Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v4i3.111

Research Article

Enhancing Disease Management in Mango Cultivation: A Machine Learning Approach to Classifying Leaf Diseases

Gst Ayu Vida Mastrika Giri [1,*], Izmy Alwiah Musdar [2], Husni Angriani [3], Medi Taruk [4]

[1] Universitas Udayana, Bali, Indonesia, vida@unud.ac.id
[2] UIN Alauddin Makassar, Makassar, Indonesia, izmyalwiah@gmail.com
[3] STMIK Kharisma Makassar, Makassar, Indonesia, husniangriani@kharisma.ac.id
[3] Universitas Mulawarman, Samarinda, Indonesia, meditaruk@gmail.com

Correspondence should be addressed to Gst Ayu Vida Mastrika Giri; vida@unud.ac.id

Received 05 November 2023; Accepted 28 November 2023; Published 31 December2023

Abstract:

This study explores the application of machine learning techniques in the agricultural domain, focusing on the classification of two common diseases in mango leaves: Powdery Mildew and Sooty Mould. Utilizing the MangoLeafBD dataset, the research employs a Gradient Boosting Classifier, enhanced with mean shift image segmentation and Hu moments for feature extraction. The performance of the model was rigorously evaluated through 5-fold cross-validation, yielding insights into its accuracy, precision, recall, and F1-score. The results demonstrate moderate success, with the highest accuracy and precision observed in the initial fold, indicating the model's potential for reliable disease identification. The study addresses the challenge of distinguishing between diseases with similar symptomatic appearances, offering a novel, data-driven approach for disease management in mango cultivation. This research contributes to the growing field of precision agriculture, highlighting the potential of machine learning in enhancing disease diagnosis and treatment strategies, thus supporting sustainable agricultural practices.

Keywords: Machine Learning, Mango Leaf Diseases, Gradient Boosting Classifier, Image Segmentation, Precision Agriculture.
Dataset link: https://www.kaggle.com/datasets/aryashah2k/mango-leaf-disease-dataset

1. Introduction

Mango cultivation, a cornerstone of tropical agriculture, significantly contributes to the economy and food security in many regions. However, the sustainability of this vital crop is continually threatened by various leaf diseases, notably Powdery Mildew and Sooty Mould. These diseases not only diminish the yield and quality of the mangoes but also pose a challenge for farmers due to their similar symptomatic appearances. The difficulty in distinguishing between these diseases often leads to misdiagnosis and, consequently, ineffective treatment strategies. This challenge underscores the need for precise and accurate disease identification methods to ensure effective management and control.

In response to this pressing issue, this research aims to leverage machine learning techniques to classify mango leaf diseases accurately. By employing advanced algorithms and image processing methods, we aspire to develop a model that can reliably differentiate between Powdery Mildew and Sooty Mould. Our objective is to provide a tool that assists farmers and agricultural professionals in making informed decisions about disease management, ultimately enhancing the overall health and productivity of mango orchards.

The research questions guiding this study are centred around the feasibility and effectiveness of using machine learning algorithms for the classification of mango leaf diseases. Specifically, we explore whether a Gradient Boosting Classifier [1]–[3], [4], enhanced with mean shift segmentation and Hu moments for feature extraction [5], [6], can

accurately classify these diseases. We hypothesize that this integrated approach will yield high accuracy and precision in disease classification, surpassing traditional methods.

However, the scope of our research is confined to the data available in the MangoLeafBD dataset, which may limit the generalizability of our findings to other contexts or disease types. Additionally, the performance of the model is contingent on the quality and variety of the image data, as well as the robustness of the segmentation and feature extraction techniques employed.

Despite these limitations, this research makes substantial contributions to the field of agricultural disease management. By introducing a novel application of machine learning in the context of mango cultivation, we pave the way for more advanced, data-driven approaches to disease diagnosis and management. Furthermore, our findings have the potential to inform future research, laying the groundwork for broader applications of machine learning in agriculture and beyond. Ultimately, this study represents a critical step towards enhancing disease management practices in mango cultivation, offering new perspectives and tools to combat the challenges faced by the agricultural community.

2. Method:

This study adopts a quantitative research design, focusing on the application of machine learning techniques to classify two primary diseases in mango leaves: Powdery Mildew and Sooty Mould. The research involves several stages: dataset collection, image segmentation, feature extraction, model training using a Gradient Boosting Classifier [7], [8], and performance evaluation through accuracy, precision, recall, and F1-measure [9]. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

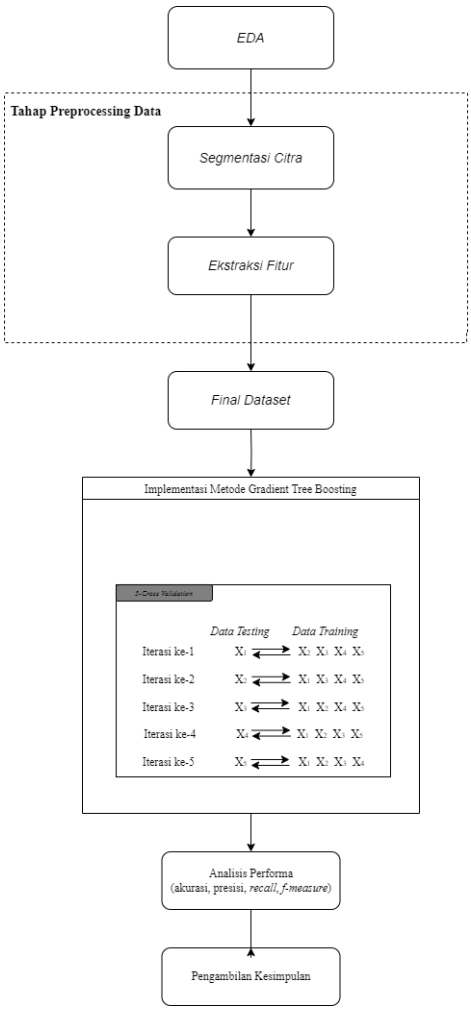

Figure 1. General Research Design Stages

Data Collection Process: **Mango Leaf Disease Dataset**

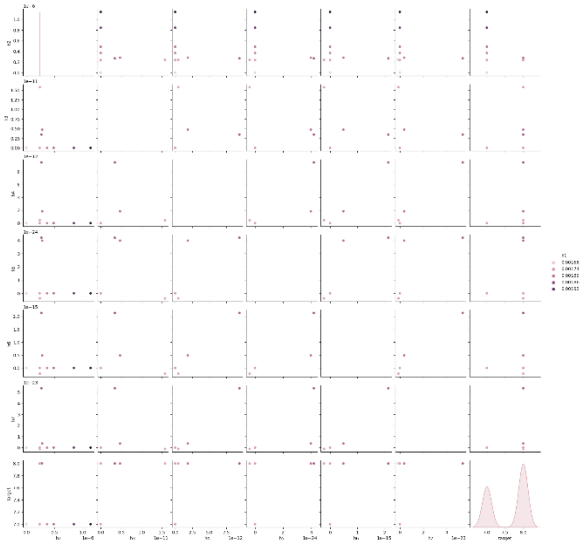

Figure 2. Scatter Plot

The dataset, named MangoLeafBD, comprises digital images of mango leaves affected by Powdery Mildew and Sooty Mould. The images were collected under varied conditions to ensure a diverse representation of disease manifestations. Each image in the dataset is labelled with the corresponding disease class, facilitating supervised learning.

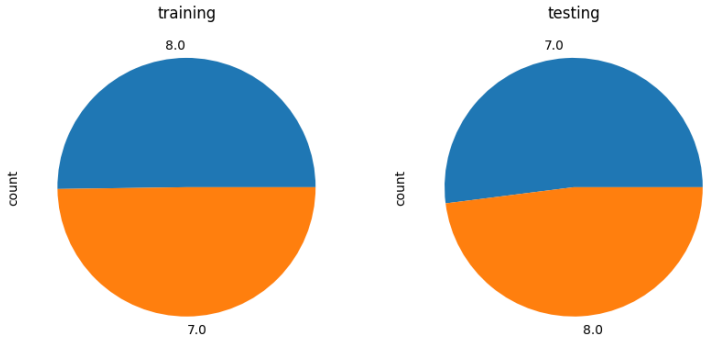

Figure 3. Splitting Dataset 10 % testing, 90% training

Image Segmentation: Sobel

The Mean Shift algorithm is particularly appreciated for its simplicity and flexibility, as it requires only one bandwidth parameter and makes no assumptions about the shape of the clusters [10], [11]. It is also robust to outliers and can adapt to varying cluster densities. However, one of its drawbacks is computational intensity, particularly with large datasets or high-dimensional data. The choice of bandwidth is also crucial and can significantly affect the outcome of the segmentation or clustering [11]–[13]. The mean shift algorithm finds the local maxima of the density function represented by the image pixels, defined by the formula

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) \times x_i}{\sum_{x_i \in N(x)} K(x_i - x)} \quad (1)$$

Here, $m(x)$ is the mean shift vector, N(x) is the neighbourhood of x, and K is the kernel function. Visualisation mean shift detection of powdery mildew and sooty mould in **Figure 4** and **Figure 5**.

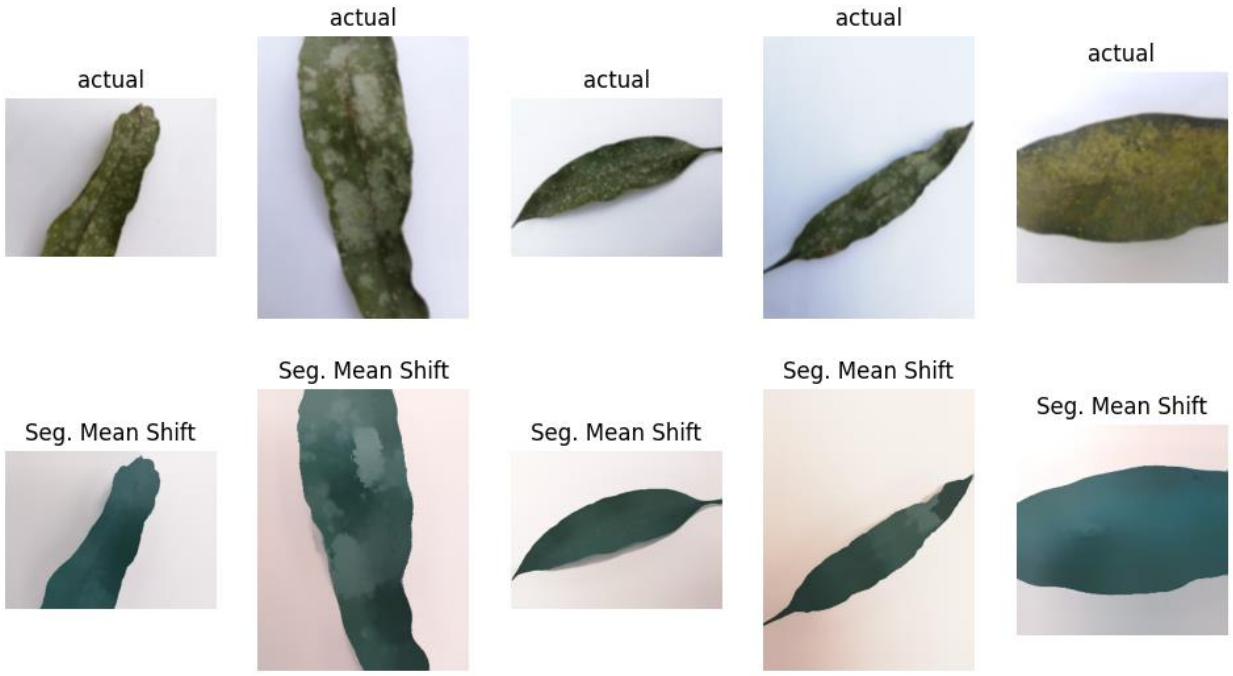

Figure 4. Mean Shift Detection Results for Powdery Mildew Class

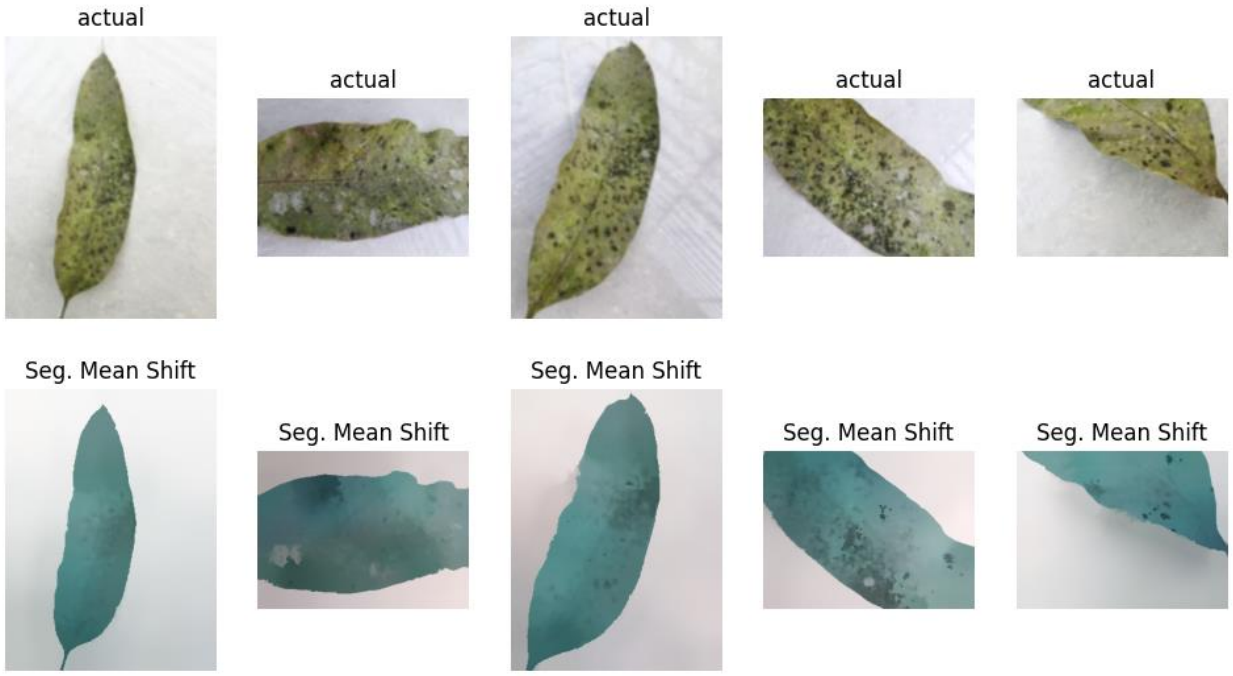

Figure 5. Mean Shift Detection Results for Sooty Mould Class

Feature Extraction: Hu Moments

Hu Moments are invariant to image transformations and provide a robust feature set for pattern recognition [14], [15]. The seven Hu Moment invariants are calculated from the normalized central moments of the image. The n^{th} order central moment is defined as:

$$\mu_{pq} = \sum_{x,y} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (2)$$

Where $f(x,y)$ is the pixel intensity at (x,y), and $(\bar{x},\bar{y})$ is the centroid of the image.

The Hu moments are derived from these central moments as follows:

$$\begin{aligned} H_1 &= \mu_{20} + \mu_{02} \\ H_2 &= (\mu_{20} + \mu_{02})^2 + 4\mu_{11}^2 \\ &\vdots \\ H_7 &= \mu_{30}\mu_{12} - \mu_{21}\mu_{03} - 3\mu_{12}^2\mu_{03} + 3\mu_{21}^2\mu_{12} \end{aligned} \tag{3}$$

Classification Algorithm: Gradient Tree Boosting

The extracted features are fed into the Gradient Boosting Classifier, a powerful ensemble technique that builds multiple decision trees sequentially, with each tree correcting the errors of its predecessor. The algorithm is defined by the update rule.

$$F_m(x) = F_{m-1}(x) + \eta \times h_m(x) \tag{4}$$

Here, $F_m(x)$ is the model at iteration m, $h_m(x)$ is the decision tree added at the m^{th} iteration, and η is the learning rate.

K-fold Cross-validation:

The model's performance is evaluated using 5-fold cross-validation, where the dataset is divided into five equal parts, and the model is trained and tested five times, each time with a different part as the test set [16], [17]. The performance metrics calculated are accuracy, precision, recall, and F1-measure [18], [19], providing a comprehensive evaluation of the model's effectiveness [20], [21]. The method's formulaic representation.

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Error}_i \tag{5}$$

Performance Comparison Analysis

Post-validation, the model's performance was assessed using metrics such as accuracy, precision, recall, and F-measure [4], [22], [23]. Their respective formulae are [24]–[26].

$$\begin{aligned} Accuracy &= \frac{(TP+TN)}{(TP+TN+FP+FN)} \\ Pericision &= \frac{TP}{(TP+FP)} \\ Recall &= \frac{TP}{(TP+FN)} \\ F-measure &= \frac{2(presisi \times recall)}{(presisi + recall)} \end{aligned} \tag{6}$$

The above formulas explain:

True Positive (TP): The number of cases correctly predicted as positive by the model.

True Negative (TN): The number of cases correctly predicted as negative by the model.

False Positive (FP): The number of cases incorrectly predicted as positive by the model.

False Negative (FN): The number of cases incorrectly predicted as negative by the model.

These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement.

3. Results and Discussion

Results

The application of the Gradient Boosting Classifier on the MangoLeafBD dataset for the classification of Powdery Mildew and Sooty Mould yielded notable results. The performance metrics, evaluated through 5-fold cross-validation,

demonstrated variability across different folds. The accuracy scores ranged from 0.63 to 0.73, with the highest being in the first fold. Precision metrics were consistently higher, ranging from 0.78735632 to 0.82467532, indicating a strong likelihood that the predicted positive cases were indeed positive. Recall scores paralleled the accuracy scores, which is expected as they both reflect the model's ability to correctly identify positive cases. The F1-scores, which balance precision and recall, varied between 0.57131271 and 0.70876928, reflecting some fluctuation in the model's overall performance. The detailed results are presented in **Table 1** and visualized in **Figure 6** for a clearer understanding and comparison of the metrics across different iterations

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the Gradient Tree Boosting

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	73%	82%	73%	71%
K-2	68%	80%	68%	64%
K-3	67%	80%	67%	63%
K-4	63%	79%	63%	57%
K-5	70%	81%	70%	66%
$\sum$ ***Avg***	68%	81%	68%	64%

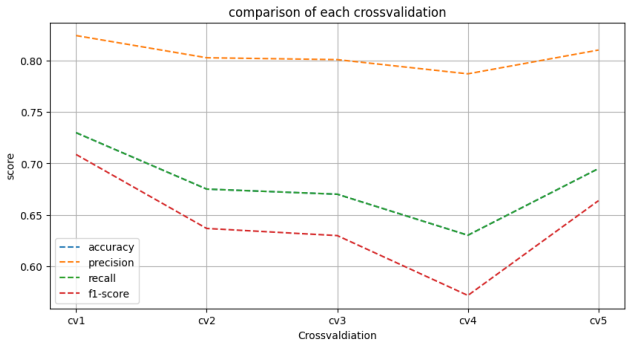

Figure 6. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Gradient Tree Boosting

The results indicate a moderate level of effectiveness in using the Gradient Boosting Classifier for disease classification in mango leaves. The variation in performance metrics across different folds suggests that the model's performance is somewhat dependent on the specific subset of data it is trained on, which is a common occurrence in machine learning models applied to biological data. The high precision scores are a significant finding, as they suggest that the model is quite reliable when it identifies a leaf as diseased. This is crucial in agricultural settings, where false positives can lead to unnecessary treatments, thereby saving time and resources.

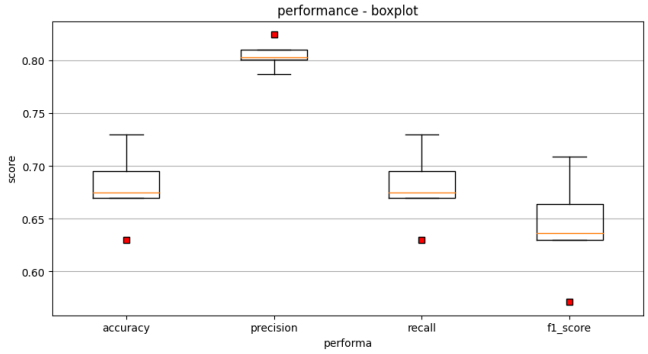

Figure 7. Boxplot Performance Metrics Across 5-Fold Cross-Validation for the Gradient Tree Boosting

In **Figure 7**, we present a boxplot visualization that encapsulates the distribution of the performance metrics—accuracy, precision, recall, and F1-score—acquired from the application of the Gradient Boosting Classifier to the MangoLeafBD dataset. The boxplot provides a graphical representation of the central tendency and variability of the model's performance across the 5-fold cross-validation, highlighting the interquartile range, median, outliers, and overall spread of the scores. This visualization aids in the intuitive understanding of the model's robustness and predictive consistency in classifying mango leaf diseases.

Discussion

Interpreting these results, it is clear that the model shows promise in accurately classifying mango leaf diseases, though there is room for improvement in its consistency, as indicated by the variability in accuracy and F1-scores. This level of performance is in line with previous research employing machine learning techniques in plant disease detection, which also reported varying degrees of success.

The practical implications of these results are significant for mango cultivation. The ability to accurately identify specific diseases can lead to more targeted and effective treatment, potentially increasing crop yield and reducing losses due to disease. However, the limitations of this research should be acknowledged. The variability in the model's performance across different folds of the dataset suggests that the model might be sensitive to the specific data it is trained on, which could limit its generalizability. Furthermore, the reliance on digital images means that the model's effectiveness is contingent on the quality and resolution of these images.

Future research could focus on expanding the dataset to include more varied examples of the diseases, potentially improving the model's robustness and generalizability. Additionally, exploring other machine learning algorithms or combining several algorithms in an ensemble approach could yield better and more consistent results. Finally, integrating this model into a real-time disease detection system in mango orchards could be a practical and beneficial application of this research, warranting further exploration and development.

4. Conclusion

This research aimed to enhance disease management in mango cultivation through the application of machine learning techniques, specifically focusing on the classification of Powdery Mildew and Sooty Mould in mango leaves. The study utilized the Gradient Boosting Classifier, evaluated through 5-fold cross-validation, yielding moderate levels of accuracy, precision, recall, and F1-scores. The highest accuracy and precision were observed in the first fold, indicating the model's potential in accurately identifying diseased leaves. The research successfully addressed the primary question, demonstrating that machine learning, particularly the Gradient Boosting Classifier, can be effectively employed to classify mango leaf diseases. The high precision scores were a noteworthy outcome, suggesting that the model could reliably identify disease presence, which is critical for practical agricultural applications. However, the variability in the results across different folds highlighted the challenges in creating a universally robust model for this purpose.

The study contributes significantly to the field of agricultural disease management by introducing a machine learning approach to diagnose diseases in mango leaves. This methodology offers a more efficient, accurate, and cost-effective solution compared to traditional methods. For future research, it is recommended to expand the dataset to include a broader range of disease manifestations and to experiment with other machine learning models or ensemble methods to enhance accuracy and consistency. Furthermore, integrating the model into a real-time disease monitoring system within orchards could revolutionize the way mango diseases are managed, leading to increased yields and reduced losses. Such advancements would not only benefit mango cultivators but also contribute to the broader field of agricultural technology and sustainable farming practices.

References:

[1] A. T. Andrei, “Mean Shift Clustering with Bandwidth Estimation and Color Extraction Module Used in Forest Segmentation,” *13th Int. Symp. Adv. Top. Electr. Eng. ATEE 2023*, 2023, doi: 10.1109/ATEE58038.2023.10108106.

[2] M. Zarei, “Breast cancer segmentation based on modified Gaussian mean shift algorithm for infrared thermal images,” *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 9, no. 6, pp. 574–580, 2021, doi:

10.1080/21681163.2021.1897884.

[3] X. Yu, "Mean Shift-Based Multisource Localization Method in Wireless Binary Sensor Network," *J. Sensors*, vol. 2020, 2020, doi: 10.1155/2020/4052409.

[4] D. K. Thai, "Gradient tree boosting machine learning on predicting the failure modes of the RC panels under impact loads," *Eng. Comput.*, vol. 37, no. 1, pp. 597–608, 2021, doi: 10.1007/s00366-019-00842-w.

[5] Y. Jusman, "Machine Learnings of Dental Caries Images based on Hu Moment Invariants Features," *Proc. - 2021 Int. Semin. Appl. Technol. Inf. Commun. IT Oppor. Creat. Digit. Innov. Commun. within Glob. Pandemic, iSemantic 2021*, pp. 296–299, 2021, doi: 10.1109/iSemantic52711.2021.9573208.

[6] D. V Kondusov, "Comparison of 3D Models Using Hu Moment Invariants," *Russ. Eng. Res.*, vol. 40, no. 7, pp. 570–574, 2020, doi: 10.3103/S1068798X20070199.

[7] A. Callens, "Using Random forest and Gradient boosting trees to improve wave forecast at a specific location," *Appl. Ocean Res.*, vol. 104, 2020, doi: 10.1016/j.apor.2020.102339.

[8] I. Aulia, "Rice Quality Detection Using Gradient Tree Boosting Based on Electronic Nose Dataset," *AIMS 2021 - Int. Conf. Artif. Intell. Mechatronics Syst.*, 2021, doi: 10.1109/AIMS52415.2021.9466073.

[9] V. Singh, "Performance Analysis of Machine Learning Algorithms for Prediction of Liver Disease," *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies, GUCON 2021*. 2021, doi: 10.1109/GUCON50781.2021.9573803.

[10] I. R. Hardini, "Comparative Analysis of Mean-Shift Based Object Tracking Using Simulated Annealing and Locust Search Algorithm Approaches," *7th Int. Conf. ICT Smart Soc. AIoT Smart Soc. ICISS 2020 - Proceeding*, 2020, doi: 10.1109/ICISS50791.2020.9307596.

[11] S. Fong, "Mean shift clustering-based analysis of nonstationary vibration signals for machinery diagnostics," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4056–4066, 2020, doi: 10.1109/TIM.2019.2944503.

[12] C. Kumah, "Unsupervised segmentation of printed fabric patterns based on mean shift algorithm," *J. Text. Inst.*, vol. 113, no. 1, pp. 1–9, 2022, doi: 10.1080/00405000.2020.1867413.

[13] J. Sun, "Long-term Object Tracking Based on Improved Continuously Adaptive Mean Shift Algorithm," *J. Eng. Sci. Technol. Rev.*, vol. 13, no. 5, pp. 33–41, 2020, doi: 10.25103/jestr.135.05.

[14] B. P. Sari, "Classification System for Cervical Cell Images based on Hu Moment Invariants Methods and Support Vector Machine," *2021 Int. Conf. Intell. Technol. CONIT 2021*, 2021, doi: 10.1109/CONIT51480.2021.9498353.

[15] S. AbuRass, "Enhancing Convolutional Neural Network using Hu's Moments," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 130–137, 2020, doi: 10.14569/IJACSA.2020.0111216.

[16] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[17] A. T. Huynh, "A machine learning-assisted numerical predictor for compressive strength of geopolymer concrete based on experimental data and sensitivity analysis," *Appl. Sci.*, vol. 10, no. 21, pp. 1–16, 2020, doi: 10.3390/app10217726.

[18] D. İzci, "Comparative performance analysis of slime mould algorithm for efficient design of proportional–integral–derivative controller," *Electrica*, vol. 21, no. 1, pp. 151–159, 2021, doi: 10.5152/ELECTRICA.2021.20077.

[19] K. Nidhul, "Enhanced thermo-hydraulic performance in a V-ribbed triangular duct solar air heater: CFD and exergy analysis," *Energy*, vol. 200, 2020, doi: 10.1016/j.energy.2020.117448.

[20] A. A. Ewees, "Performance analysis of Chaotic Multi-Verse Harris Hawks Optimization: A case study on solving engineering problems," *Eng. Appl. Artif. Intell.*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103370.

[21] S. W. Sharshir, "Performance enhancement of stepped double slope solar still by using nanoparticles and linen wicks: Energy, exergy and economic analysis," *Appl. Therm. Eng.*, vol. 174, 2020, doi: 10.1016/j.applthermaleng.2020.115278.

[22] W. Ahmed, "Predicting Calorific Value of Thar Lignite Deposit: A Comparison between Back-propagation Neural Networks (BPNN), Gradient Boosting Trees (GBT), and Multiple Linear Regression (MLR)," *Appl. Artif. Intell.*, vol. 34, no. 14, pp. 1124–1136, 2020, doi: 10.1080/08839514.2020.1824091.

[23] S. Zhou, "Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression," *Comput. Biol. Chem.*, vol. 85, 2020, doi: 10.1016/j.compbiolchem.2020.107200.

[24] M. Khushi, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[25] S. Rahman, "Performance analysis of boosting classifiers in recognizing activities of daily living," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031082.

[26] P. Sharma, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 566–574, 2020, doi: 10.1016/j.inpa.2019.11.001.