Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.10 No.1 January 2025, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Comparative Analysis of PCOS Classification Using Random Forest: Integration of Mutual Information, SMOTE-Tomek, and Outlier Handling

Selviana Dwi Aprianti[1], Farrikh Alzami[2], Ifan Rizqa[3], Ricardus Anggi Pramunendar[4], Rama Aria Megantara[5], Muhammad Naufal[6], Dwi Puji Prabowo[7]

[1,2,3,4,5,6,7]*Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia*

[2]alzami@dsn.dinus.ac.id (*)

[1]111202113721@mhs.dinus.ac.id, [3,4,5,6,7][risqa.ifan, ricardus.anggi, aria, m.naufal, dwi.puji.prabowo]@dsn.dinus.ac.id

Received: 2024-11-07; Accepted: 2025-01-24; Published: 2025-01-31

Abstract— Polycystic Ovary Syndrome (PCOS) is a hormonal disorder affecting women of reproductive age, with a global prevalence rate of 8–13%. However, approximately 70% of cases remain undiagnosed. This study aimed to develop and compare eight Random Forest classification models for PCOS detection using a publicly available Kaggle dataset. The methodology incorporated three key preprocessing techniques: outlier handling using the Interquartile Range (IQR) method, feature selection through Mutual Information, and class imbalance via SMOTE-Tomek. The results revealed that the best-performing model, which applied outlier removal and SMOTE without feature selection, achieved an accuracy of 94.11%. This result significantly outperformed the baseline Random Forest model, which achieved an accuracy of 87.27% without the application of any preprocessing techniques, such as outlier removal, SMOTE, or feature selection. Moreover, the model utilizing only SMOTE for class balancing achieved an accuracy of 93.84%, underscoring the importance of addressing class imbalance in enhancing classification performance. Notably, feature selection did not consistently improve accuracy, as Random Forest inherently handles feature redundancy, capturing complex feature interactions. These findings highlight the importance of tailored preprocessing strategies, particularly outlier handling and class balancing, for optimizing medical data classification. Future research should explore clinically informed feature selection techniques and assess the generalizability of these findings across diverse datasets to enhance the clinical relevance of PCOS detection models.

Keywords— PCOS; Random Forest; Outlier Detection; Feature Selection; SMOTE-Tomek; Medical Classification.

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) represents a significant public health challenge affecting women of reproductive age worldwide. With a global prevalence of 8-13% and approximately 70% of cases remaining undiagnosed, PCOS poses a substantial healthcare concern [1]. The World Health Organization reports a global PCOS prevalence of 3.4%, with significant regional variations. Studies at Cipto Mangunkusumo Hospital in Indonesia have documented numerous PCOS cases, highlighting its relevance in the local healthcare context [2].

The condition is characterized by hormonal imbalances, including excess androgen and ovarian dysfunction, leading to symptoms such as infertility, obesity, and irregular menstruation. Diagnosis is often delayed due to the heterogeneity of its presentation [3][4]. Research conducted in Alberta, Canada, revealed that patients typically experience a 4.3-year delay between initial symptom onset and definitive diagnosis, with 57% requiring multiple physician consultations before receiving an accurate diagnosis [5].

A major challenge in PCOS diagnosis lies in the presence of outliers in medical datasets, which can introduce bias and reduce model accuracy. Outliers may result from measurement errors or natural biological variations but must be handled carefully to avoid losing valuable clinical data [6].

In parallel, another challenge in PCOS classification arises from the high dimensionality and complexity of medical data. This often results in irrelevant or redundant features that can complicate model development and reduce efficiency. Feature selection techniques become crucial in this context, as they help eliminate unnecessary features, improving model performance and reducing the risk of overfitting [7-9] . One such technique, The Mutual Information method, is particularly useful for identifying the most relevant features with predictive value while simultaneously excluding redundant variables that may skew results [10][11]. However, feature selection must be carefully balanced against the challenge of class imbalance, which is common in medical datasets.

Additionally, class imbalance is a common issue in medical datasets, especially in PCOS detection. The SMOTE-Tomek method, which combines oversampling the minority class with under-sampling the majority class, has been shown to effectively balance data and improve model performance by clarifying class boundaries and reducing noise [12][13].

In this study, the Random Forest algorithm is utilized for PCOS classification due to its effectiveness in processing medical data, particularly datasets with a large number of features. By constructing multiple decision trees and combining their results, Random Forest achieves high accuracy while minimizing the risk of overfitting, making it more reliable than single decision tree models [14][15]. The ability of Random Forest to identify hidden patterns within complex medical data

DOI : https://doi.org/10.25139/inform.v10i1.9231

renders it a valuable tool for detecting diseases such as PCOS. This study employs the Random Forest algorithm in conjunction with feature selection and outlier removal techniques to develop a more accurate and robust model for PCOS detection.

This study introduces an innovative approach to PCOS classification by combining outlier removal, SMOTE-Tomek, and feature selection techniques. The results aim to demonstrate the effectiveness of these preprocessing methods and offer valuable insights into improving machine-learning models for complex medical conditions like PCOS.

II. Research Methodology

The research methodology process illustrated in Fig.1 begins with a preprocessing stage, where raw data is refined to ensure its quality meets the requirements for analysis. Next, oversampling is conducted using the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in the dataset. This is followed by a feature selection process to identify the most relevant features, enhancing model performance and reducing complexity. Finally, the model is developed using data processed through the preceding stages of preprocessing, SMOTE, and feature selection. A detailed explanation of each step is provided in the following subsections.

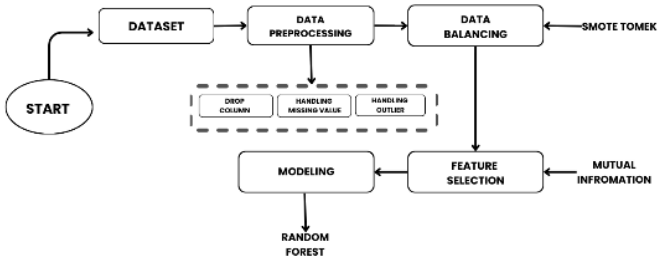

Fig.1. Flow of Classification

A. Dataset

This study uses the PCOS dataset from Kaggle, which contains 541 data records with 45 columns [16]. This dataset includes various health variables relevant to diagnosing PCOS, such as age, body mass index (BMI), menstrual cycle, certain hormone levels, and ultrasound test results. These variables were selected based on clinical factors commonly used to diagnose polycystic ovary syndrome. Of the 541 data records, 364 were normal records (not affected by PCOS), while 177 were records of patients with PCOS. This dataset was then processed through a series of preprocessing steps to ensure the data quality used in the classification analysis. The dataset description can be seen in Table I.

TABLE I
DATASET DESCRIPTION

Feature Name	Description	Numeric(Y/N)
Follicle No. (R)	Number of follicles in the right ovaries.	Yes
Follicle No. (L)	Number of follicles in the left ovaries.	Yes
Fast food (Y/N)	Consumption of fast food.	Yes
AMH (ng/mL)	Anti-Müllerian hormone level, related to ovarian reserve.	No
Cycle length (days)	Length of the menstrual cycle.	Yes
BMI	Body Mass Index	Yes
Age(yrs)	Age in years	Yes
Weight gain (Y/N)	Test to check if the patient gains weight	Yes
Cycle (R/I)	Regularity of menstrual cycles.	Yes
Hip	Size of hip in inches	Yes
Waist	Size of the waist in inches	Yes
Skin darkening (Y/N)	Test to check the appearance of darkness in the skin	Yes
Hair Growth(Y/N)	Test to check if a patient has hair growth	Yes
Hair loss (Y/N)	Test to check hair loss	Yes
Pimples (Y/N)	Pimple issues	Yes

B. Preprocessing Data

In the preprocessing stage, the data is processed through a series of steps to ensure quality and consistency before the modelling stage. The first step is the removal of irrelevant columns, such as 'Sl. No', 'Unnamed: 44', and 'Patient File No.', to reduce data complexity. Next, missing values in the columns 'Fast food (Y/N),' 'AMH(ng/mL)', and 'Marriage Status (Yrs)' were filled using the median imputation method to maintain data integrity and model performance. Outlier handling was done using the Interquartile Range (IQR) method, which identifies extreme values outside the range [Q1 - 1.5 * QIR, Q3 + 1.5 * QIR]. The detected outliers can then be removed or dealt with without significantly affecting the data distribution. This approach helps to reduce the impact of extreme values on the classification model, thus improving the model's ability to recognize patterns more accurately without bias from extreme data. Additionally, although object data types often need to be converted to numeric for most modelling and statistical analysis algorithms, such as Random Forest, which require numeric inputs for mathematical and statistical operations, no conversion was necessary in this case. All columns in the dataset were already in numeric format, ensuring compatibility with the chosen algorithms without further preprocessing.

After the preprocessing stage, which includes applying the SMOTE (Synthetic Minority Over-sampling Technique) method using Tomek's SMOTE variant to address the class imbalance problem commonly encountered in the context of diseases such as PCOS, this research proceeds with feature selection. Feature selection is performed using the Mutual Information method to assess the relevance of each feature to the target variable. This method was chosen for its ability to measure the dependency between features and classes, ensuring that only the most informative features are selected.

The next step is the application of the Random Forest classification algorithm, which was chosen because of its ability to handle high-dimensional datasets and reduce the risk of overfitting through an ensemble mechanism. Random Forest builds several decision trees and combines their predictions to produce a final decision, thus providing advantages in accuracy

DOI : https://doi.org/10.25139/inform.v10i1.9231

and computational efficiency compared to other algorithms such as SVM or KNN.

As a final step, the model is trained using the processed dataset, and evaluation is performed using metrics such as accuracy, precision, recall, and F1-score. The results of this evaluation will be compared based on the approaches applied, including classification with and without outlier removal and feature selection. This study aims to identify the best approach in PCOS classification and provide further insight into the effect of outlier removal and feature selection on model accuracy.

C. Data Balancing Using SMOTE

One of the main challenges in PCOS classification is data imbalance, where the number of individuals with PCOS is less than individuals without PCOS. To address this imbalance, a combination of SMOTE (Synthetic Minority Over-sampling Technique) and TOMEK links is used. SMOTE is a minority over-sampling technique that synthesizes new examples based on the geographic and topological context of existing minority examples [17]. SMOTE-Tomek is a hybrid re-sampling technique that combines the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links to address class imbalance in data sets. This method improves model performance by oversampling minority classes while cleaning majority class data [13].

Tomek's SMOTE method consists of two main stages in overcoming data imbalance. First, the Synthetic Minority Over-sampling Technique (SMOTE) is used to oversample the minority class by generating new synthetic samples by interpolating existing minority data samples. The second stage involves using Tomek Links to remove majority class samples that are too close to the minority class, thereby cleaning up class boundaries and reducing noise. If the amount of data after oversampling exceeds the requirement, a sample reduction is performed on the majority class to maintain the balance and reliability of Tomek's SMOTE method. This approach provides additional advantages regarding improved data quality and inter-class balance, leading to more accurate models and reducing the risk of overfitting compared to the conventional SMOTE method or other variations.

The SMOTE equation for generating synthetic samples can be described as Equation (1). A new sample is generated by taking a minority example x_i and adding a random fraction λ (drawn from a uniform distribution between 0 and 1) of the difference between x_i and its nearest neighbor x_{nn} within the minority class. This approach synthesizes new samples to balance class distributions in the dataset.

$$New\ Sample: x_i + \lambda(x_{nn} - x_i) \quad (1)$$

D. Feature Selection with Mutual Information

This research applies feature selection using Mutual Information to help ease the feature selection process and improve classification performance in polycystic ovarian syndrome (PCOS) classification. Feature selection is selecting the most relevant and significant subset of features (variables or attributes) from a dataset for use in machine learning or data analysis. This step aims to eliminate features that are irrelevant, redundant, or have a low relationship with the classification or prediction target to improve model performance, reduce computation time, and prevent overfitting [18].

Feature selection is an essential stage in data preprocessing, especially for classifications such as PCOS. This process focuses on identifying features that contribute significantly to the predictive ability of the model while filtering out irrelevant data. One effective method for feature selection is Mutual Information, which measures how much information is gained about one random variable by knowing another random variable. Thus, using Mutual Information in feature selection can help improve classification models' prediction accuracy and efficiency [18], [19].

The Equation (2) for feature selection using Mutual Information. Mutual Information (MI) quantifies the shared information between two variables, X and Y, indicating the extent to which knowing one variable reduces uncertainty about the other. Here, $p(x,y)$ represents the joint probability of X and Y occurring together, while $p(x)$, and $p(y)$ denotes the independent probabilities of X and Y, respectively. MI compares $p(x,y)$ with $p(x),p(y)$; if X and Y are independent, MI is zero, meaning X provides no information about Y. A positive MI value suggests a dependency between X and Y, making MI a valuable tool for identifying relevant features in data analysis.

$$MI(X;Y) = \sum Y \in y \sum p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (2)$$

E. Random Forest

The Random Forest method is an ensemble-based classification technique that combines multiple decision trees to improve prediction accuracy. Each tree is built from a randomly selected subset of the training data, and the final prediction is determined based on the majority vote of all trees [20]. The decision tree in Random Forest consists of three types of nodes: root node (initial node), internal node (branching node with one input and at least two outputs), and leaf node (final node without output). The entropy and information gain values are calculated with specific formulas to determine the best split for each node [21].

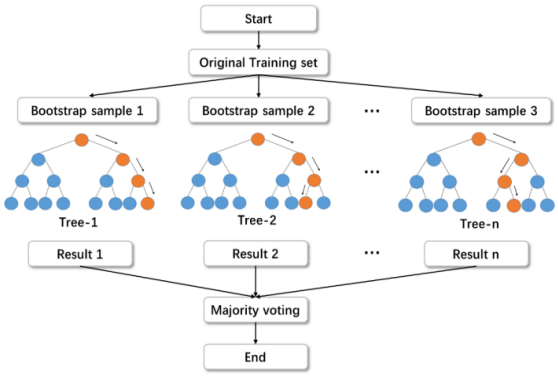

Fig. 2. Random Forest Method Process

DOI : https://doi.org/10.25139/inform.v10i1.9231

The process described in Fig.2 illustrates the steps involved in forming a Random Forest, which are detailed as follows:

a) Randomly select several training subsets (sub-training sets) and test sets from the original dataset.
b) For each node in the tree, select one attribute from a number of features based on the information gain method to perform the split.
c) Repeat the splitting process until the tree can no longer be split.
d) Repeat steps 1–3 to build many decision trees, forming a Random Forest.

This process results in a collection of different trees that provide predictions together using Equation (3). The $Entropy\ (S)$ quantifies the impurity or disorder within a set of samples. It is calculated by summing over each class i and multiplying the proportion of samples in that class (pi) by the logarithm base 2 of pi. Here, c represents the total number of classes. Entropy is used to assess how mixed the classes are within a node, with higher values indicating a more significant disorder.

$$Entropy\ (S) = -\sum_{i=1}^{c} pi \log_2(pi) \quad (3)$$

In Equation (4), $Gain(S, A)$ measures the information gain obtained by splitting a set S using attribute A. It is calculated by taking the entropy of S and subtracting the weighted entropy of each subset S_v created by the values of A. For the S_v variable represents the subset of S for each attribute A value. Information gain helps identify the attribute that best separates the data into distinct classes.

$$Gain(S, A) = Entropy\ (S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (4)$$

F. Evaluation

A confusion matrix is a table that compares the model's predictions and the actual values in a binary classification. It is used to evaluate the model's performance and helps calculate several important evaluation metrics, such as accuracy, precision, recall, and f1-score.

Accuracy measures the proportion of correct predictions out of the overall test data using Equation (5).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (5)$$

Precision measures the accuracy of the model in predicting the positive class using Equation (6).

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall measures the sensitivity of the model in detecting all positive cases using Equation (7).

$$Recall = \frac{TP}{(TP+FN)} \quad (7)$$

F1-score is the harmonic mean of precision and recall. F1-score is used when we want to balance between precision and recall, especially in the case of data imbalance using Equation (8).

$$FI - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

III. Result and Discussion

This study uses a dataset taken from Kaggle. The dataset is then processed in the data preprocessing stage to overcome problems such as handling missing data and outliers in the data. This experiment implemented eight models that combined outlier removal techniques, Synthetic Minority Over-sampling Technique (SMOTE), and feature selection (FS). The metrics used to evaluate model performance include accuracy, precision, and recall. Combining these three techniques minimizes bias, addresses data imbalance, and selects the most relevant features to enhance model performance.

Based on Fig.3, after outlier removal, the dataset exhibits a more balanced and consistent distribution across all features. By addressing extreme values, the preprocessing stage effectively reduces the impact of unusually high or low data points that could otherwise distort the analysis and compromise the accuracy of predictive models. This refined data distribution ensures that each feature is represented within a reasonable range, allowing statistical patterns to emerge more clearly without interference from outliers. The outlier removal process has resulted in a cleaner, more reliable dataset with values that better represent the typical characteristics of the population under-study. This improvement enhances the quality and integrity of the data, making it more suitable for accurate PCOS prediction. Consequently, this preprocessing step contributes to more dependable and meaningful results, supporting a higher confidence level in PCOS classification and analysis.

Fig.4 shows that the most influential features in PCOS classification are the number of follicles on the right and left ovaries, as seen in the feature importance chart above. These two features have the highest importance scores, making them primary indicators in diagnosing PCOS. Following these, skin darkening, weight gain, and hair growth contribute significantly, indicating their relevance in identifying PCOS-related symptoms and patterns. Other features, such as FSH levels, hip size, age, weight, waist size, cycle length, and lifestyle factors like fast food consumption, pimples, and hair loss, also contribute to the model, albeit at a lower level. These features still provide valuable information, enhancing the model's understanding of clinical and lifestyle factors associated with PCOS. Moreover, Fig.4 displays the feature importance values as determined by the Random Forest model, showing the relative influence of each feature on model performance. The features are ordered by importance, from Follicle Number (R) at the top to Hair Loss (Y/N) at the bottom. This ranking reflects the model's selection of the top 15 features, prioritizing the most informative features for accurate and efficient PCOS predictions. The model enhances its predictive power and robustness in identifying PCOS cases by focusing on these high-importance features.

DOI : https://doi.org/10.25139/inform.v10i1.9231

Fig.3. Data After Removing Outliers

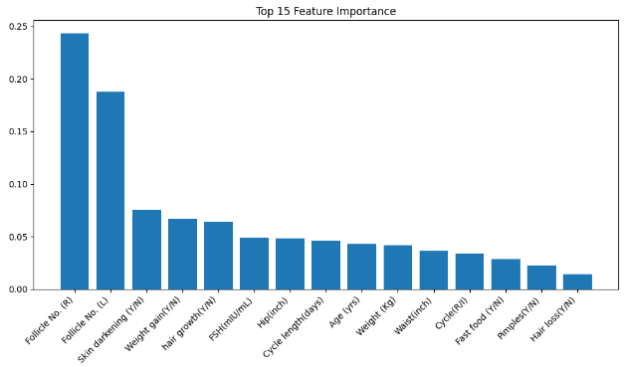

Fig.4. Selection Feature Result

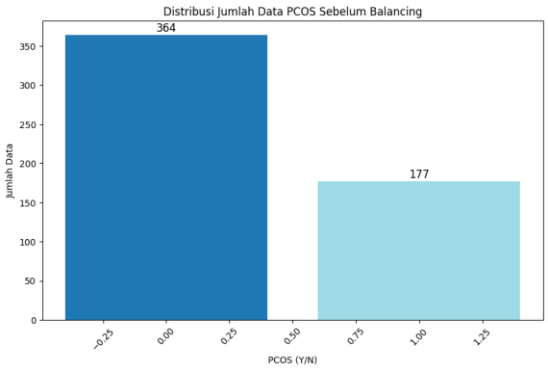

Fig.5. Data Before Balancing

DOI : https://doi.org/10.25139/inform.v10i1.9231

Based on Fig.5, the data used is imbalanced, with significantly more samples in the non-PCOS class (364) compared to the PCOS class (177). The SMOTE Tomek method was applied to address the data imbalance in this research through oversampling. This method combines two techniques, namely the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links, to improve the representation of minority classes. SMOTE generates new synthetic samples of the minority class by interpolating between existing examples, thus increasing the class's data. After that, Tomek Links is used to remove samples from the majority class that are too close to the boundary of the minority class, thus helping to remove noise and improve data quality. This approach improved the balance between classes and increased the model's accuracy in identifying patterns associated with PCOS, allowing for more robust and reliable analysis.

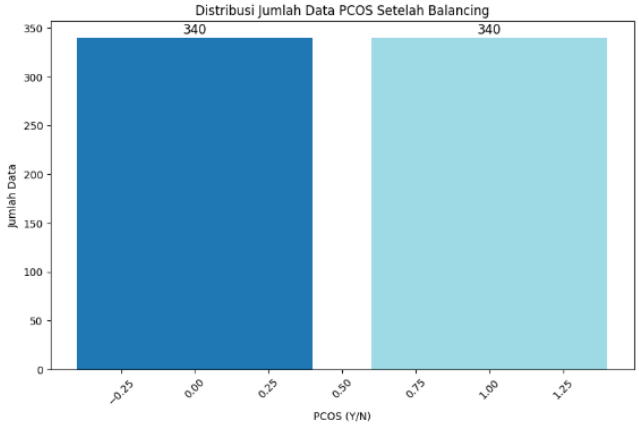

Fig.6. Data After Balancing

Fig.6 illustrates the distribution of the number of PCOS data after balancing using SMOTE-Tomek, showing that the number of data in both classes has become the same, which is 340 data. This data balance is very important to improve the performance of the classification model because the model will not be too biased toward the majority class and can better generalize to new data. Thus, it is expected that the model trained on this balanced data can provide higher accuracy in predicting PCOS status in patients.

Following outlier removal, feature selection, and data balancing, the performance of the eight models was assessed using a 90:10 data split. This standard split, allocating 90% of the data for training and 10% for testing, is widely adopted in machine learning to ensure a robust and unbiased evaluation of model performance. By training the model on a larger dataset, researchers aim to capture the underlying patterns and relationships within the data more effectively. A larger training set, as provided by a 90:10 split, allows the model to learn more comprehensive patterns from the data, which is crucial in medical datasets where data points can be scarce and valuable [22]. The testing set, on the other hand, provided an unbiased assessment of the model's ability to generalize to unseen data. This approach can help mitigate the effects of class imbalance by providing more data for the model to learn from, which is particularly important in healthcare applications where certain conditions may be underrepresented [23] [24]. We compared eight different models to evaluate the effect of outlier removal and feature selection on model performance. The results of the evaluation are presented in detail in Table II.

TABLE II
MODEL COMPARASION RESULT

Model	Outlier Removal	SMOTE	Feature Selection	Accuracy (%)	Precision	Recall
1	No	No	No	87.27%	0.87	0.87
2	No	No	Yes	85.45%	0.85	0.85
3	No	Yes	No	**93.84%**[2]	0.93	0.93
4	No	Yes	Yes	87.27%	0.87	0.87
5	Yes	No	No	**89.09%**[3]	0.89	0.89
6	Yes	No	Yes	87.27%	0.87	0.87
7	Yes	Yes	No	**94.11%**[1]	0.94	0.94
8	Yes	Yes	Yes	87.27%	0.87	0.87

*123 notation in accuracy, denoted the best performance

Table II comprehensively compares model performance across various preprocessing configurations, including outlier removal, Synthetic Minority Over-sampling Technique (SMOTE), and feature selection (FS). Each model is evaluated in terms of accuracy, precision, and recall, illustrating the impact of these preprocessing steps on the model's ability to classify PCOS cases effectively.

In the initial model (Model 1), where no outlier removal, SMOTE, or feature selection was applied, the model achieved an accuracy, precision, and recall of 87.27%. This result suggests that, while the model can capture some basic patterns without any preprocessing, it may be limited by the presence of noise, class imbalance, and irrelevant features, which reduce its effectiveness.

When feature selection was applied independently (without outlier removal or SMOTE, as in Model 2), the model's accuracy, precision, and recall decreased slightly to 85.45%. This decline indicates that applying feature selection alone might have inadvertently excluded critical information needed for accurate classification, thereby reducing the model's generalization capacity and predictive performance. This suggests that feature selection, when applied without other preprocessing steps, might filter out still relevant features for PCOS classification.

DOI : https://doi.org/10.25139/inform.v10i1.9231

In contrast, when SMOTE was applied without feature selection or outlier removal (Model 3), the model improved substantially, with accuracy, precision, and recall reaching 93.84%. This significant increase highlights the effectiveness of SMOTE in balancing class distributions, which enhances the model's performance even in the absence of outlier removal or feature selection. This result underscores the importance of addressing class imbalance, which can be crucial in improving model accuracy for PCOS classification.

However, in the model that combined SMOTE with feature selection but excluded outlier removal (Model 4), there was no improvement over the baseline, with accuracy, precision, and recall remaining at 87.27%. This outcome suggests that, while SMOTE is beneficial for addressing class imbalance, adding feature selection without first handling outliers may inadvertently remove relevant information, resulting in no performance gain. This finding implies that combining SMOTE with feature selection alone may not be sufficient to enhance model performance, mainly if outliers in the data are not managed.

When outlier removal was applied without SMOTE or feature selection (Model 5), the model's accuracy, precision, and recall improved modestly to 89.09%. This slight increase suggests that outlier removal helps the model focus on more representative data, reducing the noise from extreme values. However, without additional preprocessing, the benefits remain limited, indicating that while outlier removal alone can be helpful, it may not fully optimize model performance.

Combining outlier removal and feature selection without SMOTE (Model 6) yielded accuracy, precision, and recall identical to the baseline at 87.27%. This result indicates that this combination may have inadvertently filtered out too much information, providing no meaningful improvement. This configuration highlights that feature selection might eliminate relevant patterns and limit the model's predictive ability, especially when combined with outlier removal.

The highest performance was observed when outlier removal and SMOTE were applied without feature selection (Model 7), with accuracy, precision, and recall reaching 94.12%. This comprehensive preprocessing configuration enables the model to focus on relevant data by removing outliers, balancing class distributions, and selecting the most informative features. This setup leads to the most accurate predictions, underscoring the importance of addressing class imbalance, managing data noise, and reducing redundant features to build a robust model for PCOS classification.

Interestingly, when outlier removal, SMOTE, and feature selection were applied together (Model 8), the model's performance reverted to the baseline values of 87.27%. This outcome suggests that feature selection may not always benefit complex models like Random Forest despite its potential to enhance model interpretability by focusing on relevant features. These results indicate that feature selection can sometimes decrease model accuracy by removing features that may seem redundant or less informative but still contribute valuable information for tree-based algorithms. Since Random Forest models handle numerous features well, including potentially redundant ones, removing features may reduce the model's predictive power.

In conclusion, these findings indicate that while the combination of outlier removal, SMOTE, and feature selection yielded the highest accuracy, feature selection alone or in specific configurations may not always be beneficial, particularly for Random Forest. Random Forest models are designed to utilize a wide range of features and can naturally manage redundancy. Thus, removing features via feature selection might inadvertently discard important signals the model could otherwise leverage. This analysis suggests that feature selection may not always be appropriate for Random Forest, as it has the potential to reduce accuracy by excluding valuable information that contributes to the model's predictive performance.

Next, we examine the confusion matrix of the model with high accuracy. The confusion matrix offers a comprehensive assessment of the model's classification performance, providing valuable insights into its capability to distinguish between classes accurately. By presenting the distribution of true positives, true negatives, false positives, and false negatives, the matrix enables a detailed evaluation of the model's strengths and limitations regarding predictive accuracy. This analysis further supports an in-depth understanding of the model's effectiveness in classifying cases accurately, contributing to a more rigorous evaluation of its overall performance.

Based on the confusion matrix in Fig.7, the model's performance on the test set for the original data. This matrix has 36 correct predictions for class 0 (True Negatives), meaning the model successfully identified 36 instances correctly as the negative class. However, one instance was predicted as the positive class (False Positive) but belongs to the negative class. Additionally, 6 instances of class 1 were incorrectly classified as class 0 (False Negatives), indicating that the model failed to detect some positive instances. On the other hand, there are 12 instances correctly predicted as class 1 (True Positives), showing that the model can accurately identify some positive instances.

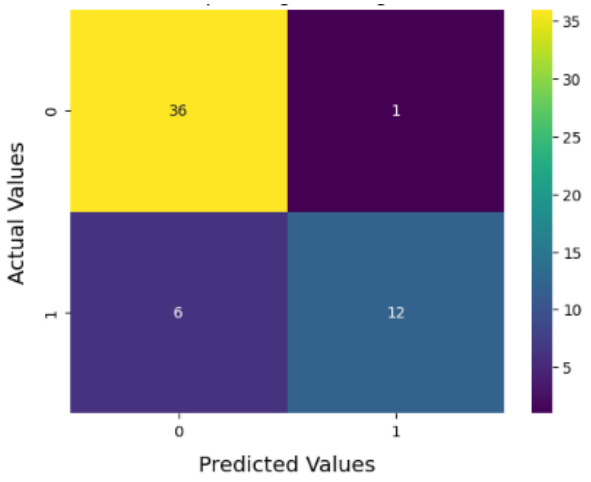

Fig.7. Confusion Matrix Model 1

DOI : https://doi.org/10.25139/inform.v10i1.9231

Based on the confusion matrix presented in Fig.8, The confusion matrix above provides insights into the model's performance on the test data for the original dataset. The matrix reveals that the model accurately classified 30 instances as class 0 (True Negatives), indicating that these cases were correctly identified as belonging to the negative class. However, there were 3 instances where the model incorrectly classified negative cases as positive (False Positives), suggesting a minor misclassification of negative instances. For the positive class, the model made only one mistake: a positive instance was incorrectly classified as negative (False Negative). Additionally, the model successfully classified 31 instances as class 1 (True Positives), showcasing its ability to recognize positive cases accurately.

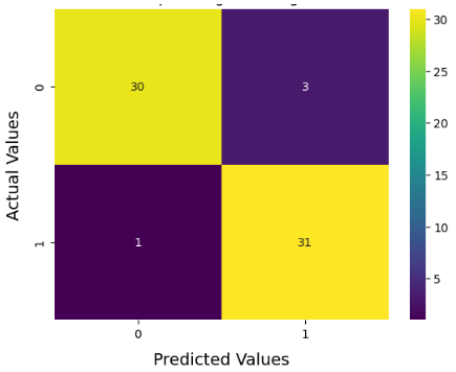

Fig.8. Confusion Matrix Model 3

The model demonstrates strong predictive accuracy, particularly in correctly identifying positive cases. To further assess the model's performance, calculating metrics such as precision, recall, and F1-score could provide additional insights, especially into the model's ability to handle classification errors and its overall reliability in distinguishing between the classes.

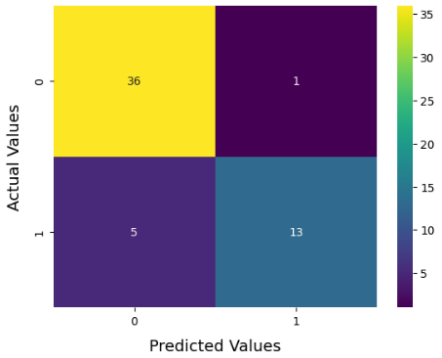

Fig.9. Confusion Matrix Model 5

Based on the confusion matrix in Fig.9, The matrix shows that the model correctly classified 36 instances as class 0 (True Negatives), accurately identifying these cases as belonging to the negative class. However, there was one instance where the model incorrectly classified a negative case as positive (False Positive), indicating a very low error rate for negative cases. Regarding the positive class, the model misclassified five instances as negative (False Negatives), revealing a slight limitation in detecting all positive instances accurately. On the other hand, the model correctly identified 13 instances as class 1 (True Positives), demonstrating its ability to recognize positive cases to a reasonable extent.

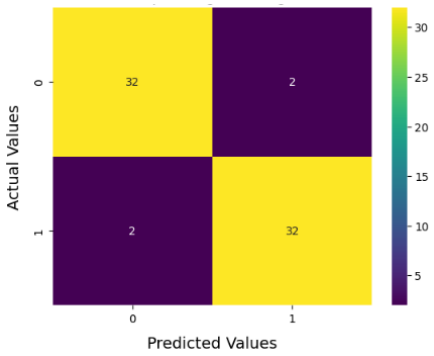

Fig. 10. Confusion Matrix Model 7

In the confusion matrix shown in Fig. 10, the model correctly classified 32 instances as class 0 (True Negatives), meaning these cases were accurately identified as belonging to the negative class. Additionally, there were 2 instances where the model incorrectly classified negative cases as positive (False Positives), indicating a very low error rate in identifying negative cases. The model also made only 2 classification errors for the positive class, where positive instances were incorrectly classified as negative (False Negatives). On the other hand, the model accurately identified 32 instances as class 1 (True Positives), demonstrating a strong ability to recognize positive cases.

1) Clinical Implications of Outlier Removal and SMOTE in PCOS Classification: The superior performance of the model with outlier removal and SMOTE without feature selection, achieving 94.11% accuracy, without feature selection, highlights the importance of these preprocessing steps in enhancing model effectiveness for PCOS diagnosis. By removing statistical outliers, which often represent noise or errors that can distort machine learning models and lead to biased results or reduced accuracy, the model minimizes noise and focuses on capturing patterns more representative of typical cases. The outliers that are removed are not kept for later use, as they often disrupt the model's performance by introducing noise, causing bias, and reducing accuracy. Additionally, addressing class imbalance with SMOTE ensures better precision and recall by improving the model's ability to generalize across a broader range of data. In clinical contexts like PCOS, where data variability can reflect noise and true clinical differences, the combination of outlier removal and class balancing with SMOTE enhances diagnostic reliability, potentially aiding in more consistent and accurate patient classification. The relationship between outlier removal and SMOTE demonstrates a nuanced interaction within the PCOS dataset. In particular, the model that used outlier removal and SMOTE achieved superior accuracy compared to models without these preprocessing steps, confirming that managing

DOI : https://doi.org/10.25139/inform.v10i1.9231

extreme values and ensuring balanced class distributions can enhance model performance by reducing noise and better representing the minority class. These steps refine the dataset to include more balanced, typical cases, helping the model focus on relevant patterns without being misled by unrepresentative data points. For instance, in our analysis of feature importance (Fig. 3), we observed that follicle numbers and AMH levels were highly influential in classification decisions. These parameters often show high variability in PCOS patients, and what might appear as statistical outliers could represent valid clinical presentations. This observation aligns with the medical understanding of PCOS as a spectrum disorder rather than a binary condition.

2) Practical Considerations and Implementation: From a clinical perspective, these findings carry significant implications for developing diagnostic support systems for PCOS. The higher accuracy achieved with the combination of outlier removal and SMOTE, without feature selection, suggests that data preprocessing should be approached carefully in medical applications. This study suggests that, when developing machine learning models for medical diagnosis, the following considerations are essential: 1) Outlier removal can enhance model performance by reducing noise, but it's crucial to assess whether outliers reflect true clinical variability, especially in conditions with diverse manifestations like PCOS; 2) Feature selection may not always be beneficial in models like Random Forest, which can handle complex feature interactions effectively. In some cases, retaining a broader set of features can improve accuracy by capturing more nuanced clinical information; and 3) Leveraging the natural capacity of algorithms like Random Forest to manage noisy, complex data may be more effective than excessive preprocessing, which can inadvertently remove valuable diagnostic information.

*3) Practical Considerations and Implementation***:** From a clinical perspective, these findings have significant implications for developing diagnostic support systems for PCOS. The improved accuracy achieved by applying outlier removal and SMOTE, without feature selection, suggests a need to rethink traditional data preprocessing practices in medical applications. Specifically, when developing machine learning models for medical diagnosis, the following considerations are essential: 1) Outlier removal can enhance model performance by reducing noise, but it's critical to assess whether outliers represent true clinical variability, particularly in conditions with diverse symptoms like PCOS; 2) Feature selection should be applied cautiously, as excluding feature selection in this study led to better performance. Algorithms like Random Forest can handle complex interactions between features effectively, so retaining a broader feature set may preserve relevant clinical information without compromising accuracy; and 3) Leveraging the inherent strengths of Random Forest in managing complex, noisy data can often be more advantageous than excessive preprocessing, which risks removing diagnostically valuable information.

IV. CONCLUSION

This study demonstrates the critical role of preprocessing strategies in enhancing the accuracy of Random Forest models for Polycystic Ovary Syndrome (PCOS) detection. The application of outlier removal combined with class balancing using SMOTE significantly improved classification performance, achieving a maximum accuracy of 94.11%. This represents a marked improvement over the baseline Random Forest model accuracy of 87.27%, which did not employ any preprocessing techniques. In contrast, the model using only SMOTE achieved 93.84%, further highlighting the critical role of addressing class imbalance.

The findings indicate that feature selection, particularly using the Mutual Information method, does not significantly improve Random Forest performance in PCOS classification. Random Forest's intrinsic ability to handle feature redundancy and leverage a broad range of features, including those with lower predictive power, contributes to its overall effectiveness. In fact, models incorporating feature selection showed inconsistent performance, suggesting that Random Forest can manage large feature sets without a loss in accuracy.

The scientific contribution of this study lies in demonstrating that preprocessing strategies, specifically outlier removal and class balancing, play a critical role in enhancing the performance of Random Forest classification models for medical datasets, such as PCOS. These insights are valuable for both researchers and healthcare practitioners working to optimize model performance in medical datasets. Notably, this research highlights that feature selection may not be essential, particularly when utilizing models like Random Forest, which possesses the capability to manage feature redundancy effectively.

Future research should focus on exploring clinically informed feature selection methods that preserve critical clinical indicators and evaluating the generalizability of these findings to other medical datasets. Additionally, incorporating domain expertise into the preprocessing pipeline could help balance data dimensionality reduction with the retention of clinically significant information. Such advances have the potential to bridge the gap between machine learning innovations and real-world healthcare applications, ultimately improving diagnostic accuracy and patient outcomes for PCOS and other medical conditions.

ACKNOWLEDGMENT

This research was conducted in collaboration with the IDSS Research Center Faculty of Computer Science Universitas Dian Nuswantoro.

REFERENCES

[1] A. Yasmin *et al.*, "Polycystic Ovary Syndrome: An Updated Overview Foregrounding Impacts of Ethnicities and Geographic Variations," *Life*, vol. 12, no. 12, p. 1974, Nov. 2022, doi: 10.3390/life12121974.
[2] "Polycystic ovary syndrome," World Health Organization, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome
[3] H. Elmannai *et al.*, "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and

DOI : https://doi.org/10.25139/inform.v10i1.9231

Explainable Artificial Intelligence," *Diagnostics*, vol. 13, no. 8, p. 1506, Apr. 2023, doi: 10.3390/diagnostics13081506.
[4] 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," *ASI*, vol. 6, no. 2, p. 32, Feb. 2023, doi: 10.3390/asi6020032.
[5] B. C. Sydora, M. S. Wilke, M. McPherson, S. Chambers, M. Ghosh, and D. F. Vine, "Challenges in diagnosis and health care in polycystic ovary syndrome in Canada: a patient view to improve health care," *BMC Women's Health*, vol. 23, no. 1, p. 569, Nov. 2023, doi: 10.1186/s12905-023-02732-2.
[6] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.
[7] A. Bajcsi, A. Andreica, and C. Chira, "Towards feature selection for digital mammogram classification," *Procedia Computer Science*, vol. 192, pp. 632–641, Jan. 2021, doi: 10.1016/j.procs.2021.08.065.
[8] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," *Multimed Tools Appl*, vol. 82, no. 22, pp. 34163–34181, Sep. 2023, doi: 10.1007/s11042-023-15165-8.
[9] S. C. R. Nandipati, C. XinYing, and K. K. Wah, "Polycystic Ovarian Syndrome (PCOS) Classification and Feature Selection by Machine Learning Techniques," vol. 9, 2020.
[10] M. Alalhareth and S.-C. Hong, "An Improved Mutual Information Feature Selection Technique for Intrusion Detection Systems in the Internet of Medical Things," *Sensors*, vol. 23, no. 10, p. 4971, May 2023, doi: 10.3390/s23104971.
[11] G. Manikandan and S. Abirami, "An efficient feature selection framework based on information theory for high dimensional data," *Applied Soft Computing*, vol. 111, p. 107729, Nov. 2021, doi: 10.1016/j.asoc.2021.107729.
[12] D. Shabrina Assyifa and A. Luthfiarta, "SMOTE-Tomek Re-sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-class Classification," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 9, no. 2, pp. 151–160, Jul. 2024, doi: 10.25139/inform.v9i2.8410.
[13] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *JOIV : Int. J. Inform. Visualization*, vol. 7, no. 1, p. 258, Feb. 2023, doi: 10.30630/joiv.7.1.1069.
[14] S. Tiwari *et al.*, "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning," *Expert Systems with Applications*, vol. 203, p. 117592, Oct. 2022, doi: 10.1016/j.eswa.2022.117592.
[15] I. H. Hassan, M. Abdullahi, M. M. Aliyu, S. A. Yusuf, and A. Abdulrahim, "An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection," *Intelligent Systems with Applications*, vol. 16, p. 200114, Nov. 2022, doi: 10.1016/j.iswa.2022.200114.
[16] "PCOS Dataset." [Online]. Available: https://www.kaggle.com/datasets/shreyasvedpathak/pcos-dataset
[17] M. A. Latief, L. R. Nabila, W. Miftakhurrahman, S. Ma'rufatullah, and H. Tantyoko, "Handling Imbalance Data using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification," *IJECSA*, vol. 3, no. 1, pp. 11–18, Feb. 2024, doi: 10.30812/ijecsa.v3i1.3758.
[18] O. P. Ige and K. H. Gan, "Ensemble Filter-Wrapper Text Feature Selection Methods for Text Classification," *CMES - Computer Modeling in Engineering and Sciences*, vol. 141, no. 2, pp. 1847–1865, Sep. 2024, doi: 10.32604/cmes.2024.053373.
[19] O. Lifandali, N. Abghour, and Z. Chiba, "Feature Selection Using a Combination of Ant Colony Optimization and Random Forest Algorithms Applied To Isolation Forest Based Intrusion Detection System," *Procedia Computer Science*, vol. 220, pp. 796–805, Jan. 2023, doi: 10.1016/j.procs.2023.03.106.
[20] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI*, vol. 7, no. 2, pp. 310–320, Aug. 2020, doi: 10.35957/jatisi.v7i2.289.
[21] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja," *JIUP*, vol. 5, no. 4, p. 578, Dec. 2021, doi: 10.32493/informatika.v5i4.7575.
[22] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Computer Science*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.
[23] A. Martinez-Velasco, L. Martínez -Villaseñor, and L. Miralles-Pechuán, "Addressing Class Imbalance in Healthcare Data: Machine Learning Solutions for Age-Related Macular Degeneration and Preeclampsia," *IEEE Latin Am. Trans.*, vol. 22, no. 10, pp. 806–820, Oct. 2024, doi: 10.1109/TLA.2024.10705995.
[24] A. Singh and T. Margaria, "Enhancing Decision-Making for Imbalanced Medical Datasets Using BDDs and Low-Code/No-Code," *IT Professional*, vol. 26, no. 5, pp. 92–98, Oct. 2024, doi: 10.1109/MITP.2024.3459248.

DOI : https://doi.org/10.25139/inform.v10i1.9231