Recursive Journal of Informatics
Vol. 2, No. 1, March 2024
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Application of C4.5 Algorithm Using Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) for Diabetes Prediction

Dela Rista Damayanti[1], Aji Purwinarko[2]
[1,2]Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Diabetes is the fourth or fifth leading cause of death in most developed countries and an epidemic in many developing countries. Early detection can be a preventive measure that uses a set of existing data to be processed through data mining with a classification process.
Purpose: Investigate the efficacy of integrating the C4.5 algorithm with Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) for improving the accuracy of diabetes prediction models. By employing SMOTE, the study aims to address the class imbalance issue inherent in diabetes datasets, which often contain significantly fewer instances of positive cases (diabetes) than negative cases (non-diabetes). Furthermore, by incorporating PSO, the research seeks to optimize the decision tree construction process within the C4.5 algorithm, enhancing its ability to discern complex patterns and relationships within the data.
Methods/Study design/approach: This study proposes the use of the C4.5 classification algorithm by applying the synthetic minority oversampling technique (SMOTE) and particle swarm optimization (PSO) to overcome problems in the diabetes dataset, namely the Pima Indian Diabetes Database (PIDD).
Result/Findings: From the research results, the accuracy obtained in applying the C4.5 algorithm without the preprocessing process is 75.97%, while the results of the SMOTE application of the C4.5 algorithm are 80%. Meanwhile, applying the C4.5 algorithm using SMOTE and PSO produces the highest accuracy, with 82.5%. This indicates an increase of 6.53% from the classification results using the C4.5 algorithm.
Novelty/Originality/Value: This research contributes novelty by proposing a hybrid approach that combines the C4.5 decision tree algorithm with two advanced techniques, Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO), for the prediction of diabetes. While previous studies have explored the application of machine learning algorithms for diabetes prediction, few have examined the synergistic effects of integrating SMOTE and PSO with the C4.5 algorithm specifically.

Keywords: Diabetes, Data Mining, C4.5 Algorithm, SMOTE, PSO
Received January 12, 2023 / **Revised** January 17, 2023 / **Accepted** March 28, 2023

INTRODUCTION

Diabetes is a chronic disease caused by metabolic disorders because the pancreas fails to produce enough insulin, which causes an increase in blood sugar levels or hyperglycemia [1]. Diabetes is a non-communicable disease, but it can be life-threatening for the sufferer in the long term. The disease is the fourth or fifth leading cause of death in most developed countries and an epidemic in many developing countries [2]. In 2019, according to the International Diabetes Federation (IDF), Indonesia was ranked seventh out of the top ten countries, with 10.7 million people with diabetes. It is estimated to increase to 13.7 million in 2030 [3]. Currently, no specific drug can cure diabetes mellitus, so continuous monitoring and control are needed [4]. Therefore, early disease detection, control, and treatment are essential preventive measures [5]. Along with technological developments, data mining can allow early detection.

Data mining is a scientific discipline that can generate knowledge from processing big data generated by various fields [6]. This technology is used in various fields such as education, the stock market, marketing, and especially in the health sector [7]. There are six data mining classes: anomaly detection, association rule learning, clustering, classification, and regression [8]. One technique that is often used in data mining

[1]*Corresponding author.
Email address: delarista@students.unnes.ac.id (Damayanti)
DOI: 10.15294/rji.v2i1.64928

is classification. Classification performs data analysis from a set of data provided by taking each instance and assigning the instance to a specific class [9]. A decision tree is an algorithm widely used in classification [10]. The decision tree method can be used to classify a case by forming a decision tree consisting of roots and leaves, which are used as prediction classes [11]. The C4.5 algorithm is included in the decision tree method [12].

The C4.5 algorithm is included in developing the iterative dichotomiser three (ID3) algorithm with the same working principle but has better results [13]. Its advantages, namely being the ability to process numerical data (continuous) and categorical (discrete), handle missing attribute values and produce rules that are easy to interpret [14]. However, in real life, other problems will arise in the data set, which often has an unbalanced amount of data in each class and will also cause problems in classification. The data imbalance affects the classification results, which favor the majority class compared to the minority class [15]. In this case, many researchers use synthetic minority oversampling (SMOTE) techniques [16]. The SMOTE process duplicates minor data to the majority of data by generating synthetic or artificial data based on the k-nearest neighbor (KNN) [17]. In addition, the classification process is also weak in dealing with the problem of irrelevant features in the data, which results in a decrease in accuracy [18]. Attributes of datasets with high dimensions and noise are generally handled by feature selection. This is done because, under these conditions, it can reduce the ability of the classification process [19]. In swarm intelligence, PSO is one of the new technologies as an optimization technique to optimize a subset of features inspired by social behavior in nature [20].

Based on the above problems, this research focuses on optimizing the C4.5 algorithm by using SMOTE to overcome data balance and PSO to select relevant attributes to produce classification accuracy in disease prediction. The proposed research is entitled "Application of the C4.5 Algorithm Using Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) for Predicting Diabetes Patients".

THE PROPOSED METHOD

C4.5 Algorithm

The C4.5 algorithm is an algorithm that produces rules and decision trees to improve the accuracy of the prediction process. J. Ross Quinlan introduced the C4.5 algorithm in 1993 to develop the ID3 algorithm for decision tree formation [11]. The C4.5 algorithm covers the weaknesses of ID3 by handling missing values in training data, pruning decision trees, and handling attributes with discrete and continuous values [21]. It is proven that in the learning process, the C4.5 algorithm can perform discretization automatically [22]. In addition, the C4.5 algorithm can provide accuracy and good performance on several classification techniques [23], as well as models that are easily accessible [24].

Synthetic Minority Oversampling Technique

Chawla first proposed SMOTE in 2002 to overcome the problem of class imbalance in datasets using the oversampling method [25]. In overcoming the existing problems, the SMOTE method uses the principle of generating synthetic data for the minority class to have a balanced amount with the majority class. Instead of duplicating data on the minority class, which results in overfitting, this method randomly builds synthetic data based on the KNN value [26]. Data with a numerical scale and synthetic data are not the same as data on a categorical scale. Numerical data is calculated for proximity using Euclidean distance, while categorical data uses mode values [27].

Particle Swarm Optimization

In 1995, the PSO algorithm was introduced by Kennedy and Eberhart as an optimization technique inspired by observations of animals' social behavior, such as birds' flocks of fish and flock theory [28]. PSO is a population-based search algorithm with random initialization and interaction between population members [29]. The PSO search process is updated from iteration to iteration using a population (swarm) of individuals (particles). This algorithm applies many optimizations and feature selection problems [30]. In search of the optimal solution, each particle changes the direction of its search based on the previous best experience (pbest) and the best experience of all other members (gbest). Position and velocity are factors of the state of the particle in the search space [31].

METHOD

Data Collection

The first step in this research is data collection. Researchers search for data as objects in research experiments through literature studies in journals, articles, and scientific works. From the literature study, there are sources of data acquisition used in research. Based on the literature study, the researcher decided that the data used in this study used the PIDD dataset obtained from Kaggle. This dataset is the same as the articles in the literature study to compare the accuracy results.

Data Processing

The data processing stage consists of several stages, namely the preprocessing stage, the data sharing stage, the classification stage, and the evaluation stage using a confusion matrix. This study aims to overcome the problem of datasets in the classification process using the C4.5 algorithm by applying SMOTE and PSO in the preprocessing stage. However, before applying SMOTE and PSO, the best method for overcoming the dataset's missing value problem was compared using the C4.5 algorithm. Two experiments were carried out at the preprocessing stage regarding handling missing values. The first experiment was carried out by ignoring the missing value in the dataset, and then the second was carried out by handling the missing value by using the mean value. The C4.5 algorithm classification flowchart is shown in Figure 1.

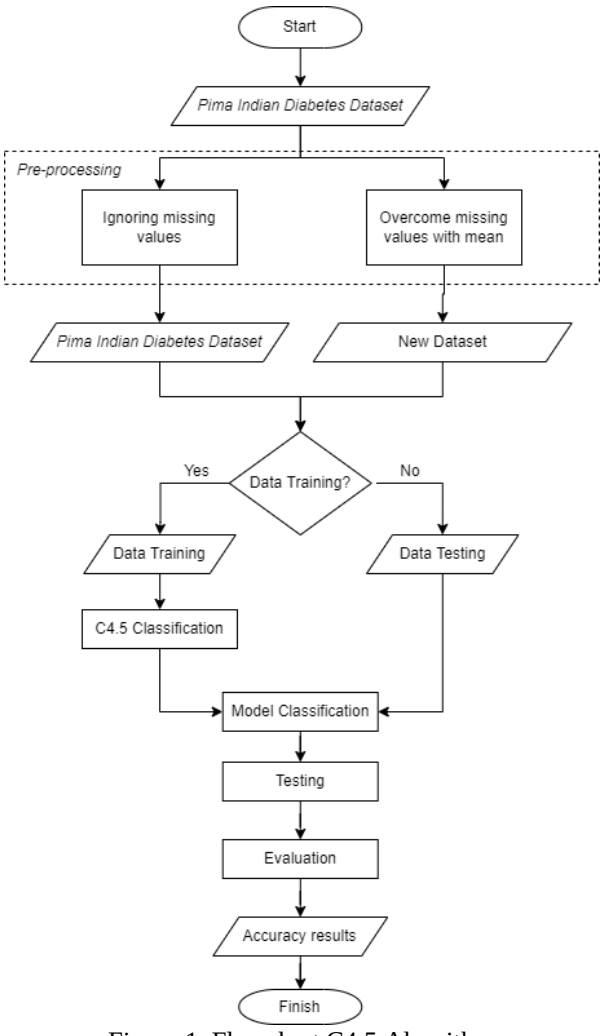

Figure 1. Flowchart C4.5 Algorithm

Based on the classification process using the C4.5 algorithm, the preprocessing stage was tested by applying two experimental methods: ignoring the missing value and overcoming the missing value by using the mean value. The process begins by retrieving the PIDD dataset for the classification process. Next is the preprocessing stage by dealing with the missing value, which produces the original PIDD data, and overcoming the missing value is replaced by using the mean value, which produces a new dataset. Using

the PIDD and the new dataset, the data division process is carried out by splitting the data into two parts with a proportion of 80% for training data and 20% for testing data. The training data's classification process is continued using the C4.5 algorithm, which produces a classification model.

Furthermore, testing is carried out using data testing and produces a confusion matrix which will then be calculated to obtain accuracy. Based on the accuracy obtained from the classification results using the C4.5 algorithm from the two methods in the preprocessing stage, a comparison of the accuracy results was conducted. The highest accuracy value will be used in the C4.5 algorithm classification process with the application of SMOTE and PSO. The research flowchart with the method proposed in this study is shown in Figure 2.

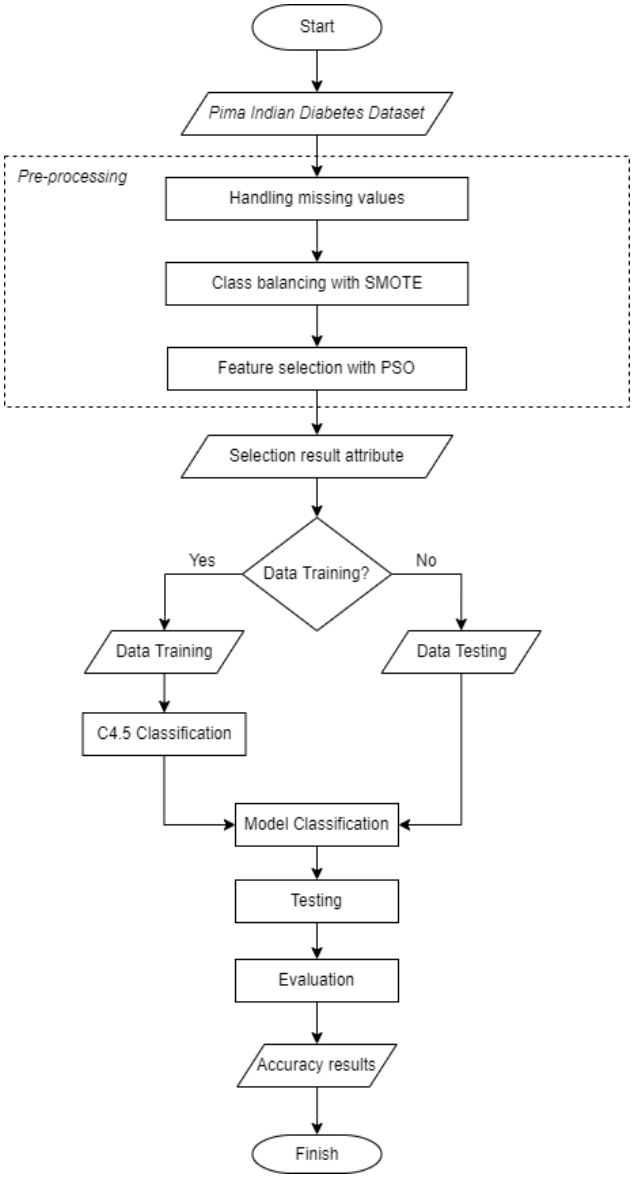

Figure 2. Flowchart of the Proposed Method

Based on the two classification processes, the C4.5 algorithm with the application of SMOTE and PSO has the same step flow as the classification process using the C4.5 algorithm. The addition process is carried out at the preprocessing stage by adding data balancing using SMOTE and feature selection using PSO after the data is finished handling missing values. Based on the implementation of SMOTE and PSO produces a new dataset that has balanced data and selected features.

RESULT AND DISCUSSION

Result

Data Collection

The researcher searched data as objects in a research experiment through a literature study in the data collection technique. This study uses a public dataset from the Kaggle website, the Pima Indian Diabetes Database (PIDD). PIDD has a total of 768 cases, all which patients are women aged at least 21 years with nine attributes. There are eight numeric value attributes and one nominal type of class attribute. The class attribute has two values, namely '0' and '1', where the label '0' is the class for the negative test for diabetes and the label '1' is the class for the positive test for diabetes. The amount of data on the class attribute consists of 500 (65.1%) cases in class '0' and 268 (34.9%) cases in class '1', which indicates a data imbalance problem. The PIDD attributes of the dataset are described in detail in Table 1.

Table 1. Pima Indian Diabetes Database

Num.	Attribute	Description	Attribute Type
1	Pregnancies	Patient pregnancy rate	Numerical
2	Glucose	Blood sugar level	Numerical
3	Blood Pressure	Blood pressure (mm Hg)	Numerical
4	Skin Thickness	Skinfold thickness (mm)	Numerical
5	Insulin	2 hours serum insulin (mu U/ml)	Numerical
6	BMI	Weight (kg)/height (m)2	Numerical
7	Diabetes Pedigree Function	Diabetic Descendants	Numerical
8	Age	Patient age	Numerical
9	Outcome	Non-diabetic or diabetic (within 5 years)	Nominal

Data Processing

Missing Value Handled

The existence of missing data in a dataset causes problems in finding knowledge and affects the level of accuracy. Therefore, it is necessary to handle missing values so that no information is lost. The missing value in PIDD is marked with a value of 0. The handling of missing values in this study is to fill in the values using the mean value. The mean is a value obtained from the average value of all records in one attribute. The mean value of each attribute to fill in the missing value is shown in Table 2.

Table 2. Mean Value to Fill in Missing Value

Attribute	Mean Value
Glucose	121
Blood Pressure	72
Skin Thickness	29
Insulin	155
BMI	32

Data Balancing with SMOTE

The SMOTE stage balances the data by oversampling, which generates synthetic data based on the KNN value so that the data becomes balanced. The initial PIDD dataset had 768 records, with 500 labeled '0' or negative for diabetes and 268 labeled '1' or positive for diabetes. Therefore, it is necessary to balance the data by creating synthetic data for the minority class. In this PIDD, the minority class is class '1' or positive diabetes. The synthetic data produced by SMOTE is 232 data in class 1, so the total data after the application of SMOTE there are 1000 new data.

Feature Selection with PSO

The list of attributes and optimal solutions in the form of selected features is represented using the value one after 100 iterations, as shown in Table 3.

Table 3. Featured and gbest

Num	Attribute	Representation
1	Pregnancies	1
2	Glucose	1
3	BloodPressure	1
4	SkinThickness	1
5	Insulin	0
6	BMI	1
7	DiabetesPedigreeFunction	1
8	Age	1
	gbest	82.5

The implementation of feature selection in this study is used to select the best features, which are then applied in the classification process using the C4.5 algorithm. The PSO algorithm is applied because this algorithm works by choosing the best feature set in PIDD.

Splitting Data

Splitting data in the data mining process divides data into two parts: training data and testing data. The splitting data ratio in this study refers to the research conducted by Azad et al. (2021), with the best accuracy results in a comparison ratio of 80% for training data and 20% for testing data. The splitting data method in this study uses the data splitting method by utilizing the sklearn library. Data distribution is done randomly by determining the consistency of randomization (random state) with a certain amount.

Data Mining

The data mining stage is carried out by researching the C4.5 algorithm and combining SMOTE and PSO methods. This stage has three data mining processes. Each method used in this study will be implemented on the PIDD dataset. Next, a comparison is made based on the accuracy of each method.

The first stage in the data mining process is to classify using the C4.5 algorithm on the dataset without overcoming the problem of data imbalance with SMOTE and without feature selection with PSO. The application of the C4.5 algorithm obtains accurate results, shown in Table 4.

Table 4. C4.5 Algorithm Accuracy Results

Algorithm	Accuracy Results
C4.5	75.32%
C4.5 + Missing Value Handling	75.97%

The resulting accuracy in applying the C4.5 algorithm without handling the missing value in classifying the PIDD is 75.32%, while by handling the missing value, the accuracy is 75.97%. The accuracy obtained has proven that the preprocessing stage, handling missing values using the mean, can provide good accuracy results in classifying PIDD datasets. Therefore, handling missing values by using the mean value is applied in this study. Meanwhile, this result can still be improved by overcoming the problem of data imbalance in the dataset using the SMOTE method.

The second stage is applying the classification using the C4.5 algorithm by overcoming the data imbalance problem with SMOTE. The results of the oversampling process with SMOTE form 232 synthetic data for class '1' or positive diabetes with a total of 1000 new data. The application of the C4.5 algorithm with class balancing using SMOTE obtains accurate results, which can be seen in Table 5.

Table 5. C4.5 Algorithm Accuracy Results on SMOTE Results Data

Algorithm	Accuracy Results
C4.5 + SMOTE	80%

The accuracy generated by applying the C4.5 algorithm with SMOTE results obtained a higher accuracy than the C4.5 algorithm without the SMOTE process. The accuracy of this algorithm's application results is 80%, an increase of 4.03% from the results of the previous method. The accuracy obtained has proven that the C4.5 algorithm with the SMOTE process can classify the PIDD dataset well and improve accuracy results. This result can still be improved by selecting features on the dataset using PSO.

The last stage is to apply classification using the C4.5 algorithm to the data resulting from the application of SMOTE and feature selection using PSO. The final process in this research is combining the application of SMOTE and PSO as feature selection in the dataset before the classification process using the C4.5 algorithm. The best features are searched through the PSO algorithm from the results of oversampling using SMOTE. PSO produces the most significant features for optimizing the classification process in the C4.5 algorithm. The application of the C4.5 algorithm with the combination of SMOTE and PSO obtains accurate results, which can be seen in Table 6.

Table 6. C4.5 Algorithm Accuracy Results with SMOTE and PSO

Algorithm	Accuracy Results
C4.5 + SMOTE + PSO	82.5%

The classification process with the C4.5 algorithm using SMOTE has an accuracy of 80%. While the accuracy generated by the combination of SMOTE and PSO in the classification using the C4.5 algorithm produces an accuracy of 82.5%. From this comparison of accuracy, there is an increase in accuracy of 2.5%

in applying the C4.5 algorithm using SMOTE and PSO for predicting people with diabetes in the PIDD dataset. The application of PSO increases accuracy because PSO chooses the best feature set in the dataset that works on a population basis. Each particle in the population searches for the global optimum solution step by step in each iteration based on the experience of the particles and groups. From the iteration results that reach the optimum solution, we get the gbest and selected features with the highest accuracy level.

Discussion

This research focuses on handling the data imbalance problem and feature selection on the dataset to optimize the classification algorithm. From the problems that have been described, the SMOTE and PSO methods are applied to overcome the problems contained in the dataset. The ability of the C4.5 algorithm to classify the PIDD dataset is seen by comparing the accuracy of the C4.5 algorithm before and after applying SMOTE and PSO as feature selection. This study was conducted to determine the ability of the C4.5 algorithm in classifying datasets by applying SMOTE to overcome the problem of data imbalance and PSO as a feature selection algorithm to find the most significant features.

In this study, we managed to record the accuracy results of each data mining process, namely C4.5, C4.5 - handling missing values, C4.5-SMOTE, and C4.5-SMOTE-PSO that have been carried out. The accuracy results are shown in Table 7.

Table 7. Accuracy Results of Each Method

Algorithm	Accuracy
C4.5	75,32%
C4.5 – Missing Value Handling	75,97%
C4.5 - SMOTE	80%
C4.5 - SMOTE - PSO	82,5%

Table 7 shows that there is an increase in the accuracy of each method used in the data mining process. The classification process using C4.5 by ignoring the missing value in the dataset produces an accuracy of 75.32%. In comparison, the classification process using C4.5 by handling missing values using the mean value produces an accuracy of 75.97%. There is an increase in accuracy of 0.65% due to the effect of handling missing values using the mean value so that there is no information on missing attributes. Furthermore, the application of SMOTE in the C4.5 algorithm classification process resulted in an accuracy of 80%, with an increase of 4.03% from the process that only applied the C4.5 algorithm with missing value handling.

Meanwhile, the proposed method in this study has the highest accuracy results among the methods that have been tested. The application of SMOTE and PSO on the C4.5 algorithm produces an accuracy of 82.5%. Obtain increasing accuracy after PSO is applied because of the influence on the attributes used in the classification process. The attributes used in the classification process are selected based on the results of the best experiences of particles and their flocks. These are carried out in stages until maximum iterations achieve the highest fitness value. Based on this description, it can be concluded that the application of SMOTE and PSO to the C4.5 algorithm contributes to increasing accuracy by 6.53% compared to the application using the C4.5 algorithm.

The comparison accuracy of this study is also compared with other studies to prove that the method applied in this study has advantages over existing methods. Researchers compare the accuracy results obtained from this study using the same dataset, namely PIDD. The results of the comparison can be seen in Table 8.

Table 8. Comparison of Research Accuracy

Researcher	Method	Result
Noviandi (2018)	C4.5	70.32%
Choubey *et al.* (2020)	PSO-C4.5	74.7826%
Azad *et al.* (2021)	SMOTE-CFS-GA-C4.5	80.1932%.
Proposed Method	SMOTE-PSO-C4.5	82.5%

Suppose the method used in this study is compared with previous studies. In that case, the accuracy produced in this study is 82.5% which indicates it is better than several previous studies using the PIDD dataset, as shown in Table 8. [32] research in predicting diabetes using the C4.5 decision tree algorithm

obtained an accuracy of 70.32%. [33] in a study to detect diabetes using the PIDD dataset using two approaches, namely the first approach with a classification process through several algorithms, one of which is the C4.5 decision tree. The second approach uses the PSO algorithm for feature reduction before the dataset classification uses the first approach. From the results of classification and feature reduction using the PSO algorithm, an accuracy of 74.7826% is obtained on the PIDD dataset. [34], in their research, showed the results of the SMOTE and CFS-GA combination process in conducting the classification process using the C4.5 decision tree algorithm resulted in an accuracy of 80.1932%.

Based on the accuracy of previous studies, it can be concluded that the proposed method by applying SMOTE and PSO on yahoo C4.5 is better because it has higher accuracy than the previous method. The application of the SMOTE method to overcome the problem of data imbalance in the dataset and feature selection using PSO to select significant features in the classification process using the C4.5 algorithm is an advantage in this study. The classification process carried out on PIDD by applying the selected method can improve the performance of C4.5 so that it becomes more optimal and obtains better accuracy.

CONCLUSION

The problem of class imbalance in the PIDD dataset is handled by oversampling using SMOTE. The results of SMOTE make the data have a balanced number of classes so that the classification is no longer in favor of the majority class. From the beginning, PIDD with a total of 768 data to 1000 data with classes '0' and '1' has a balanced condition. Furthermore, the feature selection in the dataset resulting from SMOTE is applied by PSO to find the best feature set for further classification. PSO searches several features in the dataset to determine the most relevant features. There are eight attributes in the initial dataset before the feature selection process. After PSO is applied, the selected features become seven attributes with 1 class attribute. After the feature selection process with the most optimal feature results, splitting the data into training data and testing data with a proportion of 80% for training data and 20% for testing data to be implemented on the C4.5 algorithm. The selected feature set makes the classification performance using the C4.5 algorithm more optimal. The evaluation results using the confusion matrix in the form of accuracy obtained from the classification process using the C4.5 algorithm obtained an accuracy of 75.97%. Then the SMOTE and PSO methods were applied as feature selection in the C4.5 algorithm, and an accuracy of 82.5% was obtained. It was concluded that the application of SMOTE and PSO on the C4.5 algorithm increased the accuracy of making predictions by 6.53%.

REFERENCES

[1] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput Struct Biotechnol J*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[2] D. Atlas, "International diabetes federation," *IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation*, vol. 33, no. 2, 2015.

[3] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res Clin Pract*, vol. 157, p. 107843, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[4] M. S. Diab, S. Husain, and A. Jarndal, "On Diabetes Classification and Prediction using Artificial Neural Networks," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, IEEE, Nov. 2020, pp. 1–5. doi: 10.1109/CCCI49893.2020.9256621.

[5] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Inform Med Unlocked*, vol. 15, p. 100178, 2019, doi: 10.1016/j.imu.2019.100178.

[6] M. A. Muslim, A. J. Herowati, E. Sugiharti, and B. Prasetiyo, "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease," *J Phys Conf Ser*, vol. 983, p. 012062, Mar. 2018, doi: 10.1088/1742-6596/983/1/012062.

[7] P. Mayadewi and E. Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," *SESINDO 2015*, vol. 2015, 2015.

[8] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of various decision tree algorithms for classification in data mining," *Int J Comput Appl*, vol. 163, no. 8, pp. 15–19, 2017.

[9] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," *Oriental Journal of Computer Science and Technology*, vol. 8, no. 1, pp. 13–19, 2015.

[10] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," in *2017 International Conference on Signal Processing and Communication (ICSPC)*, IEEE, Jul. 2017, pp. 264–268. doi: 10.1109/CSPC.2017.8305851.
[11] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
[12] A. P. Wibawa, M. Guntur, A. Purnama, M. F. Akbar, and F. A. Dwiyanto, "Metode-metode klasifikasi," in *Prosiding Seminar Ilmu Komputer dan Teknologi Informasi*, 2018.
[13] U. Pujianto, A. L. Setiawan, H. A. Rosyid, and A. M. M. Salah, "Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement," *Knowledge Engineering and Data Science*, vol. 2, no. 2, p. 58, Dec. 2019, doi: 10.17977/um018v2i22019p58-71.
[14] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4. 5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 13–19, 2014.
[15] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl Soft Comput*, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.
[16] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced Big Data classification: outcomes and challenges," *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 105–120, Jun. 2017, doi: 10.1007/s40747-017-0037-9.
[17] M. Mustaqim, B. Warsito, and B. Surarso, "Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan," *Register*, vol. 5, no. 2, pp. 116–127, 2019.
[18] H. W. Nugroho, T. B. Adji, and N. A. Setiawan, "Random forest weighting based feature selection for c4. 5 algorithm on wart treatment selection method," *Int. J. Adv. Sci. Eng. Inf. Technol*, vol. 8, no. 5, pp. 1858–1863, 2018.
[19] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, "Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis," *J Phys Conf Ser*, vol. 983, p. 012063, Mar. 2018, doi: 10.1088/1742-6596/983/1/012063.
[20] C.-L. Huang and J.-F. Dun, "A distributed PSO–SVM hybrid system with feature selection and parameter optimization," *Appl Soft Comput*, vol. 8, no. 4, pp. 1381–1391, Sep. 2008, doi: 10.1016/j.asoc.2007.10.007.
[21] M. H. Dunham, *Data mining: Introductory and advanced topics*. Pearson Education India, 2006.
[22] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 194–202. doi: 10.1016/B978-1-55860-377-6.50032-3.
[23] A. Nurzahputra and M. A. Muslim, "Peningkatan Akurasi Pada Algoritma C4. 5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit," *Prosiding SNATIF*, pp. 243–247, 2017.
[24] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Human talent prediction in HRM using C4. 5 classification algorithm," *International Journal on Computer Science and Engineering*, vol. 2, no. 8, pp. 2526–2534, 2010.
[25] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
[26] Q. Gu, X.-M. Wang, Z. Wu, B. Ning, and C.-S. Xin, "An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification," *Journal of Digital Information Management*, vol. 14, no. 2, pp. 92–103, 2016.
[27] S. Susanti, "Klasifikasi Kemampuan Perawatan Diri Anak dengan Disabilitas Menggunakan SMOTE Berbasis Neural Network," *Jurnal Informatika*, vol. 6, no. 2, pp. 175–184, 2019.
[28] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, IEEE, pp. 1942–1948. doi: 10.1109/ICNN.1995.488968.
[29] Y. Shi, "Particle swarm optimization," *IEEE connections*, vol. 2, no. 1, pp. 8–13, 2004.
[30] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Syst Appl*, vol. 35, no. 4, pp. 1817–1824, Nov. 2008, doi: 10.1016/j.eswa.2007.08.088.
[31] B. Liu, L. Wang, and Y.-H. Jin, "An effective hybrid PSO-based algorithm for flow shop scheduling with limited buffers," *Comput Oper Res*, vol. 35, no. 9, pp. 2791–2806, Sep. 2008, doi: 10.1016/j.cor.2006.12.013.

[32] N. Noviandi, "Implementasi algoritma decision tree c4. 5 untuk prediksi penyakit diabetes," *Indonesian of Health Information Management Journal (INOHIM)*, vol. 6, no. 1, pp. 1–5, 2018.
[33] D. K. Choubey, P. Kumar, S. Tripathi, and S. Kumar, "Performance evaluation of classification methods with PCA and PSO for diabetes," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, p. 5, Dec. 2020, doi: 10.1007/s13721-019-0210-8.
[34] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimed Syst*, vol. 28, no. 4, pp. 1289–1307, Aug. 2022, doi: 10.1007/s00530-021-00817-2.