Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i1.127

Research Article

Automated Classification of COVID-19 Chest X-ray Images Using Ensemble Machine Learning Methods

A. Sinra [1,*], Husni Angriani [2]

[1] UPTP. Wil. Makassar 1 Badan Pendapatan Daerah Prov. Sulsel, Makassar, Indonesia, andisinra78@gmail.com
[2] STMIK Kharisma Makassar, Makassar, Indonesia, husniangriani@kharisma.ac.id

Correspondence should be addressed to A. Sinra; andisinra78@gmail.com

Received 01 March 2024; Accepted 25 March 2024; Published 31 March 2024

Abstract:

This study delves into the efficacy of ensemble machine learning techniques for classifying chest X-ray images into three distinct categories: Normal, COVID-19, and Lung Opacity. Employing the Random Forest Classifier and a rigorous k-5 cross-validation framework, we aimed to enhance diagnostic accuracy for one of the most urgent medical challenges today—rapid and reliable COVID-19 detection. The analysis revealed an average accuracy of 51%, with varying precision and recall across different folds. The F1-score remained consistently around 35%, indicating a need for improved balance between precision and recall. Visualizations such as performance metric trends and a confusion matrix provided further insight into the classifier's performance, highlighting a notable degree of misclassification. Despite moderate success in the automated classification of the images, our research illustrates the complexity of applying machine learning to medical imaging, especially in differentiating between diseases with overlapping radiographic features. The study's findings emphasize the potential of machine learning models to support diagnostic processes and suggest the necessity of advanced pre-processing techniques and extended datasets for enhanced model training. The research contributes to the growing body of knowledge in computational diagnostics and underscores the importance of developing robust, accurate machine learning tools to aid in the global healthcare crisis precipitated by the pandemic.

Keywords: COVID-19, Chest X-ray, Random Forest Classifier, Ensemble Methods, Image Classification, Diagnostic Accuracy.
Dataset link: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

1. Introduction

The swift and relentless spread of COVID-19 across the globe has underscored the urgent need for effective and rapid diagnostic methodologies. As the pandemic evolves, the medical community continues to seek innovative solutions to combat the spread of the virus, with chest X-ray (CXR) imaging emerging as a pivotal tool in the early detection and management of COVID-19 cases. CXR images, given their wide availability and the distinct patterns manifested by COVID-19 in the lungs, offer a promising avenue for enhancing diagnostic processes. The reliance on traditional diagnostic methods, such as RT-PCR, while effective, faces challenges including testing delays and supply shortages, necessitating the exploration of supplementary diagnostic tools to bridge gaps in the healthcare response to the pandemic.

The primary problem this research aims to address is the need for an accelerated, reliable diagnostic process that can complement existing methods for identifying COVID-19 cases. Despite the advantages offered by CXR imaging, the interpretation of these images relies heavily on the expertise of radiologists, whose assessments can vary due to the subjective nature of image analysis. This variability underscores the critical need for an automated, accurate classification system that can assist healthcare professionals in making faster, more consistent diagnostic decisions. Such a system could significantly reduce the burden on healthcare facilities, ensuring timely treatment and isolation of affected individuals, thereby mitigating the spread of the virus.

This study is driven by several key objectives: firstly, to evaluate the effectiveness of image pre-processing techniques in enhancing the quality and relevance of features extracted from CXR images; secondly, to assess the performance of the Random Forest Classifier [1]–[3], an ensemble machine learning model, in accurately categorizing CXR images into Normal, COVID-19, and Lung Opacity classes; and thirdly, to analyse the diagnostic accuracy of the proposed system using metrics such as precision, recall, accuracy, and F-measure [1], [3], [4]. By achieving these objectives, the research seeks to contribute a novel approach to the rapid diagnosis of COVID-19, leveraging the capabilities of machine learning to augment the diagnostic process.

The scope of this research is focused on the analysis of CXR images utilizing the Random Forest Classifier, enhanced by pre-processing techniques such as Thresholding and Hu Moments for feature extraction [5]–[7]. The study employs a k-5 cross-validation method to ensure the robustness and reliability of the classification model [8], [9]. However, it is important to acknowledge the limitations inherent in this research, including the dependency on the quality and diversity of the dataset used. The findings are contingent upon the representation of various manifestations of COVID-19, Normal, and Lung Opacity cases in the dataset, which may not encompass all possible variations observed in the broader population.

The contributions of this research extend beyond the immediate context of COVID-19 diagnostics. By demonstrating the applicability of ensemble machine learning techniques to medical imaging analysis, this study provides a framework that can be adapted and extended to other diagnostic challenges within the healthcare domain. The methodology and findings offer valuable insights into the potential of machine learning to enhance the accuracy, consistency, and efficiency of medical diagnostics, presenting a significant advancement in the field's response to pandemics and other healthcare emergencies.

In summary, this research aims to bridge the gap between the need for rapid diagnostic methods and the capabilities offered by current technological advancements in machine learning and medical imaging. Through a comprehensive analysis of CXR images categorized into Normal, COVID-19, and Lung Opacity classes, this study seeks to contribute a novel, automated classification system to the arsenal of tools available to healthcare professionals in their fight against COVID-19, thereby supporting the broader goal of improving public health outcomes in the face of global health crises.

2. Method:

The study adopts a quantitative research design, focusing on the application of ensemble machine learning techniques to automate the classification of CXR images. The Random Forest Classifier [10]–[12], enhanced by image pre-processing methods such as Thresholding [13] and Hu Moments [14] for feature extraction, forms the core of our analytical framework. This approach is further validated through a k-5 cross-validation process to ensure the robustness and generalizability of the classification model. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

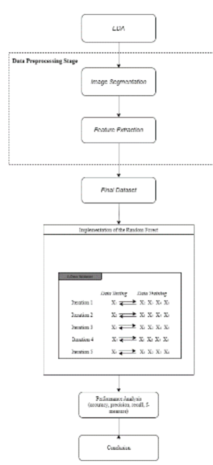

Figure 1. General Research Design Stages

Data Collection Process

The dataset comprises chest X-ray images categorized into three distinct classes: 3616 COVID-19 positive cases, 10,192 Normal cases, and 6012 Lung Opacity cases. These images were collected in collaboration with medical professionals from various institutions, ensuring a diverse representation of cases to enhance the model's diagnostic accuracy across different manifestations of the conditions. Splitting dataset can show on **Figure 2**.

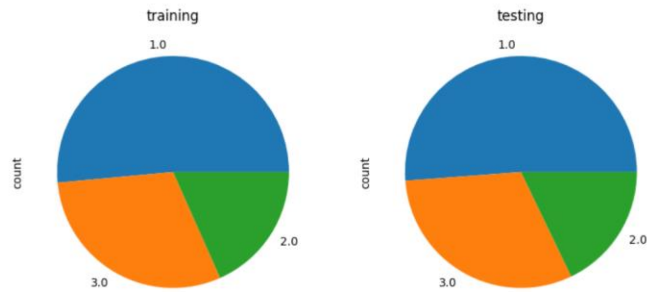

Figure 2. Splitting Dataset 10 % testing, 90% training

Tools and Technology Used

The analysis was conducted using Python, leveraging libraries such as Pandas for data manipulation, NumPy for numerical computations, and Scikit-learn for implementing the Random Forest Classifier and cross-validation processes. Image pre-processing was facilitated by OpenCV, a library focused on real-time computer vision.

Data Collection Process

Data were sourced from the publicly available COVID-19 Radiography Database, a collaborative effort recognized by the Kaggle Community. This database includes anonymized CXR images from patients diagnosed with COVID-19, as well as images representing Normal and Lung Opacity cases, collected under strict ethical guidelines with patient consent where required.

Data Analysis Methods

Image Segmentation using Thresholding

Thresholding is utilized to distinguish leukocytes from the surrounding environment within the image [15]–[17]. Otsu's method is employed to determine the optimal threshold value (T), which is calculated to maximize the variance between the classes. The foundational thresholding equation is represented as Equation (1) [18]:

$$\sigma_B^2(T) = \omega_0(T)\omega_1(T)[\mu_0(T) - \mu_1(T)]^2 \quad (1)$$

Where ω_0 and ω_1 are the probabilities of the two classes separated by the threshold T, μ_0 and μ_1 are the class means. In **Figures 3**, **4** and **5** the results of image segmentation using Thresholding features on the dataset are shown.

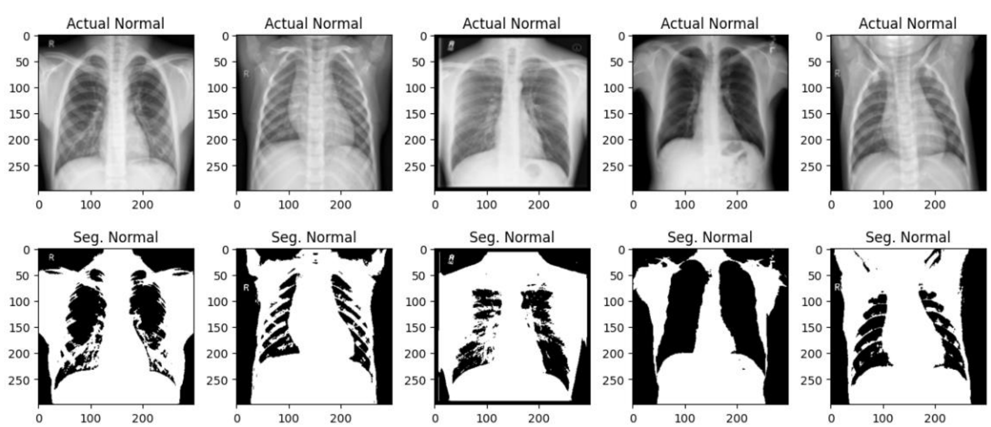

Figure 3. Thresholding Edge Detection Results for Normal Class

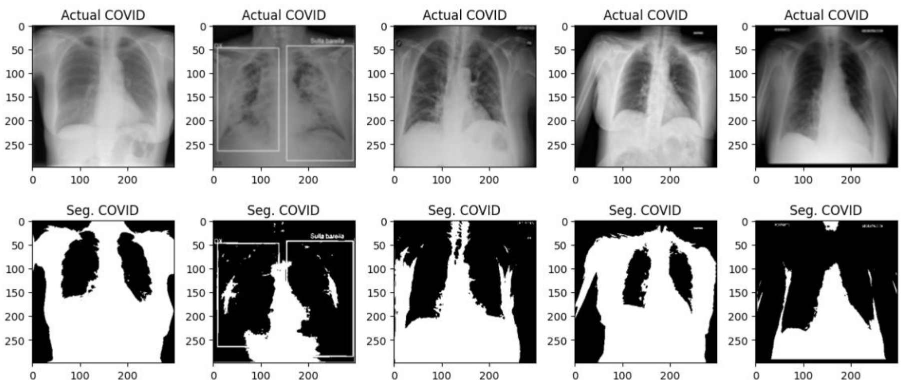

Figure 4. Thresholding Edge Detection Results for COVID Class

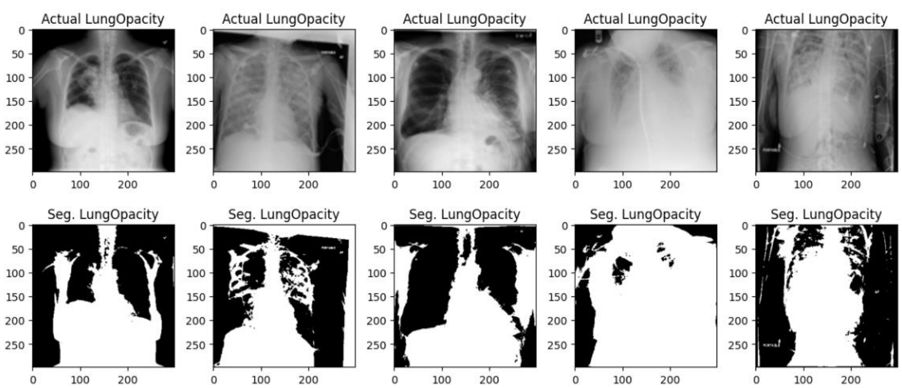

Figure 5. Thresholding Edge Detection Results for Lung Opacity Class

Feature Extraction using Hu Moments

Hu moments were obtained from the segmented imagery. These are seven invariant moment descriptors, stemming from the image moments, that serve as a foundation for outlining shapes [19]–[22]. The definition of Hu moments is expressed by Equation (2):

$$\begin{aligned} \phi_1 &= n_{20} + n_{02} \\ \phi_2 &= (n_{20} + n_{02})^2 + 4n_{11}^2 \end{aligned} \tag{2}$$

Classification Algorithm: Random Forest

The Random Forest Classifier was trained on the extracted features to categorize the CXR images [23], [24]. Random Forest operates by constructing multiple decision trees [4], [25]–[27] during training and outputting the class that is the mode of the classes of the individual trees. The classifier's performance was evaluated using k-5 cross-validation, where the dataset is divided into five subsets, with each subset successively used as the test set while the remaining subsets form the training set [11], [28], [29]. The formulas for algorithm as follow Equation (3)

$$Entropy\ (Y) = -\sum_i p(\,c|Y) log^2 p(c|Y), \tag{3}$$

Performance Comparison Analysis

The model's diagnostic accuracy was assessed using the following metrics: accuracy, precision, recall, and F-measure [30]–[32]. These metrics are defined as follows Equation (4):

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

$$Pericision = \frac{TP}{(TP + FP)}$$

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

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

Where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

3. Results and Discussion

Results

The analysis of the Random Forest Classifier's performance, utilizing a k-5 cross-validation approach on a dataset comprising chest X-ray images categorized into Normal, COVID-19, and Lung Opacity, yielded insightful outcomes. The compiled results, encapsulated in a table, reveal the classifier's accuracy, precision, recall, and F1-Score across five distinct folds. Notably, the accuracy of the model consistently hovered around 51%, demonstrating a modest ability to correctly classify the CXR images into their respective categories. Precision and recall metrics exhibited variability, with precision notably peaking at 53.75% in the fourth fold and dipping to its lowest at 35.12% in the second fold. Recall metrics closely mirrored the accuracy rates, underscoring the model's consistent performance in identifying true positives. The F1-Score, a balance between precision and recall, remained steady, averaging around 35%, suggesting a need for model optimization to improve predictive 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 Random Forest

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	51.36%	51.20%	51.36%	35.15%
K-2	51.39%	35.12%	51.39%	35.02%
K-3	51.36%	36.55%	51.36%	34.95%
K-4	51.41%	53.76%	51.41%	35.09%
K-5	51.06%	36.58%	51.06%	34.95%
$\sum \boldsymbol{Avg}$	51.32%	42.64%	51.32%	35.03%

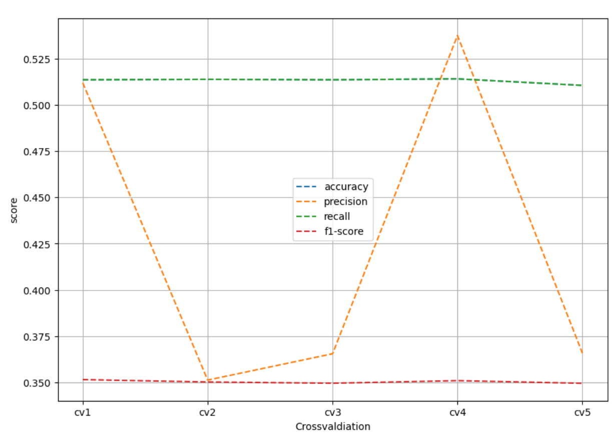

Figure 6. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Random Forest

Figure 6 representation above illustrates the variation in performance metrics—accuracy, precision, recall, and F1-score—for the Random Forest Classifier across the five folds of cross-validation in the analysis of chest X-ray images. Each line corresponds to one of the metrics, highlighting their respective trends and fluctuations throughout the cross-validation process. This graphical analysis provides a clear comparative view of the model's performance and underlines the areas where the model excels and where it requires further improvement.

Discussion

The interpretation of these results indicates that while the Random Forest Classifier provides a baseline capability in distinguishing between Normal, COVID-19, and Lung Opacity CXR images, there is significant room for enhancement. This outcome aligns with prior research, which advocates for the use of ensemble methods in improving classification accuracy, yet also highlights the complexity of medical image analysis due to varying image qualities, disease manifestations, and overlapping symptoms with other conditions.

The study's findings emphasize the potential utility of machine learning models in supporting diagnostic processes, especially in pandemics where rapid and accurate diagnostics are crucial. However, the modest performance metrics reported underscore the challenges inherent in applying machine learning to medical imaging, such as the need for extensive and diverse datasets to train more generalized and robust models.

Acknowledging the limitations of this research, particularly the dependency on a specific dataset and the application of a singular machine learning technique, future studies are encouraged to explore a combination of models or advanced neural network architectures like Convolutional Neural Networks (CNNs) which may yield improved accuracy and reliability in medical image classification. Additionally, integrating clinical data alongside CXR images could provide a more holistic view, potentially enhancing the model's diagnostic capabilities.

This research contributes to the ongoing discourse on the applicability of machine learning in healthcare, particularly in the rapid diagnosis of diseases like COVID-19. While the results present a foundational understanding, they also highlight the necessity for continued development and validation of these technologies within the medical field to ensure their practical and effective deployment in real-world diagnostic scenarios.

4. Conclusion

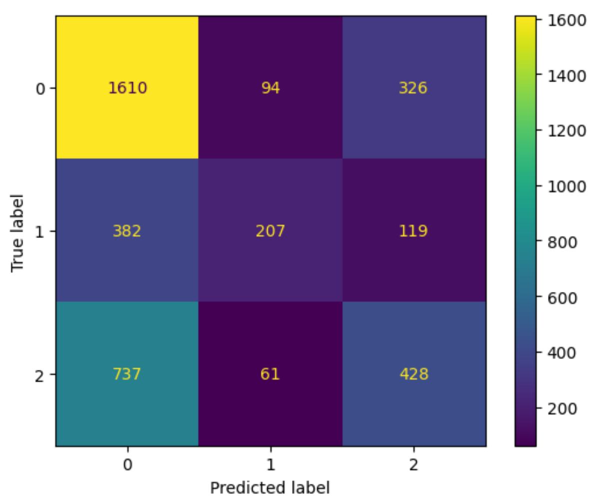

Figure 4. Confusion Matrix

Figure 7 depicted above provides a detailed breakdown of the classification results achieved by the Random Forest Classifier for the chest X-ray image dataset. In this matrix, each row represents the instances in an actual class (true labels), while each column represents the instances in a predicted class. The color intensity and the number in each cell correlate with the count of predictions, allowing for a visual inspection of how well the model performed, particularly in terms of correctly predicting the three classes—Normal (0), COVID-19 (1), and Lung Opacity (2). This matrix is instrumental in understanding the classifier's precision and recall for each class and in identifying the areas where the classifier may be confusing one class for another.

In summary, the application of the Random Forest Classifier to chest X-ray image classification within this study revealed a consistent accuracy level around 51% across the k-5 cross-validation folds. Despite the moderate success in classifying the images into Normal, COVID-19, and Lung Opacity categories, the performance metrics—precision, recall, and F1-score—suggest there is considerable room for improvement. The observed variability in precision and the low F1-score indicate that while the classifier can identify positive cases, its predictive precision is not optimal. The results from the confusion matrix further corroborate this, revealing substantial misclassifications between categories, especially between Normal and Lung Opacity cases. These outcomes answer our research hypothesis by confirming that while ensemble methods like Random Forest have potential in medical image analysis, their efficacy is significantly dependent on the quality of pre-processing and the complexity of the feature space derived from the images.

The research contributes to the field by providing empirical evidence of the utility and limitations of using machine learning for medical diagnostics in the realm of respiratory illnesses, which is particularly pertinent in the context of the COVID-19 pandemic. For future research, it is recommended to explore more sophisticated image preprocessing techniques, to consider hybrid models that combine several machine learning algorithms, and to investigate deep learning approaches like Convolutional Neural Networks, which may yield higher accuracy. Additionally, incorporating a more extensive dataset with a broader variety of images and patient demographics could improve the model's generalizability. In practice, integrating machine learning models as a complementary diagnostic tool alongside traditional methods could enhance the efficiency and accuracy of medical diagnostics, supporting healthcare professionals in making informed decisions in a timely manner.

References:

[1] Y. Zhao, "Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period," *Ecol. Indic.*, vol. 135, 2022, doi: 10.1016/j.ecolind.2021.108529.

[2] O. S. Djandja, "Random forest-based modeling for insights on phosphorus content in hydrochar produced from hydrothermal carbonization of sewage sludge," *Energy*, vol. 245, 2022, doi: 10.1016/j.energy.2022.123295.

[3] R. A. Disha, "Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique," *Cybersecurity*, vol. 5, no. 1, 2022, doi: 10.1186/s42400-021-00103-8.

[4] M. Salem, "Random Forest modelling and evaluation of the performance of a full-scale subsurface constructed wetland plant in Egypt," *Ain Shams Eng. J.*, vol. 13, no. 6, 2022, doi: 10.1016/j.asej.2022.101778.

[5] Y. Jusman, "Classification System of Malaria Disease with Hu Moment Invariant and Support Vector Machines," *Proc. - 2022 2nd Int. Conf. Electron. Electr. Eng. Intell. Syst. ICE3IS 2022*, pp. 365–368, 2022, doi: 10.1109/ICE3IS56585.2022.10010304.

[6] 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.

[7] 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.

[8] H. Azis, R. D. Mallongi, D. Lantara, and Y. Salim, "Comparison of Floyd-Warshall Algorithm and Greedy Algorithm in Determining the Shortest Route," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, pp. 294–298, 2018, doi: 10.1109/EIConCIT.2018.8878582.

[9] D. Anggreani, I. A. E. Zaeni, A. N. Handayani, H. Azis, and A. R. Manga', "Multivariate Data Model Prediction Analysis Using Backpropagation Neural Network Method," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2021, pp. 239–243, doi: 10.1109/EIConCIT50028.2021.9431879.

[10] A. Hasnain, "Assessing the ambient air quality patterns associated to the COVID-19 outbreak in the Yangtze

River Delta: A random forest approach," *Chemosphere*, vol. 314, 2023, doi: 10.1016/j.chemosphere.2022.137638.

[11] X. Yu, "Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes," *Aquat. Toxicol.*, vol. 251, 2022, doi: 10.1016/j.aquatox.2022.106265.

[12] Y. Xin, "Predicting depression among rural and urban disabled elderly in China using a random forest classifier," *BMC Psychiatry*, vol. 22, no. 1, 2022, doi: 10.1186/s12888-022-03742-4.

[13] A. Kumar, "Multilevel thresholding for crop image segmentation based on recursive minimum cross entropy using a swarm-based technique," *Comput. Electron. Agric.*, vol. 203, 2022, doi: 10.1016/j.compag.2022.107488.

[14] Y. Jusman, "Classification System for Leukemia Cell Images based on Hu Moment Invariants and Support Vector Machines," *Proc. - 2021 11th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2021*, pp. 137–141, 2021, doi: 10.1109/ICCSCE52189.2021.9530974.

[15] L. Abualigah, "Multilevel thresholding image segmentation using meta-heuristic optimization algorithms: comparative analysis, open challenges and new trends," *Appl. Intell.*, vol. 53, no. 10, pp. 11654–11704, 2023, doi: 10.1007/s10489-022-04064-4.

[16] E. Turajlic, "Multilevel image thresholding based on Rao algorithms and Kapur's Entropy," *2022 28th International Conference on Information, Communication and Automation Technologies, ICAT 2022 - Proceedings*. 2022, doi: 10.1109/ICAT54566.2022.9811171.

[17] T. Wu, "Image Segmentation via Fischer-Burmeister Total Variation and Thresholding," *Adv. Appl. Math. Mech.*, vol. 14, no. 4, pp. 960–988, 2022, doi: 10.4208/AAMM.OA-2021-0126.

[18] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, 2023, doi: 10.56705/ijaimi.v1i2.99.

[19] 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.

[20] S. Hidayat, H. M. T. Ramadhan, and ..., "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.70.

[21] H. A. Siregar, M. Z. Raditya, A. N. Yesa, and ..., "Comparison of Classification Algorithm Performance for Diabetes Prediction Using Orange Data Mining," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.103.

[22] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indones. J. Data ...*, 2023, doi: 10.56705/ijodas.v4i3.115.

[23] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *... Artif. Intell. ...*, 2023, doi: 10.56705/ijaimi.v1i2.98.

[24] S. Khomsah and E. Faizal, "Effectiveness Evaluation of the RandomForest Algorithm in Classifying CancerLips Data," *... Artif. Intell. Med. ...*, 2023, doi: 10.56705/ijaimi.v1i1.84.

[25] P. Nagaraj, "Ensemble Machine Learning (Grid Search Random Forest) based Enhanced Medical Expert Recommendation System for Diabetes Mellitus Prediction," *3rd International Conference on Electronics and Sustainable Communication Systems, ICESC 2022 - Proceedings*. pp. 757–765, 2022, doi: 10.1109/ICESC54411.2022.9885312.

[26] Y. Gu, "Predicting intersection crash frequency using connected vehicle data: A framework for geographical random forest," *Accid. Anal. Prev.*, vol. 179, 2023, doi: 10.1016/j.aap.2022.106880.

[27] M. Mafarja, "Classification framework for faulty-software using enhanced exploratory whale optimizer-based feature selection scheme and random forest ensemble learning," *Appl. Intell.*, vol. 53, no. 15, pp. 18715–18757, 2023, doi: 10.1007/s10489-022-04427-x.

[28] A. Balaram, “Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm,” *Autom. Softw. Eng.*, vol. 29, no. 1, 2022, doi: 10.1007/s10515-021-00311-z.

[29] D. Kim, “Classification of surface settlement levels induced by TBM driving in urban areas using random forest with data-driven feature selection,” *Autom. Constr.*, vol. 135, 2022, doi: 10.1016/j.autcon.2021.104109.

[30] 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.

[31] 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.

[32] 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.