EARLY DETECTION OF BREAST CANCER USING THE K-NEAREST NEIGHBOUR (K-NN) ALGORITHM

Refli Tiarma Ariani Panggabean[1], Ledy Octavia[2], Noormala Dwi[3], Aripin[4]
[1,2,3,4]Dian Nuswantoro University
Building I Floor 1 Jalan Nakula No.5-11, Semarang, Central Java, Indonesia
E-mail : reflitiarmaariani@gmail.com

ABSTRACT- Cancer is one of the Non-Communicable Disease groups whose growth and development are high-speed. One type of cancer is breast cancer (carcinoma mammae). Breast cancer is the leading cause of death for women. The first breast cancer cells can grow into tumors as large as 1 cm, spanning 8-12 years. The prevalence rate of breast cancer in Indonesia is 50 per 100,000 female population. The method used in this study uses the K-Nearest Neighbor (K-NN) algorithm by comparing k values, namely 3, 5, and 7. The dataset used was obtained from the UCI Machine Learning Repository with the Number of datasets after preprocessing, namely 653 data with a class consisting of benign tumors (benign) and malignant tumors (malignant). The variables used in this study take into account the variables of clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, cell nucleus size, chromatin, normal cell nucleus, and mitosis. The results of the most influential classification for training and testing are using k = 3 with an accuracy of training and testing at a proportion of 70:30 of 83.8074% and 75%; the ratio of 80:20 is 84.6743% and 74.8092%; the percentage of 90:10 is 84.0136% and 84.6154%. Using the value of k = 3, the resulting gap between training and testing is similar.

Keywords: breast cancer, k-nn, proportion.

1. INTRODUCTION

Cancer is a class of PTM (non-communicable diseases) whose growth and development are swift. The growth and development of cancer can disrupt the body's metabolic processes and spread between cells and tissues. Cancer growth cannot be controlled by cells or tissues[1]. One type of cancer is breast cancer (carcinoma mammae). Breast cancer is the leading cause of death for women. The first breast cancer cells can grow into tumors as large as 1 cm, spanning 8-12 years. These cancer cells can spread through the bloodstream throughout the body, where the spread cannot be known when it occurs. These breast cancer cells hide and can become active in a malignant tumor[2].

According to 2015 Agency For Research On Cancer data, breast cancer is the highest type of cancer globally, with an incidence of 38 per 100,000 female population. The prevalence of cancer in Indonesia, based on data from the 2018 Basic Health Research (RISKESDAS), is increasing. The prevalence rate of breast cancer in Indonesia is 50 per 100,000 female population. The spread of breast cancer in provinces in Indonesia that have the highest prevalence is in the first position, namely the Province of the Special Region of Yogyakarta (DIY) at 2.4%, followed in the second and third positions, namely the Provinces of East Kalimantan 1.0% and West Sumatra 0.9%. Characteristics of people living with breast cancer in Indonesia based on female gender is 2.2 per 1000 population and male is 0.6 per 1000 population[3][4].

Preventing breast cancer can be done by screening or early detection. Early detection of breast cancer often occurs in the form of complaints of swelling or lumps in the breast and armpit areas. Early detection aims to predict whether cancer is classified as benign (benign) or malignant (malignant). In the peaceful class, the tumor does not have the properties of being cancerous or does not damage the surrounding tissue, whereas, in cancer, the tumor will spread and damage the surrounding tissue. At the malignant level, breast cancer originates from the epithelial cells lining the breast lobes or ducts. With optical cell development, breast cancer initially begins as a cell hyperplasia. Then the cells invade the stoma by developing into carcinoma in situ[5][6]. The risk of death from cancer with a high level of malignancy is much higher; therefore, early detection is needed to avoid the risk of getting cancer. According to the Ministry of Health of the Republic of Indonesia, in 2017, around 43% of these deaths could have been prevented if patients had been detected early to avoid the risk of causing cancer.[7].

Early detection of breast cancer can be done by observing the thickness of the clump, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, cell nucleus size, chromatin, normal cell nucleus, and mitosis. From these variables, appropriate analysis is needed to determine whether the detected cancer is benign or malignant[8][9].

In this research, data analysis will be carried out using the variables mentioned using data mining algorithms with machine learning.[10][11].

The method or algorithm used in this study is the K-Nearest Neighbor (KNN) algorithm. K – Nearest Neighbor is a method of classifying objects based on the level of similarity or similarity in the training data sample (training data) through distance classification and the Number of nearest neighbors. The closest Neighbor is determined by the matrix function, which is based on the value of k, which is defined from the point in the data sample[12][13][14].

Previous research has been carried out (Helmi Imaduddin, 2021) diagnosing breast cancer with a comparative study using the Decision Tree algorithm and Support Vector Machine. The dataset used is Coimbra Breast Cancer secondary data obtained from the UCI Machine Learning Repository with a total of 116 data.[15].

Therefore, the authors use the K-Nearest Neighbor Algorithm to determine the accuracy of the breast cancer dataset used.

2. RESEARCH CONTENT

This study uses the Matlab simulation application version R2021 9.10. Data obtained from *UCI Machine Learning Repository*. This study implements the K-Nearest Neighbor Algorithm to determine the classification of breast cancer (carcinoma mammae). The flow of this research can be seen in Figure 1 below.

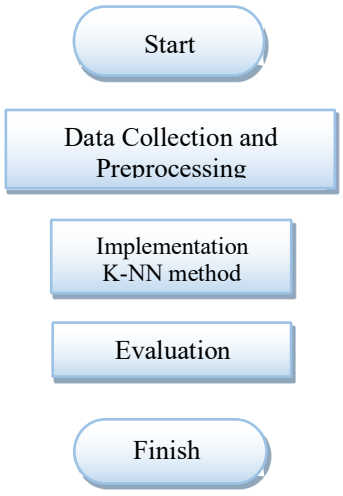

Figure 1. Research Flow

2.1 Data Collection and Research Variables

The data used in this study was obtained from UCI Machine Learning Repository. The dataset used is a risk dataset for breast cancer obtained from the Wisconsin State of 699 data then preprocessing the data into 653 datasets with classes consisting of benign tumors and malignant tumors using variables of clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, nucleus size, chromatin, normal cell nucleus, mitosis, and grade: 2 for benign cancer, 4 for malignant cancer. The distribution of the dataset has been done randomly. The Number of datasets for a peaceful class is 415 data, and the Number of datasets with a malignant type is 238 data.

Table 1. Data Proportion

Data Proportion	Training Data	Test Data
70:30	457	196
80:20	522	131
90:10	588	65

The proportion of training data and test data is done by dividing them into 3 data proportions. The first proportion is 70% training data and 30% testing data, with a lot of training data totaling 457 and testing data totaling 196. The second proportion is 80% training data and 20% testing data, with a lot of training data totaling 522 and testing data totaling 131. The third proportion, namely 90% training data and 10% testing data, with a lot of training data totaling 588 and testing data totaling 65. The data proportions can be seen in Table 1. The data proportions will be used in this study.

2.2 Research Methods

The concept of the K-NN classification method is to find the shortest Distance between the evaluated data and the closest K "neighbors" in the training data. The Euclidean concept is used in calculating the Distance. The most Number of classes with the Distance of the data becomes the class where the evaluation data is located. The K-NN method is quite effective and performs well in processing large data[16][17]. This research was conducted using a comparison of k values to find the best accuracy or performance value. The k values that have the best performance usually use odd-numbered k values. The k values used are 3, 5, and 7. In this study, k = 1 is not used because this value will produce a rigid classification because there is no choice, so that the resulting accuracy value will be the same in each data proportion.[18][19]. After the accuracy of each k value results come out, the accuracy values will be presented in a table, namely tables 2, 3, and 4.

In the K-NN method calculating the Distance can be done by Euclidean formulas[20][21].
The following is a distance search formula using the Euclidean formula[22].

1. Determine the value of the initial N data as the initial knowledge base of the system.
2. Determine the nearest k (Neighbor) value
3. Prepare training data in the form of a criterion value for new data whose status is unknown.
4. Determine the status of each training data for the data being tested (test data) based on equation (1):

$$d(x, y) = ||xy|| = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \quad \text{.....(1)}$$

d(x,y)= Euclidean distance
x= Data 1
y= Data 2
i= feature too -
d= Distance
n= Number of features

2.3 Results and Discussion

Based on the method *K-Nearest Neighbor*which has been carried out using a comparison of the value of k, the results are obtained in the table that has been presented. The results obtained for each value of k are different but do not have a huge difference (gap).

Table 2. Classification results with k=3

Proportion	Training	Testing
70:30	83.8074 %	75%
80:20	84.6743%	74.8092%
90:10	84.0136%	84.6154%

Research that has been carried out using the k-nearest neighbor method with a total of 653 data with k of 3 obtained results of accuracy for training with a proportion of 70:30 83.8074%; with the ratio of 80:20 bringing training accuracy of 84.6743%, with the balance of 90:10 received a training accuracy of 80.0136%. While the accuracy of testing with a proportion of 70:30 obtains an accuracy of 75%, with a ratio of 80:20, got a test accuracy of 74.8092%, and testing accuracy with a balance of 90:10 brings an accuracy of 84.6154%. From the results that have been obtained, it can be observed that the results of training and testing with k of 3 are good enough because the accuracy values of testing and training have a gap that is not too far away. Accuracy results are also presented in a visual form, as shown in Figure 1 below.

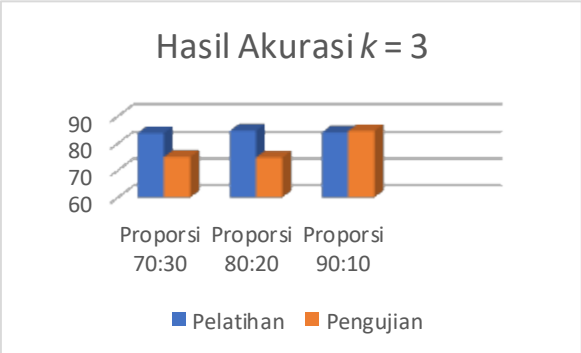

Figure 1. Graph of k = 3 Accuracy

It is clear from the visual data presented in a graphical form that the resulting gap is pretty close between each data proportion. This also shows that there is no gap between the training and testing accuracy results because the training data have higher accuracy than the results of the test data.

Table 3. Classification results with k = 5

Proportion	Training	Testing
70:30	79.6499 %	73.9796%
80:20	79.6935%	70.9924%
90:10	80.4422%	80%

Research that has been carried out using the k-nearest neighbor method with a total of 653 data with k of 5 obtained accuracy results for training with a proportion of 70:30 79.6499%; the ratio of 80:20 obtains training accuracy of 79.6935%; the percentage of 90:10 obtains a training accuracy of 80.4422%. While the accuracy of the test with the proportion of 70:30 obtains an accuracy of 73.9796%, the ratio of 80:20 obtains a test accuracy of 70.9924%; and testing accuracy with a balance of 90:10 brings an accuracy of 80%.

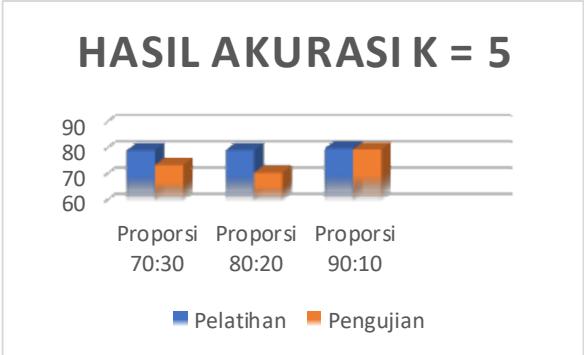

Figure 2. Graph of k = 5 Accuracy

Accuracy results using a value of k = 5 are also presented as visual data, namely by using a graph and the importance of k = 3.

Furthermore, the last experiment used a value of k = 7. Therefore, for accuracy, results using a value of k = 7 can be seen in table 4 below.

Table 4. Classification results with k=7

Proportion	Training	Testing
70:30	77.0241%	76.0204%
80:20	78.3525%	71.7557%
90:10	78.9116%	78.4615%

Research that has been carried out using the k-nearest neighbor method with a total of 653 data with k = 5 obtained accuracy results for training with a proportion of 70:30 77.0241%; with the proportion of 80:20 obtained training accuracy of 78.3525%; the ratio of 90:10 brings a training accuracy of 78.9116%. While the accuracy of the test with the proportion of 70:30 obtains an accuracy of 76.0204%, the balance of 80:20 obtains a test accuracy of 71.7557%, and testing accuracy with a ratio of 90:10 brings an

accuracy of 78.4615%. Accuracy results are also visually presented using a graph like Figure 3 below.

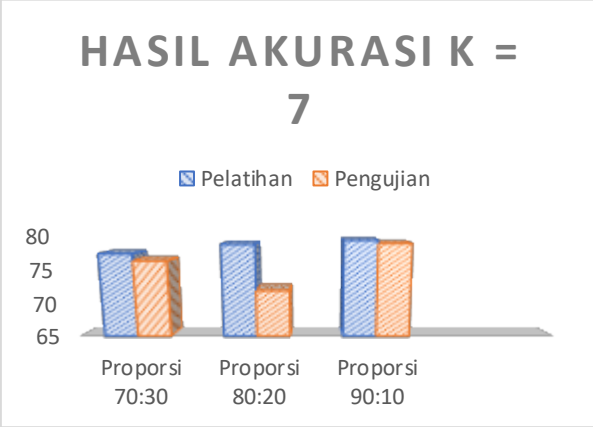

Figure 3. Graph of k = 7 Accuracy

From the graph, the higher the proportion used, the higher the accuracy value will be. This is seen in the balance of data of 90:10.

3. CLOSING

Based on the research that has been done, the classification results obtained for testing and training using values by 3 proportions of 70:30, namely 83.8074% and 75%; the proportion of 80:20 is 84.6743% and 74.8092%; the ratio of 90:10 is 84.0136% and 84.6154%. The results obtained using a k value of 5 proportions 70:30 are 79.6499% and 73.9796%; the proportion of 80:20 is 79.6935% and 70.9924%; the ratio of 90:10 is 80.4422% and 80%. The results obtained using a k value of 7 proportions 70:30 are 77.0241% and 76.0204%; the proportion of 80:20 is 78.3525% and 71.7557%; the ratio of 90:10 is 78.9116% and 78.4615%. From comparing the three k values, in an experiment using the Breast Cancer dataset from the Wisconsin state, the most optimal results were obtained by using a value of k = 3 because it produced the highest accuracy value compared to using values of k = 5 and 7.

BIBLIOGRAPHY

[1] S. Ketut, "Breast Cancer: Diagnostics, Risk Factors, and Staging," *Ganesh Med. J.*, vol. 2, no. 1, pp. 2–7, 2022.

[2] R. Rukinah and S. Luba, "Knowledge of Women of Reproductive Age About Breast Cancer Prevention," *J. Ilm. Healthy. Sandi Husada*, vol. 10, no. 1, pp. 248–252, 2021, doi: 10.35816/jiskh.v10i1.597.

[3] E. Marfianti, "Increasing Breast Cancer Knowledge and Breast Self-Examination Skills (BSE) for Early Detection of Breast Cancer in Semut Jatimulyo Dlingo," *J. Civil Service and Lestari*, vol. 3, no. 1, pp. 25–31, 2021, doi: 10.20885/jamali.vol3.iss1.art4.

[4] A. Rizka, MK Akbar, and NA Putri, "CARCINOMA MAMMAE SINISTRA T4bN2M1 METASTASIS PLEURA," *AVERROUS J. Kedokt. and Health. Malikussaleh*, vol. 8, no. 1, p. 23, 2022, doi: 10.29103/averrous.v8i1.7006.

[5] G. Ramadhan and FD Adhinata, "SMOTE Technique and Gini Score in Breast Cancer Classification," vol. 9, no. 2, pp. 125–134, 2021.

[6] KD Nugroho and U. Sucipto, "Phenomenological Studies: The Impact of Ignoring Cancer Symptoms for Clients and Families," *J. Malang Nursing*, vol. 5, no. 1, pp. 46–54, 2020, [Online]. Available: http://library.gpntb.ru/cgi-bin/irbis64r/62/cgiirbis_64.exe?C21COM=S&I21DBN=RSK&P21DBN=RSK&S21FMT=fullwebr&Z21ID=&S21STN=1&S21REF=10&Z21MFN=856891.

[7] P. Studies *et al.*, "The Effect of Extension on Knowledge Re," vol. 6, no. 1, pp. 529–537, 2020.

[8] N. Sari, R. Wulanningrum, and T. Informatics, "Implementation of the K-Nearest Neighbor Algorithm for Image Identification of Orchids," vol. 1, no. 2, pp. 177–184, 2021.

[9] Normah, B. Rifai, S. Vambudi, and R. Maulana, "Sentiment Analysis of Vtuber Development Using the SMOTE-Based Support Vector Machine Method," *J.Tek. computer. AMIK BSI*, vol. 8, no. 2, pp. 174–180, 2022, doi: 10.31294/jtk.v4i2.

[10] A. Khairi, "FOR CLASSIFICATION OF PRE-PROSPERABLE COMMUNITIES IN SAPIKEREP VILLAGE IN SUKAPURA DISTRICT 1 Introduction," vol. 2, no. 3, pp. 319–323, 2021.

[11] Y. Yuliska and KU Syaliman, "Improving the Accuracy of K-Nearest Neighbor on the Pekanbaru City Air Pollution Standard Index Data," *IT J.Res. Dev.*, vol. 5, no. 1, pp. 11–18, 2020, doi: 10.25299/itjrd.2020.vol5(1).4680.

[12] S. Margaretta, I. Arwani, and DE Ratnawati, "Implementation of the K-Nearest Neighbor Algorithm in Databases Using SQL Language," *...Inf. and Computing Science. e-ISSN*, vol. 4, no. 7, pp. 2043–2052, 2020.

[13] D. Kartini, A. Farmadi, M. Muliadi, D. Turianto Nugrahadi, and P. Pirjatullah, "Comparison of K Values in the Classification of Pneumonia in Toddlers Using K-Nearest Neighbor," *J. Computing*, vol. 10, no. 1, pp. 47–54, 2022, doi: 10.23960/komputasi.v10i1.2965.

[14] S. Sumarlinda and W. Lestari, "Application of K-Nearest Neighbor (KNN) for

Cardiovascular Disease Classification," *Pros. Monday. Nas. Technol. ...*, no. 55, pp. 259–262, 2022, [Online]. Available: http://ojs.udb.ac.id/index.php/Senatib/article/download/1897/1487.

[15] H. Imaduddin, BA Hermansyah, and FA Salsabilla B, "Comparison of Support Vector Machine and Decision Tree Methods in the Classification of Breast Cancer," *Cybersp. J. Educator. Technol. inf.*, vol. 5, no. 1, p. 22, 2021, doi: 10.22373/cj.v5i1.8805.

[16] A. Made, S. Indra, I. Bagus, and G. Dwidasmara, "Implementation Of The K-Nearest Neighbor (KNN) Algorithm For Classification Of Obesity Levels," vol. 9, no. 2, pp. 277–285, 2020.

[17] MR Andryan, M. Fajri, T. Informatics, US Karawang, T. Timur, and J. Barat, "Comparison of the performance of the xgboost algorithm and the support vector machine (SVM) algorithm for diagnosing breast cancer," vol. 6, no. 1, pp. 1–6, 2022.

[18]J. Scientific and W. Education, "No Title," vol. 8, no. 12, p. 445–452, 2022.

[19]O. Magic *et al.*, "TEXT CLASSIFICATION USING THE K-NEAREST ALGORITHM," vol. 5, no. 3, pp. 8237–8249, 2018.

[20] NN Dzikrulloh and BD Setiawan, "Application of the K-Nearest Neighbor (KNN) Method and the Weighted Product (WP) Method in Recruiting Prospective Teachers and New Administration Employees with Technology Insight (Case Study: Muhammadiyah Vocational High School 2 Kediri)," *Development Technol. inf. and Computing Science.*, vol. 1, no. 5, pp. 378–385, 2017.

[21] NLGP Suwirmayanti, "Application of the K-Nearest Neighbor Method for Car Selection Recommendation Systems," *Techno. Com*, vol. 16, no. 2, pp. 120–131, 2017, doi: 10.33633/tc.v16i2.1322.

[22] F. Kurnia, S. Kom, J. Kurniawan, and IF St, "Classification of Poor Families Using the K-Nearest Neighbor Method Based on Euclidean Distance," no. November, pp. 230–240, 2019.