Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230730

Machine learning models for COVID-19 diagnosis based on medical images and audio

Xiaotong Wang

The Department of Foreign Language, Nanjing University, Nanjing, 210000, China

211108029@smail.nju.edu.cn

Abstract. The COVID-19 pandemic presents a significant danger to human health, with far-reaching consequences for the global economy and political dynamics. It presents complex challenges in terms of rapid and accurate diagnosis, where machine learning holds potential to enhance diagnostic speed and precision while reducing time and resource burdens. Consequently, this research employs CT images and cough audio recordings as training data to create machine learning models for data classification, with the goal of aiding in the diagnosis of COVID-19. Using the Edge Impulse platform, a Convolutional Neural Network, MobileNetV2, is customized for efficient image recognition. On the audio front, the preprocessing phase encompassed three distinct feature extraction techniques: Mel Frequency Cepstral Coefficients, Mel-Filterbank Energy (MFE), and Mel spectrogram. Subsequently, model frameworks were meticulously adjusted to accommodate the classification requirements. The results of this effort are highly encouraging. In the domain of CT image recognition, the top-performing model achieved a remarkable accuracy of 93.98%. In the concurrent task of categorizing cough audio data, the best performance reached 81.8%. These findings underscore the capability of this approach as an effective supplementary tool for medical diagnostics. In the face of COVID-19's persisting impact, such machine learning advancements could significantly aid in swift and reliable diagnoses, ultimately contributing to the global battle against the pandemic.

Keywords: Machine Learning, COVID-19, Diagnosis.

1. Introduction

COVID-19, an abbreviation for "Coronavirus Disease 2019," is a global infectious disease known for its respiratory manifestations, such as fever, cough, and difficulty breathing. Its impact has extended beyond health, affecting sectors such as politics, education, and the economy. According to the World Health Organisation, as of August 2, 2023, the confirmed cases have reached 768,983,095 and the confirmed deaths have reached 6,953,743[1]. During the diagnosis in COVID-19, it faces several challenges including the diversity of clinical symptoms, ranging from mild cold-like symptoms to severe respiratory distress, and even asymptomatic cases. Moreover, due to limitations of diagnostic tools, current testing methods for diagnosing COVID-19, such as nucleic acid tests and antibody assays, have shortcomings regard to specificity and sensitivity, which impact the accuracy of diagnostic outcomes. Therefore, utilizing machine recognition in conjunction with medical imaging and other information to enhance diagnostic accuracy is necessary.

There have been previous studies using machine learning to tackle COVID-19. In terms of the training data, tomography (CT), CXR (chest X-ray), and lung ultrasound (LUS) are most commonly seen in the previous research. X-ray exams are quick for urgent screenings. They're cost-effective during resource-limited outbreaks. CT excels with high resolution, revealing detailed anatomy. It builds 3D images from various angles, aiding lung analysis. In terms of methods, 10 well-known Convolutional Neural Network (CNN) architectures were utilised the most in the previous research, including AlexNet, SqueezeNet, GoogleNet, VGG-19, VGG-16, ResNet-101, ResNet-50, ResNet-18, MobileNet-V2, and Xception. Among these architectures, Xception achieved the highest performance but lacked sensitivity to some extent. Conversely, ResNet-101 exhibited the best sensitivity, albeit with specificity needed to improve [2]. H. Panwar and colleagues developed a pre-trained convolutional neural network using CT and X-ray images, utilizing the ImageNet dataset. The most impressive result was achieved using VGG19, attaining an accuracy rate of 95.61% [3]. Jaiswal et al. opted for pre-trained models like DenseNet201, VGG16, ResNet152V2, and InceptionResNetV2. Their best-performing model was DenseNet201, which achieved an accuracy of 96.25% [4]. Ismael and their team utilized X-ray images to create a pre-trained CNN model, leveraging ImageNet weights for deep feature extraction. They employed SVM as a classifier, and the most favorable outcome was with ResNet50+SVM, achieving an accuracy of 94.74% and a recall rate of 91.00% [5].

This paper's primary contributions are as follows: It centers on classifying COVID-19 medical images with the goal of enhancing pneumonia identification's accuracy and efficiency. This paper employed publicly available lung CT datasets and cough audio datasets for model training. The proposed models were constructed through a fusion of CNN and transfer learning applied to image datasets. Additionally, the study involved a comparison of the detection performance of the MobileNetV2 network across various width multiplier settings. In the realm of audio data, different pre-processing method were experimented including MFCC, MFE and Mel Spectrogram methods. This culminated in emphasizing the potential and future prospects of COVID cough data in pneumonia detection.

2. Method

2.1. Dataset description and preprocessing

2.1.1. CT Image Dataset

The COVID-CT-Dataset includes 349 CT images marked as positive and 463 non-COVID-19 CT images, displaying clinical COVID-19 findings from 216 patients. A seasoned radiologist at Tongji Hospital in Wuhan, has validated the dataset's credibility. This radiologist possesses extensive expertise gained from diagnosing and treating a significant number of COVID-19 cases during the outbreak period from January to April. These CT images exhibit varying dimensions. The smallest, average, and largest heights measure 153, 491, and 1,853, respectively. The smallest, average, and largest widths are 124, 383, and 1,485, respectively. All these CT images are in RGB format [6].

2.1.2. Cough Audio Dataset and preprocessing

The COVID Cough audio dataset includes 9min9s audio for COVID-19 label and 5min7s for normal label in total. 20% was moved to test set for result evaluation. This work tried three different feature extract techniques to pre-processing the data, changing the audio regnition work into an image classification [7]. These are the following three extract techniques:

2.1.2.1. Mel Spectrogram

Mel spectrogram represents how the energy of several frequency components in an audio signal changes over time. It's obtained by applying a series of transformations to the audio signal, including windowing, FFT, Mel filterbank, and logarithmic compression. In a Mel spectrogram, the horizontal axis represents time, the vertical axis stands for frequency, and the intensity and color indicates the energy level of each frequency element.The auditory frequency scale makes lower frequencies more pronounced, which is crucial for understanding speech. Simultaneously, it makes higher frequencies less prominent. These higher frequencies primarily consist of noise and sounds such as fricatives, which are less vital for comprehending speech and do not require precise representation [8].

2.1.2.2. Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are a set of coefficients extracted from the Mel spectrogram, precisely engineered to capture the spectral attributes in the audio signal,with more concise and significant manner.This method is more focused on the spectral shape of the audio signal and less concerned with the temporal changes, unlike Mel spectrograms.

This feature extract technique includes seven parts. The first step is pre-emphasis, followed by framing. Subsequently, multiplying each frame by a Hamming window to reduce spectral leakage. Then applied the Fast Fourier Transform (FFT) and applying a set of triangular filters on the spectrum to compute the energy in different frequency bands according to the Mel scale. Next step is Discrete Cosine Transform (DCT), which is transforming the log-compressed filter bank energies using DCT to decorrelate the coefficients and obtain the MFCCs. Finally, computing first-order (delta) and second-order (delta-delta) differences between consecutive frames' MFCCs to capture the dynamic information [9].

2.1.2.3. Mel-Filterbank Energy (MFE)

MFE features encompass several steps to understand the characteristics of audio signals. Initially, the audio is segmented into brief intervals, the Fast Fourier Transform (FFT) is employed on each interval to expose its frequency constituents. Subsequently, a group of filters, aligned with the Mel scale to better emulate human pitch perception, gauges the energy across various Mel-frequency bands. Afterward, the logarithm of these energy values is taken to match human perception. The resulting MFE features provide insight into the intensity of various frequencies within each audio section, proving particularly valuable for analyzing non-vocal sounds like music or diverse environmental noises.

2.2. Model Architecture

2.2.1. CT Image Classification Model

In terms of model architecture, this article chooses MobilNetV2 as the classifier and set the training cycle as 100, learning rate 0.001. MobileNetV2 is a mobile-friendly architecture that utilizes depthwise separable convolutions. This approach divides the typical convolution into two layers: depthwise convolutions and pointwise convolutions. As a result, computations and parameters are greatly reduced, yet expressive power is preserved. Additionally, the architecture employs an inverted residual structure with linear bottlenecks. This helps decrease computational complexity while retaining high accuracy [10].

To create smaller, more computationally efficient models, a width multiplier (α) was introduced. This factor scales the width of a neural network, particularly in depthwise separable convolutions. Ranging between 0 and 1, common values for α include 1, 0.75, 0.5, or 0.25. When applied to a network layer, α scales the number of input channels (M) to αM and the output channels (N) to αN. Essentially, this scaling uniformly reduces the network's width at each layer [11]. This work test four different width multiplier to see the influence on the final results.

Transfer learning is also utilized in this context. This approach is chosen for its ability to utilize knowledge that already learned before and benefit the similar research. This practice results in significant efficiency improvements, eliminating the need for the model to start learning from scratch for each distinct task.

2.2.2. Cough Audio

As the paper experimented three different feature extract techniques, the parameters and architectures were set different in order to reach the best result of each model. All the training cycles were as 100, learning rate 0.005. Because the number of the audio is relatively small and the dataset is imbalanced, data augmentation was applied. Random noise was added to each spectrogram and random blocks were masked in the frequency and time axis.

In terms of neural network conducting, reshape layer is the first layer after input data. Subsequently, 2D Convolutional / Pooling Layer applies convolutional filters to the input, which extracted features from the input data and reduced the spatial dimensions of the feature maps while retaining essential information. This layer usually followed by dropout layer, which is a regularization method that, in each training iteration, randomly deactivates a portion of input units by setting them to zero. This strategy combats overfitting by diminishing the dependency on particular neurons and promotes the network to acquire more resilient features. 2D Convolutional / Pooling Layer and dropout layer can be used twice or more in order to extract deeper feature. Following this step, the Flatten Layer transformed the multi-dimensional feature maps into a single-dimensional vector, readying the data for input into a dense layer.But the Dense Layers are not necessary. When utilised MFE as the feature extract method, in the classification part, this paper added a Dense Layer with 128 neurons. Rectified Linear Activation (ReLU) is used as the activation function. The structures can be found in Figure 1.

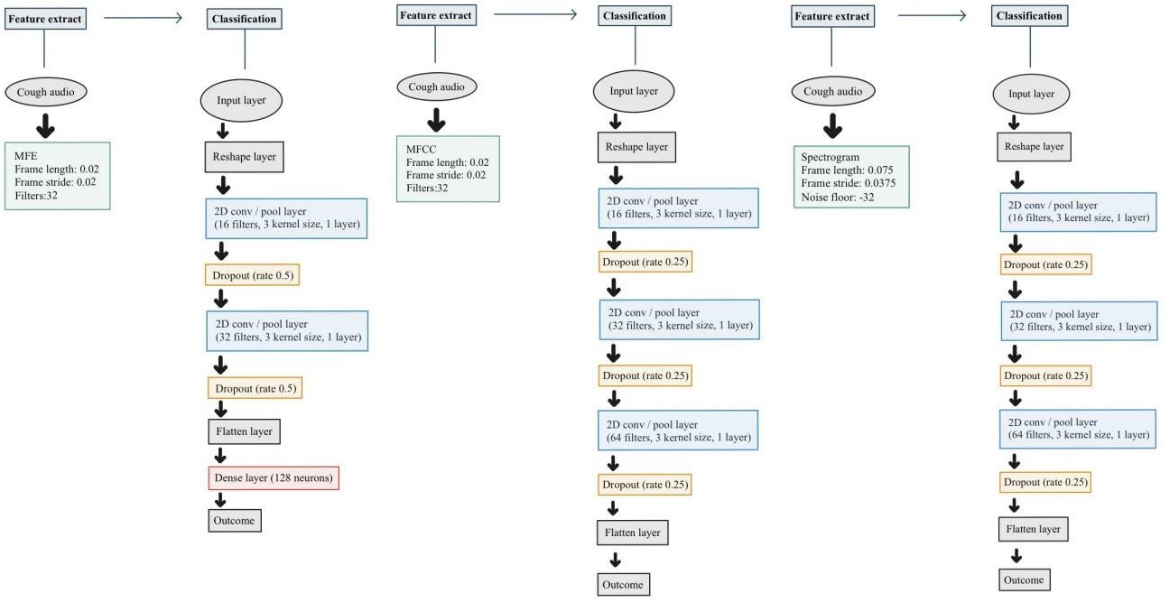

Figure 1. Describing the feature extract method and the classification model architecture (Photo/Picture credit: Original).

2.2.3. Implementation details

When assessing the effectiveness of a model, several common metrics are used to assess how well the model shows in terms of its classification task.

Accuracy, the ratio of correctly predicted instances to the total instances in the dataset, is the most commonly employed evaluation metric in practical applications, whether for binary or multi-class classification tasks. Confusion Matrix, which is a table showing the counts of true positives, true negatives, false positives, and false negatives.

Precision is the ratio of true positives to the sum of true positives and false positives.

$$\text{Precision} = TP/ (TP + FP) \tag{1}$$

Where:

True Positives (TP): The number of instances that are actually positive and were correctly predicted as positive.

False Positives (FP): The number of instances that are actually negative but were incorrectly predicted as positive.

Recall (Sensitivity or TPR): It measures the ability of the model to correctly identify positive instances.

$$\text{Recall} = TP / (TP + FN) \qquad (2)$$

Where:

False Negatives (FN): The number of instances that are actually positive but were incorrectly predicted as negative.

F1-Score balances both precision and recall, which is particularly useful when classes are imbalanced.

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \qquad (3)$$

3. Results and discussion

3.1. CT Images

According to the metrics shown in Figure 2, Figure 3 and Figure 4, this paper uses R language to realize the data visualization. As the figures indicate, the MobileNetV2 160 0.5 variant demonstrates the lowest FPR of 3.60% while maintaining a good balance between other metrics, including accuracy, F1-score, and recall. The MobileNetV2 160 1.0 variant has the highest FPR at 9.30% and a comparatively lower COVID-19 Precision of 0.89. The MobileNetV2 160 0.75 variant offers a lower FPR of 6.50% and maintains a balanced set of metrics, including a good COVID-19 Recall of 0.93. The MobileNetV2 160 0.35 variant has an FPR of 6.60% and a lower accuracy of 88.72%. Given the data above, the MobileNetV2 160 0.5 and MobileNetV2 160 1.0 appear to perform better. Due to the low tolerance for errors in real-world medical diagnosis, the model's accuracy still requires further enhancement. Additionally, broader validation with real-world data is necessary to validate its performance under more diverse circumstances.

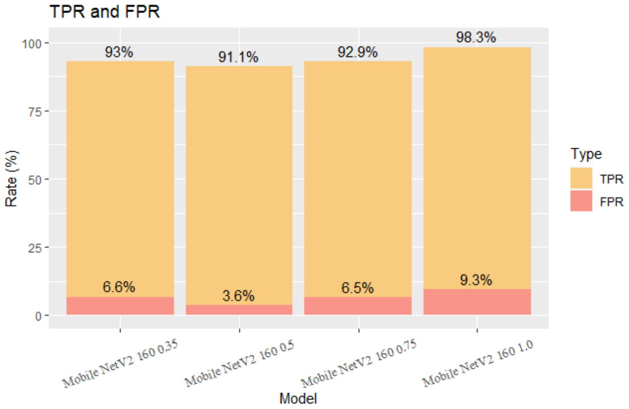

Figure 2. True positive rate and false positive rate of different models (Photo/Picture credit: Original).

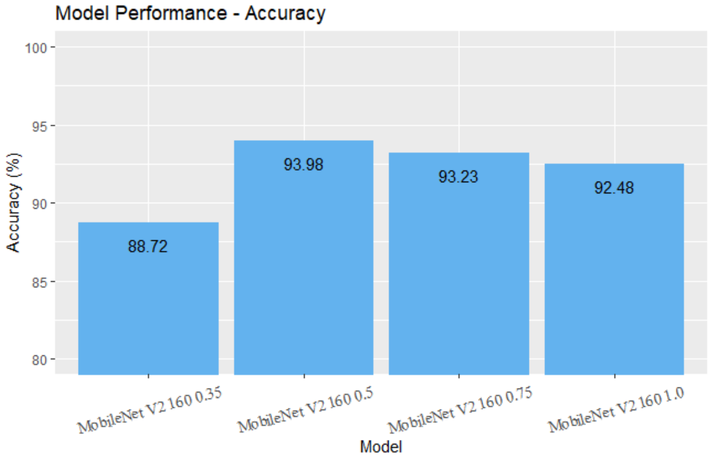

Figure 3. Accuracy of different models (Photo/Picture credit: Original).

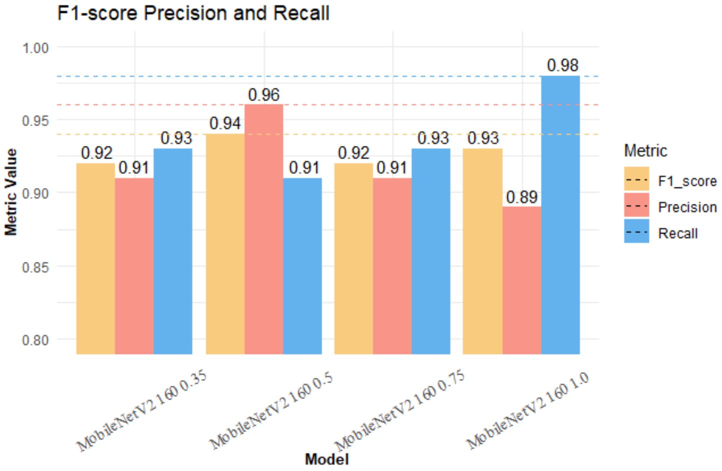

Figure 4. F1-score, precision and recall of different models (Photo/Picture credit: Original).

3.2. Cough Audio

Overall, based on the metric, the MFE technique seems to outperform the other two techniques (MFCC and MEL Spectrogram) in most aspects. It has a higher true positive rate, lower false positive rate, higher accuracy, and higher F1-score. This suggests that MFE is potentially a more suitable technique for COVID-19 classification using the given dataset. While different techniques have varying levels of performance shown in Table 1, all three techniques—MFE, MFCC, and MEL Spectrogram—show a reasonable balance between sensitivity and specificity. The given data suggests a positive potential for using cough audio to detect COVID-19. Nonetheless, additional investigation, validation, and real-world trials are imperative to confirm the dependability and precision of these methods when applied in practical healthcare scenarios.

Table 1. Performance metric of different models.

Extraction technique	FPR	Accuracy	F1-score	Precision	Recall
MFE	23.80%	81.80%	0.83	0.78	0.87
MFCC	29.00%	71.50%	0.71	0.71	0.72
MEL Spectrogram	34.00%	75%	0.74	0.70	0.83

4. Conclusion

This research investigates machine learning classification tasks based on CT images and cough audio recordings in order to promote the diagnosis of COVID-19. In the image classification task, the MobileNetV2 model was employed with various width multipliers compared. In the audio classification task, the model architecture was adjusted based on different feature extraction methods, and the performance of models using different extraction methods was compared. The experimental results demonstrate that MobileNetV2 performs well in the recognition of CT images, especially when the width multiplier is set to 1.0. Concerning audio data, feature extraction using MFE is found to be the most suitable for this task, yielding the best learning results. In the future, more comprehensive datasets related to COVID-19, encompassing both images and audio, along with additional medical expert annotations, will be essential. Strengthening image recognition can enhance diagnostic accuracy and assist in subsequent disease monitoring and precise lesion size delineation. Further advancements in the construction of new models for audio classification tasks are required to achieve more accurate predictions.

References

[1] World Health Organization 2023 https://www.who.int/

[2] Ozturk T et al 2020 Automated detection of covid-19 cases using deep neural networks with X-ray images Comput Biol Med p 103792

[3] Panwar H et al 2020 A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images Chaos Solitons Fractals vol 140 p 110190

[4] Jaiswal A et al 2020 Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning J Biomol Struct Dyn vol 0 no 0 pp 1-8

[5] Ismael AM and Şengür A 2021 Deep learning approaches for COVID-19 detection based on chest X-ray images Expert Syst Appl vol 164 p 114054

[6] Zhao J Y et al 2021 COVID-CT-Dataset: a CT scan dataset about COVID-19

[7] Coswara - A Database of Breathing Cough and Voice Sounds for COVID-19 Diagnosis https://arxivorg/abs/200510548

[8] ShiXuan E et al 2023 Can Knowledge of End-to-End Text-to-Speech Models Improve Neural Midi-to-Audio Synthesis Systems? In ICASSP 2023-2023 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) pp 1-5

[9] Muda L et al 2010 Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques arXiv preprint arXiv:10034083

[10] Sandler M et al 2018 Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition pp 4510-4520

[11] Howard A G et al 2017 Efficient convolutional neural networks for mobile vision application arXiv preprint arXiv:170404861