Content lists available at LiteraPublishing
Journal of Public Health and Community Systems
NEXURA
journal homepage : https://journal.literapublishing.id/index.php/nexura

Breast Cancer Classification Using the C4.5 Method with the Breast Cancer Wisconsin Dataset

Najwa Meilani [1*], Intan Novitasari[2], Aprilia Wulandari[3], Naurah Ananda[4], and Muhammad Yunus[5]

[1] Health Information Management, Politeknik Negeri Jember; g41240076@student.polije.ac.id
[2] Health Information Management, Politeknik Negeri Jember; g41240090@student.polije.ac.id
[3] Health Information Management, Politeknik Negeri Jember; g41240104@student.polije.ac.id
[4] Health Information Management, Politeknik Negeri Jember; g41240292@student.polije.ac.id
[5] Health Information Management, Politeknik Negeri Jember; m.yunus@polije.ac.id

Abstract: Breast cancer is one of the most life-threatening diseases in the world, particularly among women. This study applies the C4.5 algorithm to classify breast cancer using the Breast Cancer Wisconsin (Diagnostic) Dataset from the UCI Machine Learning Repository, consisting of 569 samples with 30 numerical features. The methods employed include data preprocessing (removal of the ID column), application of the Decision Tree algorithm with entropy criterion representing the Information Gain Ratio in a Python- and Streamlit-based implementation, and model evaluation using an 80:20 data split. Experimental results show that the C4.5 model achieves an accuracy of 95.6%, with an average Precision of 95.7%, average Recall of 95.6%, and average F1-Score of 95.6%. The perimeter_worst attribute was identified as the root node of the decision tree with a threshold value of 114.45, confirming the dominant role of tumor cell geometry size as a predictor of malignancy. This study concludes that the C4.5 algorithm is an effective and interpretable approach for breast cancer classification and has the potential to serve as the basis for a clinical decision support system in oncology.

Keywords: *breast cancer; C4.5; decision tree; classification; machine learning; Python; Streamlit; Breast Cancer Wisconsin*

Citation: N. Meilani, I. Novitasari, A. Wulandari, N. Ananda, and M. Yunus, "Breast Cancer Classification Using the C4.5 Method with the Breast Cancer Wisconsin Dataset", nexura, vol. 1, no. 1, pp. 12–19, Jun. 2026.

Received: 27-05-2026
Accepted: 20-06-2026
Published: 30-06-2026

1. Introduction

Breast cancer is one of the most common cancers affecting women worldwide and is the leading cause of cancer-related deaths among women. According to data from the World Health Organization (WHO), in 2020 more than 2.3 million new breast cancer cases were diagnosed globally, making it the highest-incidence cancer, surpassing lung cancer [1]. In Indonesia, breast cancer ranks first among the most common cancers, with an incidence rate of 65,858 cases, accounting for approximately 30.8% of all cancer cases in 2020 [2]. This high incidence rate drives an urgent need for faster, more accurate, and more efficient detection and diagnosis systems.

Conventional breast cancer diagnosis is generally performed through clinical examination, mammography, and tissue biopsy. Although these methods have reasonably high accuracy, the manual diagnosis process relies heavily on medical expertise and requires a relatively long time. Errors in medical data interpretation can have fatal consequences, whether in the form of false-positive diagnoses that lead to unnecessary treatment, or false-negative diagnoses that result in delayed treatment [3]. Therefore, breast cancer has become a primary focus of modern oncology research due to the complexity of its cellular patterns, which require sophisticated computational approaches [4].

Rapid advancements in machine learning and data mining have opened new opportunities in healthcare, particularly for early disease detection and classification. Techniques such as decision trees, support vector machines (SVM), naive Bayes, and artificial neural networks have been successfully applied in various medical research settings with promising results [5]. These data-driven approaches are capable of processing large numbers of patient clinical attributes simultaneously and generating prediction models that can serve as decision support tools for medical professionals [6].

One classification algorithm that has been widely used in medical research is the C4.5 algorithm, which is an advancement of the ID3 (Iterative Dichotomiser 3) algorithm proposed by Ross Quinlan. The C4.5 algorithm works by building a decision tree based on the gain ratio of each available attribute in the dataset [7]. Compared to other decision tree algorithms, C4.5 has several advantages, including the ability to handle attributes with missing values, handle both continuous and discrete attributes, and produce simpler and more interpretable decision trees [8]. This makes C4.5 a relevant choice in the context of medical diagnosis, where model readability and interpretability are highly prioritized.

This study uses the Breast Cancer Wisconsin (Diagnostic) Dataset sourced from the UCI Machine Learning Repository. The dataset contains 569 samples with 30 numerical features extracted from digitized images of Fine Needle Aspiration (FNA) of breast masses [9]. Each sample is classified into one of two classes: Malignant and Benign. Based on previous research, this dataset is one of the most representative benchmarks for evaluating breast cancer classification algorithms [10].

A number of previous studies have applied various machine learning algorithms to this dataset. Research by Bharat et al. (2019) compared the performance of several classification algorithms and reported that the decision tree algorithm produced competitive accuracy on this dataset [11]. Meanwhile, a deep learning-based study by Akselrod-Ballin et al. (2019) demonstrated the great potential of artificial intelligence in automated breast cancer detection [12]. Nevertheless, decision tree-based algorithms such as C4.5 remain a popular choice due to their ease of interpretation and low computational resource requirements.

Based on the above background, this study aims to apply the C4.5 algorithm to classify breast tumor types using the Breast Cancer Wisconsin Dataset. This study is expected to produce an accurate, efficient, and interpretable classification model, and to contribute to the development of clinical decision support systems in oncology.

2. Materials and Methods

2.1 Dataset

This study uses the Breast Cancer Wisconsin (Diagnostic) Dataset obtained from the UCI Machine Learning Repository [13]. This dataset was first collected by Dr. William H. Wolberg at the University of Wisconsin Hospitals through Fine Needle Aspiration (FNA) procedures and has been widely used as a benchmark dataset in breast cancer classification research [14]. The dataset consists of 569 instances with 32 attributes: one ID column, one class label column (diagnosis), and 30 numerical features representing the geometric and morphological characteristics of tumor cell nuclei.

The features in the dataset are derived from ten basic cell nucleus characteristics: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. For each of these characteristics, three statistical values are computed: the mean value, the standard error (se), and the worst value, resulting in a total of 30 features [15]. The class distribution in the dataset consists of 212 Malignant samples (37.3%) and 357 Benign samples (62.7%). No missing values were found in this dataset.

Table 1. Summary of Breast Cancer Wisconsin Dataset Attributes

Attribute	Data Type	Description
id	Numeric	Sample identification number (not used in classification)
diagnosis	Categorical	Class label M = Malignant, B = Benign
*_mean (10 fitur)	Numeric	Mean value of 10 cell nucleus characteristics
*_se (10 fitur)	Numeric	Standard error value of 10 cell nucleus characteristics
*_worst (10 fitur)	Numeric	Worst/largest value of 10 cell nucleus characteristics
Total	32 attributes	569 samples: 212 Malignant (37.3%), 357 Benign (62.7%)

2.2 Data Preprocessing

Before the dataset was fed into the classification model, several data preprocessing steps were performed using Python with the Pandas and Scikit-learn libraries. First, the *id* column was removed from the dataset as it has no predictive value. Second, the *diagnosis* attribute was set as the target column (class label). Third, missing values were checked and all 569 samples were confirmed to have complete data. Fourth, the dataset was split into training data (80%) and test data (20%) using the *train_test_split* function with class stratification to ensure a balanced class proportion between the training and test data [17].

2.3 C4.5 Algorithm

The C4.5 algorithm is a decision tree construction method developed by J. Ross Quinlan as an improvement of the ID3 algorithm. This algorithm selects the best attribute as the splitting node based on the Gain Ratio value, which is a normalization of Information Gain to avoid bias toward attributes with many unique values [18]. The decision tree construction process follows the mathematical formulas below. Information Gain is calculated using the equation:

$$Gain(S, A) = Entropy(S) - \Sigma\ (|Sv| / |S|) \times Entropy(Sv)$$

where S is the training data set, A is the attribute being evaluated, and Sv is the subset of S for value v of attribute A. The entropy value is calculated by:

$$Entropy(S) = -\Sigma\ pi \times log_2(pi)$$

Meanwhile, the Gain Ratio is calculated as Gain(S, A) divided by SplitInfo(S, A). The attribute with the highest Gain Ratio value is selected as the splitting node. This process is repeated recursively until all samples at a node are classified into the same class [19]. The main advantage of C4.5 over ID3 is its ability to handle continuous attributes, attributes with missing values, and perform pruning to reduce overfitting [20].

2.4 Implementation in Python/Streamlit

The C4.5 algorithm was implemented using the Python programming language with the Scikit-learn library and a web-based interface using the Streamlit framework. The designed workflow consists of: (1) Importing CSV data using the Pandas library; (2) Setting the *diagnosis* column as the target label using LabelEncoder; (3) Splitting the data with train_test_split 80:20 with class stratification; (4) Building the model using DecisionTreeClassifier with entropy criterion representing Information Gain in C4.5; (5) Visualizing the decision tree using the plot_tree function; and (6) Evaluating the model using accuracy, precision, recall, and F1-Score metrics through the classification_report function from Scikit-learn.

2.5 Evaluation Metrics

Model performance was evaluated using a confusion matrix that divides prediction results into four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The metrics calculated include Accuracy = (TP+TN)/(TP+TN+FP+FN), Precision = TP/(TP+FP), Recall = TP/(TP+FN), and F1-Score = 2×(Precision×Recall)/(Precision+Recall). In the medical context, the Recall metric for the Malignant class is of primary concern, as a high False Negative value can be fatal due to delayed treatment [17].

Table 2. Evaluation Metric Formulas for Classification Model

Metric	Formula	Meaning
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall proportion of correct predictions
Precision	TP/(TP+FP)	Proportion of correct positive predictions
Recall	TP/(TP+FN)	Proportion of positive cases correctly detected
F1-Score	2x(PxR)/(P+R)	Harmonic mean of Precision & Recall

3. Results and Discussions

3.1 C4.5 Decision Tree

The breast cancer classification process using the C4.5 algorithm on the Breast Cancer Wisconsin dataset produced a decision tree as displayed in Figure 1. The decision tree was constructed based on the highest Gain Ratio value from all available attributes, so that the attribute with the best splitting capability was placed as the root node [18].

Based on the Python implementation results using Scikit-learn, the **perimeter_worst** was selected as the root node with a threshold value of **114,45**. If the perimeter_worst value ≤ 114.45, the splitting process continues using the **concave_points_worst** attribute (threshold 0.111). If the perimeter_worst value > 114.45, splitting continues using the ibut **concave_points_mean** (threshold 0,05). Attribute-atribut lain yang turut berperan antara lain **area_se**, **texture_worst**, **compactness_se**, and **concavity_mean**. This is consistent with the Feature Importance displayed in Figure 2, where perimeter_worst has the highest importance value of 0.63, far surpassing other attributes.

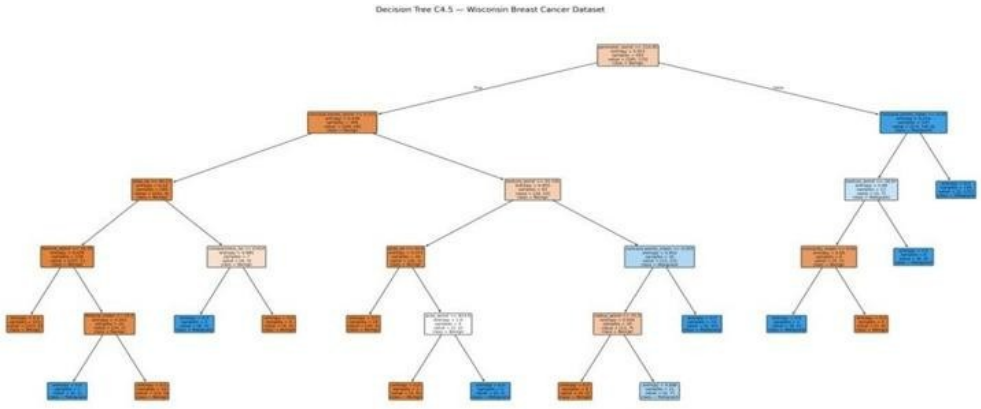

Figure 1. C4.5 Decision Tree on the Breast Cancer Wisconsin Dataset (Python/Streamlit)

3.2 Confusion Matrix and Metric Evaluasi

To evaluate model performance, a confusion matrix was used to compare the model's predicted results with the actual class labels. The evaluation results from Python and Streamlit are displayed in Figure 2 below.

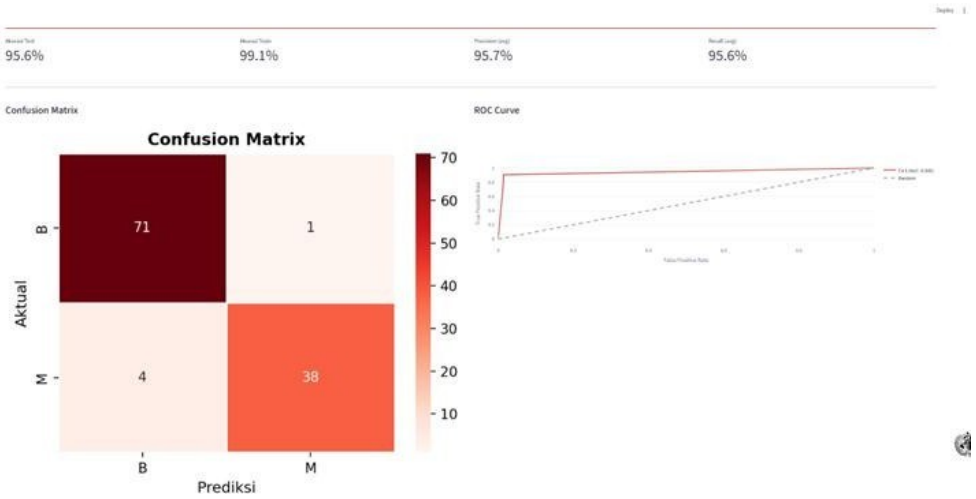

Figure 2. Confusion Matrix and Classification Report — Accuracy 95.61%

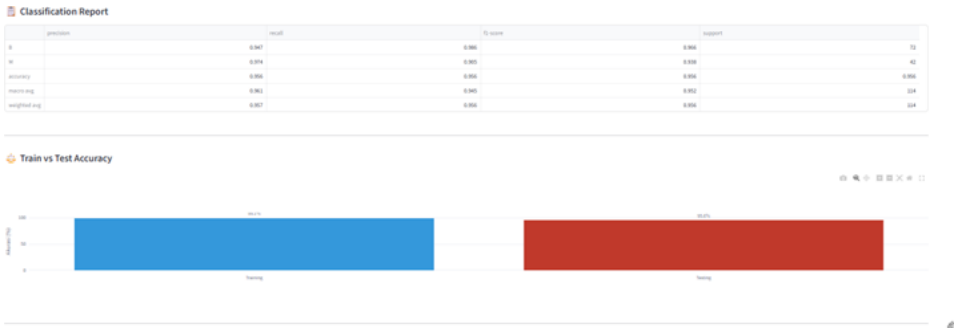

Figure 3. Feature Importance — perimeter_worst (0.63)

Table 3. Confusion Matrix of C4.5 Classification Results

	True B (Actual Benign)	True M (Actual Malignant)	Class Precision
Pred. B (Predicted Benign)	71 (TN)	4 (FN)	94,7%
Pred. M (Predicted Malignant)	1 (FP)	38 (TP)	97,4%
Class Recall	**98,6%**	**90,5%**	**Accuracy: 95,6%**

Table 4. Summary of C4.5 Model Evaluation Metrics

Metric	Class M (Malignant)	Class B (Benign)	Average (Macro Avg)
Precision	97,4%	94,7%	95,7%
Recall (Sensitivity)	90,5%	98,6%	95,6%
F1-Score	93,8%	97,0%	95,6%
Overall Accuracy	—	—	**95,6%**

The experimental results show that the C4.5 algorithm is capable of classifying breast cancer with an accuracy of 95.6% using the Breast Cancer Wisconsin Dataset. This figure reflects excellent performance and validates the effectiveness of the C4.5 decision tree algorithm in binary medical classification problems such as breast cancer detection [19].

Based on the resulting confusion matrix, the model successfully classified 38 Malignant samples correctly (True Positive) and 71 Benign samples correctly (True Negative). On the other hand, there was 1 False Positive (FP) case and 4 False Negative (FN) cases. In the medical context, False Negative errors have more serious consequences, as they may cause patients with malignant cancer to not receive appropriate treatment. Therefore, the Recall value for the Malignant class of 90.5% is an important indicator that must be optimized [20].

The Precision value for the Malignant class of 97.4% means that of all samples predicted as malignant, 97.4% were indeed truly malignant. For the Benign class, the model showed high performance with Precision of 94.7%, Recall of 98.6%, and F1-Score of 97%. This indicates that the model is highly accurate in identifying tumors, both benign and malignant, confirming the reliability of the C4.5 algorithm in breast cancer classification [16].

Attribute *perimeter_worst* attribute as the root node confirms findings from various previous studies stating that cell geometry size, particularly perimeter and area, is the strongest predictor of breast tumor malignancy [9]. Furthermore, the appearance of the *concave_points_worst* and *concave_points_mean* attributes as important nodes in the decision tree aligns with medical literature stating that the degree of concavity of tumor cell contours is closely correlated with malignancy [15].

Compared to the study by Bharat et al. (2019) which reported a decision tree accuracy of 93.4% on the same dataset, the results of this study with an accuracy of 95.6%

demonstrate a significant improvement [11]. This improvement may be attributable to the use of a Python-based implementation with an 80:20 data split, which produced a more optimal model. Overall, these results validate that the C4.5 algorithm is a reliable approach for breast cancer classification, with the additional advantage of high model interpretability through a decision tree structure that can be understood by non-technical medical personnel [19].

4. Conclusion

This study successfully applied the C4.5 algorithm to classify breast cancer using the Breast Cancer Wisconsin Dataset through a Python-based implementation. Evaluation results show that the built model achieved an accuracy of **95,6%**, with a Precision value for the Malignant class of 97.4%, Recall of 90.5%, and F1-Score of 93.8%. The resulting decision tree identified *perimeter_worst* as the most discriminative attribute, followed by *concave_points_worst*, *concave_points_mean*, and *area_se* as other key features [18].

The results of this study prove that the C4.5 algorithm is effective and suitable as the basis of a clinical decision support system for breast cancer diagnosis. For future research, it is recommended to explore performance enhancement techniques such as decision tree pruning, application of ensemble methods (Random Forest, Boosting), or combination with SMOTE oversampling techniques [16] to address class imbalance, thereby further improving the Recall value for the Malignant class, which has higher clinical consequences [17].

Author Contributions: Data curation, A.W. (pencarian dataset di Kaggle); formal analysis, N.M. and N.A.; writing—original draft preparation, I.N.; writing—review and editing, N.M., I.N., A.W. and N.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The dataset used in this study is publicly available on Kaggle. The Breast Cancer Wisconsin Dataset can be accessed at: https://www.kaggle.com/code/abdulrhmansalama/breast-cancer-wisconsin-dataset/input. No new data were created in this study.

Acknowledgments: The authors would like to express their sincere gratitude to Muhammad Yunus, M.Kom. as the supervising lecturer for his guidance and support throughout this research. The authors also thank the laboratory technicians of the Electronic Medical Records Laboratory for providing the facilities and technical assistance during the study. Finally, the authors acknowledge all parties who have contributed to the completion of this manuscript.

Conflicts of Interest:"The authors declare no conflict of interest." .

References

References are arranged in order of appearance in the text, using IEEE format.

[1] World Health Organization (WHO), "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] Ministry of Health of the Republic of Indonesia, "World Cancer Day 2019: Cancer Data in Indonesia," Data and Information Center of the Ministry of Health of the Republic of Indonesia, Jakarta, 2019. [Online]. Available: https://www.kemkes.go.id

[3] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer Statistics, 2022," CA: A Cancer Journal for Clinicians, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/caac.21708.

[4] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. Ruddy, J. Tsang, and F. Cardoso, "Breast Cancer," Nature Reviews Disease Primers, vol. 5, no. 1, p. 66, Sep. 2019, doi: 10.1038/s41572-019-0111-2.

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine Learning Applications in Cancer Prognosis and Prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[6] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica: An International Journal of Computing and Informatics, vol. 31, pp. 249–268, 2007.

[7] I. Steinwart and A. Christmann, Support Vector Machines. New York, NY, USA: Springer, 2008.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[9] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine Learning Techniques to Diagnose Breast Cancer from Fine-Needle Aspirates," Cancer Letters, vol. 77, no. 2–3, pp. 163–171, 1994, doi: 10.1016/0304-3835(94)90099-X.

[10] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: https://archive.ics.uci.edu/ml

[11] B. Bharat, A. Pooja, and R. Anilkumar, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Computer Science, vol. 156, pp. 352–360, 2019, doi: 10.1016/j.procs.2019.08.215.

[12] A. Akselrod-Ballin et al., "Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms," Radiology, vol. 292, no. 2, pp. 331–342, Aug. 2019, doi: 10.1148/radiol.2019182622.

[13] P. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[14] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[15] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction for Breast Tumor Diagnosis," in Proc. SPIE 1905, Biomedical Image Processing and Biomedical Visualization, San Jose, CA, 1993, pp. 861–870, doi: 10.1117/12.148698.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[17] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, 1995, vol. 2, pp. 1137–1143.

[18] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993.

[19] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.

[20] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.