Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

Classification of COVID-19 Cough Sounds using Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction and Support Vector Machine

Muhammad Meftah Mafazy[1], Mohammad Reza Faisal[2], Dwi Kartini[3], Fatma Indriani[4], Triando Hamonangan Saragih[5]

[1,2,3,4,5]Computer Science, Faculty Mathematics and Natural Science,
Lambung Mangkurat University, Banjarmasin, Indonesia
E-mail: meftah.mafazy@gmail.com[1], reza.faisal@ulm.ac.id[2*], dwikartini@ulm.ac.id[3], f.indriani@ulm.ac.id[4], triando.saragih@ulm.ac.id[5]

ARTICLE INFO

History of the article:
Received July 13, 2023
Revised August 24, 2023
Accepted August 29, 2023

Keywords:
COVID-19
Feature Extraction
Classification
Audio Cough
SVM

Correspondece:
E-mail: reza.faisal@ulm.ac.id

ABSTRACT

Many studies have been conducted to detect COVID-19, such as swabs, rapid antigens, and using X-ray images. However, these methods have the disadvantage of requiring sampling through physical contact with the patient. One way to avoid physical contact is to use audio through coughing with the aim of reducing COVID-19 transmission. Audio feature extraction such as Mel Frequency Cepstral Coefficient (MFCC) has often been used in audio classification research, such as music genre classification and so on. This study aims to compare the performance of audio feature classification through cough sounds for early detection of COVID-19 using Linear-based Support Vector Machine and Radial Basis Function (RBF). The dataset used is the COVID-19 cough audio dataset, before being classified, the audio data is processed into spectrograms and then feature extraction is carried out. From the research results, the highest AUC is 0.572 in linear kernel-based SVM classification when using parameter C = 0,5. Meanwhile, when using the RBF kernel, the highest AUC is 0.560 when using parameter C = 1.

INTRODUCTION

Coronavirus 2019 (COVID-19) is a new disease symptom that has never been found to infect humans before. The virus that causes COVID-19 is Sars-CoV-2 (Putri, 2020). Currently, many studies have conducted COVID-19 detection, several ways to detect are swabs, rapid antigens, and using x-ray media. this collection of methods has the disadvantage of requiring physical contact sampling with the patient (Yanti, Ismida, & Sarah, 2020). Like the research conducted by (Nugroho, 2021) namely classifying x-ray images using the KNN algorithm with Haralick Features & Histogram of Oriented Gradient feature extraction, the average accuracy is above 90% with an estimated k value of 10.

One of the symptoms of COVID-19 is coughing so that through coughing sounds can be done for early detection of COVID-19. The purpose of early detection of COVID-19 is to reduce the spread and physical contact between patients and medical personnel. Classification of cough through tuberculosis screening with audio data using MFCC and ZCR feature extraction and the Logistic Regression algorithm obtained an AUC result of 0.86, then Sequential Forward Selection (SFS) feature selection was carried out to get an increase in AUC results of 0.94 (Pahar et al., 2021). Then, research conducted by (Ritwik, Kalluri,

& Vijayasenan, 2021) detected COVID-19 using SVM and MFCC feature extraction, this study used the DiCOVA dataset, obtained an AUC of 0.734. In addition, there is also research conducted by (Han et al., 2021) diagnose COVID-19 through sound and other symptoms using SVM and Zero Crossing Rate feature extraction, this study uses data from the "COVID-19 Sounds App" application, after preprocessing the data, 828 audio data were obtained, then an AUC of 0.79 was obtained.

MFCC feature extraction is commonly used in research to detect COVID-19 through audio recordings, such as in the following study (Bansal, Pahwa, & Kannan, 2020) applied a convolutional neural network (CNN) to perform MFCC functions to detect COVID-19, with a model accuracy of around 70.58%. However, these results are considered weak compared to research (Södergren, Nodeh, Chhipa, Nikolaidou, & Kovács, 2021) who used Random Forest and Support Vector Machine (SVM) algorithms on the DiCOVA dataset with 1040 data including MFCC feature extraction for COVID-19 detection using recorded cough audio. Improvised, SVM achieved an AUC of 85.21%, better than the previous result of 85.05%. SVM is popular in classifying COVID-19 symptoms (Chowdhury, Kabir, Rahman, & Islam, 2022) thanks to its ability to separate data layers with the best hyperplane. This approach is not only applicable to COVID-19, such as research (Danika, Raharjo, & Hidayat, 2022) which tried to detect guitar sounds using SVM, obtained a stable accuracy of 95% for the RBF kernel and while the linear-based SVM only got an accuracy of 88.75%.

Cross-validation is a data resampling method that aims to assess the generalization ability of predictive models and prevent overfitting (Berrar, 2019). K-fold cross validation is often used in COVID-19 classification research through cough audio as done by (Tena, Clarià, & Solsona, 2022) classifying COVID-19 through sound detection using multiclass SVM and 10-fold cross validation obtained an AUC of 97.29%.

Based on the explanation described above, this research wants to determine the comparison of SVM performance performance using linear kernels and RBF on COVID-19 cough sound classification using MFCC feature extracts with 10 coefficients. The parameters that will be used are based on research conducted by (Chowdhury et al., 2022) with a Gamma parameter value of 0.062 and a parameter value c multiples of 0.1 to 1. The main purpose of this comparison is to determine which kernel is most suitable for this classification task that can improve the performance of the SVM in the classification of COVID-19 cough sounds.

RESEARCH METHODS

There are research methods that will be carried out to ensure the smooth running of the research in line with the objectives, as well as to determine the SVM kernel that has the best performance in cough classification for COVID-19 detection. The flow of research methods can be seen in Figure 1.

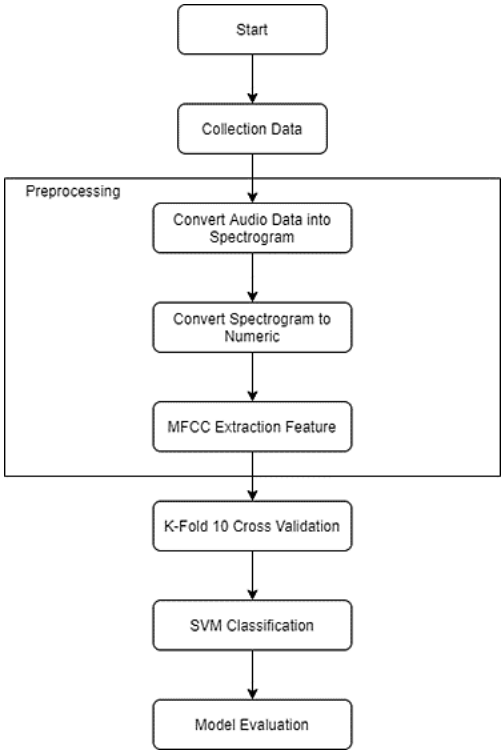

Figure 1. Research Flow

1. Data Collection

Data collection was carried out from the Kaggle site under the name COVID-19 Cough with a total dataset of 1926 audio data labeled Healthy and COVID-19 with a ratio of 1284:642 data. Examples of sound wave graphs labeled Healthy and COVID-19 can be seen in Figure 2 and Figure 3.

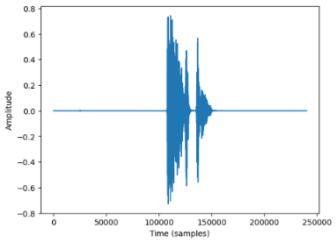

Figure 2. Cough audio labeled Healthy

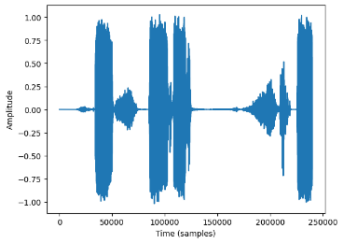

Figure 3. Audio cough labeled COVID-19

The list of initial labeled datasets can be seen in Table 1.

Table 1. Initial dataset

No	filename	label
1	00039425-7f3a-42aa-ac13-834aaa2b6b92.wav	healthy
2	0009eb28-d8be-4dc1-92bb-907e53bc5c7a.wav	healthy
....	....	....
1925	ff8363d2-016d-4738-9499-4c62480886fb.wav	COVID-19
1926	ffe0658f-bade-4654-ad79-40a468aabb03.wav	COVID-19

2. Data Preprocessing

Preprocessing is a process that makes it easier to produce values from feature extraction (Solehudin, 2018). The part that the author does is converting raw audio data into 2-dimensional data, namely

spectrograms, then MFCC feature extraction using the librosa library in the python programming language (McFee et al., 2023). The following are the stages of preprocessing audio data:

2.1 Audio Data Conversion

Preprocessing that will be done in this research is to convert audio data into visual data in the form of a spectrogram, a spectrogram is a visual representation of the frequency spectrum of an audio signal. Spectrograms in the form of images show which axis represents time and the other axis represents frequency, and the color at each point represents amplitude (Zeng, Mao, Peng, & Yi, 2019). The data that has been converted into a spectrogram can be seen in Figure 4.

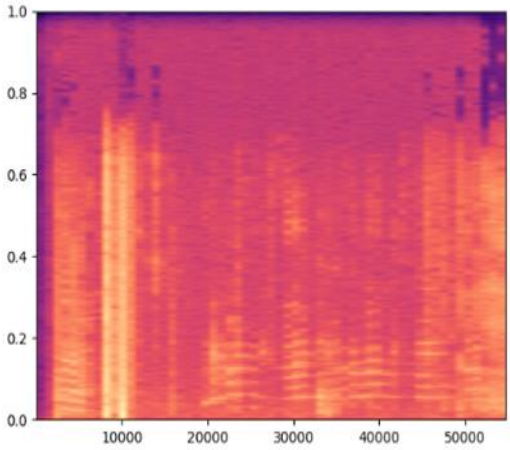

Figure 4. Spectrogram

After preprocessing into a spectrogram, feature extraction is carried out which later the spectrogram data will become numerical data. For the initial features extracted, there are 6 features such as Chroma Short Time Fourier Transform, Root Mean Square Error, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Zero Crossing Rate, and Mel Frequency Cepstral Coefficient can be seen in Table 2.

Table 2. Initial Features

No	chroma_stft	rmse	spectral_centroid	…	mfcc10
1	0,131107	0,018705	622,641	…	-4.068.914.413
2	0,396538	0,096149	1881,525	…	-3.383.172.035
….	…	…	…	…	…
1925	0,567204	0,030467	2250,516	…	29.576.611.518.859
1926	0,571835	0,153279	2543,526	…	-36.501.495.838

2.2 Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction

MFCC is a feature extraction that transforms the linear cosine of the short-time logarithmic power spectrum of a speech signal on a non-linear Mel frequency scale. (Solehudin, 2018). In the dataset obtained, the MFCC feature extraction will be carried out as many as 10 coefficients. The total number of features used in the dataset when combined with the initial 6 features amounts to 16 features. Meanwhile, examples of MFCC coefficients that successfully extracted features can be seen in Table 3.

Table 3. MFCC Feature Extraction 10 Coefficients

No	mfcc1	mfcc2	mfcc3	mfcc4	…	mfcc10
1	-5.431	23.290.40	-57.572	9.078.011	…	-4.068.914.413
2	-2.788	1.022.039	-26.285	15.891.88	…	-3.383.172.035
….	…	…	…	…	…	…
1925	-4.835.52	23.290.409	3.146.038	12.320.159	…	29.576.611.518.859
1926	-26.747.9	23.290.409	-50.933.16	32.873.756	…	-36.501.495.838

3. 10 *K-Fold Cross Validation*

Cross-validation itself is usually defined as a statistical method used to evaluate the performance of algorithms that have been made. Where the data is separated into 2 types, namely training data and testing data (Primartha, 2018). This stage divides the training data and test data using K-Fold Cross Validation with a value of k = 10. The data will be divided into 10 subsets that have the same number of classes.

4. *Support Vector Machine* (SVM) Classification

SVM itself is divided into several kernels, such as Radial Basis Function (RBF) kernel and linear kernel. RBF is a kernel concept that aims to classify data that cannot be linearly separated. Meanwhile, the linear kernel is a popular SVM kernel. Linear functions are often used to measure the similarity or dot product between two input vectors. Linear kernels can also ensure global optimization for regression or classification problems in small and large data sets and adopt predictive binary classification, i.e. the assignment of class labels to unlabeled data (Naicker, Adeliyi, & Wing, 2020). In this study, we will use RBF kernel parameters as well as linear kernel, C value, and gamma. The RBF kernel-based SVM equation is defined in equation 1 (Pambudi, 2022), while the linear kernel-based SVM is defined in equation 2 below (Huang, Zhu, Wang, & Fang, 2021).

$$\boldsymbol{K}(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \quad (1)$$

$$\boldsymbol{K}(x_i, x_j) = (x_i . x_j) \quad (2)$$

Meanwhile, the gamma value can be defined in equation 3 and equation 4.

$$gamma = \frac{1}{(2 * variance)}) \quad (3)$$

$$variance = \left(\frac{1}{N}\right) * \sum (x_i - \mu)^2 \quad (4)$$

The classification flow consists of 2 flows, namely SVM classification with default parameter initialization, and SVM classification with the specified parameter configuration defined in Table 4.

Table 4. SVM Manual Configuration Parameter Values

Parameters	Description	Values
Kernel	Transforms the original feature space into an infinite feature space.	RBF & Linear
C	Controlling between margin and misclassification in SVM models.	0,1 - 1
Gamma	Control how much influence each center point has on the mapping.	0.062

5. Evaluation

After performing the classification, the performance of linear and RBF kernel-based SVM classification was evaluated using Area Under Curve (AUC) by observing the results of each gamma and C parameter used in Table 4.

RESULTS AND DISCUSSION

Before performing classification, data collection and data preprocessing are carried out. The dataset used in this research is the COVID-19 Cough dataset. The initial dataset contains 1926 cough audio files that have different durations. This dataset has two classes consisting of 1284 healthy classes and 642 COVID-19 classes with a comparison of the percentage of healthy patient classes by 67% and the percentage of COVID-19 patient classes by 33% which is represented in the diagram in Figure 5.

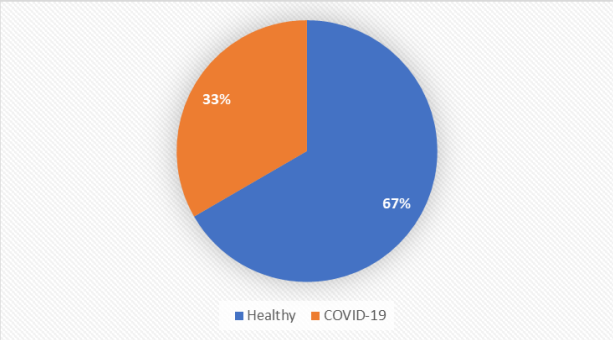

Figure 5. Comparison of Percentage of Healthy and COVID-19 Patients

After data collection, data preprocessing is carried out by converting audio into a visual representation in the form of a spectrogram, from the spectrogram, MFCC feature extraction is carried out as many as 10 coefficients. After preprocessing the data, classification is carried out with 2 schemes, classification using RBF kernel-based SVM and Linear kernel-based SVM classification with gamma and C parameters that have been determined in Table 4.

The first RBF kernel-based SVM classification study was conducted with the help of the SVC library in the python programming language as well as the gamma parameters and C values that have been set according to Table 4 and Table 5. From the classification experiment, the highest AUC was obtained of 0.560 in the default parameter and parameter configuration with a value of C = 1. While the low AUC value was obtained of 0.551 in the parameter configuration with a value of C = 0,1. The overall AUC performance in the RBF kernel-based SVM classification experiment is summarized in a graph that can be seen in Figure 6 below.

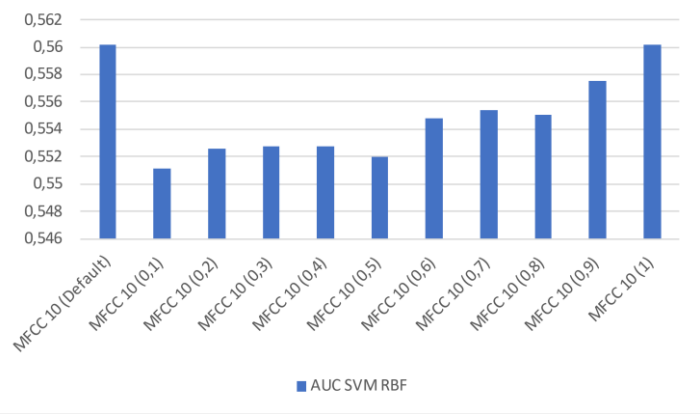

Figure 6. Comparison of AUC Graphs on RBF kernel-based SVM Classification

The second study of linear kernel-based SVM classification was conducted with the help of the SVC library in the python programming language as well as the gamma parameters and C values that have been set according to Table 4 and Table 5. From the classification experiment, the highest AUC was obtained as 0.572 in the parameter configuration with a value of C = 0.5. Meanwhile, the lowest AUC value was obtained at 0.462 in the parameter configuration with a value of C = 0.8. The overall AUC performance in the linear kernel-based SVM classification experiment is summarized in the graph that can be seen in Figure 7.

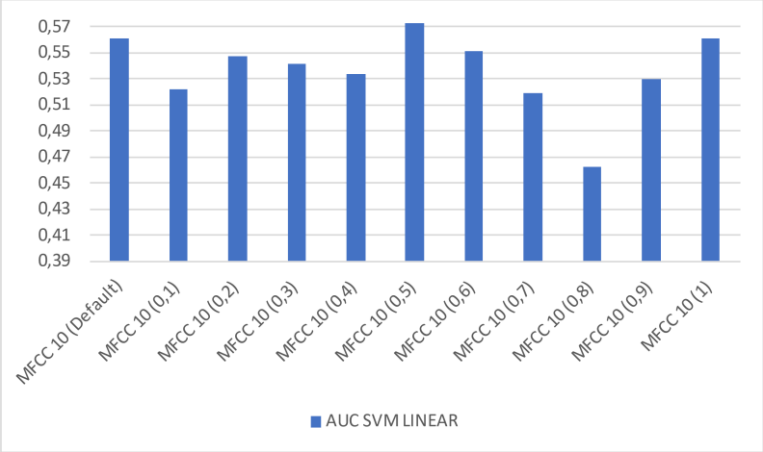

Figure 7. Comparison of AUC Graphs on Linear kernel-based SVM Classification

From the 2 classification experiments above using SVM based on RBF and Linear kernels, the best AUC is summarized which can be seen in Figure 8.

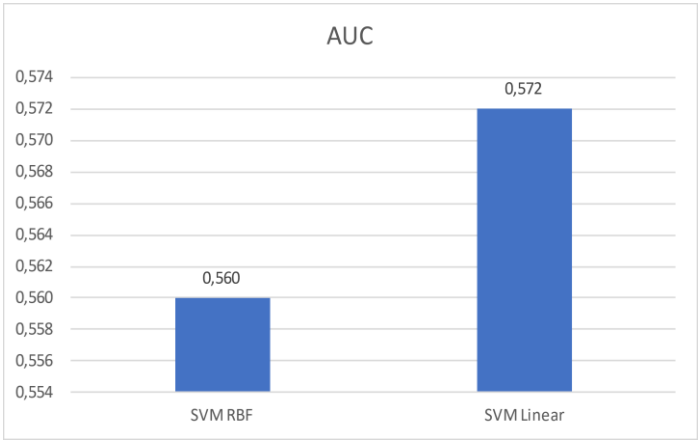

Figure 8. Comparison of AUC Graphs on SVM RBF and Linear classification

From the graph above, it can be seen that in the context of SVM classification, the highest score for area under the curve (AUC) obtained using a linear kernel has a score of 0.572. In contrast, when using the RBF kernel, the performance of the SVM classifier is slightly lower with an AUC of 0.560. Therefore, in this binomial class scenario, it can be seen that the linear kernel outperforms the RBF kernel in terms of AUC value.

CONCLUSIONS AND RECOMMENDATIONS

Early detection of COVID-19 through cough audio is carried out using the COVID-19 Cough dataset based on research conducted. This dataset has 1926 instances with 2 classes consisting of 1284 healthy and 642 COVID-19. The class percentage comparison on the dataset is around 67% healthy and 33% positive for COVID-19. From the percentage of data, it can be concluded that the COVID-19 Cough dataset includes

imbalanced data or unbalanced data. The data will be preprocessed first, such as converting audio data into visual data in the form of spectrograms, then feature extraction is carried out so that it becomes numerical data in the form of MFCC feature extraction. After preprocessing, the dataset will be stratified using Stratidied K-Fold 10 Cross Validation so that the proportion of class comparison in training data and testing data is the same. Classification in this study uses 2 SVM kernels, namely RBF kernel-based SVM and linear-based SVM. This research was conducted with the help of using the SVC library in the python programming language From the research results, the highest AUC is 0.572 in linear kernel-based SVM classification when using parameter C = 0,5. Meanwhile, when using the RBF kernel, the highest AUC is 0.560 when using parameter C = 1. Suggestions given based on this research are to perform other audio extractions such as spectral flux, spectral entropy and use feature selection algorithms such as genetic algorithms.

ACKNOWLEDGEMENT

We would like to thank LPPM Lambung Mangkurat University and the Data Science Laboratory, Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University for supporting the Program Dosen Wajib Meneliti (PDWM).

REFERENCES

Bansal, V., Pahwa, G., & Kannan, N. (2020). Cough classification for COVID-19 based on audio mfcc features using convolutional neural networks. *2020 IEEE International Conference on Computing, Power and Communication Technologies, GUCON 2020* (pp. 604–608). Institute of Electrical and Electronics Engineers Inc.

Berrar D. (2018) Cross-validation. Encyclopedia of Bioinformatics and Computational Biology, Volume 1, Elsevier, pp. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Chowdhury, N. K., Kabir, M. A., Rahman, M. M., & Islam, S. M. S. (2022). Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. *Computers in Biology and Medicine*, *145*. Elsevier Ltd.

Danika, A., Raharjo, J., & Hidayat, B. (2022). Deteksi Suara Gitar Dengan Bahan Jenis Senar Berbeda Melalui Ciri Akustik Dengan Mel-Frequency Cepstral Coefficients (MFCC) Dan Support Vector Machine (SVM) Guitar String Detection Through Acoustic Characteristics Using Mel-Frequency Cepstral Coefficients. *e-Proceeding of Engineering*, *8*, 2936–2942.

Han, J., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., et al. (2021). Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8328–8332).

Huang, B., Zhu, Y., Wang, Z., & Fang, Z. (2021). Imbalanced data classification algorithm based on clustering and SVM. *Journal of Circuits, Systems and Computers*, *30*(02), 2150036. World Scientific.

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., et al. (2023, March). librosa/librosa: 0.10.0.post2. Zenodo.

Naicker, N., Adeliyi, T., & Wing, J. (2020). Linear support vector machines for prediction of student performance in school-based education. *Mathematical Problems in Engineering*, *2020*, 1–7. Hindawi Limited.

Nugroho, C. A. (2021). Klasifikasi k-nearest neighbor chest X-ray pasien Covid-19 dengan haralick features dan histogram of oriented gradient. *MATHunesa: Jurnal Ilmiah Matematika*, *9*(1), 188–195.

Pahar, M., Klopper, M., Reeve, B., Warren, R., Theron, G., & Niesler, T. (2021). Automatic cough classification for tuberculosis screening in a real-world environment. *Physiological Measurement*, *42*(10), 105014. IOP Publishing.

Pambudi, R. (2022). Deteksi Penggunaan Masker dengan Algoritma RBF Support Vector Machine. *The Journal on Machine Learning and Computational Intelligence (JMLCI)*, *1*(2).

Primartha, R. (2018). Belajar Machine Learning Teori dan Praktik. *Bandung: Informatika Bandung*, *10*, 20–30.

Putri, R. N. (2020). Indonesia dalam Menghadapi Pandemi Covid-19. *Jurnal Ilmiah Universitas Batanghari Jambi*, *20*(2), 705. Universitas Batanghari Jambi.

Ritwik, K. V. S., Kalluri, S. B., & Vijayasenan, D. (2021). COVID-19 Detection from Spectral Features on the DiCOVA Dataset. *Interspeech* (pp. 936–940).

Södergren, I., Nodeh, M. P., Chhipa, P. C., Nikolaidou, K., & Kovács, G. (2021). Detecting COVID-19 from audio recording of coughs using Random Forests and Support Vector Machines. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Vol. 6, pp. 4256–4260). International Speech Communication Association.

Solehudin, R. (2018). *Implementasi Metode MFCC (Mel Frequency Cepstral Coefficient) Dan Naive Bayesian Untuk Klasifikasi Nada Dasar Gitar*. Universitas Komputer Indonesia.

Tena, A., Clarià, F., & Solsona, F. (2022). Automated detection of COVID-19 cough. *Biomedical Signal Processing and Control*, *71*, 103175.

Yanti, B., Ismida, F. D., & Sarah, K. E. S. (2020). Perbedaan uji diagnostik antigen, antibodi, RT-PCR dan tes cepat molekuler pada Coronavirus Disease 2019. *Jurnal Kedokteran Syiah Kuala*, *20*(3).

Zeng, Y., Mao, H., Peng, D., & Yi, Z. (2019). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, *78*, 3705–3722. Springer.