JURNAL INFOTEL
Informatics - Telecommunication - Electronics

Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

Accuracy Analysis of K-Nearest Neighbor and Naïve Bayes Algorithm in the Diagnosis of Breast Cancer

Irma Handayani[1*], Ikrimach[2]
[1,2] Universitas Teknologi Yogyakarta
[1,2]Siliwangi Street (Ringroad Utara), Jombor, Sleman, D.I.Yogyakarta, Indonesia
*Corresponding Email: irma.handayani@staff.uty.ac.id

Received 04 October 2020, Revised 25 November 2020, Accepted 29 November 2020

Abstract — In the medical field, there are many disease sufferers' records, including data on breast cancer. An extraction process to fine information in previously unknown data is known as data mining. Data mining uses pattern recognition techniques such as statistics and mathematics to find patterns from old data or cases. One of the prominent roles of data mining is classification. In the classification dataset, there is one objective attribute, which is also called the label attribute. This attribute will be searched from new data based on other attributes in the past. The number of attributes can affect the performance of an algorithm. If the classification process is inaccurate, the researcher needs to double-check each previous stage to look for errors. The best algorithm for one data type is not necessarily suitable for another data type. For this reason, the K-Nearest Neighbor and Naïve Bayes algorithms will be used as a solution to this problem. The research method used was to prepare data from the breast cancer dataset, conduct training and testing upon the data, then perform a comparative analysis. The research target is to produce the best algorithm in classifying breast cancer so that patients with existing parameters can be predicted which ones are malignant and benign breast cancer. This pattern can be used as a diagnostic measure to be detected earlier and is expected to reduce the mortality rate from breast cancer. By making comparisons, this method produces 95.79% for K-Nearest Neighbor and 93.39% for Naïve Bayes.

Keywords – classification, data mining, K-NN, Naïve Bayes, breast cancer

I. INTRODUCTION

Classification is widely used to determine decisions according to new knowledge gained from processing past data using algorithms. There is one objective attribute in the classification dataset, or it can be called the label attribute. This attribute will be searched from new data based on other attributes in the past. The number of attributes can affect the performance of an algorithm. If the classification process is inaccurate, the researcher needs to double-check each previous stage to look for errors. Data types significantly affect the performance and accuracy of an algorithm. The best algorithm for one data type is not necessarily suitable for another data type. In general, detection of the level of malignancy of breast cancer is, by the way, called prognosis. The prognosis is the medical team's "best guess" in determining whether a patient is cured of breast cancer or not. Apart from prognosis, another way is bioinformatics using data mining techniques because it has been shown to detect breast cancer's malignancy level [1]. As information technology advances, especially in artificial intelligence, machine learning techniques are being introduced to improve automatic detection capabilities. With this system's help, the possibility of misdiagnosis made by medical professionals can be avoided, and medical data can be checked in a short time and more detail [2].

Several data mining methods that are widely used for classification include the K-Nearest Neighbor and Naïve Bayes algorithms. K-Nearest Neighbor method classifies objects based on learning data close to the object according to the number of closest neighbors or the value of 'k.' Meanwhile, the Naïve Bayes method performs a classification based on probability and the Bayesian theorem with the assumption that each variable X is independent. At present, K-NN and Naïve Bayes have been widely used in problems faced

by humans [3]. In modeling areas such as detection of tumor types using Naïve Bayes [4], classification of kidney stones using K-NN [5], prediction of heart disease using Naïve Bayes [6], classification of Naïve Bayes for predicting colon cancer [7], sentiment analysis in Twitter uses Naïve Bayes [8], detection of abnormal behavior using Naïve Bayes [9] and An Improved KNN Text Classification Algorithm Based on K-Medoids and Rough Set [10]. This study aims to classify breast cancer so that patients with existing parameters can be predicted which ones are malignant and benign breast cancer. This pattern can be used as a diagnostic measure to be detected earlier and is expected to reduce the mortality rate from breast cancer.

II. RESEARCH METHOD

A. System Description

The system is designed to be able to classify cancer data using the K-NN and Naïve Bayes algorithm. The disease is then divided into two classes, namely benign and malignant. The process applied to the system is divided into three stages, including pre-processing, design classifier, and post-processing. For more details, the system flow can be seen in Fig.1.

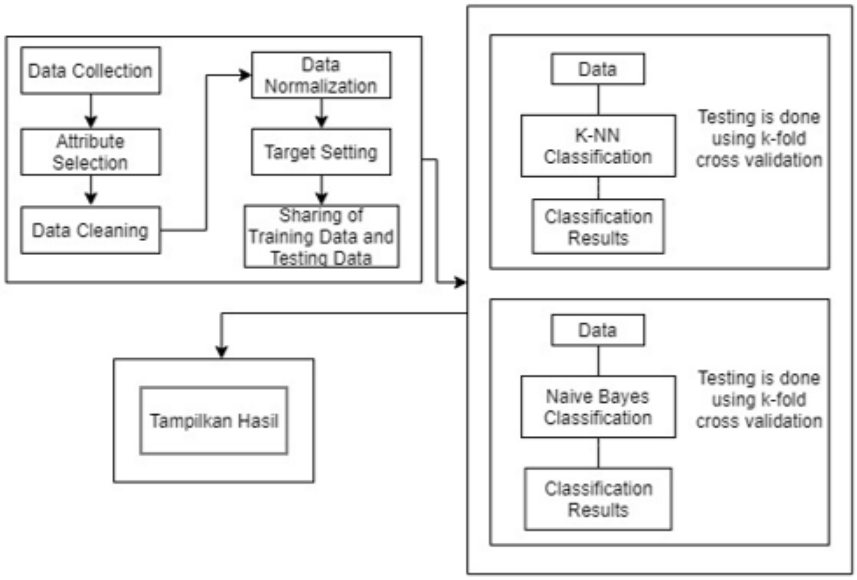

Fig.1. System Flow

The pre-processing stage is a stage that starts with the data collection process. The data collected were then grouped based on the influence on each class. After that, the data is normalized, where the data is entered into the appropriate class. Afterward, the data was distributed to two groups which are 80% for training and 20% for testing.

After the pre-processing stage is complete, the data is then entered into each classifier as knowledge. The classifier then learns from the data that has been entered and evaluated. If any of the specified attributes have not been trained, the system training process will be repeated with a different structure and function.

The third stage is the post-processing stage, where the classification results are displayed in a form that is easier to understand. The system will display whether the cancer is benign or malignant [11].

B. K-Nearest Neighbor Classifier

Classification using the K-NN algorithm is a classification method that uses learning data closest to the object.

a) Data Normalization

Data normalization is carried out so that there is a balance of data on each attribute used. Z-score normalization is a normalization method based on the mean (average value) and standard deviation. This method is fruitful if the actual minimum and maximum values of the data are unknown [12]. Z-Score is a measure of the deviation of data from its average value measured in standard deviation units. If the value is above the average, then the Z-score will be positive. While if the value is below the average, the Z-score will be negative. This Z-Score is also called the Standard Value. The benefit of standardizing the raw score values or observed values from the normal distribution into this Z Score is to allow us to calculate the probability of the score occurring in the normal distribution and also to allow us to compare two scores coming from different populations.

It should be noted that this Z score will only be helpful or meaningful if it is calculated for observations in the form of a normal distribution. Standard Normal Distribution is a normal distribution in the form of an average value of zero (0), and the standard deviation is one (1). To

find the Z Score or Standard Value, we need to know the mean and standard deviation of a population. The formula for calculating the Z Score is to subtract the observed value (raw score) from the population mean and then divide it by the standard deviation. The following is the formula for calculating the Z Score:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where:

z : z-score (standard value)

x : observed value (raw score)

μ : mean

σ : standard deviation

b) Calculation of Proximity Test Data and Training Data

Calculation of the distance between the new data and the training data 1 using the Euclidean distance.

$$d(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (2)$$

Where:

$d(\mathrm{x},\mathrm{y})$: the distance between x and data y

x_k: the value of the attribute from the test data (x), where k= 1, 2, …, n

y_k: the value of the attribute from the training data (y), where k = 1, 2, …, n

After the distance or dissimilarity (d) is calculated, then it is converted into similarity (*s*) with an interval between 0 to 1 (s ϵ [0,1]).

$$z = \frac{1}{1+d} \quad (3)$$

c) K-Fold Cross-Validation

Cross-validation is a simple form of statistical technique. Fold amount the standard for predicting error rates from data is to use 10-fold cross-validation [13]. Cross-validation is used in order to find the best parameters of one model [14]. This is done by testing the amount of error in the testing data. In cross-validation, data is divided into *k* samples of the same size. The *k* subset of data used will be used *k-1* sample training data and one remaining sample for testing data. In cross-validation, data is divided into *k* samples of the same size. The k subset of data used will be used *k-1* sample as training data and one remaining sample for testing data. This is often called k-fold validation.

d) Confusion Matrix

Confusion Matrix is a table to evaluate the performance of the identification model. Confusion Matrix shows the result of identifying the amount of correct prediction data and incorrect predictive data compared to the facts produced. Table 1 shows the Confusion Matrix [15].

$\alpha + \beta = \chi$ (1) (1)

Table 1. Confusion Matrix

Actual	Prediction	
	Negative	**Positive**
Negative	a	b
Positive	c	d

$\alpha + \beta = \chi$ (1) (1)

With:

a: many data predicted by the system with the correct results is indicated healthy, the doctor states indicated healthily.

b: many data predicted by the system with wrong results is indicated by malaria, the doctor state indicated healthily.

c: many data predicted by the system with true results is indicated wrong; the doctor stated malaria.

d: many data predicted by the system with the correct results is indicated malaria; the doctor stated indicated malaria.

There are several terms based on Table 1.

- True Positive (TP) is positive data correctly indicated on the model. Calculation TP values can be calculated using (4).

$$TP = \frac{d}{c+d} \quad (4)$$

- False Positive (FP) is positive data incorrectly indicated on the model. Calculation FP values can be calculated using (5).

$$FP = \frac{b}{a+b} \quad (5)$$

- True Negative (TN) is negative data that is correctly indicated in the model. Calculation TN value can be calculated using (6).

$$TN = \frac{a}{a+b} \quad (6)$$

e) Measurement Accuracy

Measurement of accuracy is a step to prove the level of performance of an algorithm dataset used. In this research, a confusion matrix is used as a performance measurement tool classification algorithm. A confusion matrix is a calculation that compares datasets with the results of the classification, following the actual data with the total amount of data. This matrix's final result is the level of accuracy with units of a percent (%).

This level of accuracy will be used later the researchers' reference to the classification algorithm's performance. Confusion matrix contains information comparison of classification labels with actual labels. From Table 1, the level of accuracy can be calculated from an algorithm model using (7) [16].

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Information:

a: the classification result is *positive* with the *class* actually *positive*

b: *negative* classification results with *positive* actual *class*

c: the result of the classification is *positive* with the *class* actually *negative*

d: *negative* classification results with the actual *class positive*

C. Naïve Bayes Classifier

Classification using the Naïve Bayes Algorithm is a classification method based on the Bayes Theorem assuming each other's parameter independence. Bayes' theorem provides a way to calculate the probability of a parameter's value using the value of another parameter.

Calculation of Probability and Classifier of Test Data,

- After the data is divided into training data and test data, the standard deviation and mean will be calculated for each target parameter class (Diagnosis) for each attribute.
- After the standard deviation and mean per each target parameter class (Diagnosis) per each attribute, will be used for the classification for test data 1.
- Naïve Bayes classification calculates the probability of the diagnosis parameter value based on other parameters' value. Calculation of probability using the Gaussian Naïve Bayes formula:

$$\hat{P}(x_j|c_i = \frac{1}{\sqrt{2\pi\sigma}_{ji}} exp\left(-\frac{(x_j-\mu_{ji})^2)}{2\sigma_{ji}^2}\right) \quad (8)$$

μ_{ji}: mean (average) of feature values x_j of examples for which c = c_i

σ_{ji}: standard deviation of feature values x_j of examples for which c = c_i

- After each attribute's probability is calculated, it will be multiplied into the diagnosis value's probability.

D. Testing Design

In this study, the test carried out tests the system's classification accuracy using the K-NN algorithm and the Naïve Bayes algorithm. Accuracy is measured using the k-fold cross-validation method. The results of measuring the accuracy of the two algorithms are then analyzed for comparison. In addition to the accuracy results, comparisons are also made by comparing the length of time it takes for each algorithm to perform the classification process against the test data that has been prepared. The test scheme is shown in Fig.2.

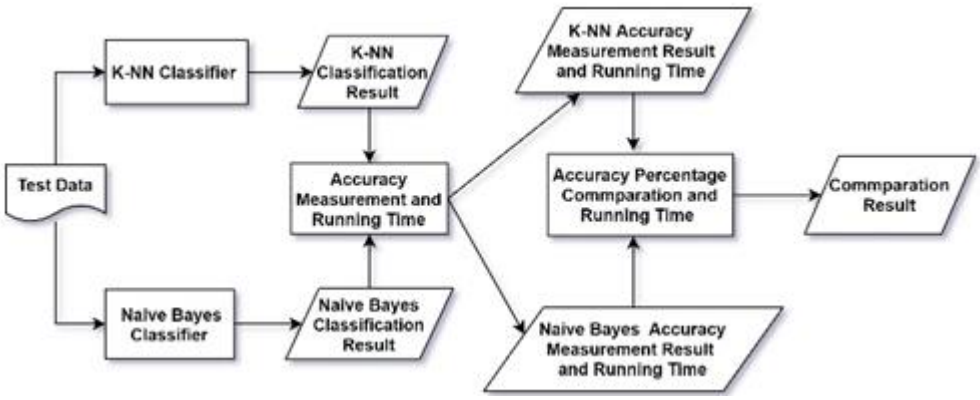

Fig.2. Test Scheme

III. RESULT

A. Cancer Data Compilation Process

The data used as training data and test data are data about breast cancer. There are 455 training data consisting of 284 data on benign cancer cases, 171 data on cases of malignant cancer.

B. Process of Arranging Attributes

The attributes used to classify are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensions based on the data obtained. The data stored for each attribute is the measurement average (mean), standard

error of measurement (se), and the minimum value (worst).

C. K-NN Classifier Process

The K-NN classification process is done by comparing the similarities between test data with training data owned by the system. If the similarity of the case value in the training data with the test data is greater, then it will be collected as a solution. Data collected as a set of solutions is as much as the value of k, so the case with k similarity value as much as k will be used as the solution set. The class diagnosis with the most frequency will be taken and displayed as a system solution [11]. Examples of cases in training data are shown in Table 2. Users classify breast cancer data with data entered into the system in test data shown in Table 3.

Table 2. Training Data

No	Radius mean	Texture mean	Perimeter mean	Area mean	Smoothness mean	..	Fractal dimension worst	Diagnosis
1	1.439512	-2.06894	1.54834	1.478442	0.591122	..	1.236755	M
2	1.932675	0.952607	1.845757	2.155258	-0.01199	..	-0.42221	M
3	-0.16762	-0.59013	-0.11347	-0.27693	1.23536	..	1.52808	M
4	0.197903	0.835859	0.201706	0.060813	0.62539	..	1.035475	M
5	-0.20823	-0.3622	-0.25309	-0.2977	-1.23742	..	-1.03081	B
6	-0.24594	0.004726	-0.31092	-0.36404	-0.36153	..	-1.15423	B
7	-0.94014	0.588465	-0.95863	-0.88874	-0.50545	..	-0.30038	B
8	-1.75415	-1.22113	-1.71539	-1.39468	0.495172	..	-0.09169	B

Table 3. K-NN Test Data

No	Attribute	Value
1	Radius mean	-0.00806
2	Texture mean	1.11105
3	Perimeter mean	0.090175
4	Area mean	-0.13386
5	Smoothness mean	1.201092
..	..	..
30	Fractal dimension worst	0.617027

The process of classifying test data in Table 3 using the training data in Table 2 is divided into several steps, namely calculating proximity, sorting the highest proximity, and determining the solution as a result of classification.

a) Proximity Calculation Process

Calculation of the distance between the new data and the training data 1 using the Euclidean distance.

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

$$= \sqrt{\begin{matrix}(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2 + (x_4-y_4)^2 + \\ (x_5-y_5)^2 + \cdots + (x_{30}-y_{30})^2\end{matrix}}$$

$$= ((1.3656 - (-0.0077))^2 + (-1.9628 - 1.0540)^2 + (1.4689 - 0.0855)^2 + (1.4026 - (-0.1269))^2 + (0.5608 - 1.1394)^2 + \ldots + (1.1733 - 0.5853)^2)^{\frac{1}{2}}$$

$$= (1.8860 + 9.1011 + 1.9138 + 2.3393 + 0.3347 + \cdots + 0.3457)^{\frac{1}{2}}$$

$$= \sqrt{72.1911}$$

$$= 8.4965$$

After calculating the closeness between the new data and the training data 1 is done, the similarity or closeness results are saved for later comparison with the results of the closeness between the new data and other training data. Calculation of the closeness between new data and other training data is done in the same way as calculating the closeness between new data and training data 1.

b) The Highest Similarity Sorting

After the closeness between the new data and all the training data has been carried out, the next step is to sort the training data based on the new data's closest proximity. The highest proximity value sorts the calculation of the new data's closeness with the training data shown in Table 4.

Table 4. Similarity Calculation

No	Data	Proximity
1	Training Data 1	8.956135
2	Training Data 2	6.297322
3	Training Data 3	3.065001
4	Training Data 4	3.993272
5	Training Data 5	7.22263
6	Training Data 6	7.303447
7	Training Data 7	8.871046
8	Training Data 8	8.678481

Based on the data from the calculation of the proximity in Table 4, the closest neighbor is taken as much as k, namely k = 4. So that the closest neighbors to be used for the next stage are Training Data 2, Training Data 3, Training Data 4, and

Training Data 5. K of the closest neighbor data is shown in Table 5.

Table 5. Data (k=4)

No	Data	Proximity	Diagnose
1	Training Data 2	6.297322	M
2	Training Data 3	3.065001	M
3	Training Data 4	3.993272	M
4	Training Data 5	7.22263	B

D. Naïve Bayes Classifier Process

The Naïve Bayes classification process is carried out by calculating the highest probability using a formula based on the Bayes Theorem. Because the available cancer data is continuous, the formula used to calculate the probability is the Gaussian Naïve Bayes Formula.

a) Data Processing

The sample data used as training data and Naïve Bayes classification test data are shown in Tables 6 and 7. The data will be randomly divided into 80% training data and 20% test data.

b) Probability Calculation and Test Data Classification

After the data is divided into training data and test data, the standard deviation and mean will be calculated for each attribute's target parameter class (diagnosis). The standard deviation and mean values are shown in Table 8 and Table 9.

c) Classification Process

After the standard deviation and mean per each target parameter class (Diagnosis) per each attribute, it will be used to classify the test data 1. Naïve Bayes classification calculates the probability of the Diagnosis parameter value based on the value of other parameters. The calculation of probability uses the Gaussian Naïve Bayes Formula.

Table 6. Naïve Bayes Training Data

No	Radius mean	Texture mean	Perimeter mean	Area mean	Smoothness mean	..	Fractal dimension worst	Diagnosis
1	17.99	10.38	122.8	1001	0.1184	..	0.1189	M
2	19.69	21.25	130	1203	0.1096	..	0.08758	M
3	12.45	15.7	82.57	477.1	0.1278	..	0.1244	M
4	13.71	20.83	90.2	577.9	0.1189	..	0.1151	M
5	12.31	16.52	79.19	470.9	0.09172	..	0.07609	B
6	12.18	17.84	77.79	451.1	0.1045	..	0.07376	B
7	9.787	19.94	62.11	294.5	0.1024	..	0.08988	B
8	6.981	13.43	43.79	143.5	0.117	..	0.09382	B

Table 7. Naïve Bayes Test Data

No	Radius mean	Texture mean	Perimeter mean	Area mean	Smoothness mean	..	Fractal dimension worst	Diagnosis
1	13	21.82	87.5	519.8	0.1273	..	0.1072	M
2	12.18	20.52	77.22	458.7	0.08013	..	0.06878	B

Table 8. Standard Deviation from Each Attribute

Attribute Diagnosis	Radius mean	Texture mean	Perimeter mean	Area mean	Smoothness mean	..	Fractal dimension worst
M	3.4360	5.1068	23.4963	344.2541	0.0075	..	0.0164
B	2.5068	2.7268	16.5437	152.9276	0.01037	..	0.0099

Table 9. Mean on Each Attribute

Diagnosis	Radius mean	Texture mean	Perimeter mean	Area mean	Smoothness mean	..	Fractal dimension worst
M	15.96	17.04	106.3925	814.75	0.1187	**..**	0.1115
B	10.3145	16.9325	65.72	340	0.1039	**..**	0.083389

Table 10. Test Data Probability Calculation

Probability Diagnosis	Radius mean	Texture mean	Perimeter mean	Area mean	Smoothness mean	..	Fractal dimension worst
M	0.080114	0.05041	0.012289	0.000803	27.37586	..	23.5136
B	0.089656	0.029352	0.010137	0.001307	3.018703	**..**	2.286222

d) System Testing

The test carried out on the system is the k-fold cross-validation test with k = 10, with data that has been previously randomized with details of 280 benign cancer cases and 170 malignant cancer cases. Then the randomized data is divided into 10 folds, with each fold containing 45 data. The division of data into folds is shown in Table 11.

Table 11. Dividing Data into Folds

Fold	Data
1	1-45
2	46-90
3	91-135
4	136-180
5	181-225
6	226-270
7	271-315
8	316-360
9	361-405
10	406-450

e) Testing the K-NN Classifier

The k-fold cross-validation test was carried out on the K-NN classifier by dividing the data in table 10. The test results using k = 10 for k-fold cross-validation are shown in Table 12.

Table 12. The Results for Each Fold

Fold	Data	Correct Prediction	Accuracy
1	1-45	44	97.8%
2	46-90	44	97.8%
3	91-135	44	97.8%
4	136-180	40	89.1%
5	181-225	43	95.6%
6	226-270	45	100%
7	271-315	43	95.6%
8	316-360	43	95.6%
9	361-405	44	97.8%
10	406-450	41	91.1%
Total	**450**	**431**	**95.79%**

f) Testing Naïve Bayes Classifier

The data used for testing the Naïve Bayes classifier uses k-fold cross-validation with data shown in table 11. The test is carried out by entering the test data one by one into the system and then recording the classification results and the system's running time to perform the classification. The test results are shown in Table 13.

Table 13. Test Results for Each Fold

Fold	Data	Correct Prediction	Accuracy
1	1-45	44	97.8%
2	46-90	44	93.4%
3	91-135	44	97.8%
4	136-180	40	93.4%
5	181-225	43	89.1%
6	226-270	45	95.5%
7	271-315	43	93.3%
8	316-360	43	91.1%
9	361-405	44	93.3%
10	406-450	41	88.8%
Total	**450**	**421**	**93.39%**

IV. DISCUSSION

Based on the calculation of accuracy using k-fold cross-validation for k = 10, the K-Nearest Neighbor algorithm's average accuracy is 95.79%. The average accuracy for k-10 is better than k = 7 with an accuracy of 95.64% and k = 5 with an accuracy of 95.38% with a confusion matrix as in Table 14 below.

Table 14. Confusion Matrix K-NN

	B	M
B	279	2
M	7	159

Based on the confusion matrix in table 14, it can be seen that the system can correctly classify 279 types of benign cancer and 159 malignant cancers. In addition, the true condition value can be seen in Table 15, where the True Positive value (TP) is the correct classification value for each class, False Positive (FP) is the classification value where the actual data is from another class but is classified into class A, for example, class B data on the original data but the classification results give M. True Negative (TN) is the classification result value of another class from the original data of another class. False Negative (FN) is the value of the classification results from the original data for class A. However, the classification results give results that are not class A. For example, the original data for class M but the classification results state class B, of each class, can be seen in Table 15.

Table 15. K-NN Classification Result Data

Class	TP	FP	TN	FN
B	279	7	159	2
M	159	2	279	7

Table 16. Confusion Matrix Naïve Bayes

	B	M
B	276	8
M	14	152

Based on the confusion matrix in Table 16, it can be seen that the system can correctly classify 276 for each type of cancer for benign, and for malignant cancer as many as 152. In addition, the value of the true condition shows the TP, TN, FP values, and FN of each class can be seen in Table 17.

Table 17. Data on Naïve Bayes Classification Results

Class	TP	FP	TN	FN
B	276	8	152	14
M	152	14	276	2

The following is a comparison of the K-Nearest Neighbor classifier results with Naïve Bayes, shown in Table 18.

Table 18. Test Result Data

No	Type of Disease	Amount of Test Data	Accuracy		Running time (second)	
			K-NN	Naïve Bayes	K-NN	Naïve Bayes
1	Benign	280	98%	95%	0.0344595909	0.0070559978
2	Malignant	170	92%	91%		
Amount		**310**	**95%**	**93%**		

V. CONCLUSION

The test results of the system using the K-NN classification method were able to classify 438 data correctly. In comparison, Naïve Bayes correctly classified 428 data using the k-fold cross-validation test with k = 10. This shows that the K-NN classification method has better accuracy than the Naïve Bayes classification method on the data used. The K-NN method gets higher accuracy because the Naïve Bayes algorithm is a parametric algorithm that assumes that each attribute of the data is independent, which is a scarce property in the real world. The average length of time required for the K-NN method to perform classification is much slower because the K-NN algorithm will calculate each training data's distance with test data. In contrast, the Naïve Bayes algorithm only needs to calculate the standard deviation and mean once for all test data.

REFERENCES

[1] G. I. Salama, M. B. Abdelhalim, and M. A. E. Zeid, "Experimental Comparison Of Classifiers For Breast Cancer Diagnosis," *in Proc. - ICCES 2012 2012 Int. Conf. Comput. Eng. Syst.*, November, 2012, pp. 180–185.

[2] E. S. Wahyuni, "Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 7, no. 1, p. 283, 2016.

[3] A. Buditjahjanto, "Determination of the Type of Heart Syndrome in Traditional Chinese Medicine with the Bayesian Network Method," *J. Infotel*, vol. 12, no. 2, pp. 32–38, 2020.

[4] F. Gemci and T. Ibrikci, "Tumor Type Detection Using Naive Bayes Algorithm on Gene Expression Cancer RNA-Seq Data Set," *in International Conference on Engineering Technologies (ICENTE'17)*, 2017.

[5] B. Saçlı *et al.*, "Microwave dielectric property-based classification of renal calculi: Application of a kNN algorithm," *Comput. Biol. Med.*, vol. 112, September, 2019.

[6] R. Shinde, S. Arjun, P. Patil, and P. J. Waghmare, "An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 637–639, 2015.

[7] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019.

[8] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," *in 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, 2016, pp. 416-419.

[9] Y. Ma, S. Liang, X. Chen and C. Jia, "The Approach to Detect Abnormal Access Behavior Based on Naive Bayes Algorithm," *in 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, Fukuoka, 2016, pp. 313-315.

[10] Y. Tan, "An Improved KNN Text Classification Algorithm Based on K-Medoids and Rough Set," *in 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, 2018, pp. 109-113.

[11] I. Handayani, "Application of K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column," *Indones. J. Inf. Syst.*, vol. 2, no. 1, p. 57, 2019.

[12] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019.

[13] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)*. Burlington: Elsevier, 2011.

[14] Trevor HastieRobert TibshiraniJerome Friedman, "The Elements of Statistical Learning" (2nd en., web version)," *Math. Intell.*, pp.269-370, 2008.

[15] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digit. Signal Process. A Rev. J.*, vol. 17, no. 4, pp. 694–701, 2007.

[16] F. Gorunescu, *Data Mining: Concept, Model and Techniques*. Heidelberg, Berlin: Springer, 2011.