JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6676

ENHANCING MACHINE LEARNING ALGORITHM PERFORMANCE FOR PCOS DIAGNOSIS USING SMOTENC ON IMBALANCED DATA

Rofiqoh Dewi[1]; Ratna Sri hayati[1]; Alfa Saleh[2*]; Dahri Yani Hakim Tanjung[1]; Abwabul Jinan[3]

Bisnis Digital[1]
Informatika[3]
Universitas Satya Terra Bhinneka, Medan, Indonesia[1,3]
http://satyaterrabhinneka.ac.id[1,3]
rofiqohdewi@satyaterrabhinneka.ac.id; ratnasrihayati@satyaterrabhinneka.ac.id;
dahritanjung@satyaterrabhinneka.ac.id; abwabuljinan@satyaterrabhinneka.ac.id

Informatika[2*]
Universitas Samudra, Aceh, Indonesia[2]
http://unsam.ac.id[2]
alfasaleh@unsam.ac.id*

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract—*Polycystic Ovarian Syndrome (PCOS) is one of the most frequently occurring endocrine disorders in women of reproductive age, distinguished by disruptions in hormonal regulation that can impact menstrual cycles, fertility, and physical appearance. Despite its high prevalence, PCOS is often diagnosed late and inaccurately, leading to inappropriate treatment and long-term health issues for patients. Machine learning can serve as an effective solution to enhance the accuracy of PCOS diagnosis. However, one of the primary challenges encountered is the class imbalance in the dataset, where the number of positive case data (PCOS) is often significantly lower than the negative case data. This imbalance can result in a biased model that is less effective in predicting the actual condition of patients. In this study, the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) method is recommended to address the issue of imbalanced data, thereby improving the performance and accuracy of the machine learning model employed. The evaluation matrix test results clearly demonstrate that the accuracy of each machine learning model improved after applying the SMOTENC method. Specifically, the accuracy of the K-Nearest Neighbors (KNN) algorithm increased from 81.6% to 89.8%, the Support Vector Machine (SVM) algorithm from 90.6% to 92.5%, the Naive Bayes algorithm from 70% to 82.3%, and the C4.5 algorithm from 99.6% to 99.7%. This research provides a substantial contribution to advancing the development of diagnostic methods thatare both more precise and efficient.*

Keywords*: imbalanced data, machine learning algorithm, PCOS, SMOTENC*

Intisari— *Sindrom Ovarium Polikistik (PCOS) merupakan salah satu gangguan endokrin yang paling sering terjadi pada wanita usia reproduktif, yang ditandai dengan gangguan pada regulasi hormon yang dapat memengaruhi siklus menstruasi, kesuburan, dan penampilan fisik. Meskipun prevalensinya tinggi, PCOS sering kali terlambat didiagnosis dan tidak akurat, sehingga menyebabkan pengobatan yang tidak tepat dan masalah kesehatan jangka panjang bagi pasien. Pembelajaran mesin dapat menjadi solusi yang efektif untuk meningkatkan akurasi diagnosis PCOS. Namun, salah satu tantangan utama yang dihadapi adalah ketidakseimbangan kelas dalam dataset, di mana jumlah data kasus positif (PCOS) sering kali jauh lebih rendah daripada data kasus negatif. Ketidakseimbangan ini dapat menghasilkan model yang bias dan kurang efektif dalam memprediksi kondisi pasien yang sebenarnya. Dalam penelitian ini, metode Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) direkomendasikan untuk mengatasi*

masalah ketidakseimbangan data, sehingga dapat meningkatkan kinerja dan akurasi model pembelajaran mesin yang digunakan. Hasil uji matriks evaluasi dengan jelas menunjukkan bahwa akurasi setiap model machine learning meningkat setelah menerapkan metode SMOTENC. Secara khusus, akurasi algoritma K-Nearest Neighbors (KNN) meningkat dari 81,6% menjadi 89,8%, algoritma Support Vector Machine (SVM) dari 90,6% menjadi 92,5%, algoritma Naive Bayes dari 70% menjadi 82,3%, dan algoritma C4.5 dari 99,6% menjadi 99,7%.

***Kata Kunci**: data tidak seimbang, algoritma pembelajaran mesin, PCOS, SMOTENC.*

INTRODUCTION

The rapid advancement of technology has significantly enhanced individuals' access to a broad spectrum of health-related information [1]. Despite this accessibility, many people including men, women, the elderly, and children continue to lead unhealthy lifestyles on a daily basis. For instance, the consumption of fast food, smoking, exposure to air pollution, and the use of food additives can adversely impact physical health and hormonal balance. Additionally, irregular lifestyles characterized by insufficient sleep, lack of exercise, and prolonged indoor work can lead to reduced mobility and may eventually trigger various health problems.

An unhealthy and irregular lifestyle can lead to various health problems, one of which is hormonal imbalance-exemplified by Polycystic Ovarian Syndrome (PCOS), a condition that frequently affects women. PCOS is represents a prevalent endocrine pathology affecting women during their reproductive years[2], marked by disturbances in hormonal homeostasis that can impact the menstrual cycle, fertility, and physical appearance. According to various medical journals, PCOS is often associated with insulin resistance[3], hyperandrogenism (elevated levels of male hormones)[4], and chronic anovulation[5]. Despite its high prevalence, the diagnosis of PCOS is often delayed and inaccurate[6], leading to inappropriate treatment and long-term health complications for patients.

Within this framework, leveraging machine learning techniques presents an effective approach to enhance the diagnostic accuracy for Polycystic Ovarian Syndrome (PCOS). Machine learning algorithms have demonstrated the capability to manage large and complex datasets [7, 8, 9, 10,11], producing predictive models that surpass traditional methods in accuracy. However, a significant challenge encountered is the class imbalance within the dataset [12,13,14,15], where the number of positive case data (PCOS) is often considerably smaller than the negative case data. This imbalance can result in biased models that are less effective in accurately predicting the patient's actual condition.

In the context of utilizing machine learning for PCOS diagnosis, various previous studies have demonstrated the potential and success of this approach.

Table 1. The Comparison with Previous Studies

Reference	Algorithm	Accuracy
[16]	LR, SVM, RF, Gradient Boost, MLP	85.00%
[17]	VM, XGBoost with LASSO + SVM-RFE	87.50%
[18]	(RF, XGB, LR, KNN) + ADASYN + BORUTA	92.00%
[19]	Linear SVM	91.60%
[20]	MLP,SVM,RBF	93.00%
[21]	BorutaShap, RF	86.00%
[22]	CatBoost	90.10%
[23]	RF	93.25%
[24]	Multi-Stacking ML	98.00%
[25]	SMOTE + LR, RF, DT, SVM and KNN	97.11%

In a comprehensive study, Dutta P. (2021) examined the use of machine learning techniques for PCOS diagnosis and highlighted that imbalanced data can significantly compromise the accuracy of predictive outcomes. They employed the SMOTE method and indicated that data optimization can enhance the quality of the resulting model, although the technique is limited to handling numerical features[25].

In response to the challenges outlined above, this study advocates the application of SMOTENC features as a means of addressing class imbalance. SMOTENC is an oversampling technique that not only increases the number of samples from minority classes but also preserves the distribution of continuous and categorical features in the dataset[26]. By employing SMOTENC, imbalanced data can be processed to achieve greater balance, Consequently, this leads to improved effectiveness and accuracy in the implementation of the machine learning algorithms.

This research is expected to optimize machine learning through the application of SMOTENC in the diagnosis of PCOS. Thus, this substantially advances the development of

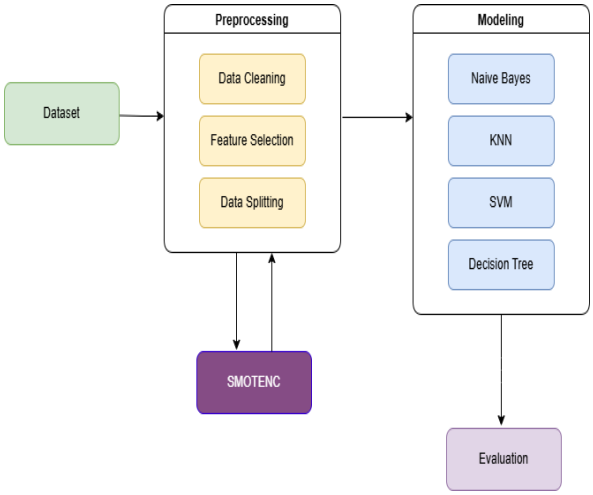
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6676

diagnostic methodologies that are both more precise and efficient.

MATERIALS AND METHODS

To facilitate each stage in the application of machine learning algorithms for PCOS diagnosis, a research methodology is required to guide the process. An illustration of the research methodology employed in this study is presented in Figure 1 below.



Source : (Research Result, 2025)

Figure 1. Research Methodology Flowchart

Dataset

In this study, a PCOS diagnosis dataset consisting of 1,000 records was utilized. This dataset includes several numerical and categorical features related to PCOS. Among these entries, 199 are diagnosed with PCOS, while 801 are diagnosed without PCOS.

Preprocessing

This stage is crucial in machine learning, as the provided dataset is often not properly structured and may contain missing, incomplete, or noisy data. During this phase, data cleaning will be performed to remove noise. After cleaning, feature selection will be conducted to identify features that significantly impact the classification process for predictive purposes, while removing those that do not. The selected features for classification include Age, Body Mass Index (BMI), menstrual cycle irregularities, serum testosterone concentration (ng/dL), and antral follicle quantification.

Subsequently, the cleaned dataset with the selected influential features will be divided into training and test data. In the present study, the dataset is partitioned in a 70:30 ratio, allocating 70% of the data for model training and the remaining 30% for validation purposes.

SMOTENC

To mitigate the class imbalance in the PCOS diagnosis dataset, the SMOTENC technique was applied to synthetically oversample instances from the minority class. Initially, the dataset contained 1,000 entries; however, after applying SMOTENC, it expanded to 1,602 entries, comprising 801 data points for individuals with PCOS and 801 data points for individuals without PCOS.

Table 2. Comparison of Oversampling method[25], [27], [28], [29]

Aspect	SMOTE-NC	SMOTE	SMOTEN
Ability to Handle Categorical Features	Specifically designed for datasets with mixed features (numerical and categorical)	Not designed for categorical features	designed for categorical features
Performance on Mixed Datasets	Effective	Less effective	Less effective
Category Distribution Preservation	Maintaining the original category distribution	Not Maintaining the original category distribution	Maintaining category distribution but can amplify noise if the original data has incorrect labels
Risk of Overfitting	Lower	Higher	Medium

Source : (Research Result, 2025)

Modeling

Naive bayes(NB)

Naive Bayes is a probabilistic machine learning algorithm commonly employed for classification, which operates based on the principles of Bayes' theorem[30]. The equations used in the Naive Bayes algorithm are as follows.

$$P(C_i|X) = \frac{\prod_{j=1}^{k} P(A_j = x_j|C_i).P(C_i)}{P(X)} \quad (1)$$

In the application of the Naive Bayes algorithm, there is always a possibility that the probability of a category is zero. To address this issue, the Laplace smoothing technique is applied.

K-Nearest Neighbour (KNN)

K-Nearest Neighbor (KNN) is a non-parametric machine learning algorithm widely used for classification, which operates by identifying the closest training examples in the feature space [31].

The equation for calculating the nearest distance in the KNN algorithm is based on the Euclidean distance formula, as shown below[32].

$$d_i = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (2)$$

Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm is a supervised learning technique widely employed for classification tasks, wherein it determines the optimal hyperplane to effectively distinguish between different data classes.[33]. The linear SVM algorithm equation used in this study is as follows.

$$f(x) = \sum_{i=0}^{N} a_i y_i x_i^T . x + \beta_0 \quad (3)$$

For problems that cannot be linearly separated, the above equation can be modified using the SVM kernel, as shown below.

$$f(x) = \sum_{i=0}^{N} a_i y_i K(x_i . x), x + \beta_0 \quad (4)$$

Decision Tree (C4.5)

C4.5 is a decision tree-based machine learning algorithm designed to generate classification models by recursively partitioning data based on attribute values[34]. The stages in the C4.5 algorithm focus on determining the entropy value using the following equation.

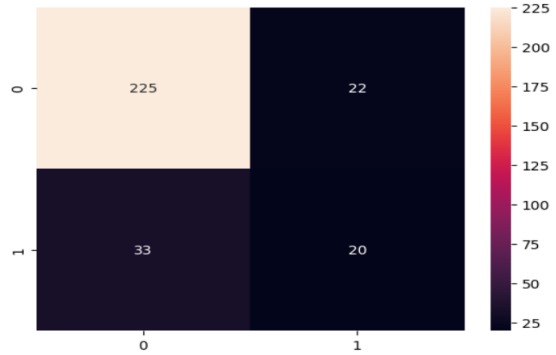
$$Entropy(S) = - \sum_{j=1}^{C} p(S,j) x \log p(S,j) \quad (5)$$

Subsequently, the gain value is calculated from the previously obtained entropy value.

$$Gain(S,T) = Entropy(S) - \sum_{Values(T_j)} \frac{|T_{S,v}|}{|T_S|} Entropy\ (S_v) \quad (6)$$

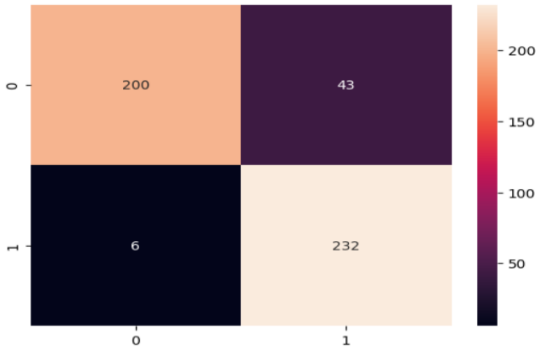
RESULTS AND DISCUSSION

This section presents the results and discussion, encompassing an evaluation of each algorithm's performance through the analysis of confusion matrix, both prior to and following the application of SMOTENC. The confusion matrices corresponding to the KNN algorithm testing are depicted in Figures 2 and 3.



Source : (Research Result, 2025)

Figure 2. Confusion Matrix Algoritma KNN



Source : (Research Result, 2025)

Figure 3. Confusion Matrix Algoritma KNN + SMOTENC

As shown in Figure 2, the results obtained from testing the KNN algorithm prior to the application of SMOTENC reveal the model's performance under imbalanced data conditions, it can be concluded that out of the total 300 data points tested, 20 were correctly classified as PCOS cases, and 225 were correctly classified as non-PCOS cases. However, 22 data points were incorrectly classified as PCOS cases, and 33 were incorrectly classified as non-PCOS cases. In Figure 3, the results of testing the KNN algorithm after applying SMOTENC show that out of 481 data points tested, 232 were correctly classified as PCOS cases, and 200 were correctly classified as non-PCOS cases. Meanwhile, 43 data points were incorrectly classified as PCOS cases, and 6 were incorrectly classified as non-PCOS cases. Table 3 provides a detailed presentation of the evaluation metrices obtained for the KNN algorithm.

Table 3. Evaluation Matrices of KNN Algorithm

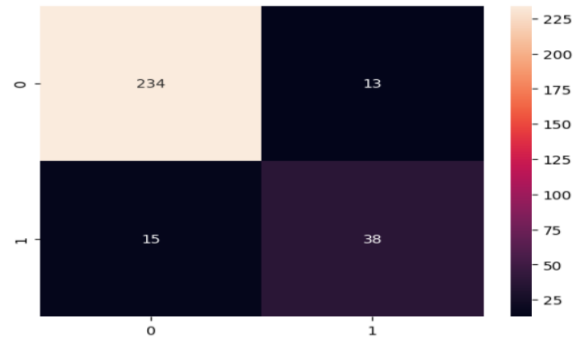
Algorithm	PCOS	Precision	Recall	F1-score
KNN	Yes	0.48	0.38	0.42
	No	0.87	0.91	0.89
KNN + SMOTENC	Yes	0.84	0.97	0.90
	No	0.97	0.82	0.89

Source : (Research Result, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

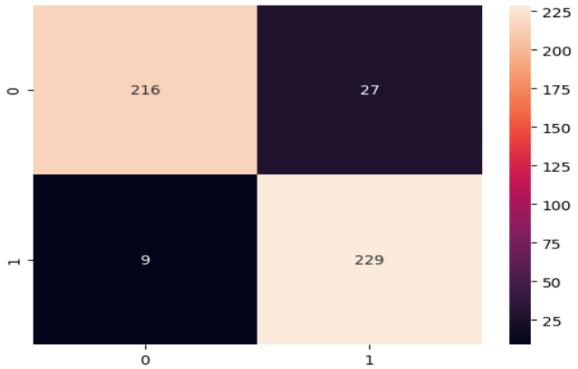
VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6676

In Table 3 above, signs of overfitting can be observed in the Precision, Recall, and F1-Score values of the KNN algorithm without the use of SMOTENC, with scores of 0.48, 0.38, and 0.42 respectively. However, after applying the SMOTENC technique to the KNN algorithm, the Precision, Recall, and F1-Score values became more consistent, reaching 0.84, 0.97, and 0.90, respectively. The confusion matrices for the SVM algorithm tests are presented in Figures 4 and 5.



Source : (Research Result, 2025)

Figure 4. Confusion Matrix Algoritma SVM



Source : (Research Result, 2025)

Figure 5. Confusion Matrix Algoritma SVM + SMOTENC

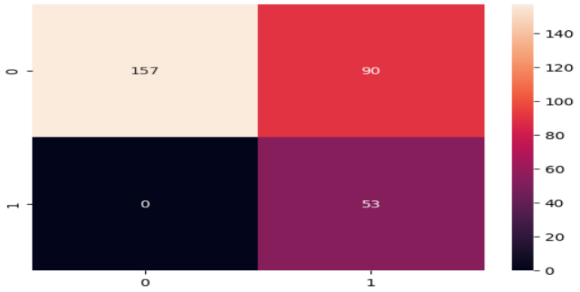
According to the test results obtained from the SVM algorithm before the implementation of the SMOTENC technique, as shown in Figure 4, out of a total of 300 data samples, 38 were correctly classified as PCOS cases, and 234 were correctly classified as non-PCOS cases. Meanwhile, 13 data samples were misclassified as non-PCOS cases (false negatives), and 15 were misclassified as PCOS cases (false positives). In contrast, Figure 5 shows the results of the SVM algorithm after the application of SMOTENC. Out of a total of 481 data samples, 229 were correctly classified as PCOS cases, and 216 were correctly classified as non-PCOS. However, 27 samples were incorrectly classified as non-PCOS cases, and 9 were incorrectly classified as PCOS cases. The evaluation matrices for the SVM algorithm are summarized in Table 4.

Table 4. Evaluation Matrices of SVM Algorithm

Algorithm	PCOS	Precision	Recall	F1-score
SVM	Yes	0.75	0.72	0.73
	No	0.94	0.95	0.94
SVM + SMOTENC	Yes	0.89	0.96	0.93
	No	0.96	0.89	0.92

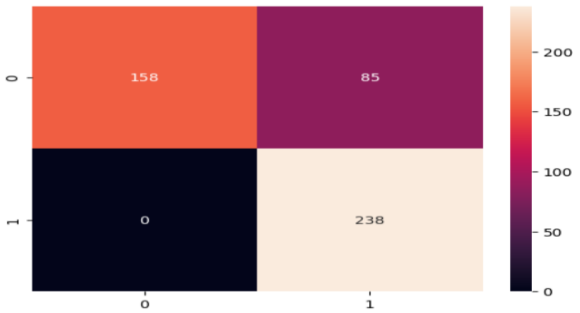
Source : (Research Result, 2025)

In Table 4 above, the Precision, Recall, and F1-Score values obtained from the SVM algorithm test without using SMOTENC are 0.75, 0.72, and 0.73, respectively. After applying the SMOTENC technique to the SVM algorithm, these values improved to 0.89, 0.96, and 0.93, respectively. The confusion matrices for the Naive Bayes (NB) algorithm tests are presented in Figures 6 and 7.



Source : (Research Result, 2025)

Figure 6. Confusion Matrix Algoritma NB



Source : (Research Result, 2025)

Figure 7. Confusion Matrix Algoritma NB + SMOTENC

Based on the results of the Naive Bayes (NB) algorithm test prior to the application of SMOTENC, as shown in Figure 6, out of a total of 300 data samples, 53 were correctly classified as PCOS cases, and 157 were correctly classified as non-PCOS cases. Meanwhile, 90 samples were misclassified as non-PCOS (false negatives), and no data were misclassified as PCOS (false positives). In Figure 7, which shows the NB algorithm results after applying SMOTENC, out of 481 data samples, 238 were correctly classified as PCOS-positive, and 158 were correctly classified as non-PCOS. In contrast, 85 samples were misclassified as non-

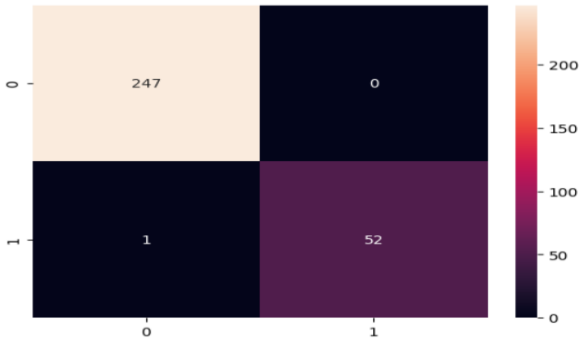
PCOS, and no samples were misclassified as PCOS. The evaluation matrices for the NB algorithm are summarized in Table 5.

Table 5. Evaluation Matrices of NB Algorithm

NB	PCOS	Precision	Recall	F1-score
No SMOTE-NC	Yes	0.37	1.0	0.54
	No	1.0	0.64	0.78
			Accuracy	70%
SMOTE-NC	Yes	0.74	1.0	0.85
	No	1.0	0.65	0.79
			Accuracy	82.3%

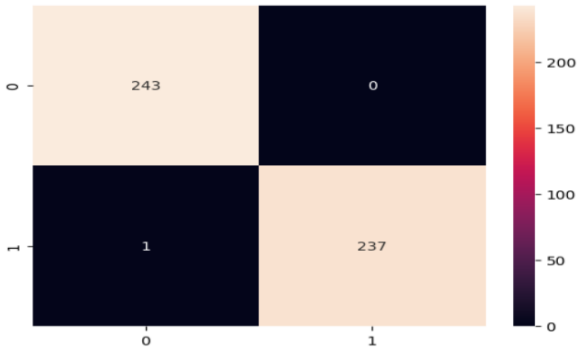
Source : (Research Result, 2025)

In Table 5 above, the precision, recall, and F1-Score values for the Naive Bayes (NB) algorithm test without SMOTENC are 0.37, 1.0, and 0.54, respectively. After applying the SMOTENC technique to the NB algorithm, these values improved to 0.74, 1.0, and 0.85. Figures 8 and 9 display the confusion matrices derived from the evaluation of the C4.5 algorithm.



Source : (Research Result, 2025)

Figure 8. Confusion Matrix Algoritma C4.5



Source : (Research Result, 2025)

Figure 9. Confusion Matrix Algoritma C4.5 + SMOTENC

As shown in Figure 8, based on the results of the C4.5 algorithm test prior to the application of SMOTENC, out of a total of 300 tested data samples, 52 samples were correctly classified as PCOS-positive, while 247 samples were correctly classified as non-PCOS. There were no samples misclassified as non-PCOS (false negatives), but there was one sample misclassified as PCOS (false positive). Next, in Figure 9, which presents the results of the C4.5 algorithm test after the application of SMOTENC, out of a total of 481 tested data samples, 237 samples were correctly classified as PCOS-positive, and 243 samples were correctly classified as non-PCOS. Similar to the previous result, no samples were misclassified as non-PCOS, but there was one sample misclassified as PCOS. . The evaluation matrices for the C4.5 algorithm are summarized in Table 6.

Table 6. Evaluation Matrices of C4.5 Algorithm

C4.5	PCOS	Precision	Recall	F1-score
No SMOTE-NC	Yes	1.0	0.98	0.99
	No	1.0	1.0	1.0
SMOTE-NC	Yes	1.0	0.99	0.99
	No	1.0	1.0	1.0

Source : (Research Result, 2025)

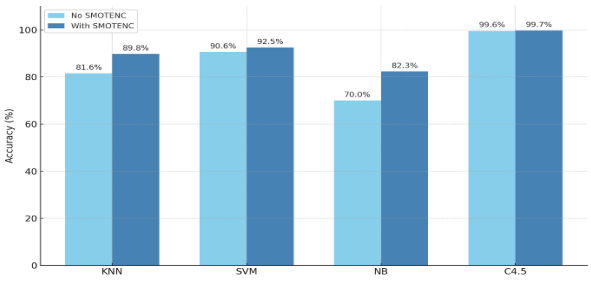
In Table 6, the precision, recall, and F1-Score values for the C4.5 algorithm test without SMOTENC were 1.0, 0.98, and 0.99, respectively. After applying the SMOTENC technique to the C4.5 algorithm, there was a slight improvement in these values to 1.0, 0.99, and 0.99. A comparison of the accuracy values for each algorithm is shown in Table 7.

Table 7. Comparison of Algorithm Accuracies

Algorithm	NO SMOTENC	SMOTENC
KNN	81.6%	89.8%
SVM	90.6%	92.5%
NB	70%	82.3%
C4.5	99.6%	99.7%

Source : (Research Result, 2025)

Based on Table 7, the accuracy of each machine learning algorithms improved in performance after using SMOTENC to address the issue of imbalanced data.



Source : (Research Result, 2025)

Figure 10. Performance Comparison of Machine Learning Algorithms Before and After Applying SMOTENC

Figure 10 presents the classification accuracy achieved by four machine learning algorithms—K-Nearest Neighbor (KNN), Support

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6676

Vector Machine (SVM), Naive Bayes (NB), and C4.5—assessed both before and after the application of SMOTENC. The principal aim of employing SMOTENC was to address class imbalance through oversampling of the minority class, thereby facilitating improved model training and predictive capability.

CONCLUSION

Imbalanced datasets can significantly compromise the efficacy of machine learning algorithms, frequently leading to predictive bias that disproportionately favors the majority class. The implementation of the SMOTENC technique as a method for addressing class imbalance has demonstrated notable effectiveness. This is evidenced by the improvement in the classification performance of several machine learning models following the application of SMOTENC. The experimental results reveal a consistent improvement in classification accuracy across all evaluated algorithms. The KNN algorithm showed a notable increase in accuracy, from 81.6% without SMOTENC to 89.8% after its application. Similarly, the SVM algorithm exhibited an improvement from 90.6% to 92.5%, indicating that the algorithm was able to leverage the more balanced dataset for better generalization.

The greatest enhancement was evident in the Naive Bayes algorithm, which exhibited an increase in accuracy from 70.0% to 82.3%. This significant gain suggests that Naive Bayes is highly sensitive to class imbalance and benefits considerably from oversampling techniques such as SMOTENC. In contrast, the C4.5 algorithm maintained high performance in both conditions, with a slight increase from 99.6% to 99.7%, indicating its robustness even under imbalanced data conditions.

Overall, these findings underscore the effectiveness of SMOTENC in improving classification outcomes, particularly for algorithms that are more vulnerable to skewed class distributions. The observed improvements underscore the importance of mitigating data imbalance as a fundamental strategy for optimizing the performance of machine learning models in the diagnosis of PCOS.

REFERENCE

[1] A. Haleem, M. Javaid, R. Pratap Singh, and R. Suman, "Medical 4.0 technologies for healthcare: Features, capabilities, and applications," *Internet of Things and Cyber-Physical Systems*, vol. 2, pp. 12–30, Jan. 2022, doi: 10.1016/J.IOTCPS.2022.04.001.

[2] Ö. Çelik and M. F. Köse, "An overview of polycystic ovary syndrome in aging women," *J Turk Ger Gynecol Assoc*, vol. 22, no. 4, p. 326, Dec. 2021, doi: 10.4274/JTGGA.GALENOS.2021.2021.0077.

[3] P. Moghetti and F. Tosi, "Insulin resistance and PCOS: chicken or egg?," *J Endocrinol Invest*, vol. 44, no. 2, pp. 233–244, Feb. 2021, doi: 10.1007/S40618-020-01351-0/METRICS.

[4] B. Meczekalski *et al.*, "Hyperthecosis: an underestimated nontumorous cause of hyperandrogenism," *Gynecological Endocrinology*, vol. 37, no. 8, pp. 677–682, 2021, doi: 10.1080/09513590.2021.1903419.

[5] M. Dapas and A. Dunaif, "Deconstructing a Syndrome: Genomic Insights Into PCOS Causal Mechanisms and Classification," *Endocr Rev*, vol. 43, no. 6, pp. 927–965, Nov. 2022, doi: 10.1210/ENDREV/BNAC001.

[6] S. Hatoum, M. Amiri, D. Hopkins, R. P. Buyalos, F. Bril, and R. Azziz, "Population-Based vs Health System and Insurer Records: Significant Underdiagnosis of PCOS," *J Clin Endocrinol Metab*, Jan. 2025, doi: 10.1210/CLINEM/DGAF037.

[7] C. Ley, R. K. Martin, A. Pareek, A. Groll, R. Seil, and T. Tischer, "Machine learning and conventional statistics: making sense of the differences," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 30, no. 3, pp. 753–757, Mar. 2022, doi: 10.1007/S00167-022-06896-6/FIGURES/1.

[8] A. Yaqoob, R. Musheer Aziz, · Navneet, and K. Verma, "Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review," *Human-Centric Intelligent Systems 2023 3:4*, vol. 3, no. 4, pp. 588–615, Sep. 2023, doi: 10.1007/S44230-023-00041-3.

[9] X. Wang, Y. Bouzembrak, A. G. J. M. O. Lansink, and H. J. van der Fels-Klerx, "Application of machine learning to the monitoring and prediction of food safety: A review," *Compr Rev Food Sci Food Saf*, vol. 21, no. 1, pp. 416–434, Jan. 2022, doi: 10.1111/1541-4337.12868.

[10] M. F. Ahmad Fauzi, R. Nordin, N. F. Abdullah, and H. A. H. Alobaidy, "Mobile Network Coverage Prediction Based on Supervised Machine Learning Algorithms," *IEEE Access*, vol. 10, pp. 55782–55793, 2022, doi: 10.1109/ACCESS.2022.3176619.

[11] A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management for production quality deep learning models: Challenges and solutions," *Journal of Systems and Software*, vol. 191, p. 111359, Sep. 2022, doi: 10.1016/J.JSS.2022.111359.

[12] M. A. Talukder *et al.*, "Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," *J Big Data*, vol. 11, no. 1, pp. 1–44, Dec. 2024, doi: 10.1186/S40537-024-00886-W/TABLES/16.

[13] M. Bourel *et al.*, "Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters," *Water Res*, vol. 202, p. 117450, Sep. 2021, doi: 10.1016/J.WATRES.2021.117450.

[14] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl Inf Syst*, vol. 65, no. 1, pp. 31–57, Jan. 2023, doi: 10.1007/S10115-022-01772-8/METRICS.

[15] L. Zhang *et al.*, "Classification of Imbalanced Data:Review of Methods and Applications," *IOP Conf Ser Mater Sci Eng*, vol. 1099, no. 1, p. 012077, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012077.

[16] Z. Zad *et al.*, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," *Front Endocrinol (Lausanne)*, vol. 15, 2024, doi: 10.3389/fendo.2024.1298628.

[17] W. Chen, J. Miao, J. Chen, and J. Chen, "Development of machine learning models for diagnostic biomarker identification and immune cell infiltration analysis in PCOS," *Journal of Ovarian Research* , vol. 18, no. 1, pp. 1–16, Dec. 2025, doi: 10.1186/S13048-024-01583-1/FIGURES/9.

[18] H. M. Emara, W. El-Shafai, N. F. Soliman, A. D. Algarni, R. Alkanhel, and F. E. Abd El-Samie, "A stacked learning framework for accurate classification of polycystic ovary syndrome with advanced data balancing and feature selection techniques," *Front Physiol*, vol. 16, p. 1435036, May 2025, doi: 10.3389/FPHYS.2025.1435036/BIBTEX.

[19] Y. A. Abu Adla, D. G. Raydan, M. Z. J. Charaf, R. A. Saad, J. Nasreddine, and M. O. Diab, "Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques," *International Conference on Advances in Biomedical Engineering, ICABME*, vol. 2021-October, pp. 208–212, 2021, doi: 10.1109/ICABME53305.2021.9604905.

[20] P. Bhardwaj and P. Tiwari, "Manoeuvre of Machine Learning Algorithms in Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis," pp. 71–84, 2022, doi: 10.1007/978-981-16-6887-6_7.

[21] I. S. Silva *et al.*, "Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models," *J Endocrinol Invest*, vol. 45, no. 3, pp. 497–505, Mar. 2022, doi: 10.1007/S40618-021-01672-8/METRICS.

[22] A. Zigarelli, Z. Jia, and H. Lee, "Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study.," *JMIR Form Res*, vol. 6, no. 3, p. e29967, Mar. 2022, doi: 10.2196/29967.

[23] S. Tiwari *et al.*, "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning," *Expert Syst Appl*, vol. 203, p. 117592, Oct. 2022, doi: 10.1016/J.ESWA.2022.117592.

[24] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, V. Bhandage, and G. K. Hegde, "A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome," *Applied System Innovation 2023, Vol. 6, Page 32*, vol. 6, no. 2, p. 32, Feb. 2023, doi: 10.3390/ASI6020032.

[25] P. Dutta, S. Paul, and M. Majumder, "An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS," Nov. 2021, doi: 10.21203/RS.3.RS-1043852/V1.

[26] F. Gurcan and A. Soylu, "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis," *Cancers 2024, Vol. 16, Page 3417*, vol. 16, no. 19, p. 3417, Oct. 2024, doi: 10.3390/CANCERS16193417.

[27] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms 2025, Vol. 18, Page 37*, vol. 18, no. 1, p. 37, Jan. 2025, doi: 10.3390/A18010037.

[28] J. Fonseca and F. Bacao, "Geometric SMOTE for imbalanced datasets with nominal and continuous features," *Expert Syst Appl*, vol. 234, Dec. 2023, doi: 10.1016/J.ESWA.2023.121053.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6676

[29] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information 2023, Vol. 14, Page 54*, vol. 14, no. 1, p. 54, Jan. 2023, doi: 10.3390/INFO14010054.

[30] B. Ravinder, S. K. Seeni, V. S. Prabhu, P. Asha, S. P. Maniraj, and C. Srinivasan, "Web Data Mining with Organized Contents Using Naive Bayes Algorithm," *2024 2nd International Conference on Computer, Communication and Control, IC4 2024*, 2024, doi: 10.1109/IC457434.2024.10486403.

[31] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–11, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[32] O. Saeful Bachri, R. Mohamad, and H. Bhakti, "Penentuan Status Stunting pada Anak dengan Menggunakan Algoritma KNN," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 3, no. 02, pp. 130–137, Nov. 2021, doi: 10.46772/INTECH.V3I02.533.

[33] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of Support Vector Machine Algorithm in Big Data Background," *Math Probl Eng*, vol. 2021, no. 1, p. 5594899, Jan. 2021, doi: 10.1155/2021/5594899.

[34] X. Zheng, W. Feng, M. Huang, and S. Feng, "Optimization of PBFT Algorithm Based on Improved C4.5," *Math Probl Eng*, vol. 2021, no. 1, p. 5542078, Jan. 2021, doi: 10.1155/2021/5542078.