [22]. In this study, the random forest algorithm will be used to classify diabetes and obtain the accuracy of the final results from using the random forest algorithm. The random forest equation formula is:

$$\boldsymbol{Entropy}\ (\boldsymbol{Y}) = -\sum \mathbf{p}\ (\mathbf{c}|\mathbf{Y})\ \mathbf{i}\ \mathbf{log2}\ \mathbf{p}\ (\mathbf{c}|\mathbf{Y}) \quad (1)$$

Where Y is the set of cases, and p(c|Y) is the proportion of Y values to class c.

$$(\boldsymbol{Y}, \boldsymbol{a}) = \boldsymbol{Entropy}\ (\boldsymbol{Y}) - \sum |\mathbf{Yv}|\ |\mathbf{Ya}|\ \boldsymbol{veValues\ Entropy}\ (\boldsymbol{Yv}) \quad (2)$$

Where Values (a) are all possible values in the case set a. Yv is a subclass of Y with class v, which is related to class a. Yes, these are all values that correspond to a.

2.6. Particle Swarm Optimization (PSO)

Kennedy and Eberhart (1995) explained that PSO is a global optimization method based on research on the behavior of flocks of birds and fish. The advantage of the PSO optimization method is that it has a concept that is easy to apply and efficient in calculations compared to other optimization techniques [23]. To get and know the results of the random forest algorithm and PSO optimization will be added to increase the accuracy of the results of the random forest algorithm.

$$Vi(t+1)=WVi(t)+c1r1(Pi-Xi)+c2r2(Pg-Xi) \quad (3)$$

$$Xi\ (t+1) = X(t) + Vi(t+1) \quad (4)$$

Description :

Vi(t)	: velocity of the i-th particle in the t-iteration
Xi(t)	: position of the i-particle in the t iteration
c1 and c2	:learning factors
W	: inertial weight
r1 and r2	: constants
Pi	: vector for the best position of the i-th particle
Pg	: best position globally

2.7. Confusion Matrix

The Confusion Matrix is a matrix that is commonly used as an evaluation that will display the performance results of the classification to determine the performance of the machine learning model used [24].

Table 1. Confusion Matrix

	Classified Negative	Classified Positive
Actual Negatives	a	b
Actual Positive	c	d

3. RESULTS AND ANALYSIS

3.1 Data collection

This study uses data on people with diabetes collected from Kaggle. This data shows that the attributes that influence diabetes are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. The classification of attributes in this study is based on the diabetes dataset to determine the accuracy value further.

Table 2. Data collection

No.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcomes
1	2	138	62	35	0	336	127	47	1
2	0	84	82	31	125	382	233	23	0
3	0	145	0	0	0	442	63	31	1
4	0	135	68	42	250	423	365	24	1
5	1	139	62	41	480	407	536	21	0
6	0	173	78	32	265	465	1159	58	0
7	4	99	72	17	0	256	294	28	0
8	8	194	80	0	0	261	551	67	0
9	2	83	65	28	66	368	629	24	0
…	…	…	…	…	…	…	…	…	…
2000	2	81	72	15	76	301	547	25	0

3.2 Preprocessing Data

The next stage is the data-cleaning process. All noise data will be cleaned at this stage, such as typos, invalid or missing [25]. This cleaning process is carried out by removing noise data as previously described. At this preprocessing stage, no data records fall into the noise data category. So that the initial data that has been collected will be used entirely with a total of 2000 data records. All initial data that has gone through the cleaning process will undergo a data transformation process, namely converting the data into a form that can be processed using Tools Rapid Miner.

Table 3. Data Transformation

No.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcomes
1	2	138	62	35	0	336	127	47	Yes
2	0	84	82	31	125	382	233	23	No
3	0	145	0	0	0	442	63	31	Yes
4	0	135	68	42	250	423	365	24	Yes
5	1	139	62	41	480	407	536	21	No
6	0	173	78	32	265	465	1159	58	No
7	4	99	72	17	0	256	294	28	No
8	8	194	80	0	0	261	551	67	No
9	2	83	65	28	66	368	629	24	No
…	…	…	…	…	…	…	…	…	…
2000	2	81	72	15	76	301	547	25	No

3.3 Random Forest Algorithm Implementation

The dataset that has been transformed will be divided using K-Fold Cross Validation (k = 10) to produce test data and training data to be used in the classification process. The implementation of the random forest algorithm is carried out using RapidMiner tools . In the diabetes dataset, there are two classes, namely "Yes Diabetes" and "No Diabetes".

Perform classification accuracy performance calculations using the Random Forest Algorithm before optimizing using Particle Swarm Optimization (PSO). The results of the test will be displayed in the form of a Confusion Matrix.

Table 4. Confusion Matrix Random Forest

Accuracy: 78.20 % +/- 2.39% (Micro Average: 78.20%)			
	Yes	True No	Class Precision
Pred. Yes	285	37	88.51%
Pred. No	399	1279	76.22%
Class Recall	41.67%	97.19%	

The Confusion Matrix in Table 4 displays an accuracy value of 78.2%. Then conduct trials on the data that has been divided using K-Fold Cross Validation (k = 10) with the random forest algorithm and adding PSO optimization. The results of testing the random forest algorithm using PSO optimization.

Table 5. Confusion Matrix Random Forest Optimization PSO

Accuracy: 82.10 % +/- 1.70% (Micro Average: 82.10%)			
	Yes	True No	Class Precision
Pred. Yes	362	36	90.95%
Pred. No	322	1280	79.90%
Class Recall	52.92%	97.26%	

The results of testing the random forest algorithm using PSO obtained an accuracy value that increased from the results of previous tests. Based on Table 5, the classification results using the random forest algorithm with PSO optimization on the RapidMiner tools produce an accuracy value of 82.1%. The following graph compares the accuracy values of the random forest and random forest algorithms with PSO optimization.

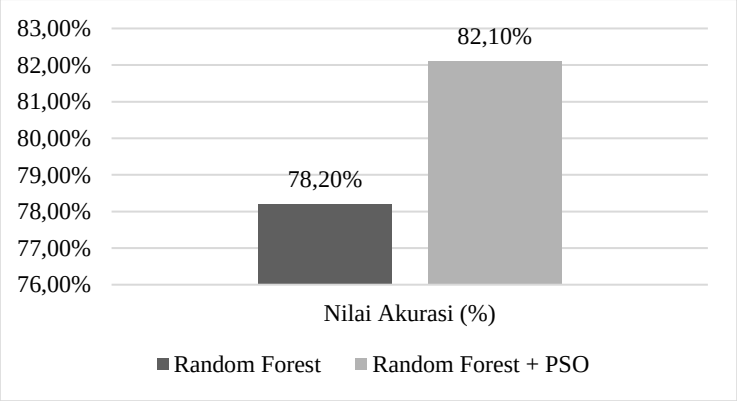

Figure 2. Accuracy of Random Forest and Random Forest + PSO

Based on the graph in Figure 2, the accuracy of the random forest algorithm with PSO optimization has an increased value of 3.9%.

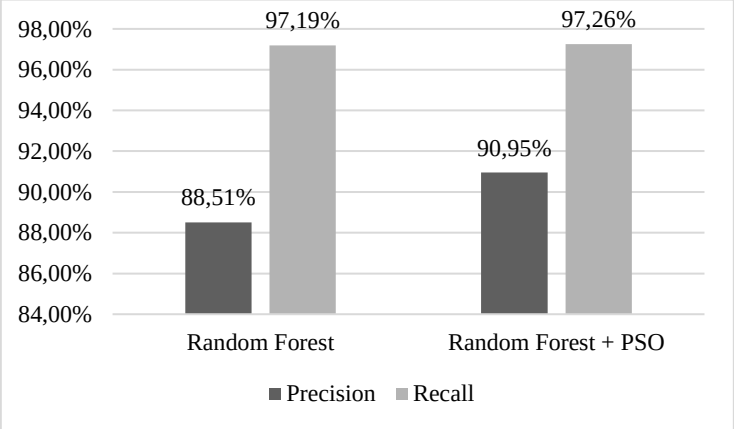

Figure 3. Comparison of Recall and Precision Random Forest with Random Forest + PSO

Recall and precision values in Figure 3, the random forest algorithm with PSO optimization also significantly increases. Random forest recall value with PSO optimization gives an increase of 2.44% and a precision of 0.07%.

4. CONCLUSIONS

Based on the results of the classification of Diabetes Disease with data sharing techniques K - Fold Cross Validation (k = 10), the application of Random Forest produces an accuracy value of 78.2%, while the random forest optimized with PSO produces an accuracy of 82.1%. The application of PSO as an attribute weighting in the Diabetes dataset is considered effective and capable of producing better accuracy than a random forest, with an accuracy increase of 3.9%. The random forest algorithm with PSO optimization has an increase in the recall value of 2.44% and 0.07% precision.

REFERENCES

[1] H. Itoh and M. Tanaka, "'Greedy Organs Hypothesis' for sugar and salt in the pathophysiology of non-communicable diseases in relation to sodium-glucose co-transporters in the intestines and the kidney," *Metab. Open*, vol. 13, p. 100169, 2022, doi: https://doi.org/10.1016/j.metop.2022.100169.

[2] A. kumar Dewangan and P. Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," *Int. J. Eng. Appl. Sci.*, vol. 2, no. 5, 2015.

[3] A. Berbudi, N. Rahmadika, A. I. Tjahjadi, and R. Ruslami, "Type 2 diabetes and its impact on the immune system," *Curr. Diabetes Rev.*, vol. 16, no. 5, p. 442, 2020.

[4] W. H. Organization, "Noncommunicable diseases country profiles 2018," 2018.

[5] N. H. Cho *et al.*, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, 2018, doi: https://doi.org/10.1016/j.diabres.2018.02.023.

[6] K. Ogurtsova *et al.*, "IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021," *Diabetes Res. Clin. Pract.*, vol. 183, p. 109118, 2022.

[7] R. Goyal and I. Jialal, "Diabetes mellitus type 2," 2018.

[8] L. C. Truong, "Reducing the Effects of Blood Sugar Infusion of Melastoma Malabathricum L. in Mus Musculus," *J. Asian Multicult. Res. Med. Heal. Sci. Study*, vol. 1, no. 1, pp. 1–10, 2020.

[9] S. A. Onikanni *et al.*, "Mitochondrial defects in pancreatic beta-cell dysfunction and neurodegenerative diseases: Pathogenesis and therapeutic applications," *Life Sci.*, p. 121247, 2022.

[10] D. Glovaci, W. Fan, and N. D. Wong, "Epidemiology of diabetes mellitus and cardiovascular disease," *Curr. Cardiol. Rep.*, vol. 21, pp. 1–8, 2019.

[11] A. D. Setyawati, P. Padila, and J. Andri, "Obesity and Heredity for Diabetes Mellitus among Elderly," *JOSING J. Nurs. Heal.*, vol. 1, no. 1, pp. 26–31, 2020.

[12] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*, vol. 4. John Wiley & Sons, 2014.

[13] R. A. Welikala *et al.*, "Automated detection and classification of oral lesions using deep learning for early detection of oral cancer," *IEEE Access*, vol. 8, pp. 132677–132693, 2020.

[14] N. Maulidah, A. Abdilah, E. Nurlelah, W. Gata, and F. N. Hasan, "Seleksi Fitur Klasifikasi Penyakit Diabetes Menggunakan Particle Swarm Optimization (PSO) Pada Algoritma Naive Bayes," *J. Ilm. Elektron. dan Komput.*, vol. 13, no. 2, pp. 40–48, 2020, [Online]. Available: http://journal.stekom.ac.id/index.php/elkom■page40

[15] F. Dany Prasetya, H. W. Nugroho, and J. Triloka, "Analisa Data Mining Untuk Prediksi Penyakit Hepatitis C Menggunakan Algoritma Decision Tree C.45 Dengan Particle Swarm Optimization," *Pros. Semin. Nas. Darmajaya*, no. April 1989, pp. 198–209, 2022, [Online]. Available: http://archive.ic

[16] S. Benbelkacem, "Random Forests for Diabetes Diagnosis," *2019 Int. Conf. Comput. Inf. Sci.*, pp. 1–4, 2019.

[17] A. Mujumdar and V. Vaidehi, "ScienceDirect ScienceDirect ScienceDirect ScienceDirect Diabetes Prediction using Machine Learning Aishwarya Mujumdar Diabetes Prediction using Machine Learning Aishwarya Mujumdar Aishwarya," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[18] J. B. Raja and S. C. Pandian, "Computer Methods and Programs in Biomedicine PSO-FCM based data mining model to predict diabetic disease," vol. 196, 2020, doi: 10.1016/j.cmpb.2020.105659.

[19] T. Islam, M. Raihan, F. Farzana, N. Aktar, P. Ghosh, and S. Kabiraj, "Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm," pp. 1–6, 2020.

[20] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, no. xxxx, 2021, doi: 10.1016/j.icte.2021.02.004.

[21] C. L. Chowdhary, S. Bhattacharya, S. Hakak, and R. Kaluri, "An Ensemble based Machine Learning model for Diabetic Retinopathy Classification," pp. 1–6, 2020, doi: 10.1109/ic-ETITE47903.2020.235.

[22] L. J. M. Ebrahem, A. A. Sani, and S. Usman, "Predictive Supervised Machine Learning Models for Diabetes Mellitus," *SN Comput. Sci.*, pp. 1–10, 2020, doi: 10.1007/s42979-020-00250-8.

[23] D. Kumar, C. Prabhat, K. Sudhakar, and T. Santosh, "Performance evaluation of classification methods with PCA and PSO for diabetes," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 8, 2020, doi: 10.1007/s13721-019-0210-8.

[24] S. S. Alaoui and B. Aksasse, *Data Mining and Machine Learning Approaches and Technologies for Diagnosing Diabetes in Women*, vol. 1. Springer International Publishing. doi: 10.1007/978-3-030-23672-4.

[25] Q. A'yuniyah *et al.*, "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," *J. Sist. Komput. dan Inform.*, vol. 4, no. 1, p. 72, 2022, doi: 10.30865/json.v4i1.4781.