Indonesian Journal of Data and Science

Volume 4 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v4i2.77

Research Article

Performance Comparison Analysis of Classifiers on Binary Classification Dataset

Rahmat Fuaji Sya [a], Rahmadani [b,*]
[1] Universitas Pancasakti, Makassar, Indonesia, -
[2] University Muslim Indonesia, Makassar, Indonesia, 13020200237@umi.ac.id

Correspondence should be addressed to Rahmadani; 13020200237@umi.ac.id

Received 10 June 2023; Accepted 18 June 2023; Published 31 July 2023

Abstract:

In this study, we compared the performance of Random Forest Classifier and Decision Tree using classification evaluation methods. The results showed that Random Forest Classifier had a higher overall accuracy rate but also produced more outliers. On the other hand, Decision Tree demonstrated consistency in classification with fewer outliers. These findings provide insights into the trade-off between accuracy and consistency when selecting the appropriate classification method. Furthermore, further research is needed to understand the impact of outliers on classification performance and to take appropriate steps in addressing them.

Keywords: Random Forest Classifier, Decision Tree, Naive Bayes, Support Vector Machine, Binary Dataset, Classification, Performance Comparison.
Dataset link: Cat and Dog

1. Introduction:

In the increasingly advanced digital era, the ability to classify and predict data has become crucial [1]–[3]. One of the methods used for classification is through the use of classification models. These models learn patterns within the data and can be used to predict the classes or labels of unknown data [4]–[7]. Machine learning-based classification models have proven to be successful in solving this problem. In this study, our aim is to compare the performance of four popular classification models: Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes. Specifically, we will compare the performance of these models in classifying binary data [5], [6] consisting of cat and dog images.

The binary classification between cats and dogs has wide-ranging applications, including image recognition, animal research, and the pet industry. However, determining the most effective classification model for this task can be a challenge. Therefore, this study aims to compare the performance of the Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes classification models in classifying binary images of cats and dogs [2], [8], [9].

The main objective of this research is to compare the performance of the Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes classification models on a binary dataset of cats and dogs. We aim to evaluate and compare the accuracy, precision, recall, and F1 score [10], [11] of these four models to determine which model performs better in classifying cat and dog images.

In this study, we have several research questions that will guide our investigation. First, we want to evaluate the performance of the Random Forest Classifier in classifying binary cat and dog images. We will measure the accuracy, precision, recall, and F1 score of this model to assess its ability to correctly predict the image classes [12], [13]. Next, we will evaluate the performance of the Decision Tree classification model in the same task. Decision Tree is a popular classification method, and we are interested in how well it can classify binary cat and dog images. We will use the same metrics, namely accuracy, precision, recall, and F1 score, to compare its performance with Random Forest.

Considering other classification methods, namely Support Vector Machine (SVM) and Naïve Bayes. SVM is a highly effective method for separating data classes by constructing an optimal hyperplane between them. We want to see how well SVM [14], [15]can classify binary cat and dog images with high levels of accuracy, precision, recall, and F1 score. Additionally, we will also evaluate the performance of Naïve Bayes, which is a probabilistic classification method based on the naive assumption of feature independence. We want to determine if Naïve Bayes can provide competitive results in the task of classifying binary cat and dog images [16], [17]. By comparing the performance of these four classification models, we hope to gain a better understanding of their abilities in classifying binary cat and dog images and make valuable contributions to the development of classification methods for similar tasks.

Furthermore, we aim to determine if there are significant differences in the performance metrics among the Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes classification models when classifying binary cat and dog images. We will use appropriate statistical tests to analyze the data and determine if these differences are statistically significant or not [18]. By answering these research questions, we will gain a better understanding of the performance of these two classification models in this particular binary classification task [13], [19]. The results of this research will provide valuable insights for practitioners and researchers in selecting the appropriate classification model for similar tasks, considering accuracy, precision, recall, and F1 score [3], [10], [20].

This study will focus on comparing the performance of the Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes classification models in classifying a binary dataset consisting of cat and dog images. The dataset used in this study will comprise properly categorized cat and dog images [17], [21]. However, it is important to note that the evaluation and performance comparison may vary depending on the dataset used. This research will be limited to measuring accuracy, precision, recall, and F1 score as performance metrics.

2. Method:

Our research design consists of 5 well-organized main stages, with the aspects illustrated in Figure 1.

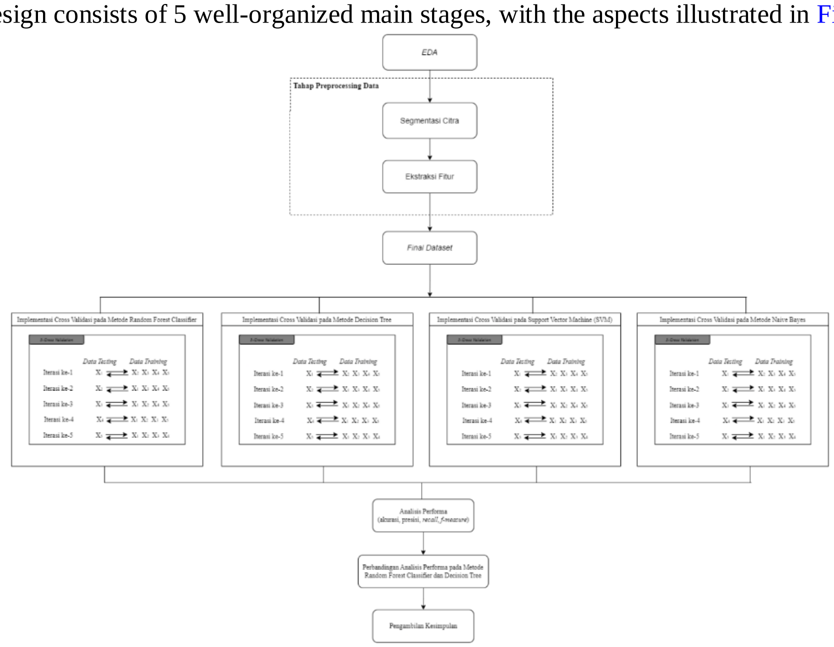

Figure 1. General Stages of Research Design

Exploratory Data Analysis (EDA)

The first step in this research is to perform Exploratory Data Analysis (EDA) to gain a comprehensive understanding of the dataset [22], [23]. Analyzing descriptive statistics, visualizations, and data distributions to gain insights into the characteristics of the dataset. Table 1 presents general information about the dataset used in this study.

Table 1. Dataset Information

Data set	Number of cases	Number of attribute	Number of classes	Attribute characteristics	Missing values
Cat-and-dog	10.032	7	2	Numeric	No

Image Segmentation with Canny

After EDA, we will perform preprocessing on the dataset. One of the preprocessing steps we will take is image segmentation using the Canny method. This method will help separate important objects from the background in the images within the dataset [24], [25]. Figure 2 illustrates the visualization results with image segmentation using Canny.

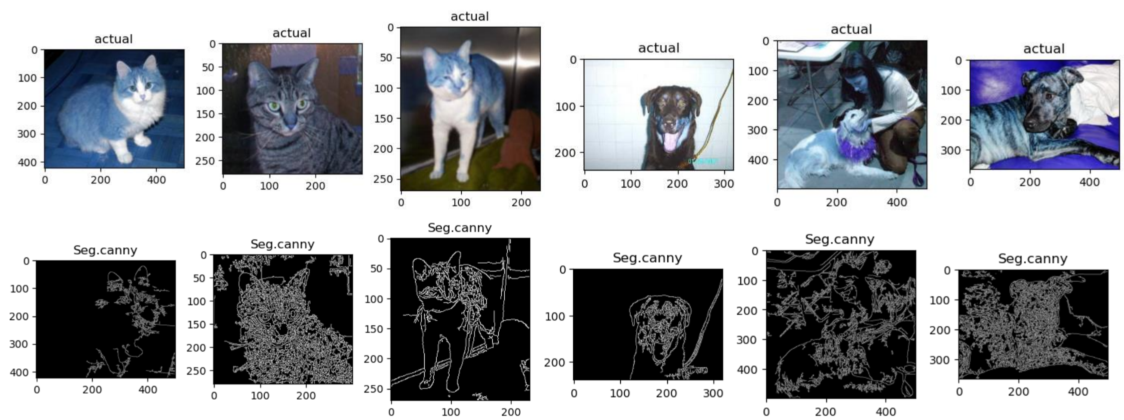

Figure 2. Visualisasi Segmentasi Canny

Feature Extraction with Humoment

Perform feature extraction using the Humoment method. This method will generate feature vectors that represent each image in the dataset [26], [27]. These feature vectors will be used as inputs in the classification model. Figure 3 illustrates the visualization of Humoment feature extraction using scatter plots and heatmaps.

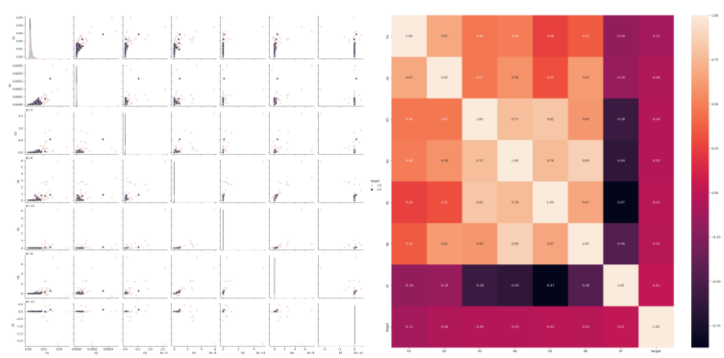

Figure 3. Visualization of Scatter Plot and Heatmap for Feature Extraction using Humoment

Random Forest Classifier

The Random Forest Classifier is a classification method that uses an ensemble of decision trees to make predictions [28], [29]. This method combines predictions from multiple decision trees to generate the final result. The main advantage of the Random Forest is its ability to overcome overfitting and produce stable predictions, as depicted in Figure 4, illustrating the concept of the Random Forest Classifier algorithm.

$$\text{Dataset } X = \{(x\_1, y\_1), (x\_2, y\_2), \ldots, (x\_n, y\_n)\}$$

$$\text{Prediction\_t} = H\_t(x), t = 1, 2, \ldots, T$$

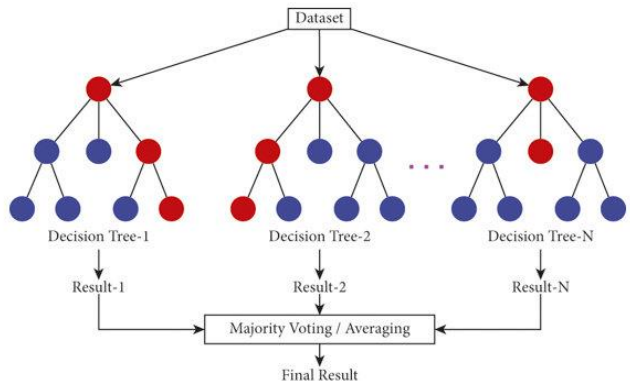

Figure 4. Random Forest Classifier Algorithm

Decision Tree

Decision Tree is a classification method that takes the form of a tree structure with nodes representing decisions or predictions [30]. At each node, the algorithm divides the data based on the most informative input variable, as depicted in Figure 5, illustrating the concept of the Decision Tree algorithm.

$$\text{Dataset } X = \{(x\_1, y\_1), (x\_2, y\_2), \ldots, (x\_n, y\_n)\}$$

$$\text{Prediction} = H(x)$$

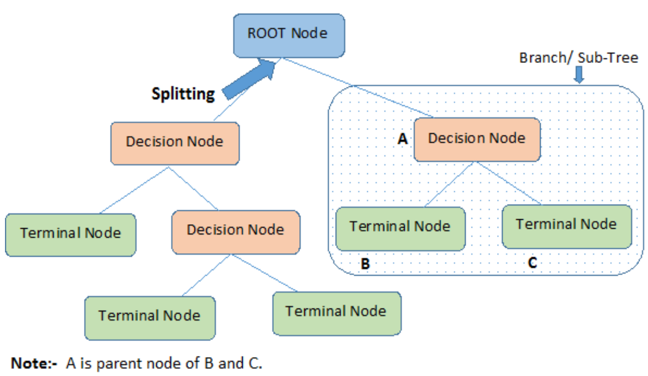

Figure 5. Decision Tree Algorithm

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning method used for classification and regression. SVM aims to find the best hyperplane that maximizes the margin between two different classes of data [10], [24]. This hyperplane is used to separate the two classes of data as far as possible and create an optimal decision boundary [9], [31], as depicted in Figure 6.

$$\text{Dataset} = (x_i, y_i)$$
$$y_i(w \cdot x_i + b) \geq 1$$

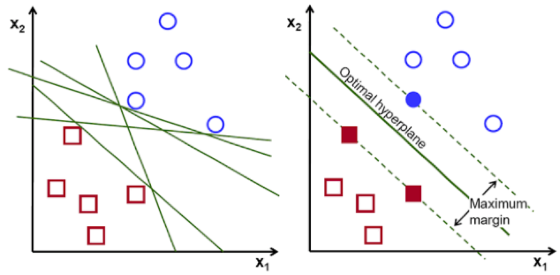

Figure 6. Algoritma Support Vector Machine

Naïve Bayes

Naïve Bayes is a probabilistic classification method based on Bayes' theorem, assuming that all features in the data are conditionally independent of each other given the class [10], [32], as depicted in Figure 7.

$$\text{Dataset } x_i = (x_{1i}, x_{2i}, \ldots, x_{mi})$$
$$p(c|x) = \frac{(p(c) * p(x_1|c) * p(x_2|c) * \ldots * p(x_m|c))}{p(x)}$$

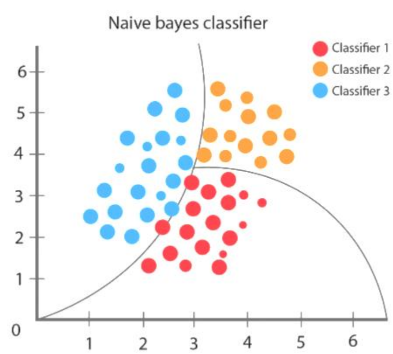

Figure 7. Naïve Bayes Algorithm

Perbandingan Analisis Performa

Comparing performance analysis between Random Forest Classifier, Decision Tree, Support Vector Machine, and Naïve Bayes [33]. We will compare the evaluation matrices obtained from the four models to see the differences in their classification performance.

The evaluation matrices used are Accuracy a measure that calculates the percentage of correct predictions out of all predictions made. Precision measures the percentage of correct positive predictions out of the total positive predictions made. Recall [28] measures the percentage of correct positive predictions out of the total amount of data that is actually positive. F-measure is a metric that combines precision and recall to provide a comprehensive value of the model's predictive quality with the evaluation metric formulation as below [34].

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

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

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

$$\text{F-measure} = \frac{2\ x\ (Precision\ \ x\ Recall)}{(Precision\ + Recall)}$$

Using these evaluation metrics, we will be able to evaluate the ability and performance of Random Forest Classifier, Decision Tree, Support Vector Machine, and Naïve Bayes in correctly classifying the data [35].

Decision Making

Based on the performance analysis and comparison between the two classification models, we will make a decision regarding the most effective and suitable model for this classification task. This decision will be based on the evaluation metrics and research objectives that have been established beforehand.

3. Results and Discussion:

Performance of four classifications, namely Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes, were evaluated based on several metrics. The metrics used to evaluate the classification performance include balanced accuracy, accuracy, weighted precision, weighted recall, and weighted F1 score. The results obtained from the evaluation are presented in Table 2.

Table 2. Performance Evaluation Results of Random Forest Classifier, Decision Tree, Support Vector Machine and Naïve Bayes.

∑ Rata-rata	Random Forest Classifier	Decision Tree	Support Vector Machine	Naïve Bayes
Balanced accuracy	0.54	0.54	0.55	0.51
Accuracy	0.55	0.54	0.55	0.51
Precision weighted	0.54	0.55	0.55	0.6
Recall weighted	0.54	0.54	0.55	0.51
F1 weighted	0.54	0.52	0.54	0.38

After analyzing the results, it can be observed that the classifications demonstrate comparable performance in terms of balanced accuracy. Both the Random Forest Classifier and Decision Tree achieved a balanced accuracy of 0.54, while the Support Vector Machine (SVM) showed a slightly higher balanced accuracy of 0.55. Naïve Bayes obtained the lowest balanced accuracy of 0.51. In terms of overall accuracy, the classifications also exhibited similar results. The accuracy of the Random Forest Classifier and Decision Tree were 0.55 and 0.54 respectively. SVM and Naïve Bayes achieved accuracies of 0.55 and 0.51 respectively.

Weighted precision, which considers the weighted average of precision across all classes, yielded varying results among the classifications. The Random Forest Classifier obtained a weighted precision score of 0.54, while the Decision Tree and SVM achieved scores of 0.55. There was a significant difference, with Naïve Bayes showing the highest weighted precision score of 0.6, indicating its ability in terms of precision. Similarly, the classifications exhibited similar performance in terms of weighted recall. The Random Forest Classifier, Decision Tree, and SVM achieved weighted recall scores of 0.54, while Naïve Bayes obtained a score of 0.51.

Weighted F1 scores, which consider both precision and recall, showed variation among the classifications. The Random Forest Classifier obtained a weighted F1 score of 0.54, the Decision Tree obtained a score of 0.52, SVM obtained a score of 0.54, and Naïve Bayes had the lowest weighted F1 score of 0.38. The lower weighted F1 score for Naïve Bayes indicates that this classification may not perform well in terms of overall precision and recall.

Overall, the results indicate that the classifications demonstrate comparable performance in terms of balanced accuracy and accuracy. However, there is variation in the scores of weighted precision, recall, and F1. Naïve Bayes shows higher precision but lower weighted F1 score compared to the other classifications. Further analysis and investigation are recommended to gain insights into the factors influencing the performance of each classification on the specific dataset.

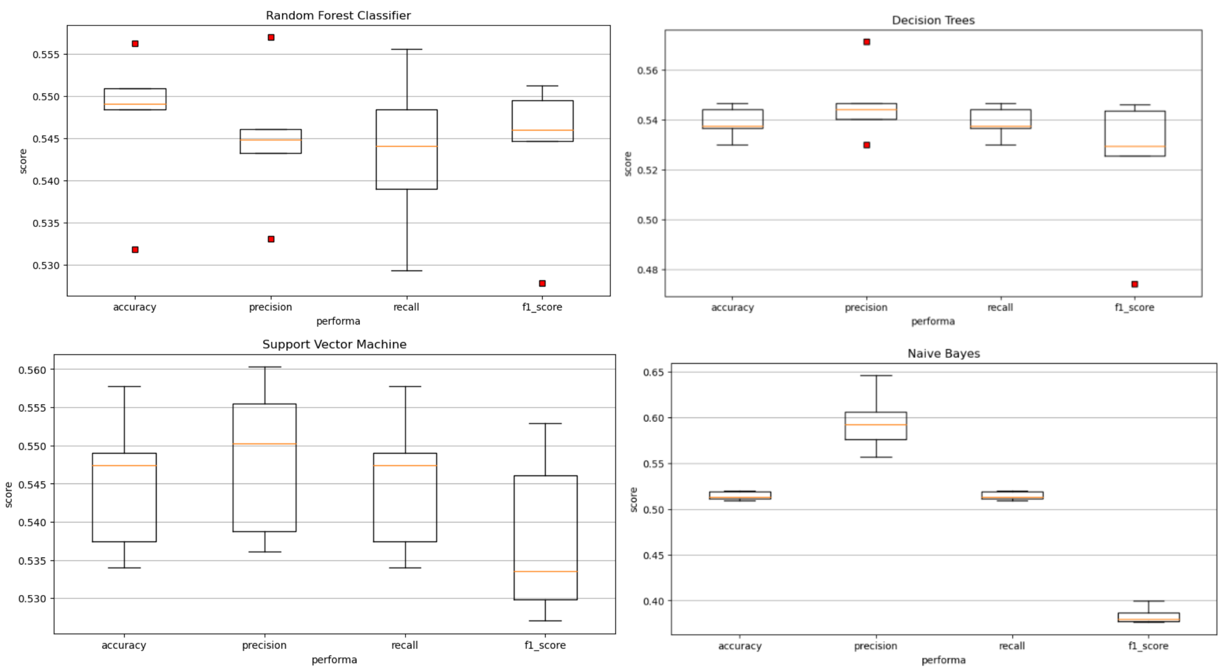

Figure 8. Comparison of Boxplot Visualization Results.

The analysis results using the boxplot method, as shown in Figure 8, indicate differences in data distribution among the classification methods: Random Forest Classifier, Decision Tree, Support Vector Machine (SVM), and Naïve Bayes. The Random Forest Classifier method exhibits a higher number of outliers, indicating higher sensitivity to unusual data, while the Decision Tree method has fewer outliers, suggesting better stability. The SVM and Naïve Bayes methods do not show any outliers, indicating consistency in data grouping. This analysis provides insights into the characteristics of each method in classifying data and can be considered when selecting an appropriate method for the desired application.

4. Conclusion

Based on the results obtained, the Random Forest Classifier has more outliers compared to the Decision Tree. The presence of these outliers can be caused by several factors, such as imbalanced datasets, noise in the data, or the influence of significant features. On the other hand, the Decision Tree demonstrates consistency in classification with a lower number of outliers. This difference indicates that the Decision Tree may be more effective in handling structured data. However, further research is needed to understand the characteristics and impact of outliers on the classification performance in both methods.

References:

[1] H. Tvedten, “Classification and laboratory evaluation of anemia,” *Schalm’s Vet. Hematol.*, 2022, doi: 10.1002/9781119500537.ch25.

[2] Nurul A’ayunnisa, Y. Salim, and H. Azis, “Analisis Performa Metode Gaussian Naïve Bayes untuk Klasifikasi Citra Tulisan Tangan Karakter Arab,” *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: 10.56705/ijodas.v3i3.54.

[3] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, “Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung,” *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[4] A. Tangkelayuk, “The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree,” *JATISI (Jurnal Tek. Inform. Dan Sist. …*, 2022, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/2048

[5] R. Ridho and H. Hendra, “Klasifikasi Diagnosis Penyakit Covid-19 Menggunakan Metode Decision Tree,” *JUST IT J. Sist. Informasi, Teknol. …*, 2022, [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it/article/view/13594

[6] A. Z. Zami, O. Nurdiawan, and ..., “Klasifikasi Kondisi Gizi Bayi Bawah Lima Tahun Pada Posyandu Melati Dengan Menggunakan Algoritma Decision Tree,” *J. Sist. …*, 2022, [Online]. Available: http://www.ejurnal.stmik-budidarma.ac.id/index.php/JSON/article/view/3892

[7] R. Fitriawanti, I. Cholissodin, and R. K. Dewi, “Klasifikasi dan Rekomendasi Jurusan Kuliah Bagi Pelajar SMA Menggunakan Algoritme Naïve Bayes -WP,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 11, pp. 4914–4922, 2018, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3129

[8] K. Gupta, N. Jiwani, and N. Afreen, “A Combined Approach of Sentimental Analysis Using Machine Learning Techniques.,” *Rev. d’Intelligence Artif.*, 2023, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=0992499X&AN=163153519&h=oW2E1RbY5Z%2FNmydQmFIt2ykqVNxazgS8gibi33%2Fq9zzG9ZcI3Zi2WHuhIBHf2LActasi%2BlBMHmR6TsmZorhDIQ%3D%3D&crl=c

[9] 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.

[10] F. T. Admojo and S. R. Jabir, “Analisis performa metode Naïve Bayesh Classifier pada Electronic Nose dalam identifikasi formalin pada tahu,” *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 1–16, 2023, doi: 10.56705/ijodas.v4i1.67.

[11] A. M. Argina, “Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes,” *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[12] …, S. Park, J. Kim, and J. Canny, “An Embedding-Dynamic Approach to Self-supervised Learning,” *Proc. …*, 2023, [Online]. Available: https://openaccess.thecvf.com/content/WACV2023/html/Moon_An_Embedding-Dynamic_Approach_to_Self-Supervised_Learning_WACV_2023_paper.html

[13] Ericha Apriliyani and Y. Salim, “Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset,” *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 47–54, 2022, doi: 10.56705/ijodas.v3i2.45.

[14] U. Barman and R. D. Choudhury, “Soil texture classification using multi class support vector machine,” *Inf. Process. Agric.*, vol. 7, no. 2, pp. 318–332, 2020, doi: 10.1016/j.inpa.2019.08.001.

[15] F. Ramzan *et al.*, “A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer’s Disease Stages Using Resting-State fMRI and Residual Neural Networks,” *J. Med. Syst.*, vol. 44, no. 2, 2020, doi: 10.1007/s10916-019-1475-2.

[16] S. Sahar, “Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung,” *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.

[17] H. Fan, Y. Mao, H. Wang, Y. Yu, X. Wu, and Z. Zhang, “Performance comparation of MEA and EDA in electrochemically-mediated amine regeneration for CO2 capture,” *Sep. Purif. …*, 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1383586623001909

[18] F. Li, Y. Qiao, H. Zhao, S. He, and ..., “Analysis of SEMI EDA standards for semiconductor equipment data acquisition,” *… Conf. Comput. …*, 2022, doi: 10.1117/12.2659134.short.

[19] M. Jena, R. K. Behera, and S. Dehuri, “Hybrid decision tree for machine learning: A big data perspective,” *Adv. Mach. Learn. Big …*, 2022, doi: 10.1007/978-981-16-8930-7_9.

[20] R. S. Patil and S. R. Kolhe, “Supervised classifiers with TF-IDF features for sentiment analysis of Marathi tweets,” *Social Network Analysis and Mining*. Springer, 2022. doi: 10.1007/s13278-022-00877-w.

[21] Y. Wang, H. Chen, Y. Fan, W. Sun, and ..., “Usb: A unified semi-supervised learning benchmark for classification,” *Adv. …*, 2022, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/190dd6a5735822f05646dc27decff19b-Abstract-Datasets_and_Benchmarks.html

[22] B. Panda, C. R. Panigrahi, and B. Pati, “Exploratory data analysis and sentiment analysis of drug reviews,” *Computación y Sistemas*. scielo.org.mx, 2022. [Online]. Available: https://www.scielo.org.mx/scielo.php?pid=S1405-55462022000301191&script=sci_arttext

[23] H. Pitroda, “A Proposal of an Interactive Web Application Tool QuickViz: To Automate Exploratory Data Analysis,” *2022 IEEE 7th Int. Conf. …*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9824068/

[24] M. Krishna Monika, N. Arun Vignesh, C. Usha Kumari, M. N. V. S. S. Kumar, and E. Laxmi Lydia, “Skin cancer detection and classification using machine learning,” *Mater. Today Proc.*, vol. 33, no. xxxx, pp. 4266–4270, 2020, doi: 10.1016/j.matpr.2020.07.366.

[25] L. Song, “Canny optimisation of the dynamic image colour automatic segmentation algorithm,” *Int. J. Comput. Appl. …*, 2022, doi: 10.1504/IJCAT.2022.127820.

[26] D. V Kondusov and V. B. Kondusova, “Shape-Based Search Modules Permitting Repeated Use of Design Knowledge,” *Russ. Eng. Res.*, 2022, doi: 10.3103/S1068798X22010105.

[27] Y. T. Ng, X. Li, J. Y. Wu, and W. F. Lu, “Hybrid Classification Method for Image-based Anomaly Detection in Manufacturing Processes,” *2022 4th Int. …*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9976593/

[28] A. Ghosh and A. Senthilrajan, “Comparison of machine learning techniques for spam detection,” *Multimedia Tools and Applications*. Springer, 2023. doi: 10.1007/s11042-023-14689-3.

[29] H. Rahman, M. O. Faruq, T. B. A. Hai, W. Rahman, and ..., “IoT enabled mushroom farm automation with Machine Learning to classify toxic mushrooms in Bangladesh,” *Journal of agriculture …*. Elsevier, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154321001691

[30] B. Kishore, A. Yasar, Y. S. Taspinar, R. Kursun, and ..., “Computer-aided multiclass classification of corn from corn images integrating deep feature extraction,” *Computational …*. hindawi.com, 2022. [Online]. Available: https://www.hindawi.com/journals/cin/2022/2062944/

[31] S. Elango, D. Haldar, and A. Danodia, “Discrimination of maize crop in a mixed Kharif crop scenario with synergism of multiparametric SAR and optical data,” *Geocarto Int.*, 2022, doi: 10.1080/10106049.2021.1920628.

[32] C. Boukhatem, H. Y. Youssef, and ..., “Heart disease prediction using machine learning,” *2022 Adv. Sci. …*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9734880/

[33] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, “A comparative analysis of machine learning

classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*. Elsevier, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772442522000569

[34] A. D. Goswami, G. S. Bhavekar, and P. V Chafle, "Electrocardiogram signal classification using VGGNet: a neural network based classification model," *International Journal of …*. Springer, 2023. doi: 10.1007/s41870-022-01071-z.

[35] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.