DOI: https://doi.org/10.52436/1.jutif.2025.6.1.4323
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPLEMENTATION OF THE RANDOM FOREST METHOD FOR CLASSIFYING LUNG X-RAY IMAGE ABNORMALITIES

Retno Supriyanti[1], M. Gus Solhan Fadlola[2], M. Syaiful Aliim[3], Yogi Ramadhani[4]

[1,2,3,4] Electrical Engineering Department, Jenderal Soedirman University, Purwokerto, Indonesia
E-mail: retno_supriyanti@unsoed.ac.id, fadlolasolhan@gmail.com, Muhammad.syaiful.aliim@unsoed.ac.id, yogi.ramadhani@unsoed.ac.id

(Article received: January 15, 2025; Revision: January 28, 2025; published: February 20, 2025)

Abstract

The Covid-19 pandemic has caused a severe global health crisis. Rapid and accurate diagnostics are essential in combating this disease. In this regard, lung X-ray images have become critical for identifying Covid-19 infections. The method used in this study is random forest, a classification method based on ensemble modeling of decision trees. The lung X-ray images used in this study were taken from a datasheet containing images from COVID-19 patients and images from non-Covid-19 patients. The data pre-processing process involves extracting features from the images using image processing techniques and statistical analysis. The random forest model is trained using the processed datasheet to classify the lung X-ray images. The model's performance is evaluated using accuracy, sensitivity, and specificity metrics. In addition, cross-validation is used to measure the reliability and generalization of the model. The study results showed that the random forest method achieved good classification performance in distinguishing COVID-19 lung X-ray images from normal ones. The resulting model provided high accuracy and good sensitivity in identifying Covid-19 cases. These results show the potential of the random forest method in supporting early diagnosis and treatment of COVID-19 disease.

Keywords: *COVID-19, X-ray Image, Random Forest, Classification, Cross-validation*

1. INTRODUCTION

Overall, a good level of public health indicates that society has an effective health system, strong social support, access to adequate health services, and an environment that supports physical and mental well-being. Unfortunately, this condition is complex to achieve now, and many people need better health services. These conditions are due to limited health service facilities, such as a lack of human resources and limited health equipment, especially in rural areas. One type of disease that requires good medical equipment is a disease related to the lungs. One example of a disease related to the lungs is COVID-19, which became a pandemic from 2019 to 2022. One way to diagnose Covid-19 is through a physical examination, commonly called an X-ray. X-ray results on the lungs infected with COVID-19 will show white spots like fog or clouds that appear due to infiltration or consolidation [1]. X-ray examination results are then manually analyzed by a doctor. Our previous research [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] we discussed implementing image processing techniques to support diagnosis in rural areas.

On the other hand, digital image processing technology in the health sector is a significant innovation in diagnosis, treatment, and medical research. Specifically in radiology, Digital Image Processing analyzes images from devices such as X-rays, CT scans, and MRIs. Image processing algorithms can improve the quality of images, helping in the early detection of diseases such as cancer. Several studies have also discussed using digital image processing technology in diagnosing COVID-19. Saleh et al. [15] used a new radionics feature technique called Auto-lesion segmentation (ALS) to diagnose COVID-19 using computed tomography imaging. The ALS technique utilizes an intensity dark channel prior based on a Deep Neural Network (ALS-IDCP-DNN) model. The methodology consists of three primary steps: data resizing, feature extraction and selection, segmentation, data augmentation, and classification. The classification was performed using a DNN-based Resnet-50 architecture. Prasad et al. [16] proposed a technique to identify positive COVID-19 cases and determine the location of the most affected COVID-19 cases for vaccine distribution to limit the impact of the disease. They proposed a cloud-based image analysis approach to use the COVID-19 vaccination distribution model (CIA-CVD). The model uses deep learning, machine learning, digital image processing, and cloud solutions to handle the increase in COVID-19 cases and prioritized vaccination distribution. Supriyanti et al. [17] used the Souvola method to segment the thorax area to develop a COVID-19 diagnosis application. The Sauvola method uses a threshold that is adjusted for each window. Stubblefield et al. [18] used transfer learning to differentiate COVID-19 from other pneumonia and healthy cases with 99.2% accuracy. They developed a CNN-based deep learning approach to automatically predict cardiovascular disease (CVD) risk in COVID-19 patients compared to normal subjects with 97.97% accuracy. The model was further validated against cardiac CT-based markers, including thoracic heart ratio (CTR), pulmonary artery to aorta ratio (PA/A), and calcified plaque. Modi et al. [19] conducted a prospective study involving 117 patients based on online and offline data collection of cough sounds of COVID-19 patients in a hospital isolation ward using smartphones. They developed web-based AI software to identify cough sounds as COVID-19 or non-COVID-19. The data was divided into three segments: training set, validation set, and test set. Pre-processing algorithms were combined with Short-Time Fourier Transform feature representation and logistic regression models. The appropriate software was used to identify vocal signatures, and K-fold cross-validation was performed. Salarabaradi et al. [20] used a new fusion method that combines dropout, data augmentation, transfer learning, and edge detection using fuzzy techniques. This fusion resulted in increased model accuracy and reduced overfitting. In addition, they applied it to CT images of 300 patients and grouped the images into two groups based on lung conditions. Bohmrah et al. [21] used an optimized hybrid DNN-ML method, combining optimized deep neural networks (DNN) models and machine learning (ML) classifiers with a compelling image preprocessing approach. They used Deep Learning (DL) models for feature extraction, namely GoogleNet, EfficientNetB0, and ResNet50, which were further fed into the Bayesian optimized ML classification method. The two main contributions of this research are Edge-based Region of Interest (ROI) extraction and using the Bayesian optimization approach to configure the optimal ML classifier architecture. Antar et al. [22] used a novel approach to detect COVID-19 virus infection in lung images, precisely the problem of infection prediction. In this method, the input image is first processed using a threshold and then resized to 128 × 128. After that, an automated tool is used to color the resized lung image. The three channels (red, green, and blue) are then separated from the colored image and further processed through image inversion and histogram equalization for further segmentation using Unet. Rojas-Zumbado et al. [23] Their research recorded a series of measurements from 252 participants aged 18 years and over who requested a SARS-CoV-2 PCR (polymerase chain reaction) test at the Zambrano-Hellion Hospital in Nuevo León, Mexico. Data for PCR results, demographics, vital signs, food intake, activity and lifestyle factors, recent medications, respiratory and general symptoms, and a thermal video session were collected in which volunteers performed a simple breath hold in four different positions. The vital signs recorded included axillary temperature, blood pressure, heart rate, and oxygen saturation. Each thermal video was divided into four scenes, corresponding to the front, back, left, and right sides, and was available in MPEG-4 format for easy inclusion into the image processing pipeline. All these variables were used to evaluate the diagnosis of COVID-19. Naidji et al. [24] used deep learning architecture to automatically detect COVID-19 and heart disease based on ECG images. They combined the EfficientNet-B0 CNN model and Vision Transformer. They created two classification schemes: binary classification to identify COVID-19 cases and multiclass classification to distinguish COVID-19 cases from normal cases and other cardiovascular diseases. From the research that has been done, the combination of image processing techniques and machine learning is a fairly robust method for developing a COVID-19 diagnosis system. One algorithm that has been proven effective in medical image classification is the Random Forest method. This machine learning method uses an ensemble of decision trees to perform classification. This method combines predictions from several randomly constructed decision trees and then produces accurate and stable prediction results. By utilizing Random Forest's ability to classify medical images, this study aims to develop a system that can automatically classify COVID-19 lung X-ray images.

2. RESEARCH METHODS

2.1. Data

The data used in this study were 200 lung images, 100 of which were images of lungs affected by COVID-19 and 100 of normal lungs. Figure 1 shows some examples of the images used in this research. Image data is taken from several databases, including the Italian Society of Medical Interventional Radiology, www.kaggle.com, and Nihcc. app. box. com, and *Radiopedia.org.*

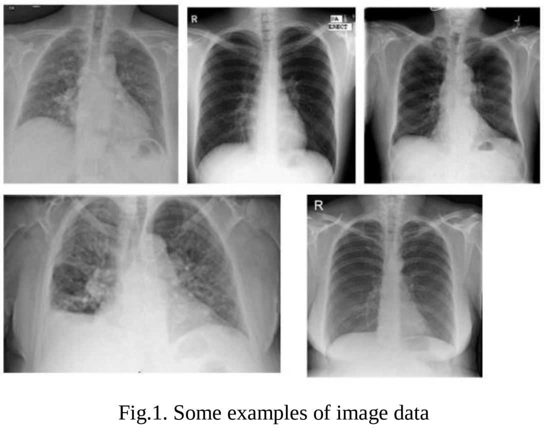

Fig.1. Some examples of image data

2.2. Texture Extraction

Texture extraction is done to obtain texture values from an image to see the difference between lungs affected by COVID-19 and normal lungs. We did this step because the image texture is formed from the distribution of intensity that occurs on the image plane curve, so each image will have a different texture value. This texture extraction uses a statistical approach by calculating the intensity histogram to produce texture features in mean values, moment 2 (standard deviation), skewness, smoothness, uniformity, and entropy. [25]. The texture value obtained indicates the difference between COVID-19 lungs and normal lungs.

2.3. Random Forest Method

The random forest method is an ensemble machine-learning technique used primarily for classification and regression tasks. It combines multiple decision trees to create a "forest" of trees, and each tree in the forest outputs a prediction (in classification, a class; in regression, a value). The final output of the random forest is determined by averaging the predictions in regression or by a majority vote in the classification. [26]. The working principle of the random forest method is as follows: (a) A decision tree splits data into smaller groups based on feature values, with each split aiming to improve prediction accuracy. For classification, the splits aim to minimize class impurity (e.g., using Gini impurity or entropy), while for regression, the goal is to minimize prediction error (e.g., using mean squared error). (b) Each tree in the random forest is trained on a random subset of the training data, created using bootstrap sampling (sampling with replacement). This introduces diversity in the trees. (c) When constructing each tree, random forests select a subset of features at each split, adding another layer of randomness and reducing the correlation between trees. (d) After training, each decision tree in the forest provides a result when making predictions. For classification problems, the final prediction is determined by a majority vote across all trees. For regression, the final result is the average of all the tree predictions. The advantages of this method are numerous, including the following. (i) High accuracy: By aggregating multiple trees, random forests often yield better predictive performance than individual decision trees. (ii) Robust to overfitting: Random forests mitigate overfitting, especially when compared to a single decision tree, as the randomness in data and feature selection leads to less prediction variance. (iii) Handles large datasets: The method is efficient for large datasets with many features. (iv) Works well with missing data: Random forests can handle missing data by averaging across the trees, which can still make predictions without those missing values.

3. RESULTS AND DISCUSSIONS

The texture extraction process begins by adjusting the object to be extracted. The input variable is the RGB image variable, the original lung image in the form of a gray image whose texture value will be calculated, as seen in Figure 2. The initial stage is to make a copy of the RGB image that was previously taken. Next, extract the image dimensions. The variable N will contain the number of rows, M will include the number of columns, and L will consist of the number of color channels in the image. The result is a vector containing the number of pixels at each color intensity level. After calculating the histogram, the result is normalized by dividing it by the total number of pixels in the image (N * M). The result produces a normalized histogram, which gives the relative distribution of color intensity across the image. From the resulting image, the texture value is then calculated.

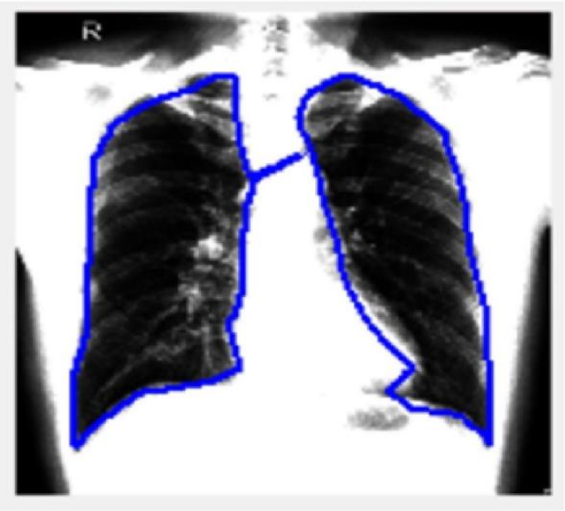

Fig.2. An example of segmented image

Based on the experiments performed by performing texture extraction on 100 COVID-19 lung images and 100 normal lung images. Figure 3 shows a comparison graph of textures between normal and COVID-19-infected lungs.

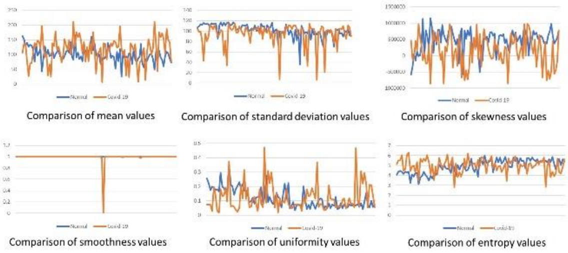

Fig.3. A comparison graph of textures between normal and COVID-19-infected lungs

Figure 3 shows that the average mean value of COVID-19 lungs is higher than that of normal lungs, where the average mean value of COVID-19 lungs is 115.679 while the average mean value of normal lungs is 100.6639. This means the COVID-19 lung

X-ray image has a higher average image intensity. COVID-19 lungs have a lower average standard deviation value of 93.8573, while the average standard deviation value of normal lungs is 101.6466. It means that normal lungs have a higher image contrast value. Normal lungs have a higher average skewness value of 470628, while COVID-19 lungs have a lower average skewness value of 132059. COVID-19 and normal lungs have an average smoothness value close to 1, indicating that both images have high contrast. The difference in the average smoothness value of COVID-19 and normal lungs is 0.00051, whereas normal lungs have a higher value. The average uniformity value in COVID-19 lungs is higher than in normal lungs. This indicates that COVID-19 lungs have fewer gray levels than normal lungs. COVID-19 lungs have a higher average entropy value of 4.9963, while the average entropy value of normal lungs is 4.90702. Comparing these values shows that COVID-19 lungs have higher image complexity. During the data training process, several variations of the number of decision trees to be formed will be used: 25, 50, 100, 150, 200, 250, and 300. The purpose of testing the classification system created is to determine the level of accuracy of the random forest method for classifying COVID-19 lung x-ray images with normal lungs based on the texture extraction values obtained. The initial stage is to divide the COVID-19 and average lung data into 2: training and test data. Then, the training and test data ratio will be determined. After that, the decision tree will be chosen. Tables 1 and 2 show the classification results of normal lungs and lungs infected with COVID-19.

The results of the classification of COVID-19 lung x-ray image data with normal lungs using the numtrees 100 variations with a ratio of each training data and test data of 60:40 obtained an accuracy of 67.50%. The classification of COVID-19 lung X-ray image data with normal lungs using the numtrees 100 variations, with a ratio of each training data set to test data of 70:30, obtained an accuracy of 76.67%. The results of the classification of COVID-19 lung X-ray image data with normal lungs using the numtrees 100 variations, with a ratio of each training data set to test data of 80:20, obtained an accuracy of 77.50%. The classification of COVID-19 lung X-ray image data with normal lungs using the numtrees 300 variation, with a ratio of each training data set to test data of 60:40, obtained an accuracy of 68.75%. The results of the classification of COVID-19 lung X-ray image data with normal lungs using the numtrees 300 variation, with a ratio of each training data set to test data of 70:30, obtained an accuracy of 80.00%. The classification of COVID-19 lung X-ray image data with normal lungs using the numtrees 300 variation, with a ratio of each training data set to test data of 80:20, obtained an accuracy of 82.50%. From the analysis of the experimental results, there are several things to consider. First, increasing the number of trees (Numtrees) in the model consistently increases accuracy. The results show that more trees give the model a better ability to understand complex patterns in the data. Second, the configuration of the dataset also has a significant impact. In this case, the proportion of 80% training data and 20% test data gave the best performance, achieving an accuracy of 82.50%. The results show that a more extensive training data distribution can give the model more information to perform good learning. However, it is essential to remember that these results are specific to this experiment and may vary depending on the unique characteristics of the dataset and the algorithm used. In conclusion, choosing the correct number of trees and distribution of training-test data is critical in improving model performance.

4. CONCLUSIONS

The following conclusions were obtained based on the discussion, system testing, and classification results. X-ray image processing of lung objects using the Otsu and Phansalkar Thresholding segmentation methods to calculate texture values can be done using MATLAB-based application programming. System testing of the application program has shown success by obtaining texture values from each data tested and showing the difference in texture values of COVID-19 lungs and normal lungs. The study showed that using Random Forest in classifying COVID-19 lung X-ray images provided adequate results. This study allows grouping COVID-19 lung x-ray images into "Covid-19 positive" and "Covid-19 negative" categories, significantly impacting the diagnosis and monitoring of COVID-19 patients. Random Forest, as an ensemble machine learning method, is effective in overcoming image variation, the diversity of clinical appearances of COVID-19, and the presence of noise in the data..

ACKNOWLEDGMENT

This research was funded by the BLU Fund of Jenderal Soedirman University through the Professor Facilitation Scheme, contract number 709/UN23/PT.01.02/2024.

REFERENCES

[1] M. Moitra, M. Alafeef, A. Narasimhan, V. Kakaria, P. Moitra, and D. Pan, “Diagnosis of COVID-19 with simultaneous accurate prediction of cardiac abnormalities from chest computed tomographic images,” *PLoS One*, vol. 18, no. 12 December, pp. 1–20, 2023, doi: 10.1371/journal.pone.0290494.

[2] R. Supriyanti *et al.*, “Support vector machine method for classifying severity of Alzheimer’s based on hippocampus object

using magnetic resonance imaging modalities," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 6, pp. 6322–6331, 2024, doi: 10.11591/ijece.v14i6.pp6322-6331.

[3] R. Supriyanti, S. L. Dzihniza, M. Alqaaf, M. R. Kurniawan, Y. Ramadhani, and H. B. Widodo, "Morphological features of lung white spots based on the Otsu and Phansalkar thresholding method," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 1, pp. 530–539, 2024, doi: 10.11591/ijeecs.v33.i1.pp530-539.

[4] R. Supriyanti, A. S. Aryanto, M. I. Akbar, E. Sutrisna, and M. Alqaaf, "The effect of features combination on coloscopy images of cervical cancer using the support vector machine method," *IAES Int. J. Artif. Intell.*, vol. 13, no. 3, pp. 2614–2622, 2024, doi: 10.11591/ijai.v13.i3.pp2614-2622.

[5] R. Supriyanti, Y. Ramadhani, and E. Wahyudi, "Hippocampus's volume calculation on coronal slice's for strengthening the diagnosis of Alzheimer's," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 21, no. 1, pp. 123–132, 2023, doi: 10.12928/TELKOMNIKA.v21i1.20746.

[6] R. Supriyanti, A. S. Aryanto, M. I. Akbar, and E. Sutrisna, "The Effect Of Combining Features On Detection Of Pre-Cervical Cancer Based On Colposcopy Images Using The Support Vector Machine Method," in *the 6 International Conference of Multidisciplinary Approaches for Sustainable Rural Development (ICMA-SURE)*, 2023.

[7] R. Supriyanti, F. F. R. Wibowo, Y. Ramadhani, and H. B. Widodo, "Comparison of conventional edge detection methods performance in lung segmentation of COVID19 patients," *AIP Conf. Proc.*, vol. 2482, no. February, 2023, doi: 10.1063/5.0111683.

[8] R. Supriyanti, E. Wahyudi, and Y. Ramadhani, "Simple tool for three-dimensional reconstruction of coronal hippocampus slice using matlab," *AIP Conf. Proc.*, vol. 2482, no. February, 2023, doi: 10.1063/5.0114412.

[9] R. F. W. Supriyanti;, Y. Ramadhani;, and H. B. Widodo, "Comparison of conventional edge detection methods performance in lung segmentation of COVID19 patients," *AIP Conf. Proc.*, vol. 2482, no. 1, 2021.

[10] R. Supriyanti, G. P. Satrio, Y. Ramadhani, and W. Siswandari, "Contour Detection of Leukocyte Cell Nucleus Using Morphological Image," in *Journal of Physics: Conference Series*, Apr. 2017, vol. 824, no. 1. doi: 10.1088/1742-6596/824/1/012069.

[11] R. Supriyanti, S. A. Priyono, E. Murdyantoro, and H. B. Widodo, "Histogram equalization for improving quality of low-resolution ultrasonography images," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 3, pp. 1397–1408, 2017, doi: 10.12928/TELKOMNIKA.v15i3.5537.

[12] R. Supriyanti, A. A. Hafidh, Y. Ramadhani, and H. B. Widodo, "MEASURING GESTATIONAL AGE AND UTERINE DIAMETER BASED ON IMAGE SEGMENTATION," *ARPN J. Eng. Appl. Sci.*, vol. 13, no. 2, 2018, [Online]. Available: www.arpnjournals.com

[13] R. Supriyanti, A. Chrisanty, Y. Ramadhani, and W. Siswandari, "Computer aided diagnosis for screening the shape and size of leukocyte cell nucleus based on morphological image," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 1, pp. 150–158, Feb. 2018, doi: 10.11591/ijece.v8i1.pp150-158.

[14] R. Supriyanti, A. Rahmadian Subhi, Y. Ramadhani, and H. B. Widodo, "A Simple Tool for Identifying the Severity of Alzheimer's Based on Hippocampal and Ventricular Size Using a Roc Curve on a Coronal Slice Image," *Asian J. Inf. Technol.*, vol. 17, no. 7, 2018.

[15] B. J. Saleh, Z. Omar, V. Bhateja, and L. I. Izhar, "Auto-Lesion Segmentation with a Novel Intensity Dark Channel Prior for COVID-19 Detection," *J. Phys. Conf. Ser.*, vol. 2622, no. 1, pp. 1–9, 2023, doi: 10.1088/1742-6596/2622/1/012002.

[16] V. K. Prasad, D. Dansana, S. G. K. Patro, A. O. Salau, D. Yadav, and M. Bhavsar, "CIA-CVD: cloud based image analysis for COVID-19 vaccination distribution," *J. Cloud Comput.*, vol. 12, no. 1, 2023, doi: 10.1186/s13677-023-00539-y.

[17] R. Supriyanti, M. R. Kurniawan, Y. Ramadhani, and H. B. Widodo, "Calculating the area of white spots on the lungs of patients with COVID-19 using the Sauvola thresholding method," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, pp. 315–324, 2023, doi: 10.11591/ijece.v13i1.pp315-324.

[18] H. X. Stubblefield J, Causey J, Dale D, Qualls J, Bellis E, Fowler J, Walker K, "COVID19 Diagnosis Using Chest X-rays and Transfer Learning," *medRxiv [Preprint]*, vol. 10, no. 09, p. 22280877, 2022, doi: 10.1101/2022.10.09.22280877.

[19] B. Modi *et al.*, "Analysis of Vocal Signatures of COVID-19 in Cough Sounds: A Newer Diagnostic Approach Using Artificial Intelligence," *Cureus*, vol. 16, no. 3, pp. 1–12, 2024, doi: 10.7759/cureus.56412.

[20] H. Salarabadi, M. S. Iraji, M. Salimi, and M. Zoberi, “Improved COVID-19 Diagnosis Using a Hybrid Transfer Learning Model with Fuzzy Edge Detection on CT Scan Images,” *Adv. Fuzzy Syst.*, vol. 2024, 2024, doi: 10.1155/2024/3249929.

[21] M. K. Bohmrah and H. Kaur, “Multiclass Chest Disease Classification Using Deep CNNs with Bayesian Optimization,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, pp. 1285–1300, 2024, doi: 10.14569/IJACSA.2024.01508125.

[22] S. Antar, H. K. H. Abd El-Sattar, M. H. Abd-Rahman, and F. F. M. Ghaleb, “COVID-19 infection segmentation using hybrid deep learning and image processing techniques,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–18, 2023, doi: 10.1038/s41598-023-49337-1.

[23] S. Rojas-Zumbado *et al.*, “Upper body thermal images and associated clinical data from a pilot cohort study of COVID-19,” *BMC Res. Notes*, vol. 17, no. 1, pp. 1–5, 2024, doi: 10.1186/s13104-024-06688-w.

[24] M. R. Naidji and Z. Elberrichi, “A Novel Hybrid Vision Transformer CNN for COVID-19 Detection from ECG Images,” *Computers*, vol. 13, no. 5, 2024, doi: 10.3390/computers13050109.

[25] R. C. Gonzales and R. . Woods, *Digital image processing*, 3rd ed. New Jersey: Prentice Hall, 2008.

[26] J. Liu, Y., Wang, Y., Zhang, “New Machine Learning Algorithm: Random Forest,” *Inf. Comput. Appl.*, vol. 7473, 2012, doi: https://doi.org/10.1007/978-3-642-34062-8_32.