

Application of the K-Nearest Neighbor Algorithm for Polycystic Ovarian Syndrome (PCOS) Classification: A Diagnostic Tool

Natasha Leslie¹, Angga Aditya Permana^{1*}, Analekta Tiara Perdana²

¹Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Banten, Indonesia

²Department of Biology, Faculty of Science, UIN Sultan Maulana Hasanuddin, Banten, Indonesia
*natasha.leslie@student.umn.ac.id, angga.permana@umn.ac.id (Corresponding Author),
analekta.tiara@uinbanten.ac.id*

Abstract. Polycystic Ovarian Syndrome (PCOS) is a common hormonal disorder affecting women of reproductive age. Early detection and accurate classification of PCOS are crucial for timely intervention and management. This study proposes the use of the K-Nearest Neighbor (KNN) algorithm for classifying PCOS based on patient symptoms and characteristics. The KNN algorithm was applied to a dataset of 72 patients, and its performance was evaluated using various training and testing ratios and different values of K. The results showed that the KNN algorithm achieved the highest accuracy of 100% with a training ratio of 90:10 and K= 11. The proposed approach demonstrates the potential of machine learning techniques for accurate PCOS classification and highlights the importance of early detection for improving patient outcomes. However, the limited dataset size and potential overfitting issues should be addressed in future research.

Keywords: K-Fold Cross Validation, K-Nearest Neighbor, Machine Learning, PCOS

1. Introduction

Menstrual cycle disorders can be a symptom of various conditions, such as polycystic ovarian syndrome (PCOS). PCOS is a condition that significantly affects women's reproductive health and is characterized by high levels of androgens, which can manifest as excessive testosterone or androgenic disorders such as hirsutism or hyperandrogenemia. Women with PCOS may also experience ovulation failure, irregular menstruation, and polycystic ovarian morphology, which refers to an excess of preantral follicles in the ovaries (Azziz et al., 2016). The normal menstrual cycle typically occurs every 21-25 days and lasts for three to seven days (Itriyeva, 2022). However, every woman may have a different menstrual cycle. One of the abnormalities in the menstrual cycle is irregular menstrual scheduling.

PCOS, or polycystic ovary syndrome, is one of the most common disorders affecting women of reproductive age and adolescent girls worldwide. The incidence of PCOS varies between 1.8% and 15%, depending on ethnicity, background, and diagnostic criteria used (Singh et al., 2023). According to Dr. Budi Santoso, a Specialist in Gynecology and Oncology, PCOS is still an underrecognized hormonal disorder in women (Anisya et al., 2019). Epidemiological data from Dr. Utari Nur Alifah estimates that more than 116 million women, or approximately 3.4% of women worldwide, have PCOS (Himpunan Endokrinologi Reproduksi dan Fertilitas Indonesia (HIFERI) & Perkumpulan Obstetri dan Ginekologi Indonesia (POGI), 2016). PCOS affects between 5% and 20% of women of reproductive age worldwide, and in 2004, PCOS had an economic impact of over \$4 billion in the United States alone. Furthermore, PCOS is associated with an increased risk of breast cancer, type 2 diabetes (T2DM), and other conditions (Singh et al., 2023).

PCOS symptoms are more common in women of reproductive age, and the complex condition can have a significant impact on women's reproductive health. According to a study by Remi.S et al. titled "Does polycystic ovary syndrome affect cognition? A functional magnetic resonance imaging study exploring working memory," hormonal disturbances in adolescent girls can affect approximately one in 10 of their lifetime, leading to menstrual dysfunction and infertility as the main causes (Soleman et al., 2016). Around 50%-80% of women with PCOS are obese, and the risk of developing related conditions is influenced by factors such as age, obesity, and family history of diabetes (McCartney & Marshall, 2016). A study by Asti Yuliadha and Rohningtyas H.S. titled "Psychoneuroimmunology of Depression in Polycystic Ovary Syndrome (PCOS)" states that approximately 40% of women with PCOS experience depression alongside physical disorders (Yuliadha & Setyaningrum, 2022). Conducting research on PCOS can help improve diagnostic criteria, leading to early detection and intervention. Considering the impact of PCOS on women's health, this research can contribute to early diagnosis, treatment, weight reduction, and the reduction of long-term complications (Azziz, 2018). It can advance medical knowledge, facilitate better patient outcomes, and enhance overall understanding of PCOS (Ndefo et al., 2013).

Various factors can influence the diagnosis of PCOS. This is due to the wide heterogeneity of PCOS symptoms (Christ & Cedars, 2023). PCOS can be detected through biochemical, clinical and ultrasonographic methods. These methods take longer and are more expensive, even though early diagnosis and treatment can reduce the possibility of PCOS (Nandipati et al., 2020). Currently, technology has advanced significantly, and one of its outcomes is Artificial Intelligence (AI). Machine learning (ML) has emerged as a powerful tool in various fields to develop intelligent predictive algorithms in the domain of AI. With machine learning, it is possible to analyze high-dimensional and multivariate data, uncover complex and dynamic relationships within the data, even in industrial environments (Wuest et al., 2016). However, the performance of these applications depends on the choice of appropriate machine learning techniques. AI can also identify patterns in medical data, such as hormone levels, to differentiate between patients with PCOS and those without it. This improved accuracy can lead to earlier and more accurate diagnoses of PCOS, increasing overall accuracy rates (Ndefo et al., 2013).

Several previous studies on PCOS classification have been conducted using various methods. Artificial Neural Network (ANN) was used to classify PCOS using Microarray Data, resulting in accuracies ranging from 50% to 100% in a study by Tiara Laksmi Basuki et al. (Basuki et al., 2020). Another study by Bedy Purnama et al., titled "A Classification of Polycystic Ovary Syndrome Based on Follicle Detection of Ultrasound Images," used Support Vector Machine (SVM) to detect PCOS through ultrasound images, achieving accuracies of 82.55% for dataset A and 78.81% using KNN-Euclidean for dataset B (Purnama et al., 2015). Lastly, Dewi R et al. conducted a study titled "Classification of polycystic ovary based on ultrasound images using competitive neural network" and utilized the Competitive Neural Network (CNN) to detect PCOS from ultrasound images, achieving the highest accuracy of 80.84% (Dewi et al., 2018).

Various goals can be achieved using different machine learning algorithms. The K-Nearest Neighbor (KNN) algorithm has been widely used in various research fields, including Health (jabbar et al., 2013; Shee & Cheruiyot, 2014; Xing & Bei, 2020), Occupations (Imandoust & Bolandraftar, 2013), and Education (Munazhif et al., 2023). These studies have yielded optimal results. The KNN method can be used for non-parametric data as it is a non-parametric classification and regression technique (Sumarlinda & Lestari, 2022). Thus, the KNN method is suitable for research with non-parametric data, such as the dataset used for classifying PCOS. KNN combined with genetic algorithms for PCOS prediction. The KNN method has several advantages, such as relatively simple training, ease of understanding, fast implementation, and effectiveness when dealing with large training datasets. A recent study was conducted in Hdaib et al. (2022) to detect PCOS using various machine learning algorithms. In terms of accuracy, the best performance was shown by linear discriminant classifier. While KNN classifier exhibited best performance in terms of sensitivity. However, KNN also has the disadvantage of biased or different K values (Lim et al., 2023). Additionally, the time frame of this study reinforces the use of the KNN method known for its ability to expedite the research process.

The purpose of conducting this research is to determine the classification results for the diagnosis of PCOS using the KNN method and to assess the accuracy level obtained from the PCOS classification results. The benefits of conducting this research are as follows: gaining a better understanding of PCOS and the K-Nearest Neighbor algorithm, advancing the overall understanding of PCOS, particularly in advancing medical knowledge in classifying PCOS based on symptoms and characteristics. The research can serve as a reference for students and researchers conducting further related studies. Additionally, the accuracy results of this research can be used as a benchmark for comparison with similar studies.

2. Materials and Method

Polycystic Ovarian Syndrome (PCOS) PCOS is a significant contributor to ovulation infertility and affects five to 20% of women of reproductive age (Azziz et al., 2016; Zhang et al., 2021). PCOS is often referred to as a hormonal disorder commonly found in fertile women (aged 15-44) (Lim et al., 2023). Nearly 70% of all individuals with PCOS go undiagnosed (Shauffee et al., 2024). There is a lack of awareness among women about this condition, despite the fact that PCOS can impact fertility and the ability to conceive in the long term (Lim et al., 2023).

Globally, the estimated prevalence of PCOS is 3.4%. In the United States, around five million reproductive-age women are affected by PCOS, with a reported prevalence of 6.5% to 8% in Europe, 15.5% of PCOS patients in Thailand, and a reported prevalence of 8.7% in Australia (Himpunan Endokrinologi Reproduksi dan Fertilitas Indonesia (HIFERI) & Perkumpulan Obstetri dan Ginekologi Indonesia (POGI), 2016). As for Indonesia, there is still no official data on the national prevalence of PCOS, but a study conducted at Cipto Mangunkusumo Hospital identified around 105 PCOS patients. Among these patients, several complaints were reported, including oligo- or amenorrhea in 94.2% of patients and hirsutism in

32.4%. The majority of patients (45.7%) were between the ages of 26 and 30 (Himpunan Endokrinologi Reproduksi dan Fertilitas Indonesia (HIFERI) & Perkumpulan Obstetri dan Ginekologi Indonesia (POGI), 2016).

According to Sun et al. (2018), the K-Nearest Neighbours (KNN) algorithm is a simple and easily applicable machine learning algorithm. It can be used for solving regression and classification problems. KNN is a type of supervised learning algorithm (Sun et al., 2018). In KNN, the computation starts by determining the number of neighbours to consider, which serves as a reference when determining the class of new data. KNN is also considered a lazy learning algorithm because it doesn't require much data for the learning process. KNN works by finding the group of K nearest (similar) objects in the training data to the object in the new or test data (Anamisa et al., 2023).

The research methods for this study include some phases: (1) data collection, (2) data processing, (3) model building and (4) model evaluation. Data is collected from Mendeley Data Website. Dataset contains 72 records of women patients with 37 patients were negative PCOS and 35 patients were positive PCOS. The features of the patients are age (y), height (m), weight (kg), body mass index (BMI) (Kg/m^2), estrous cyclicity, ovarian morphology and serum testosterone (ng/mL).

The dataset is processed to removed irrelevant features. The dataset with irrelevant features is then dropped and renamed. The drop and rename processes are run as pre-processing step of required data. Data with not a number (NaN) annotations were eliminated. In the pre-processing stage, there are several steps involved in preparing the unprocessed data before performing further processes. According to Roy et al. (2018) pre-processing is a stage to eliminate and prevent problems that may occur during data processing. Data cleaning is performed to detect inaccurate data and avoid errors during the classification process. Data cleaning is carried out by importing libraries that assist in this process. Pre-processing involves fixing missing, unnamed, or null data to prevent errors. Then, one-hot encoding is performed to convert categorical data into a format suitable for machine learning classification. Once the data is clean and ready for use, the final step is data normalization. The train-test split technique in machine learning involves dividing the dataset into two parts: training data and testing data. Classification involves making predictions about unknown classes (Rukmana et al., 2022).

PCOS classification model is created using KNN algorithm. The KNN algorithm works based on the shortest distance from the test sample to the training samples to determine its neighbours. KNN has advantages such as being easy to understand, learn, and train. It is considered simple, fast, and effective when trained with large datasets. However, KNN also has a drawback, which is the value of k-fold. There are steps involved in learning using the KNN algorithm, including (Lim et al., 2023):

1. Determining the parameter K.
2. Calculating the distance between the test data and training data. KNN uses two methods to calculate the distance: Euclidean distance and Manhattan distance. In data mining, KNN often used in grouping data. A comparison study was made between Euclidean distance and Manhattan distance. The results showed a good level of accuracy, 84.47% for Euclidean distance and 83.85% for Manhattan distance (Nishom, 2019).
3. Sorting the calculated distances.
4. Selecting the nearest K distances.
5. Assigning the appropriate class.
6. Determining the class based on most of the nearest neighbours and evaluating the data.

The model is then evaluated using cross validation and confusion matrix. The dataset distribution: 90:10, 80:20, 75:25, 70:30, 60:40, and 50:50. Aims to divide data models for training testing. The 70:30 model was used for tuberculosis and healthy microscopic slide classification, The 50:50, 60:40, 70:30, 80:20 and 90:10 models were used for classification of pneumonia (Abdullahi Ibrahim Umar, 2021). Validation is a

technique used to evaluate the results of statistical analysis and generalize them to independent datasets (Herman et al., 2020). This technique is primarily used to predict the performance of a model and estimate its accuracy when applied in real-world scenarios (Herman et al., 2020). In determining the value of K for KNN with parameter optimization, K-Fold Cross Validation plays a crucial role in determining the most optimal value of K. K-Fold involves randomly partitioning the input attributes and testing the system on these randomly selected attribute subsets (Nti et al., 2021). During cross validation, the training and testing process is repeated K times. Performing cross validation with 10 repetitions (10-Fold) has been proven to yield more stable algorithm performance, which is why many researchers utilize it (Indrayanti et al., 2017).

The Confusion Matrix is an evaluation measure that provides information on the comparison between the classification results produced by a system (model) and the true classification results (Vujović, 2021). It is a table that represents four performance measures: True Positive, True Negative, False Positive, and False Negative (Powers, 2020):

1. True Negative represents the correctly predicted negative results.
2. True Positive represents the correctly predicted positive results.
3. False Positive represents the negative data incorrectly predicted as positive.
4. False Negative represents the positive data incorrectly predicted as negative.

To facilitate understanding and remembering the explanations above, if a term is preceded by "True," it means the prediction is correct, whether it occurs or not. If a term is preceded by "False," it means the prediction is incorrect. "Positive" and "Negative" refer to the prediction outcomes generated by the model (Vujović, 2021). Once the Confusion Matrix values are obtained, further calculations can be performed to determine Recall, Accuracy, Precision, and F-1 Score.

The next step is modeling using the K-Nearest Neighbor method. Confusion Matrix will be used in this stage to predict classification values based on the executed data and adjusted with the method.

3. Result and Discussion

The research process begins with (1) data collection, (2) data processing, (3) model building and (4) model evaluation. Before running the system, the required libraries need to be imported into the system and used to execute the machine learning algorithm. After importing the libraries, the dataset is read from an .xlsx file and used as training data, displaying the first 5 rows of the data frame or dataset. Table 1 shows the first 5 rows of the dataset that will be processed in the machine learning phase.

Table 1. Head raw PCOS dataset

	Unnamed: 0	Case No.	Age	Height (m)	Estrous cyclicality	Ovarian morphology	Serum testosterone (ng/ml)
0	NaN	C1	38	1.50	Regular	Normal	0.45
1	NaN	C2	29	1.62	Regular	Normal	0.46
2	NaN	C3	26	1.67	Regular	Normal	0.50
3	NaN	C4	30	1.60	Regular	Normal	0.48
4	NaN	C5	27	1.65	Regular	Normal	0.44

In Table 1, the raw data obtained for the research is presented. The dataset still contains missing values that need to be addressed before training and testing can be performed. Therefore, pre-processing of the dataset is necessary before proceeding with the actual data processing. In the pre-processing stage, unnecessary columns are dropped, and the remaining columns are renamed. The purpose of dropping and renaming is to prepare the required data for processing while eliminating the unnecessary data. To split the data, columns with missing values or "NaN" are eliminated. Additionally, mapping is performed on the "Period Cycle" column, where "Regular" is converted to "0" to indicate consistent menstrual cycles of 28 days, "Irregular (longer)" is converted to "1" to represent irregular or longer menstrual cycles of approximately 35 to 45 days, and "oligomenorrhea" is converted to "2" to signify menstrual cycles occurring at intervals exceeding 35 days, possibly with lighter flow, indicating hormone imbalances, PCOS, thyroid disorders, or other health conditions. Mapping is also performed on the "Status" column, where "Normal" is converted to "0" to indicate no PCOS, and "Polycystic" is converted to "1" to indicate PCOS.

The first 5 rows of the dataset used are shown below.

Table 2. Head PCOS dataset

	Age	Height (m)	Weight (kg)	BMI	Period Cycle	Status	Serum testosterone (ng/ml)
0	38	1.50	55	24.4	0	0	0.45
1	29	1.62	54	20.6	0	0	0.46
2	26	1.67	50	17.9	0	0	0.50
3	30	1.60	51	19.9	0	0	0.48
4	27	1.65	60	22.0	0	0	0.44

To begin dividing the data, it is explained that the target variable "y" will use the "Status" column, while the features "X" will contain all columns except the "Status" column. The purpose of this division is to

classify the accuracy of each feature based on the target variable, which indicates the presence or absence of PCOS, and the features, which represent the symptoms of the patients based on age, height, weight, BMI, menstrual cycle, and serum testosterone.

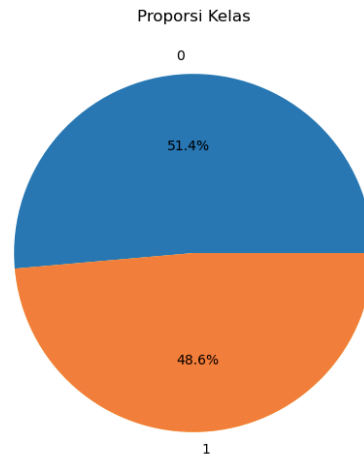


Figure 1. Illustrates the data balance or class proportions.

In Figure 1, it is shown that the dataset is already balanced, meaning that the differences between classes are not significant, and there is no need for resampling to address class imbalance.

After preparing the dataset for the research, the next step is to split the data into training and testing sets. The dataset is divided into various ratios, such as 90% training: 10% testing, 80% training: 20% testing, 75% training: 25% testing, 70% training: 30% testing, 60% training: 40% testing, and 50% training: 50% testing. The following are the results of each accuracy obtained in each data ratio (Figure 2).

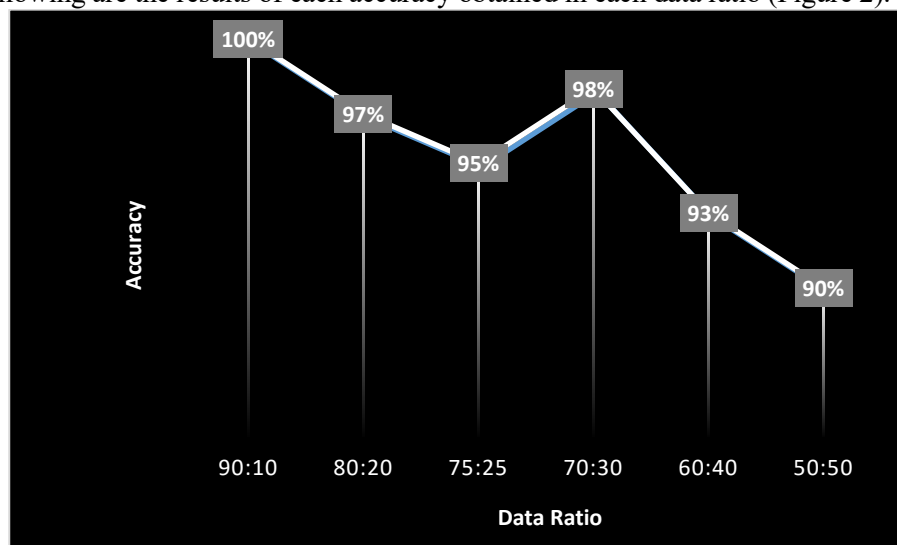


Figure 2 Accuracy of each data ratio

The process for apply KNN begins with importing the library for modeling using KNN. The KNeighborsClassifier library is used to create a KNN (K-Nearest Neighbors) model object with default parameters. The K-Fold Cross Validation is then determined as one of the processes of cross-validation. The code "k_fold = 10" defines the number of folds in k-fold cross-validation, and in the above code, 10 folds are used. "cv_scores = cross_val_score(knn, X_train, y_train, cv=k_fold)" utilizes the cross_val_score function to perform k-fold cross-validation. This function divides the dataset (X_train and y_train) into k-folds, trains and evaluates the model on each fold, and returns the accuracy scores for each fold as an array (cv_scores). Finally, the results of the accuracy obtained for each fold are printed

using f-string to display the fold index and accuracy value for that fold. It is necessary to search for the best value of K, and this search is performed using the Manhattan distance calculation. The Manhattan distance is suitable for datasets that have variations in units, such as in the diagnosis results found in the dataset. After obtaining the best value of K for each ratio, the evaluation stage is conducted.

The evaluation stage is conducted to test the model and determine the accuracy. In this study, evaluation is performed using a confusion matrix. Below are the confusion matrices that have the highest values for accuracy, precision, recall, and F1 score in each calculation.

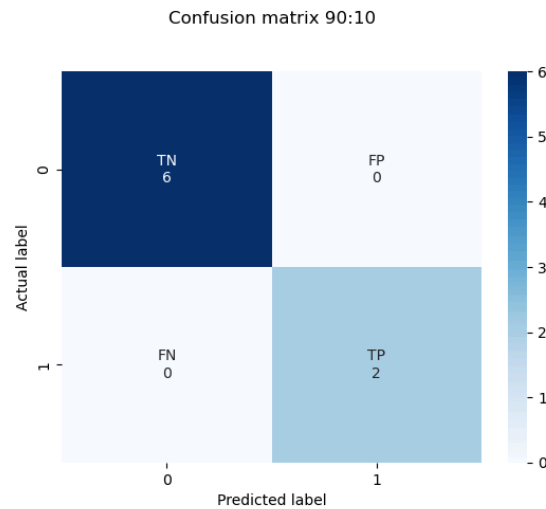


Figure 3. Confusion matrix result for the 90:10 ratio

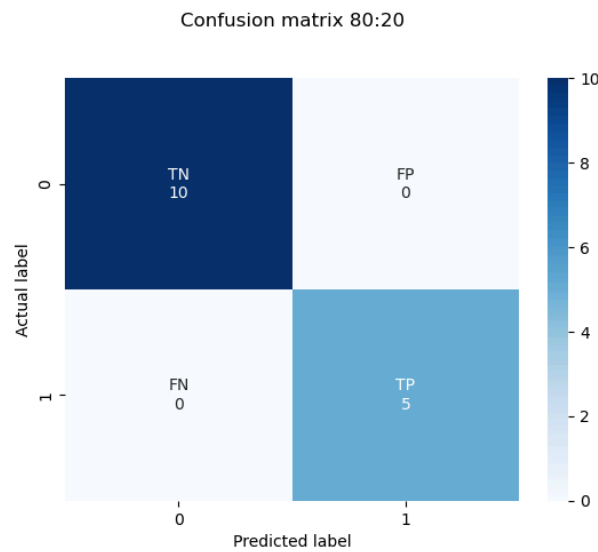


Figure 4. Confusion matrix result for the 80:20 ratio

The parameter 0 represents the "Not Affected by PCOS" status, while 1 represents the "Affected by PCOS" status, obtained from the dataset. Next, we will conduct experiments using the parameters used by KNN, specifically testing with different K values.

After conducting the evaluation, experiments will be conducted using the K parameter to assess the accuracy of KNN. The goal is to determine the impact of the number of neighbors or the K parameter on the accuracy of KNN and identify the best K value as a benchmark. There are several scenarios in the experimentation stage, namely:

1. Testing the K parameter for data training and testing ratios of 90:10, 80:20, 75:25, 70:30, 60:40, and 50:50.

2. Comparing the results of the K parameter experiments to identify the highest values for each data training and testing ratio.
3. The tested data will be presented in the form of bar graphs showing the accuracy comparison for each data testing and training ratio, the classification report graph for the best data training and testing ratio, and the percentage of PCOS classification for each parameter value, "0" and "1".
4. Based on the obtained data, training will be conducted to test the KNN method in classifying PCOS disease based on manually inputted symptoms exhibited by the patients.

Training and evaluation are performed on each fold using the subset of data, X training and y training. The prediction results are stored in y predict. Additionally, other metrics such as recall, precision, and F1-score provide more detailed insights into the best results. Table 3 below shows the experimental results of the K parameter with different data training and testing ratios. To simplify, each ratio is replaced with the corresponding term: R1 for 90:10, R2 for 80:20, R3 for 75:25, R4 for 70:30, R5 for 60:40, and R6 for 50:50.

Table 3. Results of K parameter experimentation.

K parameter	Data Ratio	Accuracy	Precision	Recall	F1-Score
	Testing : Training				
K=1	R1	0.90625	1.0	0.8182 0.8	0.9
	R2	0.8947	1.0	0.8571	0.889 0.9231
	R3	0.9259 0.96	1.0	0.9231	0.960 0.9767
	R4	0.9767	1.0	0.9545	0.9444
	R5	0.9444	1.0	0.9444	
	R6			0.9444	
K=3	R1	0.9531	1.0	0.9091	0.9524
	R2	0.9649	1.0	0.9333 1.0	0.9655
	R3	1.0	1.0	1.0 0.9545	1.0
	R4	1.0	1.0	0.8333	1.0
	R5	0.9767	1.0		0.9767
	R6	0.9167	1.0		0.9091
K=5	R1	0.9844	1.0	0.9697	0.9846
	R2	0.9825	1.0	0.9667	0.9830
	R3	0.9815 0.96	1.0	0.9643	0.9818 0.960
	R4	0.9767	1.0	0.9231	0.9767
	R5	0.9444	1.0	0.9545	0.9412
	R6		1.0	0.8889	

K=7	R1	0.9844	1.0	0.9697	0.9846
	R2	0.9333	1.0	0.9333	0.9655
	R3	0.9630 0.98	1.0	0.9286	0.9630
	R4	0.9767	1.0	0.9231	0.9804
	R5	0.9444	1.0	0.9545	0.9767
	R6		1.0	0.8889	0.94117
K=9	R1	0.9531	1.0	0.9091	0.9524
	R2	0.9649	1.0	0.9333	0.9655
	R3	0.9815	1.0	0.9643 1.0	0.9818
	R4	1.0	1.0	0.9545	1.0
	R5	0.9767	1.0	0.8889	0.9767
	R6	0.9444	1.0		0.9412
K=11	R1	0.9531	1.0	0.9091	0.9524
	R2	0.9825	1.0	0.9667 1.0	0.9830
	R3	1.0	1.0	1.0 0.9545	1.0
	R4	1.0	1.0	0.9444	1.0
	R5	0.9767	1.0		0.9767
	R6	0.9722	1.0		0.9714
K=13	R1	0.96875	1.0	0.9394	0.96875
	R2	0.9825	1.0	0.9667	0.9830
	R3	0.9630 0.98	1.0	0.9286	0.9630
	R4	1.0	1.0	0.9615 1.0	0.9804
	R5	0.9412	1.0	0.8889	1.0
	R6		1.0		0.9412

In Table 3, each data division has varying accuracy values and average cross-validation scores. Table 4 presents a compilation of the K values with the highest accuracy for each data training and testing ratio.

Table 4. Final neighbors result.

K Parameter	Data Ratio	Final Experiment Result
K=3	R3 & R4	100%

K=11	R3 & R4	100%
K=13	R5	100%

The classification results of PCOS disease from the dataset show that parameter 0 represents the "Not Affected by PCOS" status, while 1 represents the "Affected by PCOS" status. The data obtained from the dataset reveals that 51.4% of patients are not affected by PCOS, while 48.6% of patients are affected by PCOS.

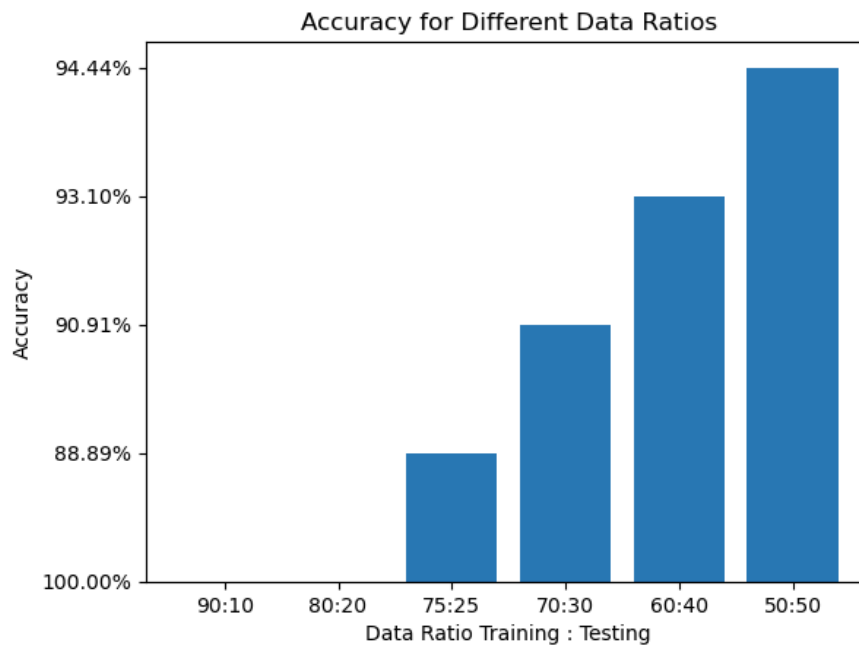


Figure 5. Results of percentage comparison between data divisions.

Figure 5 shows a graph comparing the overall accuracy percentages without performing cross-validation between data divisions of 90:10, 80:20, 75:25, 70:30, 60:40, and 50:50. The lowest accuracy percentage is observed in the 75:25 data division with 88.89%, while the highest accuracy percentages are observed in the 90:10 and 80:20 data divisions with 100%. Previous studies have also achieved high accuracy in classification using the KNN method with data divisions of 90:10 (Anggi Priliani Yulianto & Darwis, 2021; Yogaswara, 2021; Zulaikhah Hariyanti Rukmana et al., 2022) and 80:20 (Anisa & Andri, 2020; Reza et al., 2022).

Here are the results of data processing to generate the output status "The person has PCOS" or "The person does not have PCOS".

```

import numpy as np

feature_names = ['Age', 'Height (m)', 'Weight (kg)', 'BMI', 'Cycle',
                 'Serum testosterone (ng/ml)']
user_input = np.array([[21, 1.54, 56, 19.5, 2, 0.45]])

# Membuat model KNN dengan parameter K terbaik
knn_30 = KNeighborsClassifier(n_neighbors=3, metric='manhattan', weights='uniform', algorithm='auto', leaf_size=30, p=2)

# Melatih model KNN dengan parameter K terbaik
knn_30.fit(X_train_10, y_train_10)

# Melakukan prediksi pada data baru
prediction = knn_30.predict(user_input)

user_input = pd.DataFrame(user_input, columns=feature_names)
prediction = knn_30.predict(user_input)

if prediction[0] == 1:
    print("The person has PCOS")
else:
    print("The person does not have PCOS")

```

The person has PCOS

```

import numpy as np

feature_names = ['Age', 'Height (m)', 'Weight (kg)', 'BMI', 'Cycle',
                 'Serum testosterone (ng/ml)']
user_input = np.array([[21, 1.7, 60, 22.4, 0, 0.45]])

knn_30 = KNeighborsClassifier(n_neighbors=3, metric='manhattan', weights='uniform', algorithm='auto', leaf_size=30, p=2)

knn_30.fit(X_train_10, y_train_10)

# Melakukan prediksi pada data baru
prediction = knn_30.predict(user_input)

user_input = pd.DataFrame(user_input, columns=feature_names)
prediction = knn_30.predict(user_input)

if prediction[0] == 1:
    print("The person has PCOS")
else:
    print("The person does not have PCOS")

```

The person does not have PCOS

Figure 6. Final results of input data array testing.

Figure 6 indicates that the processed data results in a value of 0, which means that the patient does not have PCOS, and it also shows that the processed data results in a value of 1, indicating that the patient has PCOS.

The potential issue of overfitting, given the limited size of the dataset prevents generalizing perfect models to well fit data. We prevent overfitting using drop out or early stopping and detection in a trained model. Performance of KNN as traditional algorithm may enhanced by rebalance the training set. Besides that, improvement can occur by random oversampling, random undersampling and resemble oversampling to the training data (Shi, 2020).

4. Conclusion

In performing the classification using the K-Nearest Neighbor method with a dataset consisting of 72 data obtained from the Mendeley Dataset Website, the modeling is evaluated using the KNN model. Based on the evaluation results using the entire dataset, it can be concluded that using a data training ratio of 90% and 80%, the KNN model achieves 100% accuracy based on the class report. This indicates that the model has a high ability to learn patterns from the majority of the training data and generalize it perfectly to separate validation data.

The evaluation results using k-fold cross-validation show that the data training ratios of 75%, 70%, and 60% provide the highest accuracy, reaching 100% in terms of average accuracy. This suggests that using the appropriate K parameter in the KNN model can result in excellent performance within the range of these data training ratios. However, all of these results may be prone to overfitting due to the limited size

of the dataset, which increases the risk of overfitting during classification.

The classification results using numerical calculations in the dataset used cannot be considered as accurate references for predicting whether someone is affected by PCOS or not. This is because, especially for the "Status" column, accurate information can be obtained through imaging techniques such as ultrasound, which allow visualization of the ovaries and identification of other abnormalities.

Our results showed that KNN can be used for classifying PCOS with high accuracy and precision. This model can make early diagnosis and treatment possible to do. Among the challenges and limitations of our research such as insufficient amount of dataset, we can utilize KNN to help improve the classification performance of the model. Future research we can use hybrid models or combination models of some algorithms.

Acknowledgment

The authors would like to thank Universitas Multimedia Nusantara for the support of this research work.

References

- Abdullahi Ibrahim Umar. (2021). *APPLICATION OF ARTIFICIAL INTELLIGENCE IN MICROBIOLOGY AND CRISP* (Vol. 53, Issue February). <https://doi.org/10.1080/09638288.2019.1595750><https://doi.org/10.1080/17518423.2017.1368728><http://dx.doi.org/10.1080/17518423.2017.1368728><https://doi.org/10.1016/j.ridd.2020.103766><https://doi.org/10.1080/02640414.2019.1689076>
- Anamisa, D. R., Jauhari, A., & Ayu Mufarroha, F. (2023). K-Nearest Neighbors Method for Recommendation System in Bangkalan's Tourism. *ComTech: Computer, Mathematics and Engineering Applications*, 14(1), 33–44. <https://doi.org/10.21512/comtech.v14i1.7993>
- Anggi Priliani Yulianto, & Darwis, S. (2021). Penerapan Metode K-Nearest Neighbors (kNN) pada Bearing. *Jurnal Riset Statistika*, 1(1), 10–18. <https://doi.org/10.29313/jrs.v1i1.16>
- Anisa, C., & Andri. (2020). Penerapan Algoritma k-Nearest Neighbor untuk Prediksi Penjualan Obat pada Apotek Kimia Farma Atmo Palembang. *Bina Darma Conference on Computer Science*, 199–208.
- Anisya, V., Rodiani, & Graharti, R. (2019). Resiko Infertilitas yang dapat Dicegah melalui Penurunan Berat Badan Pada Wanita Obesitas Polycystic Ovary Syndrome: Risk of Infertility that Can be Prevented Through Weight Loss in Obese Women. *Fakultas Kedokteran, Universitas Lampung*, 9, 267–275. <http://juke.kedokteran.unila.ac.id/index.php/medula/article/view/2380>
- Azziz, R. (2018). Polycystic Ovary Syndrome. *Obstetrics & Gynecology*, 132(2), 321–336. <https://doi.org/10.1097/AOG.0000000000002698>
- Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S. E., Legro, R. S., Lizneva, D., Natterson-Horowitz, B., Teede, H. J., & Yildiz, B. O. (2016). Polycystic ovary syndrome. *Nature Reviews Disease Primers*, 2(1), 16057. <https://doi.org/10.1038/nrdp.2016.57>
- Basuki, T. L., Jondri, & Wisesty, U. N. (2020). Deteksi Polycystic Ovarian Syndrome (PCOS) Menggunakan Klasifikasi Microarray Data dengan Algoritma Artificial Neural Network (ANN) Backpropagation dan Principal Component Analysis. *E-Proceeding of Engineering*, 5(3), 8173–8181.
- Che, J., Xian, H., & Zhang, Y. (2021). Adaptive Hybrid Optimized Support Vector Regression with Lasso Feature Selection for Short-term Load Forecasting. *IAENG International Journal of Computer Science*, 48(4).
- Christ, J. P., & Cedars, M. I. (2023). Current Guidelines for Diagnosing PCOS. *Diagnostics*, 13(6), 1113. <https://doi.org/10.3390/diagnostics13061113>

- Dewi, R. M., Adiwijaya, Wisesty, U. N., & Jondri. (2018). Classification of polycystic ovary based on ultrasound images using competitive neural network. *Journal of Physics: Conference Series*, 971, 012005. <https://doi.org/10.1088/1742-6596/971/1/012005>
- Herman, I. H., Widiyanto, D., & Ernawati, I. (2020). Penggunaan K-Nearest Neighbor untuk Mengidentifikasi Citra Batik Pewarna Alami dan Pewarna Sintetis Berdasarkan Warna. *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, 504–515.
- Himpunan Endokrinologi Reproduksi dan Fertilitas Indonesia (HIFERI), & Perkumpulan Obstetri dan Ginekologi Indonesia (POGI). (2016). *Konsensus Tata Laksana Sindrom Ovarium Polikistik Himpunan Endokrinologi Reproduksi dan Fertilitas Indonesia (HIFERI) Perkumpulan Obstetri dan Ginekologi Indonesia (POGI) 2016*. 1–69.
- Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background. *Int. Journal of Engineering Research and Applications*, 3(5), 605–610.
- Indrayanti, Sugianti, D., & Karomi, M. A. Al. (2017). Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus. *IC-Tech*, 7(2), 1–6. <https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/3/2>
- Itriyeva, K. (2022). The normal menstrual cycle. *Current Problems in Pediatric and Adolescent Health Care*, 52(5), 101183. <https://doi.org/10.1016/j.cppeds.2022.101183>
- jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. *Procedia Technology*, 10, 85–94. <https://doi.org/10.1016/j.protcy.2013.12.340>
- Lim, J., Li, J., Feng, X., Feng, L., Xia, Y., Xiao, X., Wang, Y., & Xu, Z. (2023). widi. *BMC Complementary Medicine and Therapies*, 23(1), 1–15. <https://doi.org/10.1186/s12906-023-04249-5>
- McCartney, C. R., & Marshall, J. C. (2016). Polycystic Ovary Syndrome. *New England Journal of Medicine*, 375(1), 54–64. <https://doi.org/10.1056/NEJMcp1514916>
- Munazhif, N. F., Yanris, G. J., & Hasibuan, M. N. S. (2023). Implementation of the K-Nearest Neighbor (kNN) Method to Determine Outstanding Student Classes. *Sinkron*, 8(2), 719–732. <https://doi.org/10.33395/sinkron.v8i2.12227>
- Nandipati, S. C. R., Chew, X., & Wah, K. K. (2020). Polycystic Ovarian Syndrome (PCOS) Classification and Feature Selection by Machine Learning Techniques. *Applied Mathematics and Computational Intelligence*, 9(M1), 65–74.
- Ndefo, U. A., Eaton, A., & Green, M. R. (2013). Polycystic ovary syndrome: A review of treatment options with a focus on pharmacological approaches. *P and T*, 38(6), 336–355.
- Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, 4(1), 20–24. <https://doi.org/10.30591/jpit.v4i1.1253>
- Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*, 13(6), 61–71. <https://doi.org/10.5815/ijitcs.2021.06.05>
- Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. May. <https://doi.org/10.9735/2229-3981>
- Purnama, B., Wisesti, U. N., Adiwijaya, Nhita, F., Gayatri, A., & Mutiah, T. (2015). A classification of polycystic Ovary Syndrome based on follicle detection of ultrasound images. *2015 3rd International*

Conference on Information and Communication Technology, ICoICT 2015, 396–401. <https://doi.org/10.1109/ICoICT.2015.7231458>

Reza, D. A. M., Siregar, A. M., & Rahmat. (2022). Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung. *Scientific Student Journal for Information, Technology and Science*, III(1), 105–112.

Roy, S., Sharma, P., Nath, K., & Bhattacharyya, D. K. (2018). *Pre-Processing : A Data Preparation Step*. April 2020. <https://doi.org/10.1016/B978-0-12-809633-8.20457-3>

Shauffee, L. H., Jantan, H., Fatihah, U., & Bahrin, M. (2024). Polycystic Ovary Syndrome (PCOS) Prediction System Using PSO-SVM Polycystic Ovary Syndrome (PCOS) Prediction System Using PSO-SVM. *Journal of Computing Research and Innovation*, 9(1), 269–282. <https://doi.org/10.24191/jcrinn.v9i1.414>

Shee, H., & Cheruiyot, W. (2014). *Application of k-Nearest Neighbour Classification in Medical Data Mining Network Security View project*. December.

Shi, Z. (2020). Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification. *IOP Conference Series: Materials Science and Engineering*, 719(1). <https://doi.org/10.1088/1757-899X/719/1/012072>

Singh, S., Pal, N., Shubham, S., Sarma, D. K., Verma, V., Marotta, F., & Kumar, M. (2023). Polycystic Ovary Syndrome: Etiology, Current Management, and Future Therapeutics. *Journal of Clinical Medicine*, 12(4). <https://doi.org/10.3390/jcm12041454>

Soleman, R. S., Kreukels, B. P. C., Veltman, D. J., Cohen-Kettenis, P. T., Hompes, P. G. A., Drent, M. L., & Lambalk, C. B. (2016). Does polycystic ovary syndrome affect cognition? A functional magnetic resonance imaging study exploring working memory. *Fertility and Sterility*, 105(5), 1314–1321.e1. <https://doi.org/10.1016/j.fertnstert.2016.01.034>

Sumarlinda, S., & Lestari, W. (2022). Aplikasi K-Nearest Neighbor (KNN) untuk Klasifikasi Penyakit Kardiovaskuler. *Sumarlinda, Sri Lestari, Wiji*, 55, 259–262. <http://ojs.udb.ac.id/index.php/Senatib/article/download/1897/1487>

Sun, J., Du, W., & Shi, N. (2018). A Survey of kNN Algorithm. *Information Engineering and Applied Computing*, 1(1), 1–10. <https://doi.org/10.18063/ieac.v1i1.770>

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, 12(6), 599–606. <https://doi.org/10.14569/IJACSA.2021.0120670>

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1), 23–45. <https://doi.org/10.1080/21693277.2016.1192517>

Xing, W., & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. *IEEE Access*, 8, 28808–28819. <https://doi.org/10.1109/ACCESS.2019.2955754>

Yogaswara, A. R. (2021). Klasifikasi Malware Family menggunakan Metode k-Nearest Neighbor (k-NN). *Jurnal Repositor*, 3(3), 305–314. <https://doi.org/10.22219/repositor.v2i3.1313>

Yuliadha, A., & Setyaningrum, R. H. (2022). Psikoneuroimunologi Depresi pada Polycystic Ovary Syndrome (PCOS). *Smart Medical Journal*, 5(1), 38. <https://doi.org/10.13057/smj.v5i1.43238>

Zhang, X., Liang, B., Zhang, J., Hao, X., Xu, X., Chang, H.-M., Leung, P. C. K., & Tan, J. (2021). Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. *Molecular and Cellular Endocrinology*, 523, 111139. <https://doi.org/10.1016/j.mce.2020.111139>

Zulaikhah Hariyanti Rukmana, S., Aziz, A., & Harianto, W. (2022). OPTIMASI ALGORITMA K-NEAREST NEIGHBOR (KNN) DENGAN NORMALISASI DAN SELEKSI FITUR UNTUK KLASIFIKASI PENYAKIT LIVER. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 6(2), 439–445. <https://doi.org/10.36040/jati.v6i2.4722>