International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Analysis of the Accuracy Level of Breast Cancer Data Classification Using K-Means and Fuzzy C Means Methods

Firda Azmalia[1], Marliza Ganefi Gumay[2]
[1, 2]Department of Business Information Systems, Faculty of Information Systems Management, Gunadarma University, Jl. Salemba Raya No.53, Jakarta Pusat, 10440
Email address: [1]firdaazmalia(at)gmail.com, [2]marliza(at)staff.gunadarma.ac.id

Abstract— *Breast cancer is the most diagnosed cancer in women worldwide. In 2020, there are an estimated 684,996 deaths from breast cancer. Some cases of breast cancer that lead to death are caused by delays in treatment. Seeing how important it is to diagnose breast cancer early, there are many studies that discuss it. This background makes machine learning research made, regarding the diagnosis of breast cancer malignancy by classification. The algorithm used is K-Means and Fuzzy C Means. The data used is obtained from UCI Breast Cancer Wisconsin as much as 699 data, with tools for program development, namely jupyter and the python programming language. The classification of the two algorithms gives the results of the accuracy values to be compared. Furthermore, the classification performance measurement is carried out using the confusion matrix. The result of this research is that for the K-Means algorithm, the accuracy value is 98% and the FCM algorithm is 97%. The measurement of the confusion matrix also shows the error rate for K-Means of 0.0201 and FCM of 0.0301. So when viewed from the level of accuracy and error rate, the one with the best value is the K-Means algorithm.*

Keywords— *Classification, Machine Learning, K-Means, Fuzzy C Means, Confusion Matrix*

I. INTRODUCTION

Breast cancer is cancer that forms in the cells of the breast. Breast cancer can occur in both men and women, but it is much more common in women. Breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more rapidly than healthy cells and continue to accumulate, forming clumps or masses [2]. Cells can spread (metastasize) through the breast to the lymph nodes or to other parts of the body.

Breast cancer is the most diagnosed cancer in women worldwide. It is the most common cancer among both sexes and is the leading cause of cancer death in women. An estimated 2.3 million new cases indicate that one in every 8 cancers diagnosed in 2020 will be breast cancer. In 2020, there are an estimated 684,996 deaths from breast cancer.

Data on breast cancer in Indonesia based on data from WHO is cancer that has the highest number among other types of cancer. Seen in the number of cases in 2020 as many as 65,858, which is the highest presentation before cervical cancer. Meanwhile, the mortality rate from breast cancer is 22,430 which has the second presentation rate, where the highest percentage of mortality is lung cancer [6].

Until now not known with certainty the main cause of breast cancer. However, based on sources from the health department, it is stated that several risk factors can increase the occurrence of breast cancer, one of which is having a history of tumors, menstruation too young or menopause over the age of 50 years, giving birth to the first child over the age of 35 years, unhealthy eating patterns with excessive fat consumption, and obesity. Some cases of breast cancer that can lead to death are caused by delays in treatment [2].

Most patients are not aware of the signs of breast cancer so they are late for a check-up and cause treatment too late even when the patient has entered late-stage breast cancer where the risk of death is even greater [3]. Breast cancer can be found early with BSE. Breast self-examination (BSE) is a breast self-examination to detect any abnormalities in the breast [1]. The purpose of being aware of it is to detect lumps and abnormal changes in the breast early, as well as to detect cancer early. Meanwhile, in our country, healing is already difficult. In fact, detecting breast cancer at an early stage is very easy and can be done at home. The more often the woman examines the breast, the more she will recognize and the easier it will be to find abnormalities in the breast [1].

Seeing how important it is to diagnose breast cancer early, there are many studies that discuss the diagnosis of breast cancer. In 2020, Samiksha Marne, Shweta Churi and Maheshwari Marne conducted a study entitled Predicting Breast Cancer using an effective Classification with Decision Tree and K Means Clustering technique. This study predicts breast cancer using classification with Decision Tree and K Means clustering techniques. The existing datasets were grouped using K Means into benign and malignant classes, then classified using Decision Tree and the accuracy of the correct prediction results was 94.16% [4].

Another research is to compare the level of accuracy of the classification results using several machine learning techniques. This research was conducted by Sharmin Ara, Annesha Das and Ashim Dey in 2021 entitled Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. The algorithm used is Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, Decision Tree and K-Nearest Neighbors (KNN), and the highest accuracy is 96.5% with the SVM algorithm [5].

Based on the description above, this is the background for conducting research using machine learning regarding the diagnosis of breast cancer malignancy. Machine learning that is

Firda Azmalia and Marliza Ganefi Gumay, "Analysis of the Accuracy Level of Breast Cancer Data Classification Using K-Means and Fuzzy C Means Methods," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 127-131, 2022.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

carried out is in the form of classification of breast cancer data using the K Means and Fuzzy C Means algorithms, then the results of the classification of each algorithm will be analyzed to obtain the highest level of accuracy. The data used is secondary data, namely Breast Cancer Wisconsin (Dr. William H. Wolberg, 1992, Breast Cancer Wisconsin (Original) Data Set, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+ (original)), taken from data on breast cancer patients at UCI Machine Learning Wisconsin University [7].

II. Research Methods

The research was conducted using a classification technique with the K-Means and Fuzzy C Means (FCM) algorithms to group categorical data, namely breast cancer data. Furthermore, conducting a more in-depth analysis of the results of the classification for a better level of accuracy between the two algorithms used. The target of machine learning is to classify whether the existing data sets are classified as benign or malignant. The research method carried out is depicted in Figure 1.

Starting from data collection, where the data used is secondary data originating from the UCI Machine Learning Repository, namely the Wisconsin Breast Cancer data https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

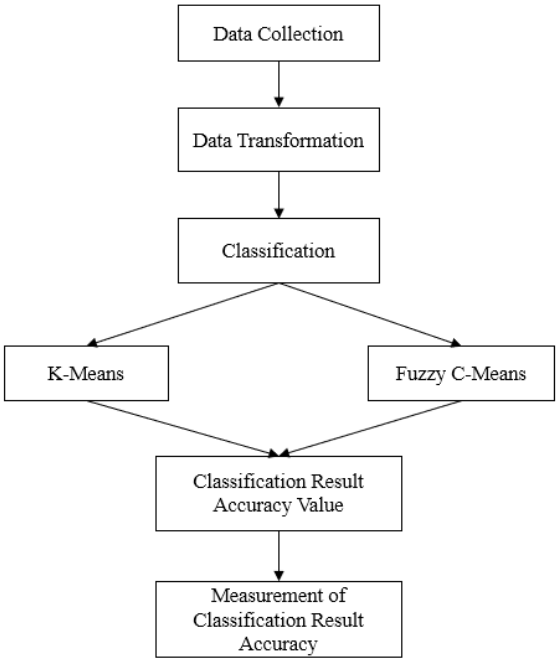


Figure 1. Research Method

Next, transform the processed data through a program built with jupyter tools and the python programming language. The data set used is divided into two parts, the attribute column and the target column as the goal of the classification results, namely benign and malignant. The attribute column is then further divided into training data and test data as needed for data classification.

The classification process is carried out using two algorithms, namely K-Means and Fuzzy C Means. Where to do first is the training model using the training data that has been determined. In the study of 699 data sets, divided into 499 training data and 200 test data. After completing the training data, then the next step is to test data and predict the results of the tests carried out. The next step is to calculate the accuracy of the predicted test results using the K-Means and FCM algorithms.

The last stage is the measurement of classification performance. Where used confusion matrix. The confusion matrix is used as a measurement tool to calculate the performance or the correctness of the classification process. Based on the confusion matrix, it will be possible to calculate the error rate, accuracy, precision, recall, and specificity of each algorithm used.

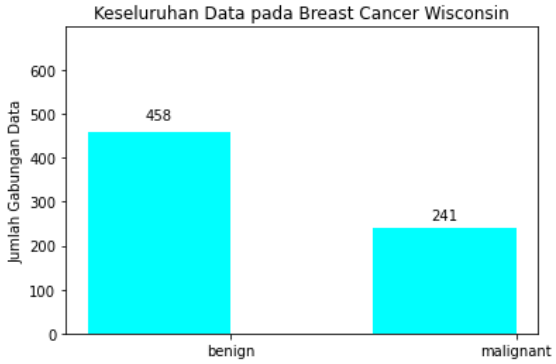
In this study, what is used as a comparison of the two algorithms used is the results of accuracy and error rate. Precision, recall and specificity are not used as comparison reference values. This is because the calculation is more for determining the type of data that has a choice, or in the sense of choosing a right or wrong decision. While the data used in this study is not included in this type of data. The best algorithm is determined from the greater the accuracy value with the smaller error rate.

III. Results and Discussion

At this stage, we will discuss the implementation and results of the classification of breast cancer data previously described using the K-Means and Fuzzy C-Means algorithms.

Data Transformation

This process is carried out in a program made to import data in csv form. Then declare and divide the data into training data and test data. The sample data is taken to be used as a comparative analysis of the two algorithms used. Where the sample data is obtained from the experimental results of data sharing several times, until the maximum results are obtained for further discussion.




```
jumlah keseluruhan data benign 458
jumlah keseluruhan data malignant 241
```

Figure 2. Number of benign and malignant datasets

Figure 2 below shows the number of breast cancer datasets as many as 699 data. A total of 699 data were divided into a total of 458 benign data and 241 malignant data.

Firda Azmalia and Marliza Ganefi Gumay, "Analysis of the Accuracy Level of Breast Cancer Data Classification Using K-Means and Fuzzy C Means Methods," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 127-131, 2022.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Data Classification

After data transformation has been carried out, then data training is carried out, where the existing dataset for training data as much as 400 data is trained so that later predictions can be successfully carried out with good results. This good result can be said if the results that fail to predict are getting smaller.

Prediction of Fuzzy C Means Test Data

Prediction of test data is included in the data classification stage to predict test data, where the test data used as described in Figure 2 is 299 data. The entire test data is further divided into 229 benign data and 70 malignant data. The results of the predicted test data are illustrated in Figure 3 for the Fuzzy C-Means algorithm.

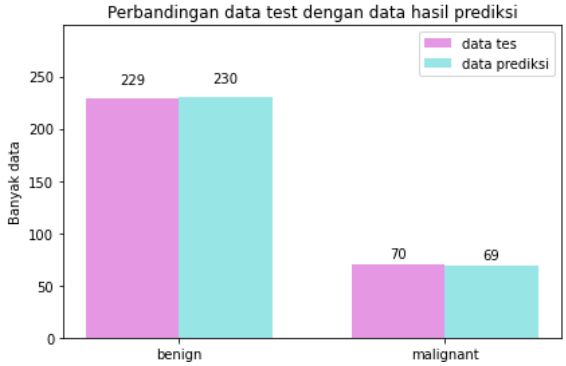


Figure 3. Comparison of test data and prediction results of Fuzzy C-Means

Figure 3 shows the form of a bar chart resulting from the prediction data from the test data that has been previously divided into benign and malignant data. Comparison of benign data obtained predictive data results of 230 from test data 229. In malignant data obtained predictive data results of 69 from test data 70.

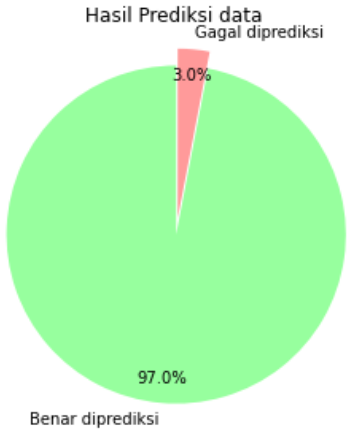


Figure 4. Pie chart of data prediction results

From the previous prediction data, it is also shown in the form of a pie chart in Figure 4. Where it is presented in the form of a percent, that is, the correct data is predicted at 97.0% and the data failed to be predicted at 3.0%.

Figure 5 shows a bar chart of the correct and failed predictions. The yellow bar chart represents right, and the pink one represents wrong predictions. In benign data, 223 data were correctly predicted, including benign data, and 7 data failed to predict. Meanwhile, for malignant data, 67 data were correctly predicted as malignant data, and 2 data failed to predict.

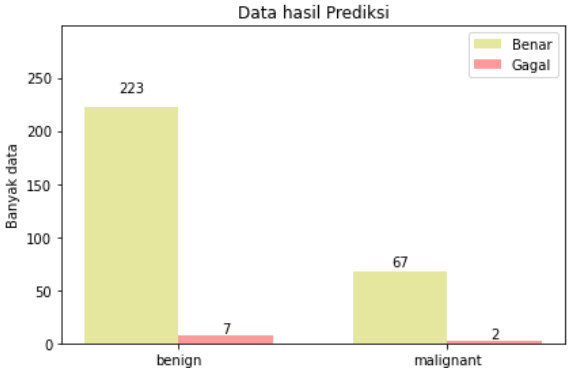


Figure 5. Predicted data bar chart

Prediction of K-Means Test Data

Calculation of data predictions is also carried out using the K-Means method. The amount of test data used remains the same as the Fuzzy C-Means, because with the aim of being able to see the comparison of the results with the same variables.

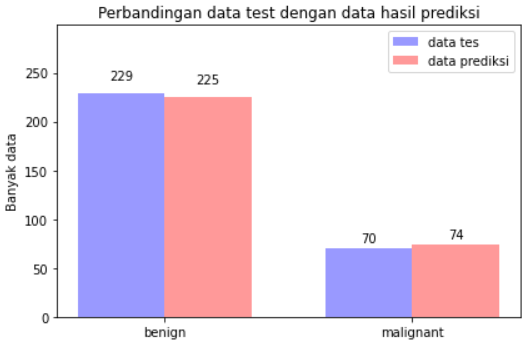


Figure 6. Comparison of test data and K-Means prediction results

In Figure 6, a comparison of test data and predictive data is shown in the form of a bar chart, with purple color representing benign data and pink representing malignant data. Predicted data using the K-Means method for benign data is 225 out of 229 data. Malignant data has predictive results of 74 from test data of 70.

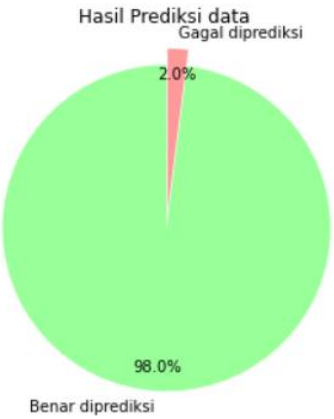


Figure 7. Pie diagram of K-Means. data prediction results

Firda Azmalia and Marliza Ganefi Gumay, "Analysis of the Accuracy Level of Breast Cancer Data Classification Using K-Means and Fuzzy C Means Methods," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 127-131, 2022.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

After predicting the data for the test data that has been determined, the results are displayed in the form of pie charts and bar charts such as Figure 7 and Figure 8.

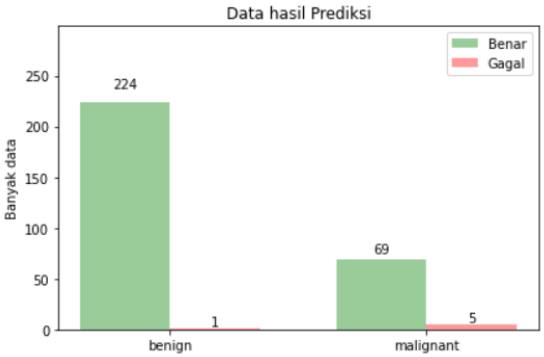


Figure 8. Bar diagram of the predicted K-Means data

The green color represents the data correctly predicted and the pink color represents the data failed to predict for both pie charts and bar charts. Figure 7 shows a pie chart in percentage form, where the correct data is predicted at 98.0% and the data that fails is predicted at 2.0%. Figure 8 explains in more detail the details of the predicted data results. The data for the correct prediction of benign, including benign, were 224 data, and 1 data was incorrectly predicted. While for malignant data, 69 data were correctly predicted as malignant data, and 5 data failed to predict.

Trial Accuracy Results

In this trial phase, the results of the accuracy of breast cancer data were obtained which had previously been predicted data. The FCM method gets an accuracy rate of 0.968996655518395 or in percent, namely 97.0% for 400 training data and 299 test data. It can be seen more clearly in Figure 9.

```
--------------------------
Akurasinya adalah =  0.9698996655518395
Akurasinya adalah =  97.0 %
Data Benar diprediksi  290
Data Salah diprediksi  9
Total data diprediksi  299
```

Figure 9. Fuzzy C-Means Test Results

Another trial stage for the K-Means method as shown in Figure 10 is that the accuracy level of 0.979933110367893 is obtained or in percent, namely 98.0% for training and test data, the same as the FCM method, which is 400 training data and 299 test data.

```
--------------------------
Akurasinya adalah =  0.979933110367893
Akurasinya adalah =  98.0 %
Data Benar diprediksi  293
Data Salah diprediksi  6
Total data diprediksi  299
--------------------------
```

Figure 10. Test Results of K-Means

Classification Performance Measurement

After knowing the accuracy results from the data classification, a confusion matrix table is then made for manual calculations and as a measure of whether the accuracy level is appropriate or not.

TABLE 1. Confusion Matrix Fuzzy C-Means

		Prediction	
		Benign	Malignant
Actual	Benign	223	2
	Malignant	7	67

1. Calculating Accuracy

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$
$$Accuracy = \frac{223 + 67}{223 + 2 + 7 + 67} = \frac{290}{299}$$
$$Accuracy = 0.9698996655518395$$

2. Calculating System Error Rate

$$Eror\ Rate = 1 - Accuracy$$
$$Eror\ Rate = 1 - 0.9698996655518395$$
$$Eror\ Rate = 0{,}0301003344481605$$

3. Calculating Precision

$$Positive\ Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$Positive\ Precision = \frac{223}{223 + 7} = \frac{223}{230}$$
$$Positive\ Precision = 0{,}9695652173913043$$
$$Negative\ Precision = \frac{\text{TN}}{\text{TN} + \text{FN}}$$
$$Negative\ Precision = \frac{67}{67 + 2} = \frac{67}{69}$$
$$Negative\ Precision = 0{,}971014492753623$$

4. Calculating Recall/Sensitivity

$$Sensitivity = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$Sensitivity = \frac{223}{223 + 2} = \frac{223}{225}$$
$$Sensitivity = 0{,}9911111111111111$$

5. Calculating Specificity

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$Specificity = \frac{67}{67 + 7} = \frac{67}{74}$$
$$Specificity = 0{,}9054054054054054$$

TABLE 2. Confusion Matrix K-Means

		Prediction	
		Benign	Malignant
Actual	Benign	224	5
	Malignant	1	69

1. Calculating Accuracy

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$
$$Accuracy = \frac{224 + 69}{224 + 5 + 1 + 69} = \frac{293}{299}$$

Firda Azmalia and Marliza Ganefi Gumay, "Analysis of the Accuracy Level of Breast Cancer Data Classification Using K-Means and Fuzzy C Means Methods," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 127-131, 2022.

$$Accuracy = 0.979933110367893$$

2. Calculating System Error Rate

$$Eror\ Rate = 1 - Accuracy$$
$$Eror\ Rate = 1 - 0.979933110367893$$
$$Eror\ Rate = 0.020066889632107$$

3. Calculating Precision

$$Positive\ Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$Positive\ Precision = \frac{224}{224 + 1} = \frac{224}{225}$$
$$Positive\ Precision = 0.9955555556$$
$$Negative\ Precision = \frac{\text{TN}}{\text{TN} + \text{FN}}$$
$$Negative\ Precision = \frac{69}{69 + 5} = \frac{69}{74}$$
$$Negative\ Precision = 0.9324324324$$

4. Calculating Recall/Sensitivity

$$Sensitivity = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$Sensitivity = \frac{224}{224 + 5} = \frac{224}{229}$$
$$Sensitivity = 0.9781659389$$

5. Calculating Specificity

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$Specificity = \frac{69}{69 + 1} = \frac{69}{70}$$
$$Specificity = 0.9857142857$$

Comparison of Classification Results

The development of a machine learning program for breast cancer data classification has been carried out. Accuracy results from the K-Means and FCM algorithms have also been obtained. Then the performance measurement for the two algorithms has been carried out as well. Where the results of the calculation of the error rate, precision, recall, and specificity can be seen in Table 3 as a comparison of the two algorithms used.

TABLE 3. Comparison of Classification Results

	Akurasi	Error Rate	Precission		Recall	Specificity
			Positif	Negatif		
Fuzzy C-Means	0.9699	0.0301	0.9696	0.9710	0.9911	0.9054
K-Means	0.9799	0.0201	0.9955	0.9324	0.9782	0.9857

Table 3 shows that the FCM accuracy value is 0.9699 and the K-Means is 0.9799. So based on the accuracy value, the K-Means algorithm is better, because the greater the accuracy value, the better the algorithm. When viewed from the error rate, where the smaller the value, the better, the K-Means algorithm is also the best algorithm with a value of 0.0201 while FCM is 0.0301.

IV. CONCLUSION

Based on the identification of the problems described in the introduction, the conclusions from the research conducted on the analysis of the level of accuracy in the classification of breast cancer data using the K-Means and Fuzzy C Means methods are:

A. Classification for breast cancer data using two methods K-Means and Fuzzy C Means was successfully carried out. Where the program development for the classification process is built using jupyter tools and the python programming language.

B. The analysis of the results of the best accuracy level of the two algorithms used was successfully carried out, along with the performance measurement of the classification performed using a confusion matrix. The best accuracy results obtained are for models with the K-Means algorithm, where the accuracy rate is 98% and with Fuzzy C Means the accuracy is 97%.

REFERENCES

[1] Astutik, Reni, Y. (2017). Payudara dan Laktasi. Jakarta: Salemba Medika.

[2] Gagan. (2017). Pengertian Kanker Payudara. [Online]. Available at: https://dinkes.bantenprov.go.id/read/berita/603/Pengertian-Kanker-Payudara.html. [Accessed 25 Maret 2021]

[3] Kemenkes RI. (2017). Deteksi Dini Kanker Payudara dengan SADARI dan SADANIS. [Online]. Available at:http://p2ptm.kemkes.go.id/kegiatan-p2ptm/subdit-penyakit-kanker-dan-kelainan-darah/deteksi-dini-kanker-payudara-dengan-sadari-dan-sadanis [Accessed 12 April 2021]

[4] Samishka, M., Shweta, C., Maheshwari, M. (2020). Predicting Breast Cancer using effective Classification with Decision Tree and K Means Clustering technique. Emerging Smart Computing and Informatics (ESCI).

[5] Sharmin, A., Annesha, D., Ashim, D. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. ICAI.

[6] WHO. (2020). *Breast Cancer*. [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [Accessed 12 April 2021]

[7] Wolberg, H. William, Dr., (1997). *Breast Cancer Wisconsin (Original) Data Set*. Madison, Wisconsin, USA: University of Wisconsin Hospitals.