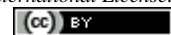


Comparison of Naive Bayes and Decision Tree Methods in Breast Cancer Classification

Daning Nur Sulistyowati^{1*}, Sri Hadian², and Nissa Almira Mayangky³

^{1,2,3} Universitas Nusa Mandiri, Indonesia

MEDINFTech is licensed under a Creative Commons 4.0 International License.



ARTICLE HISTORY

Received: 29 July 25
Final Revision: 08 December 25
Accepted: 14 December 25
Online Publication: 31 December 25

KEYWORDS

Breast Cancer, Decision Tree, Early Diagnosis, Machine Learning, Naive Bayes

CORRESPONDING AUTHOR

daningnur.dgs@nusamandiri.ac.id

DOI

10.37034/medinftech.v3i4.112

ABSTRACT

The early diagnosis of breast cancer is a critical factor in improving recovery rates and reducing cancer-related mortality. This study aims to compare the performance of two widely used machine learning algorithms in medical data classification Naive Bayes and Decision Tree in detecting breast cancer using the Breast Cancer Wisconsin (Diagnostic) dataset. The dataset consists of 569 samples with 30 numerical features and one target label. The methodology includes data preprocessing, model training, and performance evaluation using six metrics: accuracy, precision, recall, F1-score, AUC, and MCC. Naive Bayes achieved higher performance, with 96.5% accuracy, 97.6% precision, 93.0% recall, 95.2% F1-score, 0.997 AUC, and 0.925 MCC, compared to Decision Tree with 93.9% accuracy, 90.9% precision, 93.0% recall, 92.0% F1-score, 0.936 AUC, and 0.87 MCC. Confusion matrix and ROC curve analyses support these results, particularly in minimizing classification errors. While Decision Tree offers better interpretability, Naive Bayes may be more suitable for early breast cancer detection under similar dataset conditions. Future studies could explore ensemble approaches to combine the strengths of both methods.

1. Introduction

Breast cancer is one of the most common cancers among women worldwide and remains a major cause of cancer-related mortality [1], [2]. As shown in Figure 1, breast cancer can be identified in tissue samples through characteristic cellular changes.

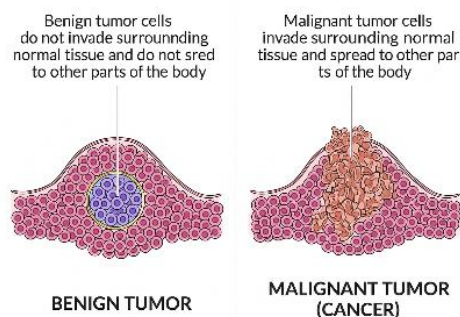


Figure 1. Visual Representation of Breast Cancer

According to the World Health Organization (WHO), breast cancer accounts for approximately 12% of all

newly diagnosed cancers globally, indicating its high public health burden [3]. Early detection is essential because it significantly increases treatment success rates and patient survival [4], [5].

With advances in digital health technology, machine learning has become widely used to support the early diagnosis of breast cancer by analyzing clinical or imaging data [6], [7]. Among the various machine learning methods, Naive Bayes and Decision Tree remain widely used in medical classification tasks due to their simplicity, fast computation, and interpretability [8], [9]. Naive Bayes offers strong performance for numerical data with probabilistic modeling and independence assumptions [10], [11]. while Decision Trees provide clear visual explanations suitable for clinical decision-making [12], [13]. This interpretability advantage is further supported by the Explainable AI (XAI) approach, which provides transparent explanations for model decisions [14], [15].

This study uses the Breast Cancer Wisconsin (Diagnostic) dataset, a widely recognized benchmark

dataset for breast cancer detection research [16]. The dataset includes numerical features extracted from microscopic images of breast tissue biopsy samples. Although many studies have compared classification methods for breast cancer, several research gaps remain. Comparisons between Naive Bayes and Decision Tree using a comprehensive set of evaluation metrics, including MCC and AUC, are still limited [17], [18].

Previous studies have shown that both Naive Bayes and Decision Tree have strong potential in breast cancer detection, but each has its strengths and weaknesses that require deeper analysis [19], [20]. In addition, several studies have suggested the use of ensemble or hybrid approaches to improve classification performance [21].

Therefore, this research focuses on conducting a detailed comparative analysis between Naive Bayes and Decision Tree for breast cancer classification using the Breast Cancer Wisconsin (Diagnostic) dataset. The contributions of this study are as follows:

1. Providing a comprehensive comparison using six evaluation metrics: accuracy, precision, recall, F1-score, AUC-ROC, and MCC.
2. Presenting preprocessing outputs and algorithm configuration to improve research transparency and reproducibility.
3. Offering empirical insight on which method performs more reliably for early breast cancer detection.

This study is expected to provide additional references for developing explainable and accurate machine-learning-based breast cancer diagnostic systems.

2. Research Method

This study adopts a quantitative experimental design aimed at evaluating and comparing the performance of two classification algorithms, namely Naive Bayes and Decision Tree, for the detection of breast cancer. The analysis is conducted using the Breast Cancer Wisconsin (Diagnostic) dataset. The overall research methodology is presented in Figure 2.

2.1. Dataset

The research process commenced with data collection using the Breast Cancer Wisconsin (Diagnostic) dataset, obtained from the UCI Machine Learning Repository (accessible at <https://doi.org/10.24432/C5DW2B>) and also available on Kaggle. This dataset is widely used as a benchmark for machine learning-based breast cancer diagnosis studies. It consists of 569 instances, each described by 30 numerical features representing cellular characteristics extracted from digitized fine needle aspirate (FNA) images of breast masses. The dataset also includes a target label for each instance, indicating whether the cells are benign or malignant. Among the samples, 357 are labeled as benign and 212 as

malignant. The features include various measurements of the cell nuclei, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.. A summary of the dataset is presented in Table 1.

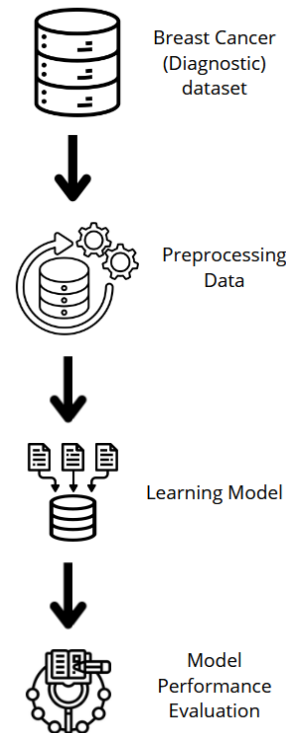


Figure 2. Research Method

Table 1. Summary of the Breast Cancer Wisconsin Dataset

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	...	Symmetry_worst	fractal_dimension_worst
842302	M	17,9	10,3	122,8	1001	...	0,4	0,11
842517	M	20,5	17,7	132,9	1326	...	0,2	0,08
843009	M	19,6	21,2	130,5	1203	...	0,3	0,08
851042	B	13,5	14,3	87,4	566,3	...	0,2	0,07
851065	B	13,0	15,7	85,6	520	...	0,3	0,08
851082	B	9,50	12,4	60,3	274	...	0,2	0,07
...
925311	B	11,2	29,3	70,6	386	...	0,1	0,05
925622	M	15,2	30,6	103,4	716,9	...	0,4	0,14

id	diag- noses	radius_ mean	texture_ mean	perimeter_ mean	area_ mean	smoothness_ worst	fractal_ dimension
926	M	20,9	25,0	143	13	0,2	0,09
125		2	9	47	929	873	
926		21,5	22,3	142	14	0,2	0,07
424	M	6	9	79	06	115	
926		20,1	28,2	131,	12	0,2	0,06
682		3	5	2	61	572	637
926	M	16,6	28,0	108,	85	0,2	0,07
954		8	3	8,1	218	82	
927		20,6	29,3	140,	12	0,4	0,12
241	M	3	1	65	087	4	
927		24,5	47,9	18	0,2	0,07	
51		7,76	4	2	1	871	039

2.2. Preprocessing Data

Data preprocessing was performed to ensure data quality before modeling.

1. Handling missing values, the dataset was checked and confirmed to contain no missing entries.
2. Feature normalization, Min-Max normalization was applied to all numerical features:

$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

3. Dataset splitting, data were split into 80% training and 20% testing.
4. Label encoding – “M” was encoded as 1 (malignant) and “B” as 0 (benign).

2.3. Learning Model

The model training and testing process employed two classification algorithms Naive Bayes and Decision Tree. The Naive Bayes model used in this study was Gaussian Naive Bayes, as it aligns well with the distribution of the numerical data in the dataset and no manually tuned hyperparameters. Meanwhile, the Decision Tree was developed using the CART (Classification and Regression Tree) algorithm, which is capable of splitting the data based on the most optimal separation criteria. Hyperparameters used criterion = “gini”, max_depth = None, min_samples_split = 2, min_samples_leaf = 1 and random_state = 42. Both models were trained using the training data and then tested on the testing data to obtain prediction results.

2.4. Performance Evaluation Model

The performance of the classification models was assessed using six widely adopted metrics in medical diagnostics: accuracy, precision, recall (sensitivity), F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and Matthews Correlation Coefficient (MCC). The calculations of accuracy, precision, recall, and F1-score are defined as shown in Equations (1)–(4), respectively.

Accuracy is calculated using Equation (1), which measures the overall correctness of the model predictions:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (1)$$

Precision, defined in Equation (2), quantifies the proportion of correctly predicted positive samples among all predicted positives

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

Recall (sensitivity), given in Equation (3), measures the proportion of actual positive samples that are correctly identified:

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

Finally, the F1-score, calculated in Equation (4), represents the harmonic mean of precision and recall:

$$F1 - Score = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (4)$$

Where TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

After computing all evaluation metrics, a comparative analysis between the two algorithms was performed to determine the method that demonstrates the best performance in breast cancer detection. This evaluation considered both quantitative metric values and qualitative aspects such as interpretability, computational efficiency, and prediction stability. The results aim to provide a recommendation for the most effective classification algorithm to be implemented in a machine learning-based early diagnosis support system for breast cancer.

3. Result and Discussion

This study aims to compare the performance of the Naive Bayes and Decision Tree methods in classifying breast cancer using the Breast Cancer Wisconsin (Diagnostic) dataset. The evaluation was conducted using several classification metrics—accuracy, precision, recall, F1-score, AUC-ROC, and Matthews Correlation Coefficient (MCC)—to provide a comprehensive overview of each model’s performance.

The evaluation results of the Naive Bayes and Decision Tree models are presented in Table 1.

Table 2. Performance Comparison of Naive Bayes and Decision Tree Model

Performance	Naive Bayes	Decision Tree
Accuracy	0.965	0.939
Precision	0.976	0.909
Recall	0.930	0.930
F1 Score	0.952	0.920
AUC	0.997	0.937
MCC	0.925	0.870

Based on the results, Naive Bayes generally demonstrated better performance, particularly in terms of precision, AUC, and MCC, although both models

yielded the same recall value. A previous study reported an accuracy of 77.14% [22], while the current study achieved an accuracy of 96.5%, indicating a significantly higher performance. This suggests that Naive Bayes is more capable of maintaining a balance between false positives and false negatives, as reflected in its higher MCC value.

The detailed analysis of Table 2 shows that the Naive Bayes model achieved an accuracy of 96.5%, indicating that it correctly classified a large majority of the samples. Its precision of 97.6% implies that among the samples predicted as malignant, almost all were truly malignant, reducing the likelihood of false alarms. The recall (sensitivity) of 93.0% demonstrates that the model successfully identified most of the actual malignant cases, equivalent to the recall of the Decision Tree model. The F1-score, which balances precision and recall, was 95.2% for Naive Bayes, compared to 92.0% for Decision Tree, confirming the superior overall classification performance. In addition, the AUC value of 0.997 for Naive Bayes, higher than 0.937 for Decision Tree, indicates that the Naive Bayes model has a better ability to discriminate between malignant and benign cases across various decision thresholds. The MCC value, 0.925 for Naive Bayes versus 0.870 for Decision Tree, further emphasizes the overall reliability of Naive Bayes, accounting for all true and false positives and negatives, which is particularly important in imbalanced datasets.

Further analysis was conducted using the confusion matrix to understand the distribution of the model's predictions relative to the actual labels. The confusion matrix illustrates how well the model correctly classifies instances (true positives and true negatives) and, conversely, how many errors occur (false positives and false negatives).

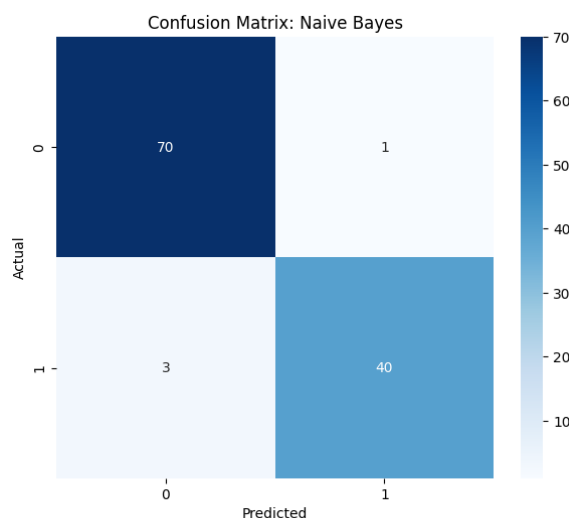


Figure 3. Confusion Matrix Naïve Bayes

Figure 3 presents the confusion matrix for the Naive Bayes model in classifying breast cancer samples. The

model correctly identified 70 negative cases (true negatives) and 40 positive cases (true positives), while misclassifying only 1 negative sample as positive (false positive) and 3 positive samples as negative (false negatives). This indicates that Naive Bayes achieves high precision and recall, with a particularly low false positive rate, which is important in clinical settings to minimize unnecessary interventions for healthy patients. At the same time, the small number of false negatives demonstrates the model's strong ability to detect malignant cases, supporting its reliability for early cancer diagnosis.

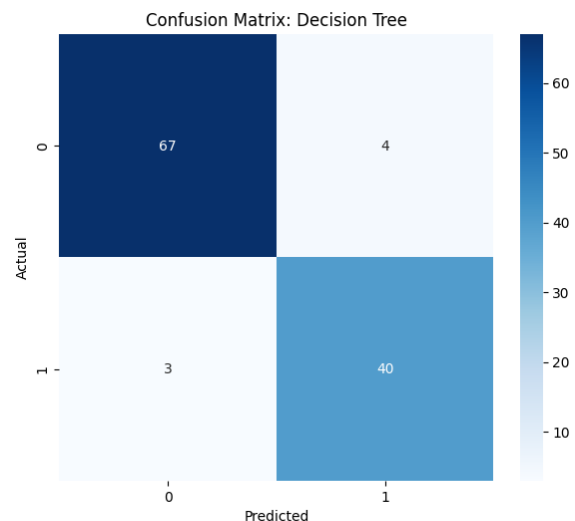


Figure 4 Confusion Matrix Decision Tree

Figure 4 shows the confusion matrix for the Decision Tree model. In this case, the model correctly classified 67 negative samples and 40 positive samples, but produced 4 false positives and 3 false negatives. Compared to Naive Bayes, the higher number of false positives suggests that Decision Tree is slightly less effective at avoiding misclassification of non-cancerous patients, although it maintains similar sensitivity in detecting actual cancer cases. This distinction highlights the superior performance of Naive Bayes in minimizing false positives, which is critical for clinical applications where unnecessary treatments must be avoided.

The final evaluation was conducted through ROC curve analysis for each model. The ROC curve illustrates the model's ability to distinguish between positive and negative classes. The closer the curve is to the top-left corner, the better the model's performance.

Figure 5 displays the ROC curves comparing the performance of the Naive Bayes and Decision Tree models in breast cancer classification. The ROC curve for Naive Bayes is closer to the top-left corner of the plot, with an Area Under the Curve (AUC) value of 1.00, suggesting a very high classification ability on this dataset. The Decision Tree ROC curve, while slightly farther from the ideal corner, also shows strong performance with an AUC of 0.94. These high AUC

values indicate that both models are effective in distinguishing between positive and negative classes, with Naive Bayes showing somewhat better performance under the conditions of this study.

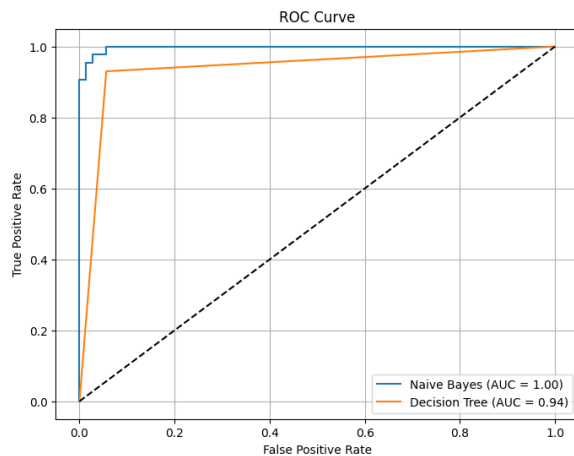


Figure 5. ROC Curve

3.1. Discussion

The results indicate that both Naive Bayes and Decision Tree models performed well in classifying breast cancer samples, with Naive Bayes showing slightly better overall performance. Naive Bayes achieved higher precision (97.6%), F1-score (95.2%), AUC (0.997), and MCC (0.925), while recall was similar for both models (93.0%). The confusion matrices and ROC curves suggest that Naive Bayes had a lower false positive rate and slightly better discriminative ability across thresholds.

Several previous studies have also compared machine learning methods for breast cancer detection. Shah et al. (2020) reported that Naive Bayes achieved 94% accuracy on the same dataset, slightly lower than the 96% obtained in this study [23]. Another study by Kadhim & Kamil (2024) showed that Decision Tree reached 91% accuracy, comparable to our 93% result [24]. Meanwhile, Zhang et al. (2024) found that variants of Naive Bayes such as Tree-Augmented Naive Bayes produced higher AUC values than standard Naive Bayes [25]. These comparisons suggest that Naive Bayes performs reliably in probabilistic classification tasks involving numerical biomedical data, although performance may vary depending on dataset characteristics and experimental settings.

Overall, the findings highlight the potential of Naive Bayes for accurate and consistent breast cancer classification, with further validation recommended on larger and more diverse datasets.

4. Conclusion

This study compared the performance of Naive Bayes and Decision Tree in classifying breast cancer using the Breast Cancer Wisconsin (Diagnostic) dataset. Naive Bayes achieved 96.5% accuracy, 97.6% precision,

93.0% recall, 95.2% F1 score, 0.997 AUC, and 0.925 MCC, outperforming Decision Tree in most metrics, while both models had similar recall. Decision Tree remains valuable for its interpretability, whereas Naive Bayes may offer advantages for early breast cancer detection systems under similar dataset conditions.

Future work could include testing additional datasets from different institutions, optimizing Decision Tree parameters, exploring other algorithms such as Random Forest, SVM, or ensemble methods, and integrating explainable AI frameworks to enhance clinical interpretability.

References

- [1] J. S. Ahn *et al.*, "Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine," *J. Breast Cancer*, vol. 26, no. 5, pp. 405–435, 2023, doi: 10.4048/jbc.2023.26.e45.
- [2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.
- [3] World Health Organization (WHO), "Breast cancer: Prevention and control," *WHO Fact Sheets*, 2022. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
- [4] L. Wang, "Mammography with deep learning for breast cancer detection," *Front. Oncol.*, vol. 14, no. February, pp. 1–16, 2024, doi: 10.3389/fonc.2024.1281922.
- [5] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," *Multimed. Tools Appl.*, vol. 84, no. 25, pp. 24079–24190, doi: <https://doi.org/10.1007/s11042-024-20011-6>.
- [6] I. P. Ramadhani, Fanny; Al-Khowarizmi; Sari, "Improving the Performance of Naive Bayes Algorithm by Reducing the Attributes of Dataset Using Gain Ratio and Adaboost," *IEEE Xplore*, 2021, doi: 10.1109/IC2SE52832.2021.9792027.
- [7] M. G. M. Lauande *et al.*, "Classification of Histopathological Images of Penile Cancer using DenseNet and Transfer Learning," *Proc. Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl.*, vol. 4, pp. 976–983, 2022, doi: 10.5220/0010893500003124.
- [8] C. Shrimali, M. Baghel, and S. Rajput, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification," *Indian J. Sci. Technol.*, vol. 8, no. 29, pp. 1–8, 2015, doi: 10.17485/ijst/2015/v8i29/84646.
- [9] M. Biswas *et al.*, "State-of-the-art review on deep learning in medical imaging," *Front. Biosci. Landmark*, vol. 24, no. 3, pp. 392–426, 2019, doi: <https://doi.org/10.2741/4725>.
- [10] H. Shah, S. Agrawal, P. Oza, and S. Tanwar, "Comparative Study on Machine Learning Algorithms for Breast Cancer Diagnosis," *Procedia Comput. Sci.*, vol. 259, pp. 1994–2003, 2025, doi: 10.1016/j.procs.2025.04.155.
- [11] J. Ma'touq and N. Alnuman, "Comparative analysis of features and classification techniques in breast cancer detection for Biglycan biomarker images," *Cancer Biomarkers*, vol. 40, no. 3–4, pp. 263–273, 2024, doi: 10.3233/CBM-230544.
- [12] A. A. Balasubramanian *et al.*, "Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology," *Cancers (Basel)*, vol. 16, no. 12, 2024, doi: 10.3390/cancers16122222.

- [13] R. R. Kadhim and M. Y. Kamil, "Comparison of machine learning models for breast cancer diagnosis," *IAES Int. J. Artif. Intell.*, vol. 12, no. 1, pp. 415–421, 2023, doi: 10.11591/ijai.v12.i1.pp415-421.
- [14] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innov.*, vol. 3, no. 5, pp. 1–22, 2024, doi: 10.1002/cai2.136.
- [15] T. Islam *et al.*, "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-57740-5.
- [16] V. Nemade, S. Pathak, and A. K. Dubey, "A Systematic Literature Review of Breast Cancer Diagnosis Using Machine Intelligence Techniques," *Arch. Comput. Methods Eng.*, vol. 29, no. 3, pp. 4401–4430, 2022, doi: 10.1007/s11831-022-09738-3.
- [17] A. Ara, Sharmin; Das, Annesha; Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021. doi: <https://doi.org/10.1109/ICAIS2203.2021.9445249>.
- [18] Q. Zhang, H. Zheng, Y. Gong, Z. Liu, and S. Chen, "Comparative Analysis of Bayesian Networks for Breast Cancer Classification: Naive Bayes vs. Tree-Augmented Naive Bayes," *2024 4th Int. Conf. Electron. Inf. Eng. Comput. Sci. EIECS 2024*, pp. 1078–1081, 2024, doi: 10.1109/EIECS63941.2024.10800362.
- [19] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in *Data Mining and Big Data*, 2020, pp. 108–117. doi: http://dx.doi.org/10.1007/978-981-15-7205-0_10.
- [20] D. C. Maryann, S. Dalal, E. M. Onyema, and P. Kumar, "A hybrid machine learning model for breast cancer prediction," *Int. J. Model. Simul. Sci. Comput.*, vol. 14, no. 4, 2023, doi: <http://dx.doi.org/10.1142/S1793962323410234>.
- [21] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.
- [22] C. W. Cahyana and A. Nurlayli, "Analisis Performa Logistic Regression, Naïve Bayes, dan Random Forest sebagai Algoritma Pendeteksi Kanker Payudara," *Inser. Inf. Syst. Emerg. Technol. J.*, vol. 4, no. 1, pp. 51–64, 2023, [Online]. Available: <https://ejournal.undiksha.ac.id/index.php/insert/article/view/62362>
- [23] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, p. 345, 2020, doi: 10.1007/s42979-020-00365-y.
- [24] A. Kaur and S. Gupta, "Unveiling Precision in Breast Cancer Prediction with Random Forest and Decision Trees," in *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, 2024, pp. 1232–1236. doi: 10.1109/ICOSEC61587.2024.10722493.
- [25] S. Liu, J. Shen, and J. Zhang, "An integrated model combining BERT and tree-augmented naive Bayes for analyzing risk factors of construction accident," *Kybernetes*, vol. 54, no. 10, pp. 5651–5675, May 2024, doi: 10.1108/K-08-2023-1605.