The Feature Extraction for Coronavirus Disease Classification Techniques

Anis Fitri Nur Masruriyah (✉ anis.masruriyah@ubpkarawang.ac.id)

University of Buana Perjuangan <https://orcid.org/0000-0002-2449-4426>

Hasan Basri

BRI Institute

Hanny Hikmayanti Handayani

University of Buana Perjuangan Karawang

Ahmad Fauzi

University of Buana Perjuangan Karawang

Research

Keywords: Coronavirus disease, Digital image processing, Discrete wavelet transform, Feature extraction

Posted Date: July 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-731100/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

The Feature Extraction for Coronavirus Disease Classification Techniques

Anis Fitri Nur Masruriyah¹, Hasan Basri², Hanny Hikmayanti Handayani³, Ahmad Fauzi⁴

^{1,3,4}Department of Informatics, University of Buana Perjuangan Karawang, 41361, Indonesia

²Department of Information System, BRI Institute, 12550, Indonesia

*Correspondence: anis.masruriyah@ubpkarawang.ac.id

Abstracts - COVID-19 has been an epidemic since the end of 2019. The number of patients with COVID-19 continues to escalate until new variants emerge. The COVID-19 detection procedure begins with detecting early symptoms, furthermore confirmed by the swab and CXR methods. The process of swab and CXR takes a relatively long time since in CXR some patients have the same symptoms as pneumonia. This study carried out the classification of COVID-19 and not COVID-19 with feature extraction techniques and classification methods. The result of this study capable to identify CXR with COVID-19 and an accuracy of 96.5%. In addition, this study even compares the classification results without using feature extraction techniques. The comparison result showed that feature extraction was able to significantly improve accuracy.

Keyword:

Coronavirus disease, Digital image processing, Discrete wavelet transform, Feature extraction

1. INTRODUCTION

Coronavirus or which is commonly known as COVID-19 initially occurred in the city of Wuhan, China at the end of 2019 [1]–[3]. The initial symptom of COVID-19 is a reduced function of the sense of smell and taste. In certain patients, there are those who experience shortness of breath and dry cough and the symptoms are similar to pneumonia [4]–[7]. The basic discrepancy between COVID-19 and pneumonia is when someone who has COVID-19 experiences symptoms in the form of fever, dry cough, and fatigue at an early stage. In addition, patients may also experience nausea, diarrhea, muscle aches, and vomiting. However, if the infection has already induced pneumonia, the patient may undergo a faster heart rate, shortness of breath, rapid and short breaths, and sweating a lot. [3], [8], [9]. Furthermore, pneumonia generally has several symptoms that occur, which are the bluish appearance of the lips and nails, experiencing delirium, coughing that produces mucus, and severe chest pain, especially when coughing. Even so, the most visible thing from the difference in pneumonia symptoms, in COVID-19 at the beginning of the attack, is that the cough does not produce phlegm. Hereafter, so that the difference can be seen clearly, it is necessary to use a swab and Chest X-Ray (CXR) technique.

Several studies have proven that in addition to the swab method, the results of a chest scan using X-Ray were more effective in helping the treatment of COVID-19. Previously, Jacobi *et al.* [10] elucidated that the occupy of Chest X-Ray (CXR) was able to show abnormal formations or various chest diseases such as COVID-19, pneumonia, cystic fibrosis,

emphysema, cancer, and more. The study described the most common manifestations and patterns of pulmonary abnormalities in CXR in COVID-19 in order to equip the medical community in its efforts to cope with this pandemic. Furthermore, using computational techniques several studies have proven the success rate in handling COVID-19 [11]–[13].

In a study conducted by Asnaoui and Chawki [11] defined that the comparison of the inception_Resnet_V2 and Densnet201 algorithms gave better results compared to other models used in the study. Accuracy results reached 92.18% for Inception-ResNetV2 and 88.09% accuracy for Densnet201 in detecting and classifying coronavirus pneumonia. Experiments were performed using a chest CXR dataset of 6087 images. Then, the confusion matrix was deployed to evaluate the performance of the model. Moreover, Narin *et al.* [12] employed CXR results of COVID-19 patients were processed applying three different binary classifications with four classes (COVID-19, normal (healthy), viral pneumonia, and bacterial pneumonia). Later on, using 5-fold cross-validation. It was found that the pre-trained ResNet50 model gave the highest classification performance (96.1% accuracy for Dataset-1, 99.5% accuracy for Dataset-2 and 99.7% accuracy for Dataset-3) among the other four models employed. In addition, Ucar and Korkmaz [13] in their research compared the proposed method with other studies. The method proposed in the study was SqueezeNet which was designed with lightweight tissue and tuned for the diagnosis of COVID-19 with Bayesian optimization additives. The result of the study was that the proposed method performed felicitous than the existing network design and obtained a higher accuracy of COVID-19 diagnosis.

Published research described algorithms capable of detecting and classifying COVID-19, but no one has yet explained feature extraction. Several studies explained that feature extraction capable to improve algorithm performance and made identification results more accurate [14]–[17]. Khalil [14] amplified that the feature extraction technique can improve the accuracy of the MR-Brain image classification. The simulation results in this study involving feature extraction were able to achieve accuracy results of up to 99.37%. Furthermore, the research conducted by Santos *et al.* [15] proposed a two feature extraction approach to detect wafer test patterns. The first approach was classical image processing and restoration combined with feature engineering. Then, the second approach was a data-based deep generative model. The two approaches were then evaluated on the synthetic and real-world datasets. The evaluation metric of the approach taken showed results that were close to the same. This research also proved the reliability of feature extraction because it was able to reduce power and space when classifying. Moreover, Liu *et al.* [16] introduced the Flexible Unsupervised Feature Extraction (FUFE) technique and it was proven that the model made with the FUFE implementation showed better effectiveness. The research also verified that FUFE was able to effectively characterize local and global geometric structures in images.

There were also reviews regarding feature extraction methods that have a major influence on image classification [17]. The scientific paper explained that the most widely used feature extraction types were Gray Level Co-Occurrence Matrix (GLCM) and Local Binary Pattern (LBP). However, it must still pay attention to the type of case to determine the appropriate feature extraction technique. So this study aimed to prove that feature extraction was able to make the classification more optimal. The feature extraction method Discrete Wavelet Transform (DWT) was chosen because it is proven to be able to reduce noise in the image, so that the characteristics of the image used capable to be identified properly [18].

The details of the discussion of this scientific paper are explained in sequence. In part II, DWT is discussed in detail along with its advantages. Furthermore, section III describes the implementation of DWT in this study. This is followed in part IV which discusses the results of the DWT implementation. Conclusions and suggestions are explained in section V.

2. METHOD

2.1. Data Description

In this study, data were obtained from [19] which consisted of 5910 CXR images of normal patients and patients with COVID-19. In Figure 1 Normal lungs showed a black or dark colour on CXR results. Meanwhile, when the lungs start to look white, it is a sign that they have begun to be covered by fluid or other damage.

2.2 Preprocess

Preprocessing was done to perform data selection as needed, some of the unused data samples are shown in Figure 2. When collecting medical related data it may be time consuming and expensive. To overcome this kind of difficulty, augmentation can be applied. Augmentation can overcome the problem of over-fitting and improve the accuracy of the proposed model.

Furthermore, class separation is carried out for normal images and COVID-19 images. Some image data are not used because the angle of image capture is not uniform. Then solve the noise problem contained in the image. Noise can be an obstacle in feature extraction, if the amount is too much, the noise will be captured and considered as information from the image. This is not good in feature extraction, several ways to deal with this problem are by removing noise or cropping it and changing the grayscale image.

Figure 1. CXR Normal (a) dan CXR Covid

Figure 2. Unselected Image Sample

2.3 Feature Extraction using Discrete Wavelet Transform (DWT)

Feature extraction was intended to find significant feature areas in the image depending on the intrinsic characteristics and the application. This research uses DWT as a feature extraction method. Wavelets are small waves that are able to group image energy and are concentrated in a small group of coefficients [18], [20]. On the other hand, the other coefficient groups contain only a small amount of energy that can be removed without reducing the information value. In 1982, Jean Morlet came up with the idea of the wavelet transform which provided a new mathematical tool for seismic wave analysis [21]–[23]. Morlet first introduced the idea of a wavelet as a subset of functions constructed from the translation and dilation of a single function called a "mother wavelet" with the notation $\Psi(t)$ according to Equation 1.

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \Psi\left(\frac{t-b}{a}\right), \quad a, b \in R, a \neq 0, \quad (1)$$

In Equation 1 'a' is a scale parameter that measures the degree of compression (scale). Furthermore, the parameter 'b' is a translation of the wavelet time location.

Based on equations 1 to 5, how the DWT works was shown in the Algorithm 1.

Algorithm 1: Discrete Wavelet Transform

Input: CXR COVID-19

Algorithm:

```

1 begin
2   Read images
3   Perform signal transformation through HPF and LPH.
4   Convert images into multi-resolution using iteration technique on approximation coefficient.
5   while true do
      if appropriate resolution level then ok
      else repeat
    end

```

Output: The coefficients Value

The DWT method performs signal transformation through a high-pass filter (HPF) and a low-pass filter (LPF) [18], [20], [21], [23], [24]. Furthermore, the data is converted into multi-resolution using iteration techniques on approximating coefficients. The iteration is carried out continuously until it reaches the appropriate resolution level. A detailed illustration of the DWT decomposition process is shown in Figure 1. Symbol $\downarrow 2$ shows down sampling by a factor of 2 and produces low-low (LL), low-high (LH), high-low (HL) and high-high (HH).

Based on Figure 3 way to produce a number of images that match the response of the filter bank, moreover the decomposition process produces four different images. Further illustration of the Level 1 decomposition process is shown in Figure 4.

Figure 3. Decomposition Discrete Wavelet Transform

Figure 4. Decomposition Discrete Wavelet Transform

In the level 1 decomposition process, LL is the result (approximate) using the scaling function ($\varphi(x, y)$). Next, LH is the vertical detail using the wavelet function ($\Psi^H(x, y)$), then HL is the horizontal detail using $\Psi^V(x, y)$ and HH is the diagonal detail using $\Psi^D(x, y)$. After that, the next level decomposition process only uses approximations as input values. The Scaling Function and Wavelet Function equations are shown in equations 2 and 3.

$$\varphi_{j,m,n}(x, y) = 2^{\frac{1}{2}}\varphi(2^j x - m, 2^j y - n) \quad (2)$$

$$\Psi^H(x, y) = 2^{\frac{1}{2}}\Psi^i(2^j x - m, 2^j y - n), i = V, D, H \quad (3)$$

Then, information from each subgroup of wavelet decomposition will be calculated using wavelet energy (Equation 4) and Shannon entropy (Equation 5) as feature space for prediction [20], [22], [24].

$$E = \sum_{n=1}^N |f(n)| \quad (4)$$

$$S = - \int_{n=1}^G h_n \log_2(h_n) \quad (5)$$

In Equation 4 energy is represented by the symbol E, then N is the total data and the function to get the data at position n is indicated by the symbol f(n). Furthermore, n in Equation 5 is the graylevel value of several subbands. Then h_n represents the nth probability value of graylevel and G is the total value of graylevel.

The coefficient values obtained from algorithm 1 were used to calculate energy and entropy, respectively. The results of the data distribution of entropy and energy values were shown in Figure 5.

Figure 5. Distribution of Entropy and Energy Values

In Figure 5, outliers appear so outlier analysis is needed to remove outliers that are too far away. The results of the analysis are shown in Figure 6 with better details of the distribution.

Figure 6. Distribution Details after Removing Outliers

3. RESULTS AND DISCUSSION

After implementing the DWT, then classification was carried out using the Support Vector Machine (SVM) algorithm. The SVM classifier was used to differentiate normal and COVID-19 positive CXR from the collected CXR images. In order for the effectiveness of feature extraction to be well known, it needs to be evaluated. The evaluation method uses data that is divided into training data and test data with K-Fold Cross Validation. In this study 10-folds cross validation was chosen because based on [25], [26] 10-folds cross validation got the best error estimate. Then, the calculation of accuracy (Equation 5) uses the confusion matrix from Table 1 to get the evaluation results.

$$\frac{\text{True Positive} + \text{True Negative}}{\text{Sum All Matrix}} \quad (5)$$

Table 1. Confusion Matrix

		Positive	Negative
Predict	Positive	8 True Positive	5 False Negative
	Negative	0 False Positive	151 True Negative

The results of the identification using the SVM method with the RBF kernel reached 96.5%. In Table II, the results of the comparison of algorithms using DWT and without DWT appear significant. In general, the accuracy results increase if the DWT is applied first. However, in the SVM algorithm the increase in accuracy shows a significant number. While in the Neural Network algorithm, the increase occurred but not as much as in other algorithms. This means that this feature extraction is able to help identify COVID-19 exposure in CXR images more efficiently.

Classification of CXR images using the SVM algorithm is shown in Figure 7. The results of this classification are based on two parameters, namely energy and entropy. In Figure 7 the number 0 shows a CXR image identified as infected with COVID-19, while the number 1 indicates a healthy CXR images.

Table 2. Result Comparassin

Algorithm	Accuracy (%) without Discrete Wavelet Transform	Accuracy (%) with Discrete Wavelet Transform
SVM	51.2	96.5

Decision Tree	53.3	83.5
Neural Network	74.7	95.3

Figure 7. SVM Classification Results.

4. CONCLUSION

This study utilizes CXR images of patients with COVID-19 and proves that the feature extraction method is able to improve the classification results of patients with COVID-19. Several studies also explain that feature extraction provides convenience in digital image processing. The results showed that the accuracy increased significantly in the digital image processing using the discrete wavelet transform technique. The improvement is significantly different in the SVM algorithm, where the accuracy without the application of discrete wavelet transform is 51.2%. Then the accuracy increased to 96.5% after DWT was applied.

Furthermore, comprehensive data is needed that displays CXR results and other medical data. It is suggested to be able to help perform the analysis and treatment of patients appropriately.

5. Availability of data and materials

The datasets used and/or analysed during current study are available from the corresponding author on reasonable request.

6. Ethics Declaration

Ethics approval and consent to participate

Not applicable.

Consent for Publication

Not applicable.

Competing Interest

The authors declare that they have no competing interest

7. Funding

See Acknowledgements.

8. Author's Contributions

All authors read and approved the final manuscript. AM carried out the idea and implementation of this research as well as interpretation of results. HB, HHH and AF made substantial contributions to the conception and design of this research as well as critically reviewing and interpreting results.

9. Acknowledgements

The Authors would like to thank the research and community service institutions of Universitas Buana Perjuangan Karawang and BRI Institute which have provided support and assistance for the implementation of this research. The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

10. REFERENCES

- [1] D. Kumar, "Corona Virus: A Review of COVID-19," *Eurasian J. Med. Oncol.*, no. March, 2020, doi: 10.14744/ejmo.2020.51418.
- [2] M. Khan *et al.*, "COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far," *Molecules*, vol. 26, no. 1, pp. 1–25, 2020, doi: 10.3390/molecules26010039.
- [3] World Health Organization, "Coronavirus disease 2019," *World Heal. Organ. Rep.*, vol. 169, no. 4, pp. 308–311, 2021, doi: 10.1007/s00112-021-01158-0.
- [4] E. Chuang, P. A. Cuartas, T. Powell, and M. N. Gong, "'We're Not Ready, But I Don't Think You're Ever Ready.' Clinician Perspectives on Implementation of Crisis Standards of Care," *AJOB Empir. Bioeth.*, vol. 11, no. 3, pp. 148–159, 2020, doi: 10.1080/23294515.2020.1759731.
- [5] T. Aven and F. Boudier, "The COVID-19 pandemic: how can risk science help?," *J. Risk Res.*, vol. 23, no. 7–8, pp. 849–854, 2020, doi: 10.1080/13669877.2020.1756383.
- [6] S. Setiati and M. K. Azwar, "COVID-19 and Indonesia," no. April, 2020.
- [7] J. E. Hollander and B. G. Carr, "History in a Crisis - Lesson for Covid-19," *N. Engl. J. Med.*, vol. 382, no. 18, pp. 1679–1681, 2020, doi: 10.1056/NEJMp2003539.
- [8] M. Allam *et al.*, "COVID-19 diagnostics, tools, and prevention," *Diagnostics*, vol. 10, no. 6, pp. 1–33, 2020, doi: 10.3390/diagnostics10060409.
- [9] K. M. Gostic, A. C. R. Gomez, R. O. Mummah, A. J. Kucharski, and J. O. Lloyd-Smith, "Estimated effectiveness of symptom and risk screening to prevent the spread of

COVID-19," *Elife*, vol. 9, pp. 1–18, 2020, doi: 10.7554/eLife.55570.

- [10] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, "Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review," *Clin. Imaging*, vol. 64, no. April, pp. 35–42, 2020, doi: 10.1016/j.clinimag.2020.04.001.
- [11] K. El Asnaoui and Y. Chawki, "Using X-ray images and deep learning for automated detection of coronavirus disease," *J. Biomol. Struct. Dyn.*, vol. 0, no. 0, pp. 1–12, 2020, doi: 10.1080/07391102.2020.1767212.
- [12] A. Narin, C. Kaya, and Z. Pamuk, "Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks," *arXiv Prepr. arXiv2003.10849*, 2020, [Online]. Available: <https://arxiv.org/abs/2003.10849>.
- [13] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Med. Hypotheses*, no. January, 2020.
- [14] M. Khalil, H. Ayad, and A. Adib, "Performance evaluation of feature extraction techniques in MR-Brain image classification system," *Procedia Comput. Sci.*, vol. 127, pp. 218–225, 2018, doi: 10.1016/j.procs.2018.01.117.
- [15] T. Santos *et al.*, "Feature Extraction from Analog Wafermaps: A Comparison of Classical Image Processing and a Deep Generative Model," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 2, pp. 190–198, 2019, doi: 10.1109/TSM.2019.2911061.
- [16] Y. Liu, F. Nie, Q. Gao, X. Gao, J. Han, and L. Shao, "Flexible unsupervised feature extraction for image classification," *Neural Networks*, vol. 115, pp. 65–71, 2019, doi: 10.1016/j.neunet.2019.03.008.
- [17] A. Humeau-Heurtier, "Texture feature extraction methods: A survey," *IEEE Access*, vol. 7, pp. 8975–9000, 2019, doi: 10.1109/ACCESS.2018.2890743.
- [18] H. Basri and R. Widyastuti, "Image Extraction of Merbau Wood and Jati Bojonegoro Wood using Discrete Wavelet Transform," *BRITech (Jurnal Ilm. Ilmu Kpmputer, Sains dan Teknol. Ter.)*, vol. 1, no. 2, 2020.
- [19] P. Govindaraj, "Classify the X Ray image which is having Corona," 2020. <https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset> (accessed Dec. 20, 2020).
- [20] F. F. Tampinongkol, Y. Herdiyeni, and E. N. Herliyana, "Feature extraction of Jabon (*Anthocephalus* sp) leaf disease using discrete wavelet transform," *Telkomnika (Telecommunication Comput. Electron. Control.)*, vol. 18, no. 2, pp. 740–751, 2020, doi: 10.12928/TELKOMNIKA.V18I2.10714.
- [21] P. J. Van Fleet, *Discrete Wavelet Transformations: An Elementary Approach with Applications*. 2011.
- [22] L. Debnath and F. A. Shah, *Wavelet transforms and their applications, second edition*. 2015.
- [23] D. Sundararajan, *Discrete Wavelet Transform: A Signal Processing Approach*. 2015.
- [24] S. Widiyanto, Y. Sukra, S. Madenda, D. T. Wardani, and E. P. Wibowo, "Texture

feature extraction based on GLCM and DWT for beef tenderness classification,” *Proc. 3rd Int. Conf. Informatics Comput. ICIC 2018*, pp. 1–4, 2018, doi: 10.1109/IAC.2018.8780569.

- [25] T. Djatna, M. K. D. Hardhienata, and A. F. N. Masruriyah, “An intuitionistic fuzzy diagnosis analytics for stroke disease,” *J. Big Data*, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0142-7.
- [26] A. F. N. Masruriyah, T. Djatna, M. K. D. Hardhienata, H. H. Handayani, and D. Wahiddin, “Predictive Analytics For Stroke Disease,” pp. 13–16, 2019.