Early Detection of COVID-19 Infection Without Symptoms (Asymptomatic) with a Support Vector Machine (SVM) Model Through Voice Recording of Forced Cough

Ni Nyoman Wahyuni Indraswari[1(✉)], I Gede Pasek Suta Wijaya[1], Arik Aranta[1], and Rani Farinda[2]

[1] University of Mataram, Mataram 83115, Indonesia
nunikyuni0300@gmail.com, {gpsutawijaya,arikaranta}@unram.ac.id
[2] Vistula University, 02-787 Warsaw, Poland
ranifarinda@gmail.com

Abstract. COVID-19 is an infectious disease caused by a coronavirus which spreads from direct human contact through droplets of mucus in the respiratory tract of an infected person. The American Centers for Disease Control and Prevention (CDC) says that asymptomatic COVID-19 patients may account for more than 50% of the transmission rate. This research uses the SVM (Support Vector Machine) model as a feature extraction processor from voice data in the training and testing process, so that it can detect asymptomatic COVID-19 from the extraction of cough voice recordings. Of the 171 subjects studied, 120 subjects (70%) for training data and 51 (30%) for test data. The data is divided into the SMOTE data and without the SMOTE data process. The results of the two data have an average performance matrix of above 80%, with accuracy for without the SMOTE data of 98.3% and for SMOTE data of 100%.

Keywords: Accuracy · Asymptomatic · Forced cough · COVID-19 · SVM Model

1 Introduction

Coronavirus disease, also known as COVID-19, is an infectious disease caused by a coronavirus. The first COVID-19 case in Indonesia was discovered on March 2, 2020. In patients with COVID-19, the most common symptoms were fever, cough, and dyspnea [1].

Many cases of COVID-19 transmission have occurred in asymptomatic people, including in the early stages of infection, so symptoms have not yet appeared. The American Centers for Disease Control and Prevention (CDC) says that asymptomatic and asymptomatic COVID-19 patients may account for more than 50% of the transmission rate. According to the CDC, 24% of people in asymptomatic transmit this virus to

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 282–297, 2022.
https://doi.org/10.2991/978-94-6463-084-8_25

others, and 35% of COVID-19 patients who do not show symptoms transmit it to others before they develop symptoms [2].

Alternative solutions for early detection of COVID-19 are still being carried out. Reporting from the Massachusetts Institute of Technology News Office, asymptomatic COVID-19 infection can be detected by recording a forced cough. Coughing is one of the most common symptoms of COVID-19. A forced cough recording is a forced cough recording that aims to reconstruct the original cough sound and can be used as a early diagnosis to find out whether or not someone is infected by COVID-19 [3].

Previous research has classified using the ANN model with recall and f1-score of the positive class below 0.7. The classification begins with the convolution of the spectrogram results of forced cough recordings. There is a possibility that the model is overfitting, seeing the ratio of positive and negative class imbalances in the dataset without the SMOTE (Synthetic Minority Over-sampling Technique) process [4]. In this study, researchers want to classify using the SVM (Support Vector Machine) model with voice recording data extracted into features.

Support Vector Machine (SVM) is a model with a high accuracy level in predicting data classification and tries to find the best hyperplane on the input. SVM is a linear classifier that was developed to work on non-linearity.

Seeing the potential for the success of this development to provide a stimulus to explore the ability of SVM on forced coughing voice data as an alternative for early detection of COVID-19 symptoms, the authors propose the idea of "Early Detection Of COVID-19 Infection Without Symptoms (Asymptomatic) With A Support Vector Machine (SVM) Model Through Voice Recording of Forced Cough".

2 Literature Review and Theoretical Base

2.1 Literature Review

Previous studies using the SVM model have been carried out in the past 5 years, especially in voice recognition cases. The correct results rate reaches more than 70%. Thus the Models have a good performance while studying the data. Also, research on the recognition of COVID-19 through forced coughing recordings has been carried out several times, and most studies recorded a high degree of accuracy.

2.2 Theoretical Base

COVID-19. Corona Virus 2019, or COVID-19, is an irresistible infection caused by the extreme intense respiratory disorder SARS-CoV-2. Side effects of COVID-19 shift, but the foremost common incorporate fever, cough, difficulty breathing, and loss of smell and taste. Indications, by and large, show up from the primary to the fourth day after introducing the virus [1]. The spread of the COVID-19 infection happens when a tainted individual is in close contact with others. The infection can spread through a contaminated nose and mouth when breathing, coughing, sneezing, singing, or talking. Other individuals can end up contaminated if the infection gets into their mouth, nose, or eyes [5].

Asymptomatic COVID-19. There are two instruments by which asymptomatic transmission can occur:

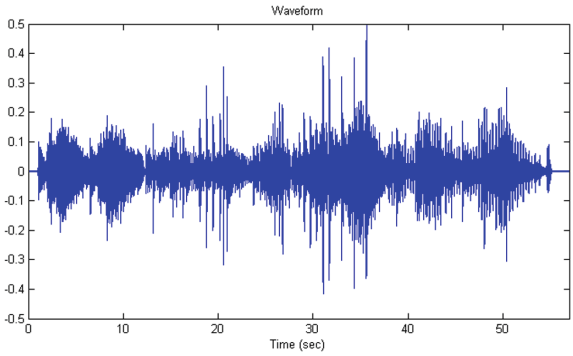

Fig. 1. Plotting audio file with WAV format.

1. Transmission from somebody who has never experienced side effects on the off chance that the tainted individual is asymptomatic throughout the contamination remains contagious.
2. Transmission from a person during the incubation period if the infected person is contagious before developing symptoms [6].

WAV Recordings. WAV (Waveform Audio File Format) is a standard digital audio file with WAV format and stores waveform data. WAV is one of the best audio file formats because of its lossless sound quality. WAV generally stores uncompressed audio in the 44.1 kHz, 16-bit stereo format, the standard format used for CD audio [7] (Fig. 1).

Support Vector Machine (SVM). Support Vector Machine (SVM) is a machine learning calculation with a supervised learning approach with a working strategy searching for a hyperplane (class separation function). This method uses a hypothesis as a linear function in feature space with high dimensions. The level of accuracy in the model that the transition process with SVM will generate is very dependent on the kernel function and the parameters used [8]. The data in a dataset is assigned the variable *xi*, while the class in the dataset is assigned the variable *yi*. The SVM method divides the dataset into two classes. The problem with classification is finding the hyperplane line separating the two groups. The solid line in Fig. 2 means the best hyperplane between the two classes. Meanwhile, red and yellow dots inside the black circle are support vectors.

Synthetic Minority Over-sampling Technique (SMOTE). The SMOTE method deals with data imbalances (the number of objects of a data class is more than that of another class) with the oversampling method. This strategy increments perceptions by expanding the number of minor class data to be equivalent to the significant class with artificial data [9]. Procedure for artificial data:

Numerical Data. Calculate the difference between the primary vector and its nearest neighbors. Multiply the difference by the random number between 0 and 1, and add this difference to the principal value of the original principal vector to obtain a new principal vector.

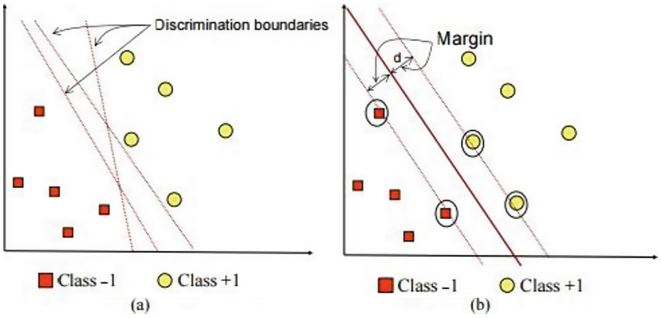

Fig. 2. The best hyperplane that separates both negative and positive classes.

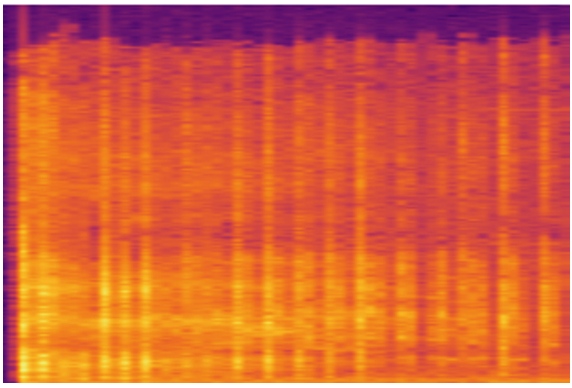

Fig. 3. Examples of non-COVID-19 Spectrograms.

Spectrogram. The spectrogram is a visualization of each formant value plotted against time and amplitude equipped with varying energy levels per unit time (bandwidth). Formant bandwidth can be used to map or identify forced coughing sound recordings because the spectrogram contains details in terms of typical formant patterns and bandwidth [10]. The spectrogram of the coughing sound will be extracted to produce features such as MFCC, spectral centroid, zero-crossing rate, chroma frequencies, and spectral roll-off and saved in CSV format. An example of a spectrogram can be seen in Fig. 3.

Mel Frequency Cepstrum Coefficients (MFCC). MFCC (Mel Frequency Cepstral Coefficients) is a feature extraction method for deciding values or vectors that can be utilized as objects or person characters. Some of the advantages of this method are [11]:

1. Able to capture the characteristics of the voice, which is exceptionally imperative for discourse acknowledgment, or in other words, can capture basic data contained within the voice signal.
2. Produce minimal data without losing essential information it contains.
3. Replicate the human hearing organ in perceiving the signal.

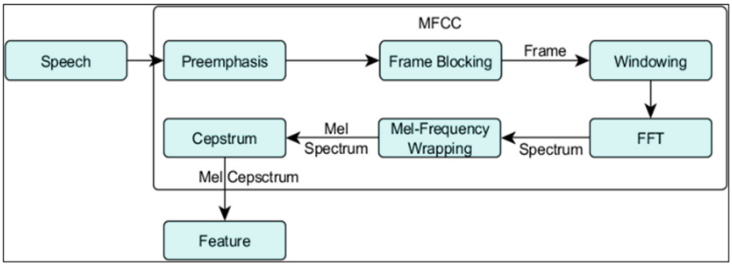

Fig. 4. MFCC Process.

The extracted feature is the cepstral coefficient which considers the perception of the human auditory system. The way MFCC works is based on the difference in frequencies that are captured by the human ear, so it can represent the sound signal as follows [12] (Fig. 4).

Biomarker. Biomarkers are molecules that show normal or abnormal processes that occur in the body as a sign of a person's health condition. The main focus of the biomarkers in this study is the signs in the coughing sound generated from the extracted features. These features can be signs that can be classified using the SVM Model [13].

3 Materials and Methodology

3.1 Materials and Tools

The tools and materials in the research carried out in the form of hardware and software, as well as the data needed during the activity, are as follows:

Tools. The tools used in this research process are divided into two parts, namely:

Hardware. The hardware that used in this research are within the specifications as shown in Table 1.

Software. The software that used in this research are within the specifications as shown in Table 2.

Materials. The tools used in this research process are divided into two parts, namely:

1. The literature comes from journals, studies, and supporting books on cough sound processing, feature extraction, activation function, machine learning, email spectrogram, MFCC, and SVM.

Table 1. Hardware Specifications.

No	Name	Specification
1	Processor	Intel(R) Core (TM) i5-1035G1 CPU @ 1.00 GHz (8 CPUs), ~1.2 GHz
2	RAM	8 GB RAM
3	GPU	Radeon 620 Series

Table 2. Software Specifications.

No	Name	Specification
1	Operating System	Windows 10 64 bit
2	Programming Language	Python 3.8
3	Microsoft Office	Office 2019
4	Text Editor	Google Colab (Colaboratory)

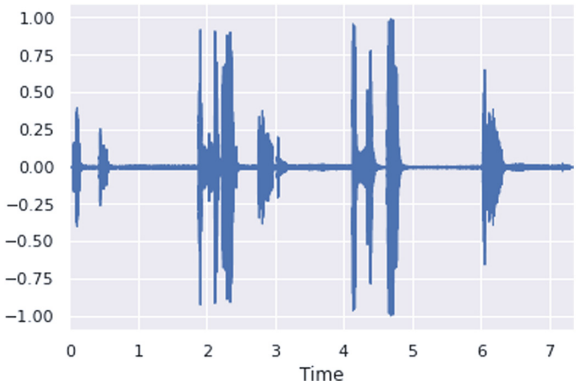

Fig. 5. Waveplot recording of a positive COVID-19 forced cough.

2. The dataset used in this study is divided into two parts: the training dataset and the test dataset. The dataset was obtained through a research team from the Indian Institute of Technology Kharagpur, which is available through Kaggle.com. A total of 120 subjects (70%) were used for training data and 51 subjects (30%) for test data. Dataset recording formats will be converted using.wav audio format with a maximum duration of 5 s (Figs. 5 and 6).

3.2 Research Flowchart

The research flowchart from literature research to writing the final report is shown in Fig. 7.

3.3 Literature Research

The material studied in this literature study is related to studies related to forced coughing sound recordings, feature extraction, oversampling, machine learning, Mel spectrogram, MFCC, and SVM, along with the accuracy of each method application and its development on the model.

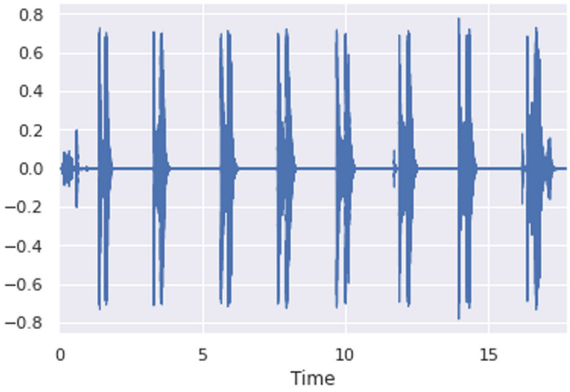

Fig. 6. Waveplot recording of a negative COVID-19 forced cough.

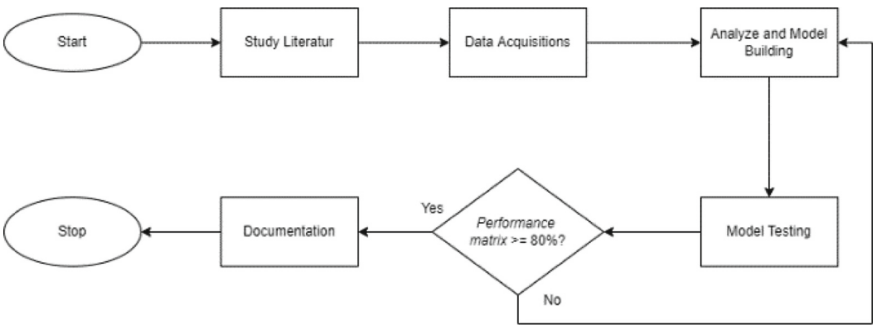

Fig. 7. Research Flowchart.

3.4 Data Acquisitions

The dataset is divided into two parts: the training dataset and the test dataset. The dataset was obtained through a research team from the Indian Institute of Technology Kharagpur, which is available through Kaggle.com. A total of 120 subjects (70%) were used for training data and 51 subjects (30%) for test data. Dataset recording formats will be converted using.wav audio format with a maximum duration of 5 s. In the model's development, an oversampling data scenario was carried out on the training dataset using the SMOTE method to obtain more data on the COVID-19 positive label.

3.5 Model Building

Analysis and Model Development is a step to understanding the model's performance to be built by analyzing the needs intensively and specifically. Figure 8 shows the flow of the model made.

Preparation Data. It started with to prepare a forced cough recording in .wav format from the existing dataset and process each sound sample to be cut into a 5 s duration. After that, the data is processed as described in the next point.

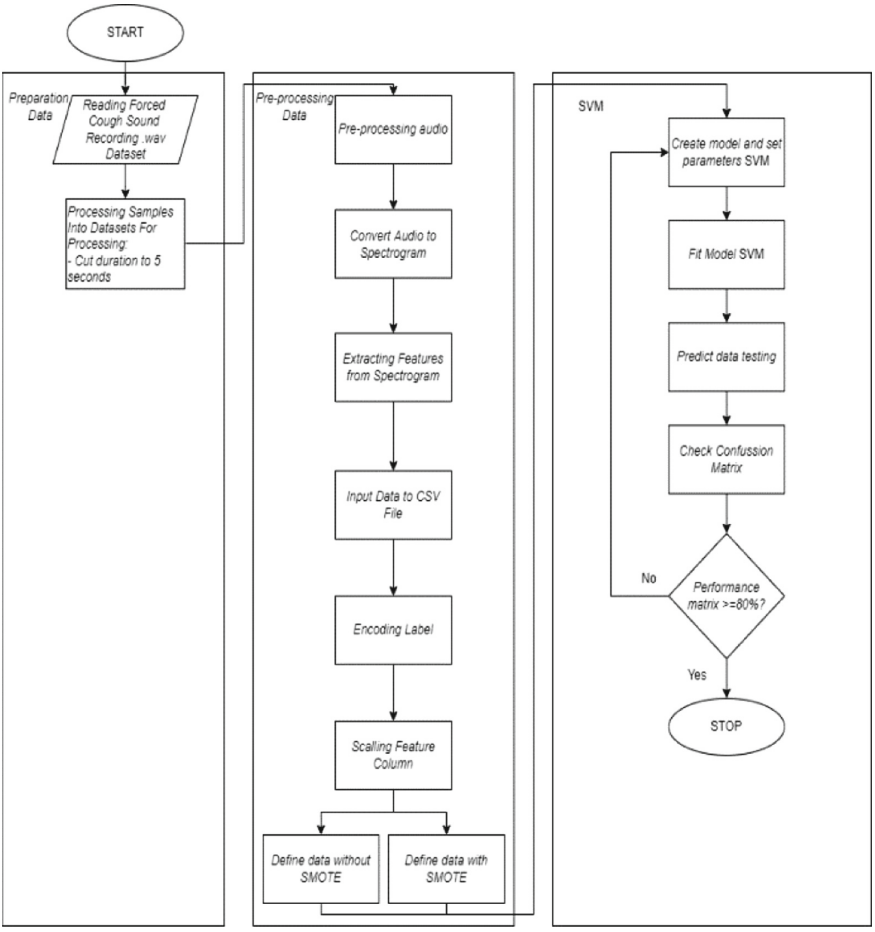

Fig. 8. Analyze model building process.

Pre-processing Data. Pre-processing data is one of the important stages in the mining process by converting raw data into more appropriate data. In this study, pre-processing was carried out to obtain data that was only the duration of the voice used to be in the image data format (image). The pre-processing data flow is made as shown in Fig. 9.

Pre-processing Audio. On this stage, the duration of the coughing sound recording data will be cut by eliminating the duration of the noise and the duration of the silence, which are part of the recorded data but are not needed to get the correct coughing sound.

Convert Audio to Spectrogram. The conversion is done by changing the sound recording from .wav format into a spectrogram image using the specgram() function as a technique used to identify coughing sound data. The sound taken is 5 s long.

Extracting Features from Spectrogram. Spectrogram image extraction aims to visualize audio and makes extracting features in audio easier, such as Mel-frequency cepstral

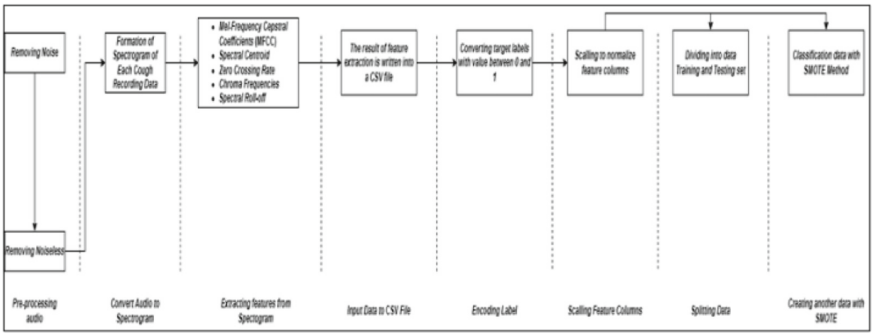

Fig. 9. Data Preprocessing Stages.

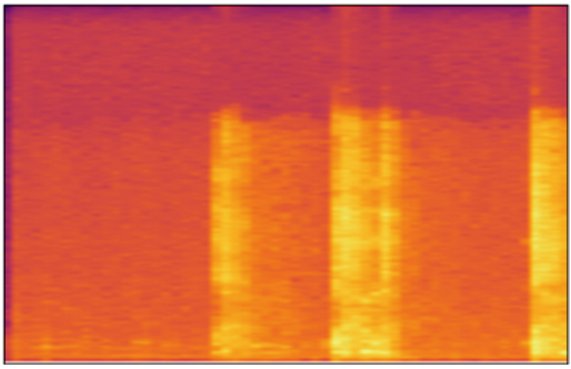

Fig. 10. Spectrogram Covid-19.

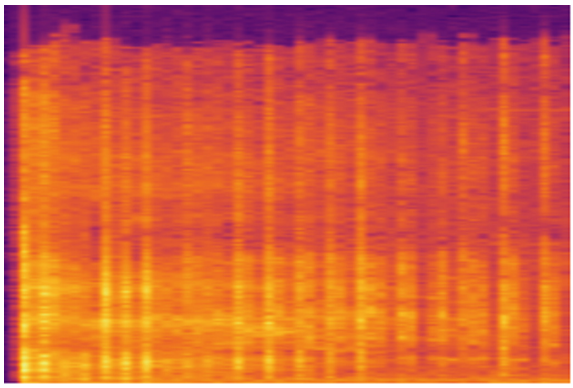

Fig. 11. Spectrogram Non Covid-19.

coefficients (MFCC), Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, and Spectral Roll-off values (Figs. 10 and 11).

Input Data to CSV File. After the features are extracted, the next step is to create column names according to the extracted feature names. Each feature value is entered into the CSV according to its respective column. The resulting dataset is then analyzed using pandas and eliminating some unnecessary columns.

Encoding Label. Encoding Labels are used to change the shape of labels from non-numeric to numerical. The labels in the dataset consist of not_covid and covid. The label is changed using the LabelEncoder() function so that it becomes 1 (for covid) and 0 (for not_covid).

Scaling Feature Columns. The scaling feature aims to create a standard range of values for better algorithm performance. Use the StandardScaler() function to subtract features by the mean, then scale to a unit variance that divides all values by the standard deviation.

Data without SMOTE. Before building the model, the cleaned data were divided into training and testing data with a comparison of 70:30 using the train_test_split() function.

Data SMOTE. The researcher aims to make other data using the SMOTE method, which will later compare the model results between SVM with data without SMOTE and SVM with SMOTE data. At the time of SMOTE implementation, this stage will add the number of minority class instances to balance the data against the majority class. In this test, the data X and y are resampled with the parameters that have been identified. The percentage value of SMOTE is the percentage of the minority class that becomes the synthesis instance.

Building SVM Model. SVM Model build using a forced cough recording in .wav format from the existing dataset and process each sound sample to be cut into a 5 s duration. After that, the data is processed as described in the next point.

Identify the model using SVC (Support Vector Classifier). The Linear SVC (Support Vector Classifier) aims to match the data and then return the best hyperplane to categorize the data entered. The best hyperplane can be assigned some features to the classifier to see the prediction class.

Defining the SVM Kernel. Kernels in SVM will automatically become RBF when not declared. The RBF kernel function is the most widely used and has a localized and limited response along the x-axis.

Fitting with the Data. After identifying the kernel, then train the classification with 2 data, namely Data without SMOTE and Data SMOTE.

Finding the score Data. Each model has a score () method after training on the data. When calling score () on the classifier, this method calculates the accuracy score by default (# of correct_preds/# of all_preds). By default, the score () method does not require actual predictions because predictions are made using X_test and the results of those predictions to calculate an accuracy score.

3.6 SVM Model Testing

After the coding process is complete, the testing process is carried out using a testing dataset to see the results of the classified forced coughing sound data as input data.

One technique that can be used to measure the performance of a model, especially in the case of classification (supervised learning) in machine learning, is the confusion matrix. The confusion matrix gives data on comparing the results carried out by the

Table 3. Confusion Matrix.

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

model with the actual results. Table 3 is a confusion matrix with four predicted and actual value combinations.

Some of the performances measured are accuracy, precision, recall, and f1 score.

Accuracy. Accuracy shows how accurate the model has been in classifying the data correctly. Accuracy is the closeness of the predicted value to the actual value.

Precision. Precision shows the ratio of true positive predictions to the overall positive predicted results. Of all the positive classes that have been correctly predicted, there are how many data are truly positive.

Recall. Recall shows the success of the model in finding information from the ratio of true positive predictions to the comprehensive data that is truly positive.

F1 Score. F1 Score is a weighted comparison of the average precision and recall value.

4 Results and Discussion

4.1 Results and Discussion of SVM Model with Data Without SMOTE

After modeling with data without SMOTE, the results of the accuracy score using training data reached 98.3% with a performance matrix, as shown in Table 4.

Furthermore, testing the SVM model using data testing, the performance matrix of this model also meets the criteria (above 80%) with a description as in Table 5.

It can be reviewed with the elaboration of the confusion matrix as follows (Fig. 12).

The results of the SVM model with data without SMOTE process, 5 out of 6 true positives and all true negatives data were detected correctly. It proves the initial hypothesis that the SVM model is one of the models with a high level of accuracy in predicting asymptomatic patients on the data. When compared with the ANN model, the results of the SVM matrix performance are much better, as shown in Table 6.

Table 4. Performance Matrix Model SVM in Data Training without SMOTE

	Precision	Recall	F1-Score	Support
0	1.00	0.85	0.92	13
1	0.98	1.00	0.99	106
Accuracy			0.98	119
Macro avg	0.99	0.92	0.95	119
Weighted avg	0.98	0.98	0.98	119

Table 5. Performance Matrix Model SVM in Data Testing without SMOTE.

	Precision	Recall	F1-Score	Support
0	1.00	0.83	0.91	6
1	0.98	1.00	0.99	45
Accuracy			0.98	51
Macro avg	0.99	0.92	0.95	51
Weighted avg	0.98	0.98	0.98	51

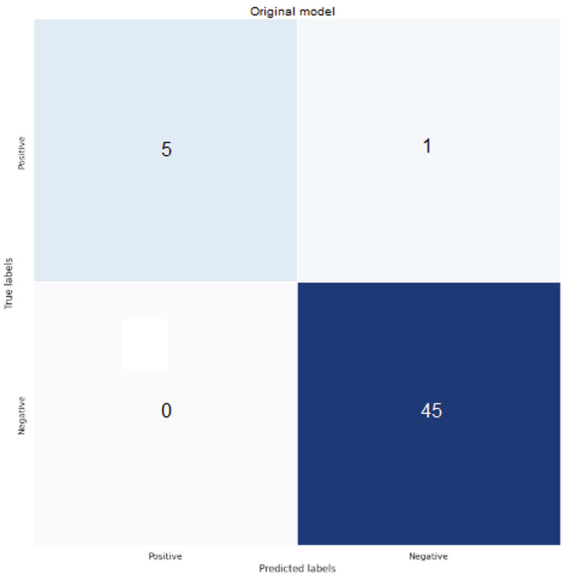

Fig. 12. Confusion matrix SVM model in data testing without SMOTE.

Table 6. Performance Matrix ANN Model in Data Testing without SMOTE.

	Precision	Recall	F1-Score	Support
0	1.00	0.33	0.50	6
1	0.92	1.00	0.96	45
Accuracy			0.92	51
Macro avg	0.96	0.67	0.73	51
Weighted avg	0.93	0.92	0.90	51

Table 7. Performance Matrix Model SVM in Data Training with SMOTE Data.

	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	105
1	1.00	1.00	1.00	151
Accuracy			1.00	256
Macro avg	1.00	1.00	1.00	256
Weighted avg	1.00	1.00	1.00	256

Table 8. Performance Matrix Model SVM in Data Testing with SMOTE Data.

	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	105
1	1.00	1.00	1.00	151
Accuracy			1.00	256
Macro avg	1.00	1.00	1.00	256
Weighted avg	1.00	1.00	1.00	256

4.2 Results and Discussion of SVM Model with Data SMOTE

After modeling with data with SMOTE, the results of the accuracy score using training data reached 100% with a performance matrix, as shown in Table 7.

Furthermore, testing the SVM model using data testing, the performance matrix of this model also meets the criteria (above 80%) with a description as in Table 8.

It can be reviewed with the elaboration of the confusion matrix as follows (Fig. 13).

The results of the SVM model with data with SMOTE process, all true positives and true negatives data were detected correctly. It proves the initial hypothesis that the SVM model is one of the models with a very high level of accuracy in predicting asymptomatic patients on the data. When compared with the ANN model, the results of the SVM matrix performance are much better, as shown in Table 9.

5 Conclusion

Based on the analysis and discussion of the results, the authors obtain conclusions that can be drawn from this study as follows:

1. The sound recording of forced coughing can be an alternative in the early detection of asymptomatic COVID-19 infection.
2. A Support Vector Machine (SVM) model with RBF kernel (default) responding localized and limited along the x-axis trained on the COVID-19 cough data set to generate a binary classification (positive or negative).

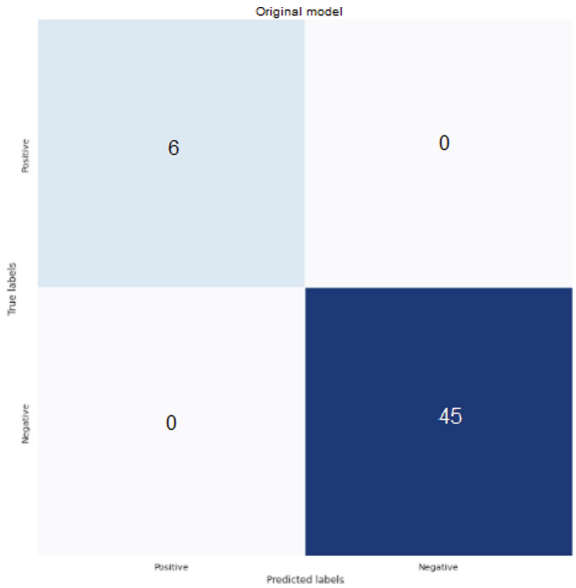

Fig. 13. Confusion matrix model SVM in data testing with SMOTE data.

Table 9. Performance Matrix Model in Data Testing with SMOTE.

	Precision	Recall	F1-Score	Support
0	1.00	0.83	0.91	6
1	0.98	1.00	0.99	45
Accuracy			0.98	51
Macro avg	0.99	0.92	0.95	51
Weighted avg	0.98	0.98	0.98	51

3. The Support Vector Machine (SVM) model with the results of data without SMOTE and SMOTE data managed to achieve an average performance matrix of above 80%, with accuracy for data without SMOTE of 98.3% and SMOTE data of 100% using forced cough recordings, meaning the SVM model becomes the right solution to generate the optimal model on a small dataset.

Acknowledgment. The preparation of this paper can not be separated from the support from various parties. On this occasion, we would like expresses our deepest gratitude to:

1. God Almighty with all His graces has given us strength in completing this paper.

2. Both beloved parents of each team member who has been helping in the form of affection, attention, enthusiasm, and prayers which never stops flowing for the success of the writers in completing this paper.

3. All lecturers, academic staff, and big family of the Department of Informatics Engineering study program of University of Mataram that always help in providing facilities, knowledge, as well as enlightenment for the writers in completing this paper.

References

1. Q&A on coronaviruses (COVID-19), World Health Organization (2020), https://www.who.int/emergencies/diseases/novel-coronavirus-2019/qanswers hub/%0Aq-a-detail/q-a-coronaviruses%0A, last accessed 2021/05/15.
2. Ulfa Rahayu: Sebesar Apa Dampak Orang Tanpa Gejala COVID-19 Berpengaruh Terhadap Penularan Wabah?. hellosehat, Jakarta, (2020).
3. B. Lange, D. Li, E. Nehoran, E. Tuzhilina, and M. Lu2: Early Detection of COVID-19 from Cough Sounds, Symptoms, and Context Machine Learning / Signal Processing Sub-Team CS 472. In: Data science and AI for COVID-19, p. 12 (2020).
4. A. K. Nisa, I. G. P. Suta Wijaya, and A. Aranta: EARLY DETECTION OF ASYMPTOMATIC COVID-19 INFECTION WITH ARTIFICIAL NEURAL NETWORK MODEL THROUGH VOICE RECORDING. no. Cdc, p. 12 (2022).
5. M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina: Discovering symptom patterns of COVID-19 patients using association rule mining. Comput. Biol. Med., vol. 131, no. January, p. 104249 (2021).
6. G. U. Kim et al.: Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19. Clin. Microbiol. Infect., vol. 26, no. 7, pp. 948.e1-948.e3 (2020).
7. Similarities and Differences between Flu and COVID-19, https://www.ucsfhealth.org/education/can-you-tell-if-its-the-flu-or-covid-19, last accessed 2021/09/06.
8. N. Melek Manshouri: Identifying COVID-19 by using spectral analysis of cough record-ings: a distinctive classification study. In: Cogn. Neurodyn., no. January (2021).
9. X. Wu, C. Chen, M. Zhong, J. Wang, and J. Shi: COVID-AL: The diagnosis of COVID-19 with deep active learning. In: Med. Image Anal., vol. 68, p. 101913 (2021).
10. C. Caihua: Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM. In: IEEE Int. Conf. Power, Intell. Comput. Syst. ICPICS 2019, pp. 173–176 (2019).
11. F. Wu, S. Sun, J. Zhang, and Y. Wang: Singing voice detection of popular music using beat tracking and SVM classification. In: IEEE/ACIS 14th Int. Conf. Comput. Inf. Sci. ICIS 2015 - Proc., pp. 525–528 (2015).
12. A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad: Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease. In: Proc. 2015 Int. Conf. Electr. Inf. Technol. ICEIT 2015, pp. 300–304 (2015).
13. COVID-19 Surveillance Group: Characteristics of COVID-19 patients dying in Italy Report based on available data on March 20 th , 2020. In: COVID-19 Surveill. Gr., pp. 4–8 (2020).