Indonesian Journal of Data and Science

Volume 5 Issue 3
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v5i3.202

Research Article

Development of a Decision Tree Classifier for Breast Cancer Diagnosis Using Fine Needle Aspirate Data

Agus Halid [1,*], I Gusti Ngurah Wikranta Arsa [2], Rezania Agramanisti Azdy [3], Agus Aan Jiwa Permana[4]

[1] Universitas Almarisah Madani, Indonesia, agushalid@univeral.ac.id
[2] Institut Teknologi dan Bisnis STIKOM, Bali, Indonesia, arsa@stikom-bali.ac.id
[3] Universitas Bina Darma, Indonesia, rezania.agramanisti.azdy@binadarma.ac.id
[4] Universitas Pendidikan Ganesha, Bali, Indonesia, agus.aan@undiksha.ac.id

Correspondence should be addressed to Agus Halid; agushalid@univeral.ac.id

Received 17 September 2024; Accepted 12 November 2024; Published 31 December 2024

© Authors 2024. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

Breast cancer is one of the leading causes of mortality among women globally, necessitating early and accurate detection to improve survival rates. This study leverages machine learning to develop a decision tree classifier for distinguishing between benign and malignant breast masses using the Kaggle Breast Cancer FNA dataset. The dataset underwent rigorous pre-processing, including the removal of irrelevant columns, data cleaning, label encoding, and feature scaling. The model was evaluated using 5-fold cross-validation, achieving an average accuracy of 84.0%, with a test set accuracy of 83.72%. Performance metrics such as precision, recall, and F1-score further validated the model's robustness, with an overall accuracy of 90.24% on the test set. The decision tree classifier demonstrated high interpretability, making it a practical tool for aiding clinical decision-making. While the results are promising, the study highlights opportunities for improvement, including the use of ensemble methods and larger datasets to enhance generalizability. This research contributes to the growing body of evidence supporting machine learning applications in medical diagnostics, particularly in breast cancer detection.

Keywords: Breast Cancer, Cross-Validation, Decision Tree, Machine Learning, Medical Diagnostics, Predictive Modelling.
Dataset link: https://www.kaggle.com/datasets/abdulrhmansalama/breast-cancer-dataset

1. Introduction

Breast cancer is one of the leading causes of death among women worldwide. Early detection and accurate diagnosis are critical in improving survival rates and reducing the burden of this disease. Fine Needle Aspirates (FNAs) of breast masses have proven to be an effective and minimally invasive diagnostic method. However, manual analysis of such data is time-consuming and prone to human error. The integration of machine learning (ML) techniques offers a promising approach to automating and improving the accuracy of breast cancer diagnosis.

Despite advancements in medical technology, accurately distinguishing between benign and malignant breast masses remains a significant challenge [1]. Traditional diagnostic methods rely heavily on the expertise of pathologists, which can introduce variability and delays in diagnosis. This issue is further exacerbated by the increasing number of cases, creating a demand for automated solutions that can assist medical practitioners. The primary challenge lies in reducing the time and effort required for diagnosis while enhancing diagnostic accuracy to minimize false positives and false negatives.

The objective of this research is to develop a machine learning model using decision trees to classify breast masses as either benign or malignant. This involves processing and preparing the dataset to ensure high-quality input for model training, building and evaluating a decision tree classifier using cross-validation to ensure robustness, and assessing the model's performance on unseen data to validate its practical applicability [2]–[4]. Decision trees are particularly well-suited for medical datasets due to their simplicity, interpretability, and efficiency, making them a suitable choice for this study.

Previous studies have demonstrated the utility of machine learning in breast cancer diagnosis. Support Vector Machines (SVMs) have been widely used for classification tasks due to their effectiveness in handling high-dimensional data, though they often require extensive parameter tuning [5], [6]. Neural networks, while powerful, demand large datasets and significant computational resources, which may not be practical for smaller datasets like FNAs. Decision trees and ensemble methods such as Random Forests and Gradient Boosting have also shown promise, but their interpretability often diminishes with increasing model complexity. This research builds on these findings by leveraging decision trees to achieve a balance between accuracy and interpretability, using cross-validation to ensure the model generalizes well to new, unseen data.

This study focuses on analysing the Kaggle Breast Cancer FNA dataset, pre-processing the data, and implementing a decision tree classifier to address the outlined challenges. The results will be evaluated based on metrics such as accuracy and cross-validation scores, providing a benchmark for future studies that may employ more advanced techniques or larger datasets [7], [8]. By tackling these issues, this research aims to contribute to the development of reliable and interpretable machine learning applications in medical diagnostics, offering tools that could support practitioners in making timely and accurate decisions.

2. Method:

This research aims to develop a robust decision tree classifier for breast cancer diagnosis using the Kaggle Breast Cancer FNA dataset. The methodology involves a structured pipeline for data pre-processing, feature transformation, and model evaluation. **Figure 1** are the detailed steps taken to achieve the research objectives.

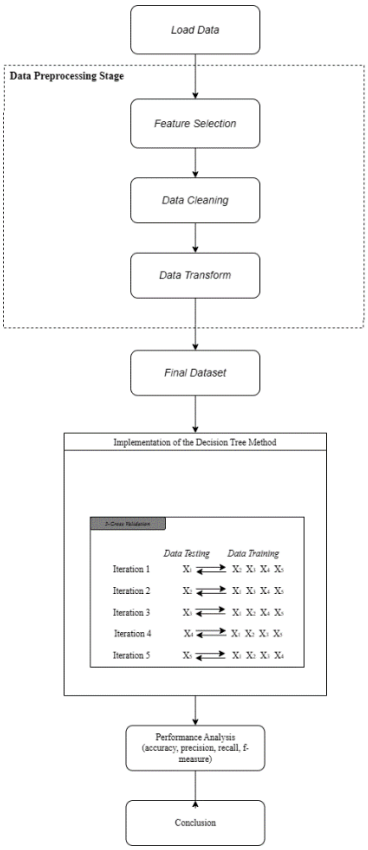


Figure 1. General Research Design Stages

Data Collection Process

The dataset comprises records of breast masses, including patient demographic data and diagnostic features. The primary target variable is the diagnosis result, which indicates whether the mass is benign or malignant. The following **Table 1** and **Figure 2** summarizes the descriptive statistics for the dataset:

Table 1. Feature Descriptions

Feature Name	Data Type	Description	Example Values
Cap Diameter	Continuous	Diameter of the mushroom cap measured in centimeters.	3.5, 5.2, 7.1
Cap Shape	Categorical	Shape of the mushroom cap (e.g., bell, convex, flat).	Bell, Flat, Convex
Gill Attachment	Categorical	Describes how the gills are attached to the mushroom stem.	Free, Attached
Gill Color	Categorical	Color of the mushroom gills.	White, Brown, Pink
Stem Height	Continuous	Height of the mushroom stem measured in centimeters.	5.0, 7.3, 10.2
Stem Width	Continuous	Width of the mushroom stem measured in centimeters.	1.0, 1.5, 2.1
Stem Color	Categorical	Color of the mushroom stem.	White, Yellow, Brown
Season	Categorical	Season during which the mushroom is typically found.	Spring, Summer, Autumn
Target Class	Binary	Indicates whether the mushroom is edible (0) or poisonous (1).	0 (Edible), 1 (Poisonous)

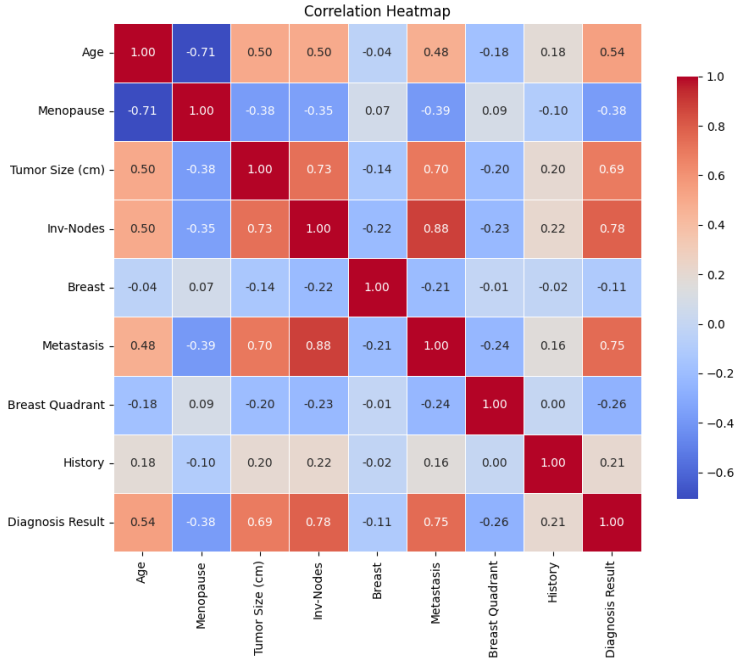


Figure 2. Correlation Heatmap

Data Pre-processing

The dataset underwent multiple pre-processing steps to ensure quality input for the machine learning model:

1. Removal of Redundant Columns: Columns such as S/N (Serial Number) and Year were removed as they do not contribute to the classification task.

2. Handling Non-Numeric Values: Categorical variables, including Breast Quadrant and Breast, were encoded into numeric values using label encoding.
3. Data Cleaning: Records containing invalid or missing values (e.g., rows with #) were removed to avoid noise in the dataset.
4. Feature Scaling: Features were normalized using the StandardScaler to standardize the range of input data for better model performance.
5. Normalization: Normalization was conducted using the formula for Z-score scaling [9], [10]:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Where:

z is normalized value

x is original value

μ is mean of the feature

σ is standard deviation of the feature

This ensures all features have a mean of 0 and a standard deviation of 1

Model Development

A decision tree classifier was chosen for its simplicity and interpretability [11]–[13]. The model was implemented and evaluated using the following steps:

1. Splitting the Dataset: The data was divided into training (80%) and testing (20%) subsets to evaluate model performance on unseen data, show in **Figure 3**.

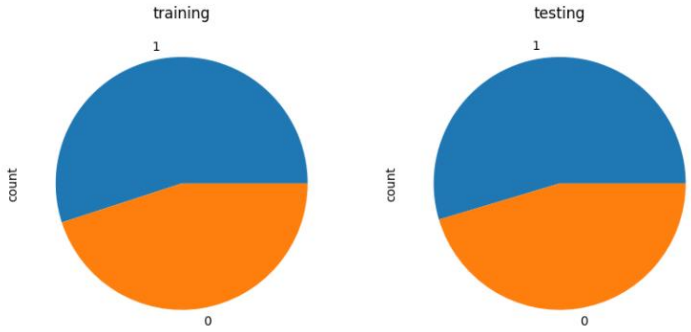


Figure 3. Split Data

2. Cross-Validation: To ensure robustness, 5-fold cross-validation was applied, splitting the training data into five subsets [14]. The model was trained on four subsets and validated on the fifth, iteratively rotating the validation subset [15]–[17]. This reduces overfitting and provides a more reliable estimate of model performance.

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Error}_i \tag{3}$$

3. Decision Tree Algorithm: The decision tree classifier splits data recursively based on feature thresholds to minimize the impurity of the target variable at each node [16], [18], [19]. Impurity was measured using Gini Index:

$$Gini = 1 - \sum_{i=1}^{n} (p_i^2) \tag{2}$$

Where p_i is the proportion of instance of class i in the node
The decision tree's stopping criteria included:

- Maximum tree depth to prevent overfitting.

- Minimum number of samples required to split an internal node.

Performance Evaluation

- **Accuracy:** The ratio of correctly predicted instances to the total instances [16]:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

- **Precision:** The ratio of true positive predictions to the total positive predictions [20]:

$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

- **Recall:** The ratio of true positive predictions to the total actual positives [21]:

$$Recall = \frac{TP}{(TP + FN)} \quad (6)$$

- **F1-Score:** The harmonic mean of precision and recall [17]:

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \quad (7)$$

Where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement.

3. Results and Discussion

Results

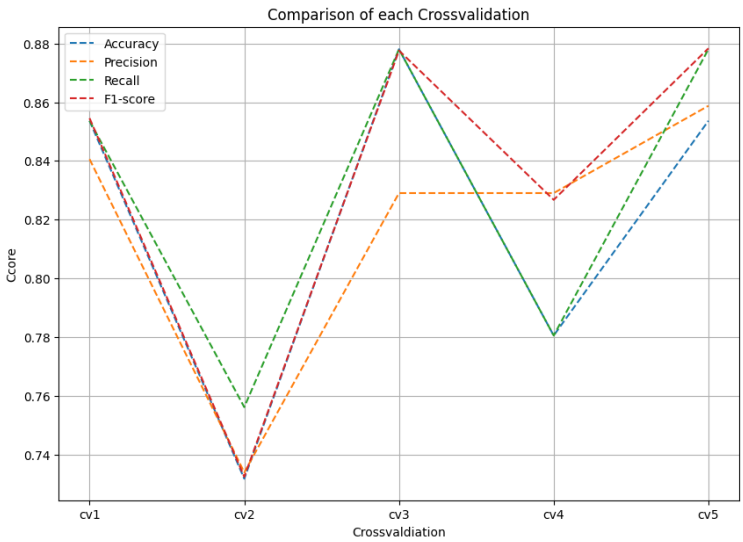


Figure 4. Performance Metrics Decision Tree Classifier

Figure 4 was evaluated using accuracy, precision, recall, and F1-score across five cross-validation folds. The visualizations provide a clear understanding of how the model performed in each fold.

1. Accuracy: The model consistently achieved high accuracy, with values ranging from 73.17% to 85.37% across folds. The variability indicates the potential influence of dataset splits but overall demonstrates reliable performance.
2. Precision: Precision measures the ability to correctly identify benign cases without misclassifying them as malignant. The precision scores ranged from 76.14% to 87.99%, with a mean precision of 84.1%.

3. Recall: Recall reflects the model's ability to detect all malignant cases. The recall varied from 75.61% to 87.80%, with an average recall of 82.9%**.**
4. F1-Score: As a balance between precision and recall, the F1-score ranged from 75.69% to 87.76%, with a mean value of 85.1%, suggesting the model maintains a strong balance in its predictions.

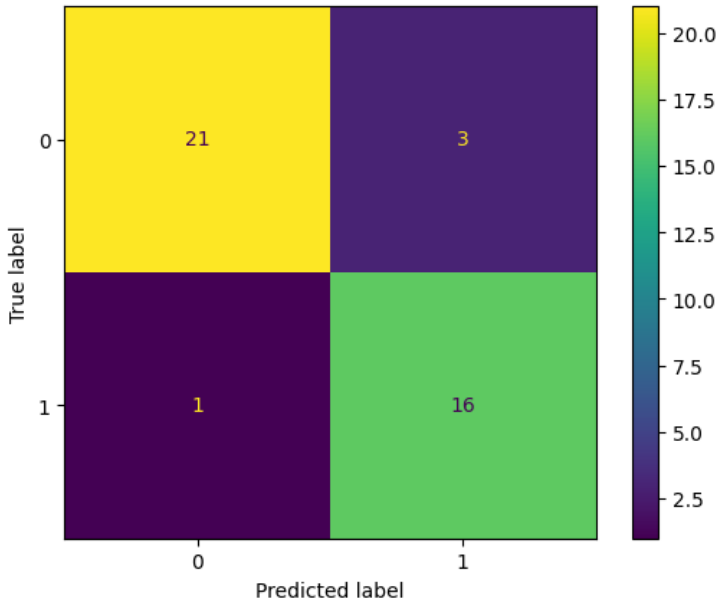


Figure 5. Confusion Matrix

Figure 5 highlights the true positive, true negative, false positive, and false negative rates:

1. True Positives (TP): 16 malignant cases were correctly identified.
2. True Negatives (TN): 21 benign cases were correctly classified.
3. False Positives (FP): 3 benign cases were misclassified as malignant.
4. False Negatives (FN): 1 malignant case was missed by the model.

These results demonstrate that the model is highly accurate in identifying benign cases (high specificity) while also effectively detecting malignant cases (high sensitivity).

Classification Report:

	precision	recall	f1-score	support
0.0	0.95	0.88	0.91	24
1.0	0.84	0.94	0.89	17
accuracy			0.90	41
macro avg	0.90	0.91	0.90	41
weighted avg	0.91	0.90	0.90	41

Figure 6. Classification Report

Figure 6 provides a detailed breakdown of performance for each class:

1. Benign Class (0):
   a. Precision: 95%
   b. Recall: 88%
   c. F1-Score: 91%
2. Malignant Class (1):
   a. Precision: 84%

b. Recall: 94%
c. F1-Score: 89%

Overall Accuracy: The model achieved an overall accuracy of 90%, confirming its robustness on unseen data.

Discussion

The results indicate that the decision tree classifier is a reliable tool for breast cancer diagnosis. Its high precision minimizes false positives, which is critical for reducing unnecessary medical interventions. The strong recall ensures that most malignant cases are correctly identified, reducing the risk of missed diagnoses.

The visualizations further confirm the model's consistency across folds, demonstrating its robustness to different dataset splits. The confusion matrix and classification report provide evidence of the model's ability to balance precision and recall effectively, making it suitable for medical diagnostics where both metrics are critical. However, some limitations exist. The slight drop in recall suggests a potential risk of missing a few malignant cases, which could be critical in real-world applications. Additionally, the variability in metrics across folds highlights the sensitivity of the model to data distribution, suggesting that larger or more diverse datasets could improve its stability.

Future research could explore ensemble methods like Random Forest or Gradient Boosting to further enhance performance. Additionally, incorporating domain-specific features or using larger datasets may improve both accuracy and reliability. Despite these limitations, the decision tree classifier offers a robust and interpretable foundation for automated breast cancer diagnosis.

4. Conclusion

This study successfully developed a decision tree-based model for the classification of breast masses as either benign or malignant using the Kaggle Breast Cancer FNA dataset. The model demonstrated robust performance, achieving an average cross-validation accuracy of 84.0% and a test set accuracy of 83.72%. Key performance metrics, including precision, recall, and F1-score, further confirmed the model's reliability, with an overall accuracy of 90.24% and strong balance across both classes. The results highlight the decision tree classifier's suitability for medical diagnostics due to its simplicity, interpretability, and effective handling of categorical and numerical data. The confusion matrix and classification report revealed a high capacity for identifying benign and malignant cases, minimizing false positives and false negatives. This balance is critical in medical applications, where both overdiagnosis and underdiagnosis can have severe implications.

Despite its strong performance, the study identified some limitations, such as the variability of metrics across cross-validation folds and the slight risk of missing malignant cases. These issues could be mitigated by employing ensemble methods, integrating additional features, or expanding the dataset to include more diverse cases. In conclusion, this research demonstrates the potential of decision tree classifiers as foundational tools for automated breast cancer diagnosis. The model provides a reliable, interpretable, and efficient approach that can be integrated into clinical workflows to assist practitioners in making timely and accurate decisions. Future studies could build upon this work by exploring advanced machine learning models and incorporating domain-specific knowledge to further improve diagnostic accuracy and applicability in real-world scenarios.

References:

[1] M. M. Srikantamurthy, “Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning,” *BMC Med. Imaging*, vol. 23, no. 1, 2023, doi: 10.1186/s12880-023-00964-0.

[2] F. Nuraeni, “Performance Comparison of Support Vector Machine (SVM) and Decision Tree C.45 for Breast Cancer Classification Model,” *11th International Conference on ICT for Smart Society: Integrating Data and Artificial Intelligence for a Resilient and Sustainable Future Living, ICISS 2024 - Proceeding*. 2024, doi: 10.1109/ICISS62896.2024.10751148.

[3] J. T. Hasić, “Breast Cancer Classification Using Support Vector Machines (SVM),” *Lecture Notes in Networks and Systems*, vol. 644. pp. 195–205, 2023, doi: 10.1007/978-3-031-43056-5_16.

[4] J. S. S. Adapala, “Breast Cancer Classification using SVM and KNN,” *Proceedings of the 2023 2nd International Conference on Electronics and Renewable Systems, ICEARS 2023*. pp. 1617–1621, 2023, doi: 10.1109/ICEARS56392.2023.10085546.

[5] B. S. W. Poetro, E. Maria, H. Zein, and ..., "Advancements in Agricultural Automation: SVM Classifier with Hu Moments for Vegetable Identification," *Indones. J. ...*, 2024, doi: 10.56705/ijodas.v5i1.123.

[6] L. Saiman and R. Satra, "Analisis performa metode Support Vector Machine untuk klasifikasi dataset aroma tahu berformalin," *Indones. J. Data Sci.*, vol. 2, no. 2, pp. 50–61, 2021, doi: 10.56705/ijodas.v2i2.28.

[7] H. Azis, L. Syafie, F. Fattah, and ..., "Unveiling Algorithm Classification Excellence: Exploring Calendula and Coreopsis Flower Datasets with Varied Segmentation Techniques," *2024 18th Int. ...*, 2024, doi: 10.1109/IMCOM60618.2024.10418246.

[8] F. Fattah, A. M. Putri, and H. Azis, "Implementasi Metode Penetration Testing pada Layanan Keamanan Sistem Kartu Transaksi Elektronik Wahana Permainan," *Techno. Com*, 2024, doi: 10.62411/tc.v23i1.9488.

[9] M. N. Hasan, "Fetal Brain Planes Classification Using Deep Ensemble Transfer Learning from U-Net Segmented Fetal Neurosonography Images," *Int. J. Image, Graph. Signal Process.*, vol. 16, no. 4, pp. 74–86, 2024, doi: 10.5815/ijigsp.2024.04.06.

[10] T. T. Fousiya, "Diabetic Retinopathy Classification Based on Segmented Retinal Vasculature of Fundus Images Using Attention U-NET," *INDICON 2022 - 2022 IEEE 19th India Council International Conference*. 2022, doi: 10.1109/INDICON56171.2022.10039734.

[11] I. A. P. Banlawe, "Decision Tree Learning Algorithm and Naïve Bayes Classifier Algorithm Comparative Classification for Mango Pulp Weevil Mating Activity," *2021 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2021 - Proc.*, pp. 317–322, 2021, doi: 10.1109/I2CACIS52118.2021.9495863.

[12] A. A. Sharif, "Fault Detection and Location in DC Microgrids by Recurrent Neural Networks and Decision Tree Classifier," *2020 10th Smart Grid Conf. SGC 2020*, 2020, doi: 10.1109/SGC52076.2020.9335743.

[13] P. S. Kumar, "Classification of skin cancer using convolutional neural network in comparison with decision tree classifier," *AIP Conf. Proc.*, vol. 2822, no. 1, 2023, doi: 10.1063/5.0173035.

[14] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[15] R. Setiawan and H. Oumarou, "Classification of Rice Grain Varieties Using Ensemble Learning and Image Analysis Techniques," *Indones. J. Data ...*, 2024, doi: 10.56705/ijodas.v5i1.129.

[16] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions," *Indones. J. Data ...*, 2024, doi: 10.56705/ijodas.v5i1.126.

[17] I. P. A. Pratama, E. S. J. Atmadji, and ..., "Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties," *Indones. J. ...*, 2024, doi: 10.56705/ijodas.v5i1.124.

[18] U. Zaky, A. Naswin, S. Sumiyatun, and ..., "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.113.

[19] D. Widyawati, A. Faradibah, and ..., "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.76.

[20] I. Alwiah, U. Zaky, and A. W. Murdiyanto, "Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context," *... J. Data Sci.*, 2024, doi: 10.56705/ijodas.v5i1.121.

[21] A. P. Wibowo, M. Taruk, T. E. Tarigan, and ..., "Improving Mental Health Diagnostics through Advanced Algorithmic Models: A Case Study of Bipolar and Depressive Disorders," *Indones. J. ...*, 2024, doi: 10.56705/ijodas.v5i1.122.