Comparison of Supervised Learning Methods for COVID-19 Classification on Chest X-Ray Image

Faisal Dharma Adhinata[1*], Nur Ghaniaviyanto Ramadhan[2], Arif Amrulloh[3], and Arief Rais Bahtiar[4]
[1−4]Department of Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto
Jawa Tengah 53147, Indonesia
Email: [1]faisal@ittelkom-pwt.ac.id, [2]ghani@ittelkom-pwt.ac.id, [3]amrulloh@ittelkom-pwt.ac.id, [4]ariefbahtiar@ittelkom-pwt.ac.id

***Abstract*—The Coronavirus (COVID-19) pandemic is still ongoing in almost all countries in the world. The spread of the virus is very fast because the transmission process is through air contaminated with viruses from COVID-19 patients' droplets. Several previous studies have suggested that the use of chest X-Ray images can detect the presence of this virus. Detection of COVID-19 using chest X-Ray images can use deep learning techniques, but it has the disadvantage that the training process takes too long. Therefore, the research uses machine learning techniques hoping that the accuracy results are not too different from deep learning and result in fast training time. The research evaluates three supervised learning methods, namely Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest, to detect COVID-19. The experimental results show that the accuracy of the SVM method using a polynomial kernel can reach 90% accuracy, and the training time is only 462 ms. Through these results, machine learning techniques can compensate for the results of the deep learning technique in terms of accuracy, and the training process is faster than the deep learning technique. The research provides insight into the early detection of COVID-19 patients through chest X-Ray images so that further medical treatment can be carried out immediately.**

***Index Terms*—Supervised Learning Methods, COVID-19 Classification, Chest X-Ray Images**

I. Introduction

THE COVID-19 virus was first identified in Hubei, China, at the end of 2019 [1]. This virus continues to spread in almost all countries globally until it becomes a new pandemic [2]. This virus spreads quickly because it is transmitted through air contaminated with the virus from droplets of infected patients [3]. This virus belongs to the SARS-CoV and MERS-CoV viruses that cause acute respiratory symptoms to the most severe, which cause death. Symptoms of this virus include fever, cough, sore throat, muscle pain, and shortness of breath [4].

One of the accurate techniques in diagnosing the COVID-19 virus is using the Real-time Polymerase Chain Reaction (RT-PCR) swab. Checking is done by taking mucus or fluid from the nasopharynx (the part between the nose and throat), oropharynx (the region between the mouth and throat), or the lungs of patients suspected of being infected with the COVID-19 virus. Other diagnostic techniques use Computed Tomography (CT) and X-Ray for early detection of COVID-19 [5]. Patients who experience COVID-19 symptoms for ten days will show signs of pneumonia in their lungs [6].

Several studies are related to the diagnosis of being infected with the COVID-19 virus using CT scan data of the patient's lungs. Some of these studies show that CT scans or X-Ray images of the patient's lungs are useful data for medical personnel to diagnose and monitor COVID-19 patients. For example, the trials of 81 COVID-19 patients prove that they can effectively detect COVID-19 in symptomatic patients [7]. In another study, chest X-Ray images can be used to estimate the long-term health of patients [8]. In addition, chest CT scan images can be used to monitor the condition of COVID-19 patients during the treatment process [9].

The field of Artificial Intelligence is often used to process image data. Several studies use CT scan data in processing COVID-19 patients' lungs. An example is using deep learning for COVID-19 detection. It results in an accuracy of 86.93% [10]. Then in another study [11], the detection of COVID-19 using the Deep Residual Network method produces a reasonably ac-

Received: Dec. 16, 2021; received in revised form: May 29, 2022; accepted: May 30, 2022; available online: Sep. 19, 2022.
*Corresponding Author

curate accuracy of 99%. This deep learning method is entirely accurate because the training process also takes a long time. Therefore, in the research, it analyzes the use of a method that the training process is not too long using machine learning.

Machine learning is a branch of Artificial Intelligence [12]. There are two types of machine learning: supervised and unsupervised. In the research, supervised learning is used because the data are labeled. Supervised learning methods include Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest. In other cases, these three methods yield 90% to 95% accuracy for classifying land cover [13]. Then in another previous study, the SVM, K-NN, and Random Forest methods can be used to see the satisfaction of Shopee users. The study results in an accuracy of 89% for the K-NN method, 83% for Random Forest, and 89.4% for SVM [14]. Based on previous research, these three methods produce reasonably good accuracy. Therefore, the research analyzes the accuracy and training speed of the three methods. The research is expected to provide insight into the use of machine learning techniques for early detection of COVID-19 through CT scan images of the lungs.

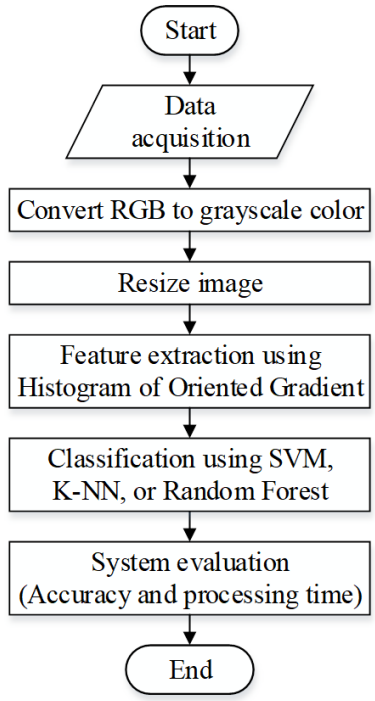

Fig. 1. Steps of research on COVID-19 detection systems.

II. Research Method

The COVID-19 disease detection system begins with acquiring CT scan data of lungs infected with COVID-19, healthy (normal) lungs, and pneumonia-infected lungs. The data obtained are pre-processed first to prepare them for the feature extraction stage. In the research, the pre-processing process is color conversion (Red, Green, Blue (RGB)) to grayscale and image resizing. Furthermore, the results of the pre-processing performed feature extraction using the Histogram of Oriented Gradient (HoG) method. The feature extraction results are used for the classification process using the supervised learning SVM, K-NN, or Random Forest methods. The research evaluates the accuracy and training speed of the three classification methods. Figure 1 shows the research stages.

A. Data Acquisition

The research uses the COVID-19 chest X-ray dataset from previous studies [15, 16]. A research team has collected this dataset from Qatar University, Qatar, and Dhaka, Bangladesh. They have collaborated with Pakistan, Malaysia, and medical personnel to create a Chest X-Ray dataset for COVID-19 patients, a healthy Chest X-Ray, and a bacterial Chest X-Ray known as pneumonia (viral pneumonia). The number of datasets used in the research is 500 in each class. The dataset format is a file with a PNG extension. The original data is still in RGB format. Then, the distribution of training data and testing data is 80% and 20%, so the number of testing data is 100 in each class. Figure 2 shows an example of the dataset used in the research.

Fig. 2. Example of the dataset in the research (a) COVID-19 (b) Normal (c) Viral Pneumonia.

B. Data Pre-Processing

Before the feature extraction process is carried out, the original dataset needs to be pre-processed to adjust the input to the feature extraction method. The research uses the HoG method as feature extraction. The HoG method uses a grayscale input image. Therefore, the original data is converted from RGB to grayscale. Then, the converted image is resized to speed up the training process. The research uses 64×64 for image resolution.

Cite this article as: F. D. Adhinata, N. G. Ramadhan, A. Amrulloh, and A. R. Bahtiar, "Comparison of Supervised Learning Methods for COVID-19 Classification on Chest X-Ray Image", CommIT Journal 16(2), 195–201, 2022.

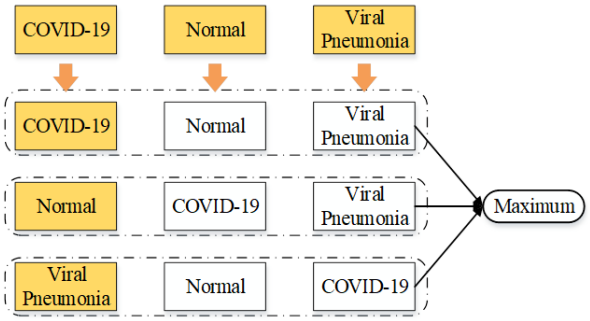

Fig. 3. Illustration of Support Vector Machine (SVM) method.

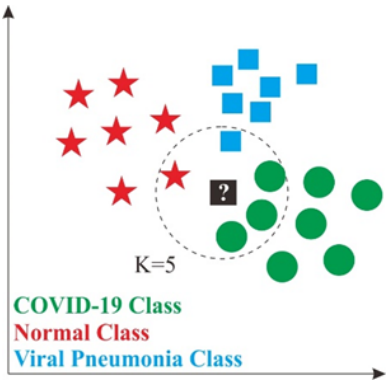

Fig. 4. Illustration of K-NN method.

C. Feature Extraction

The feature extraction used is the HoG method. Each image will have its gradient distribution. This distribution is obtained by dividing the image into small regions or cells. Each cell consists of a histogram of the gradient. The combination of histograms will be used as a characteristic of classifying the disease in the patient's lungs.

The initial stage of HoG feature extraction is to use the input image that has been converted into a grayscale. The image is calculated for the gradient value in each pixel. The following process determines the number of bins for making histograms, better known as the spatial orientation binning process. Prior to the gradient calculation process, the chest X-Ray image is divided into several cells grouped into a larger size or known as a block. The normalization process for blocks uses Rectangular Histogram of Oriented Gradient (R-HoG) for overlapping blocks [17].

D. Classification

Supervised machine learning is a technique for grouping data according to labels. In the research, images are grouped according to labels, namely COVID-19, normal, and viral pneumonia. The label classifies the data into these three lung conditions. The research uses three classification methods: SVM, K-NN, or Random Forest.

First, SVM is a supervised machine learning invented by Vapnik and his colleagues [18]. At first, SVM can only be used to classify two classes, but now it can classify more than two classes (multiclass) using the One-vs-Rest (OvR) technique [19]. This OvR stage creates one class as a positive class and another as a negative class. Then, the binary classifier is applied to each class. Figure 3 shows an illustration of the use of SVM in the research.

In the prediction of multiclass SVM, a maximum value is calculated from each class comparison [20]. Equation (1) shows the multiclass SVM formula. It shows W_{KD} as the feature vector from the generated model, X_D as the feature vector from the testing image, and s_K as class prediction results.

$$\begin{bmatrix} W_{11} & W_{12} & W_{13} & .. & W_{1D} \\ W_{21} & W_{22} & W_{23} & .. & W_{2D} \\ .. & .. & .. & .. & .. \\ W_{K1} & W_{K2} & W_{K3} & .. & W_{KD} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ .. \\ X_D \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ .. \\ s_K \end{bmatrix} \quad (1)$$

Second, the training process of K-NN uses a multidimensional vector of the feature space in each data and its label. The new feature vectors and data label classes will be matched with the stored feature and class databases in the testing phase. The selection of the K value on the K-NN dramatically affects the accuracy results. If the number of classes is even, the K value must be odd. Moreover, vice versa, if the number of classes is odd, the K value must be even [21]. Figure 4 illustrates the use of the K-NN method in the research.

If the COVID-19 class is the most common classification for the K-NN in this space, the point is designated as a COVID-19 class. Neighbors are often determined using the Euclidean distance and Eq. (2) to see whether it is near or far. It shows X_i as the training data from each class, Y_i as the testing data, and n as the number of training data.

$$dist(X,Y) = \sqrt{\sum_{i-1}^{n}(X_i - Y_i)^2} \quad (2)$$

Third, the Random Forest method is a predictive model method often used for classification [22]. Random Forest creates a random decision tree with a certain number of trees. The decision tree contains the features of each class. The decision or classification

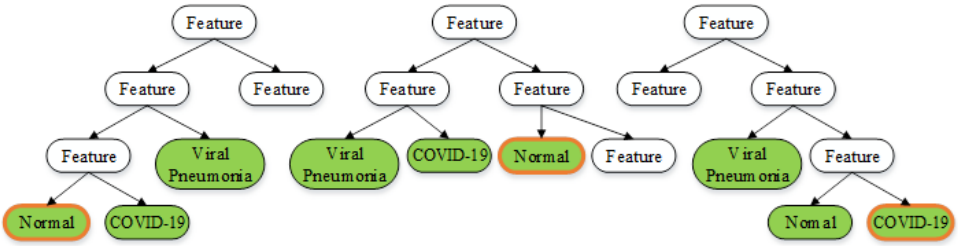

Fig. 5. Illustration of Random Forest method.

process is taken from the majority of the decision tree voting decisions. Then, Fig. 5 illustrates the use of the Random Forest method in the research.

In Fig. 5, for example, the number of trees is 3. Then, each node in the tree is a feature of each class. From the three trees, two normal classes and one COVID-19 class are obtained, so it can be concluded that the result of the Random Forest method is the normal class.

E. System Evaluation

The COVID-19 detection system is evaluated with a confusion matrix using the Python confusion_matrix library from Scikit Learn Python [23]. The evaluation matrix in the research uses recall, precision, F1-score, and accuracy. The research evaluates the parameters of the supervised machine learning, SVM, K-NN, and Random Forest methods, to get the best results. In addition, it also examines the training speed of each method.

III. Results and Discussion

The COVID-19 virus detection system through chest X-Ray images is evaluated based on the accuracy and speed of the training program. The research evaluates three supervised learning methods. There are SVM, K-NN, and Random Forest.

A. Training Data Using Support Vector Machine (SVM)

The SVM method uses a kernel that contains mathematical functions. The kernel transforms the input into a specific format, so the model output is not good when the transformation is incorrect [24]. The research conducts experiments on the three kernels: Linear, Polynomial, and RBF. Table I shows the experimental results using the SVM method.

Based on Table I, the best results are obtained using a Polynomial kernel. Image data input makes the formed features significant in number, so Linear or straight-line kernels produce the lowest accuracy compared to the other two kernels. The Polynomial kernel is also the fastest in terms of training speed, which only takes 462 ms to classify the three classes. Figure 6 shows the value of the confusion matrix using a Polynomial kernel.

TABLE I
Experiment Results Using Support Vector Machine (SVM) Method.

SVM Kernel	Accuracy	Processing time
Linear	87.00%	507 ms
Polynomial	90.00%	462 ms
RBF	89.67%	482 ms

	precision	recall	f1-score	support
1	0.85	0.88	0.86	104
2	0.87	0.84	0.85	86
3	0.97	0.97	0.97	110
accuracy			0.90	300
macro avg	0.90	0.89	0.90	300
weighted avg	0.90	0.90	0.90	300

Fig. 6. Recall, precision, F1-score, and accuracy values using Support Vector Machine (SVM).

Figure 6 shows that the recall and precision values are more than 0.8 in each class. The result makes the lowest F1-Score value 0.85. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Meanwhile, the recall ratio is the number of accurately predicted positive observations divided by the total number of observations in the actual class – yes [25]. Implementing this multiclass SVM algorithm results in high precision and recall. The high precision of SVM multiclass with a Polynomial kernel implies that it returns more relevant results than irrelevant ones. Then, high recall refers to the algorithm's ability to retrieve the most relevant results (whether irrelevant ones are also produced or not).

Cite this article as: F. D. Adhinata, N. G. Ramadhan, A. Amrulloh, and A. R. Bahtiar, "Comparison of Supervised Learning Methods for COVID-19 Classification on Chest X-Ray Image", CommIT Journal 16(2), 195–201, 2022.

TABLE II
EXPERIMENT RESULTS USING K-NEAREST NEIGHBOR (K-NN) METHOD.

SVM Kernel	Accuracy	Processing time
2	82.67%	283 ms
4	82.67%	285 ms
6	84.67%	299 ms
8	81.67%	283 ms
10	80.67%	283 ms

	precision	recall	f1-score	support
1	0.84	0.70	0.76	104
2	0.76	0.83	0.79	86
3	0.92	1.00	0.96	110
accuracy			0.85	300
macro avg	0.84	0.84	0.84	300
weighted avg	0.85	0.85	0.84	300

Fig. 7. Recall, precision, F1-score, and accuracy values using K-Nearest Neighbor (K-NN).

B. Training Data Using K-Nearest Neighbor (K-NN)

The K-NN method calculates the closest distance between the testing data and the database in the vector space. The research uses Euclidean Distance as the closest distance measurement. There are three classes in the research dataset, namely COVID-19, normal, and viral pneumonia. Therefore, the research uses an even K value experiment. Table II shows the experimental results using the K-NN method.

Based on Table II, the best results are obtained using the K = 6, with an accuracy of 84.67% and a training speed of 299 ms. In this experiment, the training speed does not change significantly with every change in the K value. Then, the results of the calculation of the confusion matrix, namely recall, precision, F1-score, and accuracy, using a value of K = 6, are shown in Fig. 7.

Figure 7 shows a recall value of only 0.7, while the precision value is 0.84. It is a high precision but a low recall system that returns very few results. However, most of its predicted labels are correct compared to the training labels. The result also affects the F1-score value, which produces a value below 0.8.

C. Training Data Using Random Forest

Classification using Random Forest produces several decision trees used to make predictions. A vote is made for the final prediction result from all the resulting decision trees. Therefore, an experiment is carried out based on the number of trees produced in the research. Table III shows the experimental results using the Random Forest method.

TABLE III
EXPERIMENT RESULTS USING RANDOM FOREST METHOD.

Number of trees	Accuracy	Processing time
20	86.00%	301 ms
40	85.67%	394 ms
60	86.33%	624 ms
80	86.33%	806 ms
100	87.00%	924 ms

	precision	recall	f1-score	support
1	0.81	0.84	0.82	104
2	0.82	0.79	0.80	86
3	0.96	0.96	0.96	110
accuracy			0.87	300
macro avg	0.87	0.86	0.86	300
weighted avg	0.87	0.87	0.87	300

Fig. 8. Recall, precision, F1-score, and accuracy values using Random Forest.

Based on Table III, the best results are achieved with 100 trees. However, the more trees there are, the slower the speed will be. On the other hand, the higher the number of trees is, the higher the accuracy value will be. Then, the results of the calculation of the confusion matrix results, namely recall, precision, F1-score, and accuracy using 100 trees, are shown in Fig. 8.

The experimental results using 100 trees in the Random Forest show quite good results. The values of precision and recall in each class are not much different. The result also affects the F1-score value. However, the result is still less than using the multiclass SVM method.

The experimental results using 100 trees in the Random Forest show quite good results. The values of precision and recall in each class are not much different. The result also affects the F1-score value. However, the result is still less than using the multiclass SVM method.

D. Analysis and Discussion

The research uses three supervised learning methods to classify cases of COVID-19 disease through patients' chest X-Ray images. The supervised machine learning method is used for the classification process that already has a label. Based on the experimental results, the SVM method with a Polynomial kernel produces the best accuracy compared to the K-NN and Random Forest methods, with an accuracy of 90%. However, the fastest method in the training process is K-NN. The training speed does not exceed 300 ms. In general, using the SVM method with a Polynomial kernel is also better than previous studies using deep learning [10], which results in an accuracy of 86.93%.

TABLE IV
EXPERIMENT RESULTS USING DEEP LEARNING MODEL.

Deep learning model	Accuracy	Processing time
ResNet50	77.67%	12 min 16 s
DenseNet201	96.34%	22 min 19 s
EfficientNetB7	33.42%	42 min 57 s
CNN	89.33%	6 min 39 s

The research is proven to compensate for deep learning techniques that produce 90% accuracy with a training speed of 462 ms.

The research also compares various deep learning methods. It tests the ResNet50, DenseNet201, EfficientNetB7, and CNN models. The test results are shown in Table IV. From the four models, all training times are more than 1 minute, in contrast to the machine learning approach with under 1 second in the training process. In terms of accuracy, the best model is obtained using the DenseNet201 model with an accuracy of 96.34%. The result is not much different from the machine learning approach, which can still produce 90% with a faster training process. The machine learning method can produce quite an optimal accuracy because the amount of data used is not too much, only 1,500 chest X-Ray image data.

IV. CONCLUSION

The COVID-19 pandemic is still not over. Several new variants of mutations even have appeared. World scientists have developed many rapid detection results from patients infected with this virus. One of the detections of COVID-19 infection is through chest X-Ray image data. Many researchers have developed this detection method using deep learning. The research evaluates the case study using a machine learning approach to achieve optimum accuracy and a fast-training process.

Based on the results of experiments and research discussions, it is concluded that the method with the best accuracy is SVM with a polynomial kernel, with an accuracy of 90%. Then, the method that produces the fastest time in the training process is the K-NN method, but the SVM method does not have too much difference in speed. Through the results of this study, machine learning methods can achieve optimal accuracy compared to deep learning because the amount of data is not too much.

Even though the resulting accuracy has reached 90%, the result is still far from 100% accuracy. It indicates that there are still detection errors. When viewed in more detail, the recall and precision values in the COVID-19 and normal classes are still below 90%. Therefore, future research is suggested to combine feature extraction and other classification methods to achieve better accuracy. In the feature extraction method, it can use layers in deep learning. Then, for classification, it can use supervised learning.

ACKNOWLEDGEMENT

The research was supported by a grant from Institut Teknologi Telkom Purwokerto in 2022. The authors are indebted to the Lembaga Penelitian dan Pengabdian Masyarakat at the Institut Teknologi Telkom Purwokerto which provided a grant to assist with the research.

REFERENCES

[1] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. W. On, W. Aslam, and G. S. Choi, "COVID-19 future forecasting using supervised machine learning models," *IEEE Access*, vol. 8, pp. 101 489–101 499, 2020.

[2] F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei *et al.*, "A new coronavirus associated with human respiratory disease in china," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020.

[3] N. H. L. Leung, "Transmissibility and transmission of respiratory viruses," *Nature Reviews Microbiology*, vol. 19, pp. 528–545, 2021.

[4] CDC, "Symptoms of COVID-19," 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[5] R. Shrestha and L. Shrestha, "Coronavirus disease 2019 (COVID-19): A pediatric perspective," *JNMA: Journal of the Nepal Medical Association*, vol. 58, no. 227, pp. 525–532, 2020.

[6] F. Pan, T. Ye, P. Sun, S. Gui, B. Liang, L. Li, D. Zheng, J. Wang, R. L. Hesketh, L. Yang, and C. Zheng, "Time course of lung changes at chest CT during recovery from Coronavirus disease 2019 (COVID-19)," *Radiology*, vol. 295, no. 3, pp. 715–721, 2020.

[7] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: A descriptive study," *The Lancet Infectious Diseases*, vol. 20, no. 4, pp. 425–434, 2020.

[8] R. Yasin and W. Gouda, "Chest X-ray findings monitoring COVID-19 disease course and severity," *Egyptian Journal of Radiology and Nuclear Medicine*, vol. 51, no. 1, pp. 1–18, 2020.

[9] S. H. Yoon, K. H. Lee, J. Y. Kim, Y. K. Lee, H. Ko, K. H. Kim, C. M. Park, and Y. H. Kim, "Chest radiographic and CT findings of the 2019

novel Coronavirus disease (COVID-19): analysis of nine patients treated in Korea," *Korean Journal of Radiology*, vol. 21, no. 4, pp. 494–500, 2020.

[10] N. Yudistira, A. W. Widodo, and B. Rahayudi, "Deteksi Covid-19 pada citra sinar-x dada menggunakan deep learning yang efisien," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, no. 6, pp. 1289–1296, 2020.

[11] Y. S. Hariyani, S. Hadiyoso, and T. S. Siadari, "Deteksi penyakit covid-19 berdasarkan citra x-ray menggunakan deep residual network," *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 8, no. 2, pp. 443–453, 2020.

[12] A. U. Zailani, A. Perdananto, N. Nurjaya, and Sholihin, "Pengenalan sejak dini siswa smp tentang machine learning untuk klasifikasi gambar dalam menghadapi revolusi 4.0," *KOMMAS: Jurnal Pengabdian Kepada Masyarakat*, vol. 1, no. 1, pp. 7–15, 2020.

[13] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, pp. 1–20, 2017.

[14] S. Watmah, S. Suryanto, and M. Martias, "Komparasi metode K-NN, support vector machine dan random forest pada e-commerce Shopee," *INSANtek*, vol. 2, no. 1, pp. 15–21, 2021.

[15] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi, R. M. B. I., and M. T. Islam, "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132 665–132 676, 2020.

[16] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, pp. 1–16, 2021.

[17] R. Y. Endra, A. Cucus, F. N. Afandi, and M. B. Syahputra, "Deteksi objek menggunakan Histogram Of Oriented Gradient (HOG) untuk model smart room," *Explore: Jurnal Sistem informasi dan telematika (Telekomunikasi, Multimedia dan Informatika)*, vol. 9, no. 2, pp. 99–105, 2018.

[18] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[19] K. B. Duan, J. C. Rajapakse, and M. N. Nguyen, "One-versus-one and one-versus-all multiclass SVM-RFE for gene selection in cancer classification," in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Valencia, Spain: Springer, April 11–13, 2007, pp. 47–56.

[20] F. D. Adhinata, A. Harjoko, and Wahyono, "Object searching on video using ORB descriptor and support vector machine," in *International Conference on Computational Collective Intelligence*. Da Nang, Vietnam: Springer, Nov. 30–Dec. 3, 2020, pp. 239–251.

[21] D. Kurniawan and A. Saputra, "Penerapan K-Nearest Neighbour dalam penerimaan peserta didik dengan sistem zonasi," *Jurnal Sistem Informasi Bisnis*, vol. 9, no. 2, pp. 212–219, 2019.

[22] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water*, vol. 11, no. 5, pp. 1–37, 2019.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] M. Awad and R. Khanna, *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer Nature, 2015.

[25] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, pp. 1–26, 2020.