Computers, Materials & Continua
DOI: 10.32604/cmc.2022.029663
Article

Tech Science Press

Face Mask Recognition for Covid-19 Prevention

Trong Hieu Luu[1], Phan Nguyen Ky Phuc[2,*], Zhiqiu Yu[3], Duy Dung Pham[1] and Huu Trong Cao[1]

[1]College of Engineering, Can Tho University, Can Tho city, 910900, Viet Nam
[2]International University-Vietnam National University, Vietnam National University, HoChiMinh City, 70000, Vietnam
[3]Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, 10607, Taiwan
*Corresponding Author: Phan Nguyen Ky Phuc. Email: pnkphuc@hcmiu.edu.vn

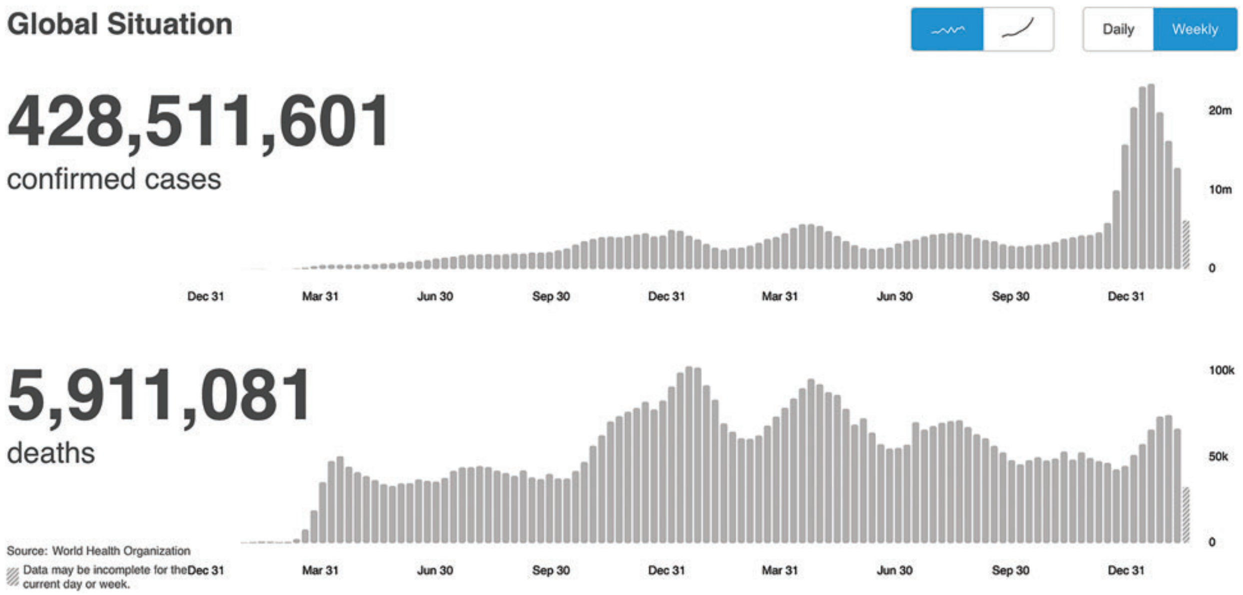
Received: 09 March 2022; Accepted: 08 May 2022

Abstract: In recent years, the COVID-19 pandemic has negatively impacted all aspects of social life. Due to ease in the infected method, i.e., through small liquid particles from the mouth or the nose when people cough, sneeze, speak, sing, or breathe, the virus can quickly spread and create severe problems for people's health. According to some research as well as World Health Organization (WHO) recommendation, one of the most economical and effective methods to prevent the spread of the pandemic is to ask people to wear the face mask in the public space. A face mask will help prevent the droplet and aerosol from person to person to reduce the risk of virus infection. This simple method can reduce up to 95% of the spread of the particles. However, this solution depends heavily on social consciousness, which is sometimes unstable. In order to improve the effectiveness of wearing face masks in public spaces, this research proposes an approach for detecting and warning a person who does not wear or misuse the face mask. The approach uses the deep learning technique that relies on GoogleNet, AlexNet, and VGG16 models. The results are synthesized by an ensemble method, i.e., the bagging technique. From the experimental results, the approach represents a more than 95% accuracy of face mask recognition.

Keywords: Face mask detection; deep learning; AlexNet; GoogLeNet; VGG16; ensemble method

1 Introduction

According to World Health Organization (WHO), Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. Depending on medical conditions, infected people suffer different types of symptoms. In most cases, people undergo mild to moderate respiratory illness and recover without requiring special treatment. However, due to the cytokine storm, i.e., excessive production of cytokines, some people have experienced severe aggravation and widespread tissue damage, which cause multi-organ failure and can lead to death. Older infected people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop severe illness and often require special treatments. Anyone can get sick with COVID-19 and become seriously ill or die at any age. According to the data of WHO, up to the time of this study, there are more than 420 million patients, and nearly 6 million people have passed away [1]. Fig. 1 shows the statistics of COVID-19 confirmed cases from December 2019 to December 2021.



Source: World Health Organization

▨ Data may be incomplete for the current day or week.

Figure 1: The statistics of COVID-19 confirmed cases from December 2019 to December 2021

According to Fig. 1, the death number had remained stable and had not decreased until the end of December 2021, when the pandemic lasted for more than two years and several treatments were proposed.

Apart from the high fatality rate, another prominent risk of this pandemic is that the patients can also continue to experience severe symptoms after their initial recovery even after receiving proper treatments. These health issues are also sometimes called "post-COVID-19 conditions" [2]. These are not limited to only elders and people with many severe medical conditions, and even young and healthy people can feel unwell for weeks to months after infection. Some common symptoms of "post-COVID-19 conditions" include fatigue, shortness of breath or difficulty breathing, cough, joint pain, chest pain, etc. Most of these symptoms derive from organ damage under the virus's effects. Although several types of vaccines have been developed and deployed in numerous countries, the appearance of new variants of COVID-19 and the attenuation of vaccine effects over time also create new challenges to slow down and prevent the spread of the pandemic. Currently, the most effective and cheapest method to protect people from infection is to stay at least one meter apart from others, wear a properly fitted mask, and wash hands with an alcohol-based rub frequently.

The major infected way of the virus is through small liquid particles from the mouth or nose when people cough, sneeze, speak, sing, or breathe. These particles range from larger respiratory droplets to smaller aerosols. Due to the infected mechanism, wearing the face mask properly can reduce up to 95% of the spread of the particles [3]. Recognizing the importance of properly face mask usage against COVID-19, this study tries to develop an effective method to classify whether a person has used the face mask correctly in the public area. By detecting people without a face mask or inappropriately wearing a face mask, the system can give a warning and show the image on the large screen so that other people can realize the potential risk and keep their distance from them. These systems are extremely helpful in the public space with a high density of people, such as supermarkets, classrooms, restaurants, or waiting rooms of airports. The system will save the headcounts as well as increase the consciousness of people of how the importance of wearing the face mask properly in the public area.

The rest of this study is organized as follows. Works related to our approach to deep learning and convolutional neural networks are reviewed in Section 2. In addition, the proposed methodology of this study is also described in this section. The experimental setup and results are presented in Section 3, while the conclusion and future works are presented in Section 4.

2 Related Works & Proposed Method

Convolution neural networks (CNN), first introduced by LeCun [4], is a deep learning model which can obtain high accuracy for the classification task. CNN principle is based on activities of the visual perception of the human brain when recognizing simple shapes. In general, CNN is constructed by using three different primal layer types, i.e., convolution layers, nonlinear layers, and filtering layers, which are stacked in several sequential orders, as shown in Fig. 2.

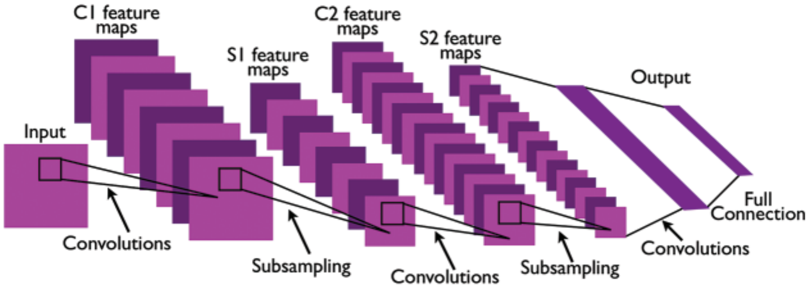


Figure 2: The CNN structure proposed by LeCun

In the proposed network, each layer is responsible for a specific function in the classification process. For example, the convolution layers will apply different filters to extract abstract features in the image content. A rectifier linear unit layer is often put behind the convolutional layer to produce the nonlinear property as well as more abstract information for subsequent layers. Apart from convolutional layers and rectifier linear unit layers, pooling layers are often employed in CNN to reduce the sample size while still retaining the most promising features for classifications. Lecun's study motivated other researchers and then created an explosion in introducing new deep learning models for image classification. Three of the most successful models adopted in this study are AlexNet, GoogLeNet, and VGG16. The structure of the AlexNet is shown in Fig. 3.

The AlexNet network was first introduced in [5]. This network won the first on The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. The AlexNet comprises several layers, including an input layer, five convolutional layers, five ReLu layers, two normalization layers, three pooling layers, three fully connected layers, a drop layer, and a softmax output layer. AlexNet takes an RGB image with the size of 224 × 224 as an input, and the output is a vector with the size of 1000 × 1. The prominent point in the AlexNet model is the employment of the dropout technique to prevent overfitting. In this technique, each node has the probability p to be selected for training in each epoch.

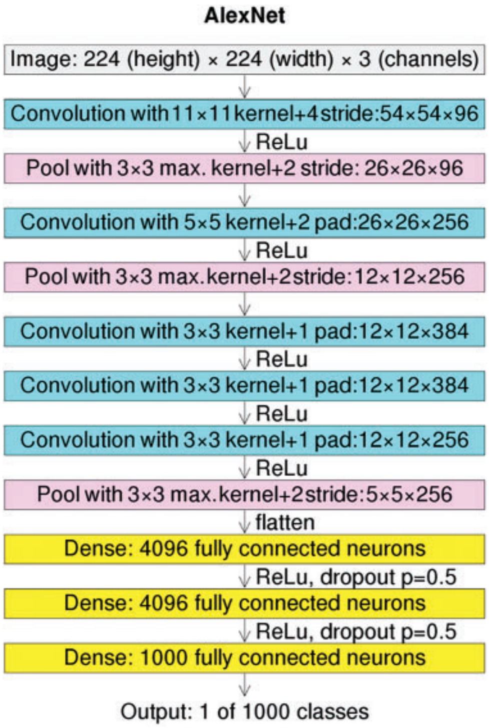

AlexNet
Image: 224 (height) × 224 (width) × 3 (channels)
Convolution with 11×11 kernel+4 stride:54×54×96
ReLu
Pool with 3×3 max. kernel+2 stride: 26×26×96
Convolution with 5×5 kernel+2 pad:26×26×256
ReLu
Pool with 3×3 max.kernel+2stride:12×12×256
Convolution with 3×3 kernel+1 pad:12×12×384
ReLu
Convolution with 3×3 kernel+1 pad:12×12×384
ReLu
Convolution with 3×3 kernel+1 pad:12×12×256
ReLu
Pool with 3×3 max.kernel+2stride:5×5×256
flatten
Dense: 4096 fully connected neurons
ReLu, dropout p=0.5
Dense: 4096 fully connected neurons
ReLu, dropout p=0.5
Dense: 1000 fully connected neurons
Output: 1 of 1000 classes

Figure 3: The AlexNet structure

Besides the AlexNet, this study also adopts the GoogLeNet [6] for the classification task. This network resulted from the combination between Google company and Cornell University, and it also won the ILSVRC in 2014. The structure of GoogLeNet is shown in Fig. 4.

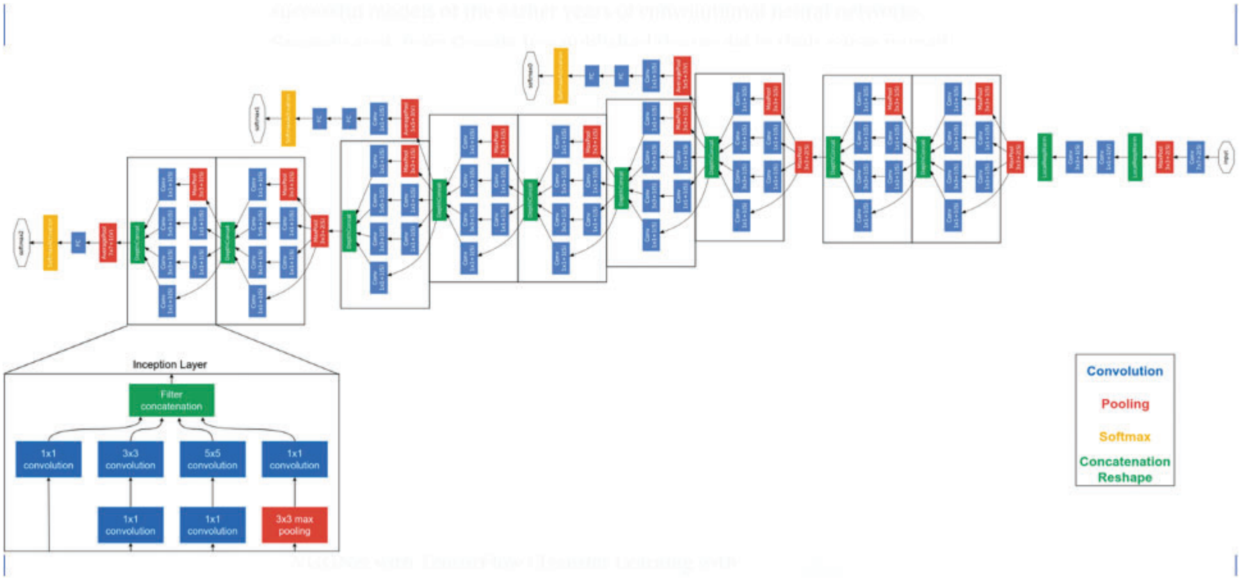


Figure 4: The GoogLeNet structure

The GoogLeNet differentiates itself from other deep learning models by using nine inceptions blocks from 22 deep layers and five pooling layers. These blocks allow the networks to conduct parallel learning. In other words, one input, whose size is 224 × 224, can be fed to several different convolutional layers to create different output features. These features then can be concatenated into an output. The parallel learning process helps the network itself to study and obtain more features than the traditional CNN approach. Moreover, the network also applies the 1 × 1 convolutional block to decrease the network size so as to quickly train the network. Fig. 5 shows the structure of an inception block in GoogLeNet. In the inception block, the outputs from the former layer are used as inputs, then fed into four different branches, and finally concatenated into one output.

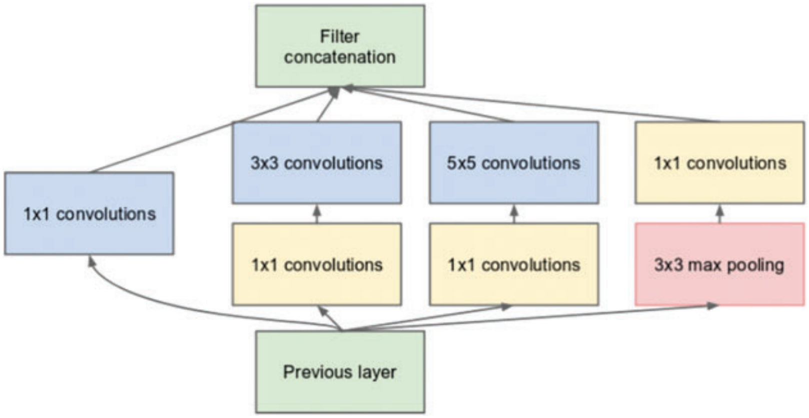


Figure 5: The intercept block structure

The last deep neural network integrated into the proposed model is VGG16. VGG16 [7] is a Convolutional Neural Network architecture first introduced in 2014 and achieved 92.7% top-5 test accuracy in ImageNet. Thestructure of VGG16 is described below in Fig. 6.

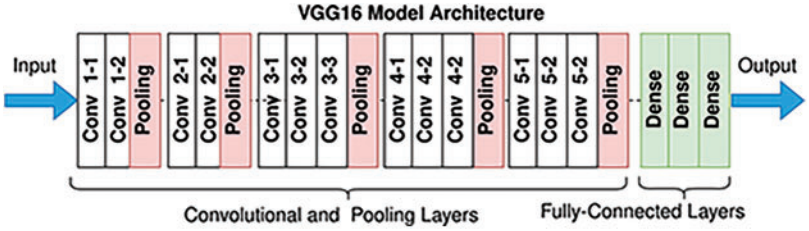


Figure 6: The VGG16 structure

VGG16 takes an RGM image with size 224 × 224 as an input. This input is passed to the whole network then the resulting output is passed to other layers to automatically extract features and make the classification. It is noted that the fourteen and fifteen layers are fully connected hidden layers of 4096 units, followed by a softmax output layer of 1000 units.

To utilize the power of these successful networks as well as to reduce the training time of these models, transfer learning and ensemble technique are applied in later steps. Generally, the main objective of transfer learning is to store knowledge gained while solving one problem and using it for a different but related problem [8]. In the case of CNN, transfer learning often relates to retraining some layers of well-trained models while freezing other layers. The selected layers for retraining are often the last layers of the model where the softmax function is applied to conduct classification. In this study, transfer learning is applied to all networks above.

In order to create the proper inputs for this network, the Viola-Jones algorithm [9] is applied to detect and extract faces from the image. The Viola-Jones algorithm basically comprises four main operations, including applying a Haar-like filter, calculating the integrated image, applying the Ada Boost algorithm, and returning the results based on a cascading classifier. The principle of the cascade classifier is given in Fig. 7.

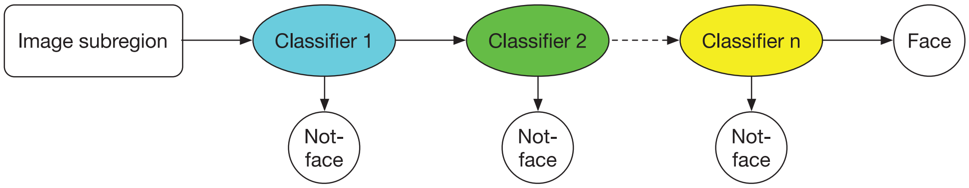


Figure 7: Principle of cascade algorithm

It is noted that in our proposed system, one raw image can have multiple faces that can belong to different label classes. As a result of that, the Viola-Jones algorithm returns several sub images of faces. Each face image must be passed to preprocessing phase for enhancing or downsizing so that it can be formatted properly, i.e., a size of 224 × 224, and sent to all networks as the inputs. Even though there are other methods for face and object detection [10–13], this study still adopts the Viola-Jones algorithm due to its robust and fast computation.

The returned output vectors from each network are then recollected for ensembling. In this study, the bagging technique of the voting method is applied at the ensembling stage for making final decisions and final classifications [14]. The final label is decided by majority rule. For example, if output vectors of AlexNet, GoogLeNet, and VGG16 are $v^A = [0.1, 0.8, 0.1]$, $v^G = [0.2, 0.7, 0.1]$, and $v^V = [0.2, 0.2, 0.6]$, respectively, according to the majority rule the image is classified as class 2 due to the agreement of AlexNet and GoogLeNet. If three networks give three different output results, then the label with the highest average weight is selected. For instance, if $v^A = [0.8, 0.1, 0.1]$, $v^G = [0.2, 0.7, 0.1]$, $v^V = [0.2, 0.2, 0.6]$, and $\overline{v} = \frac{v^A + v^G + v^V}{3} = [0.4, 0.33, 0.27]$, the face is classified as class 1. The main use of the ensemble technique is to improve the overall performance of the entire system. By combining several independent weak classifiers, the ensemble method can create a strong classifier with higher sensitivity. The diagram of the proposed system is described in Fig. 8.

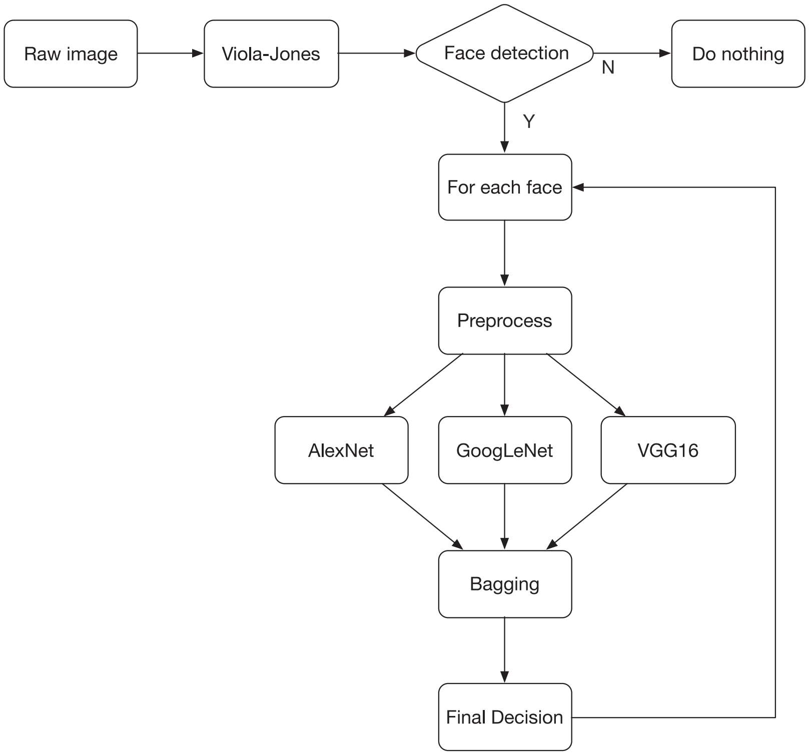


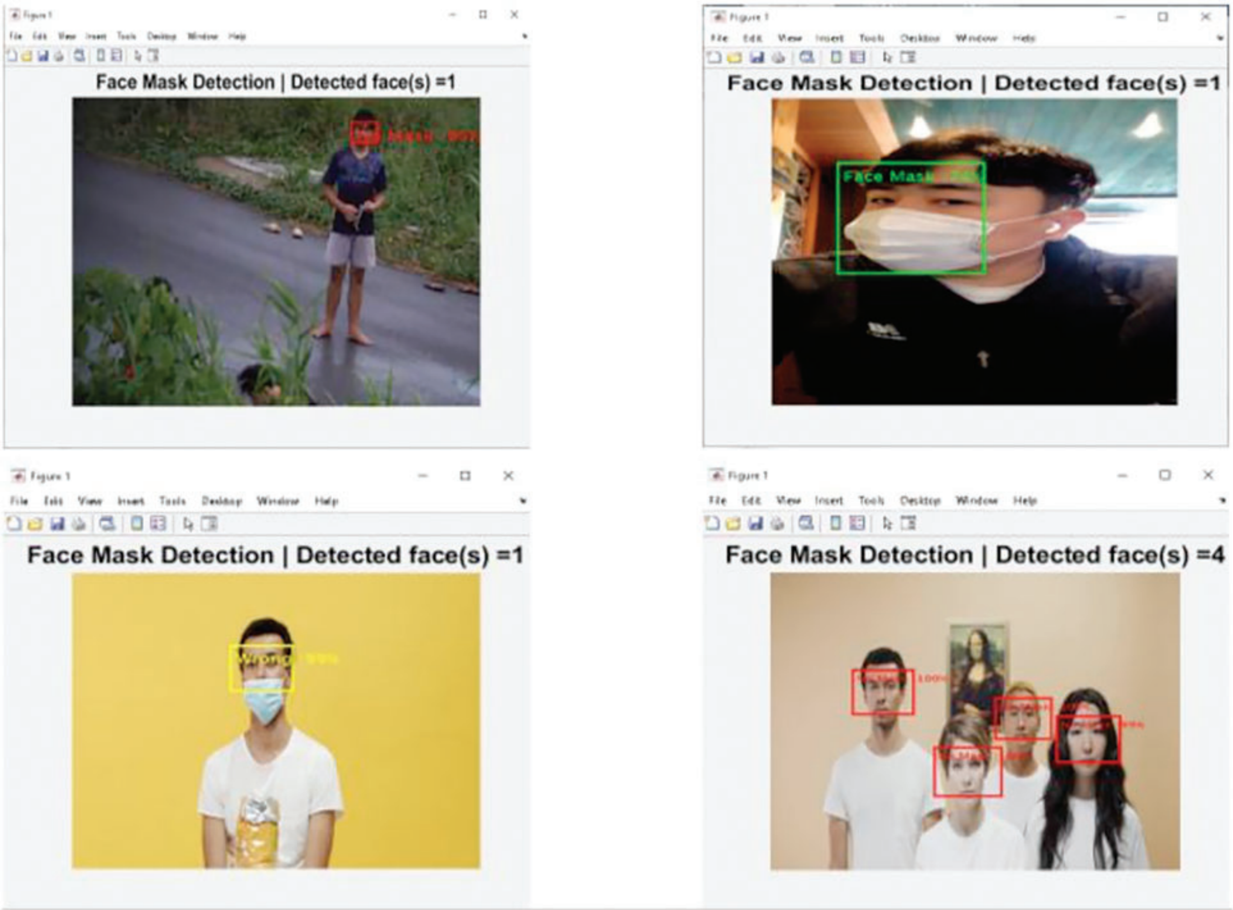
Figure 8: The proposed system structure

3 Data & Experimental Setup

The dataset, which was created by our team, includes three classes regarding three respective outputs as described above. The number of samples for each class label in the training set is shown in Tab. 1. In the training set, several mask types as well as the number of faces in the image are selected to ensure diversity. However, one image only includes one type of output labels, as shown in Fig. 9. Each image's resolution is 1280 × 960, and it will be resized to fit the input of each deep learning method. Only 80% in each class is used for training, and 20% is used for validation.

Table 1: The characteristic of the training data set

Class	Details	Number of images
C1	No mask	1200
C2	Misuse	1200
C3	Wear mask	1200



Face Mask Detection | Detected face(s) =1

Face Mask Detection | Detected face(s) =1

Face Mask Detection | Detected face(s) =1

Face Mask Detection | Detected face(s) =4

Figure 9: Images from the training data set

In this study, the stochastic gradient descent method [15] is applied for training each network with a learning rate of approximately 0.0001. The batch size is 30, and the number of epochs is 120. The division of data set into small batches with the size of 30 is to ensure loading capability. Loading the whole training data set at once is impractical for most problems due to RAM limitations. In our cases, 80% of the training data set is 3600, which requires 120 epochs of training. The selected objective function is the cross-entropy loss function. In general, the learning rate must be chosen very carefully to avoid a low learning process as well as divergence. Overfitting is prevented by the dropout technique, and the pixel values of the input image are normalized between 0 and 1.

To test the accuracy of the proposed system in reality, the testing set is designed in a different way. In the testing set, each image can have more than one type of output labels. The testing set now includes nine classes. The details of the testing set are given in Tab. 2, and images in the testing set are shown in Fig. 10.

Table 2: The details of the testing set

Class	Details	Number of images
C1	1 faces without mask	150
C2	2 faces without mask	150
C3	1 face with mask misuse	150
C4	2 faces with mask misuse	150
C5	1 face with mask	150
C6	2 faces with mask	150
C7	1 face with mask and 1 face with mask misuse	150
C8	1 face with mask and 1 face without mask	150
C9	1 face without mask and 1 face with mask misuse	150

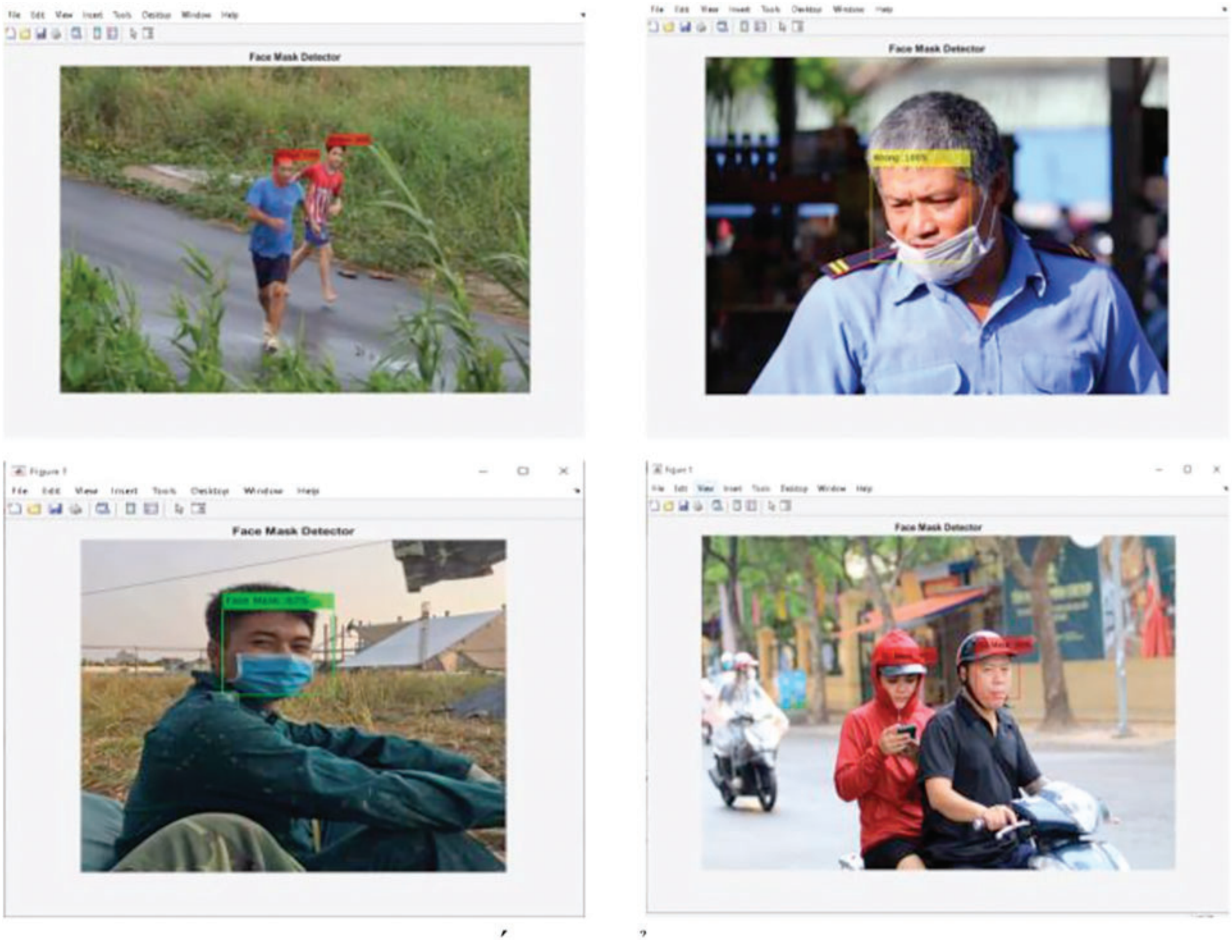


Figure 10: Images from the testing set

Tab. 3 presents the results of the proposed system in the form of a confusion matrix. Finally, Tab. 4 will summarize the performance of each network in the proposed system on the different classes of the test set. It is noted that, even though one class can be wrong classified as another class, the half of content inside the image of a class can still be accurate since the return output of the proposed system includes only “NoMask”, “Misuse”, and “WearMask”.

Table 3: The confusion result matrix

Actual label	C1	148	2	–	–	–	–	–	–	–
	C2	2	148	–	–	–	–	–	–	–
	C3	–	–	144	3	3	–	–	–	–
	C4	–	–	–	129	–	8	13	–	–
	C5	–	–	–	–	149	1	–	–	–
	C6	–	–	–	–	3	147	–	–	–
	C7	–	–	–	21	–	10	119	–	–
	C8	–	–	–	–	–	–	–	141	9
	C9	13	–	6	–	–	–	–	–	131
		C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**
							Predicted label			

Table 4: Performance of each network in the proposed system.

Class	Total sample	Alexnet	Googlenet	VGG16	Sensitivity
C1	150	98.6%	98.6%	98.6%	99.94%
C2	150	96.2%	97.4%	96%	99.65%
C3	150	94%	96.5%	94%	99.24%
C4	150	82.3%	86.1%	79.5%	92.07%
C5	150	98.6%	98.6%	98.6%	99.94%
C6	150	98.6%	98.6%	98.6%	99.94%
C7	150	77.2%	79.3%	76.1%	87.14%
C8	150	91%	94%	91%	98.21%
C9	150	83.7%	87.2%	81.1%	93.2%

During the testing phase, we also test the capability of the proposed system with some very special types of face masks, i.e., the transparent face mask, and funny face mask which are shown in Fig. 11

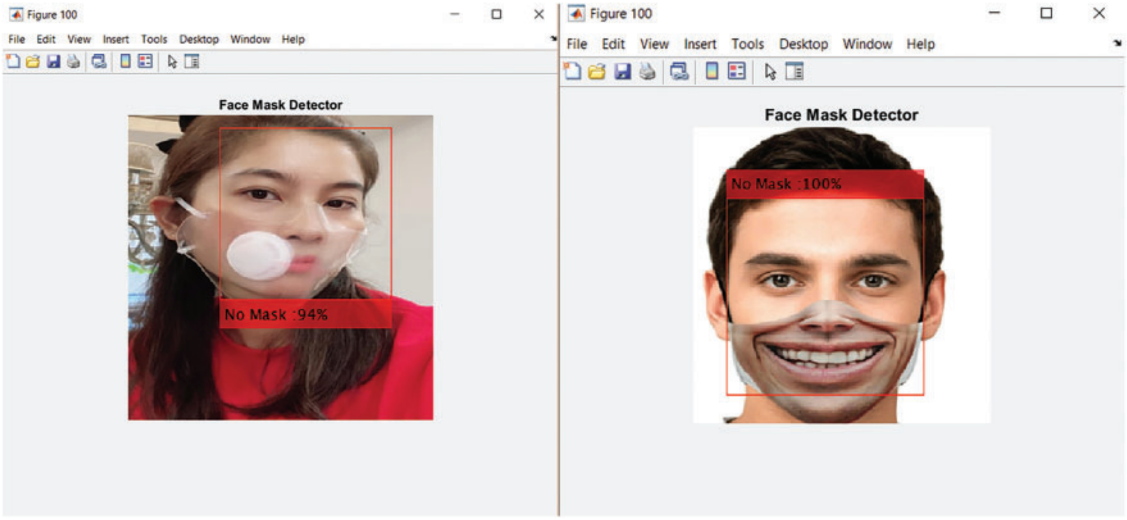


Figure 100
File Edit View Insert Tools Desktop Window Help
Face Mask Detector
No Mask :94%
Figure 100
File Edit View Insert Tools Desktop Window Help
Face Mask Detector
No Mask :100%

Figure 11: Transparent and funny face masks

Given the funny face masks, all three networks cannot perform really well. The proposed system cannot classify correctly in about 85% of test cases. This can be seen as the current limitations of our approach. In the case of transparent face masks, through carefully checking the output of each separate network, the probability that a network labeled it as the "WearMask" is not high compared to the other two classes. Each separate network often wrong assigned them to the label "Misuse" since both the mouth and nose also appear in the image while the rest of the face is covered by an exhalation valve. This phenomenon is the main explanation for the poor classification of class C7, which incidentally includes several images of transparent face masks. As explanations, these transparent face masks are often misclassified as "Misuse" with falls into C4.

It can be seen that, except in the case of transparent facemask and funny face mask, the proposed system has a very high sensitivity for the rest. It can be explained by the high sensitivity of each network. Given a class, if all networks are assumed as independent of other networks and the sensitivity of network i is p_i, with three networks probability that network i correctly classified the given class as p_i. There are four cases that the proposed system can correctly make the classification. The first case is that all three networks correctly make the classification. The rest is when one network assigns wrong labels, but the others make decisions correctly. The probability that here the sensitivity of proposed whole system when using voting process classifies a face correctly regarding probability theory is under this assumption is given by Eq. (1).

$$P\,(sys) = p_1p_2p_3 + p_1p_2\,(1 - p_3) + p_1\,(1 - p_2)\,p_3 + (1 - p_1)\,p_2p_3 \tag{1}$$

It can be seen that this probability is much higher than the probability of one network.

4 Conclusion & Future Works

Wearing face masks in an effective, simple, and cheap way can prevent the spread of the COVID virus in the public space. However, due to some reasons, not wearing a face mask and misusing a face mask still can be met very frequently in public spaces. These behaviors can spread the droplet and aerosol from person to person, which will increase the risk of virus infection. As a result of this, a system to detect such behavior and give proper warnings in public spaces is therefore required. This kind of system is extremely necessary for close public spaces such as university classes or offices. This study tries to propose a real time system to detect the proper usage of face masks in the public space. By combining Viola-Jones, three different deep learning neural networks, and the voting process, the proposed system has effectively handled normal face mask-wearing issues. However, the proposed system still has an obvious limitation on correctly labelling when it has to handle transparent and funny face masks. This limitation will be considered in future research.

Acknowledgement: The authors thank TopEdit (www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Report of WHO on Coronavirus (COVID-19) Dashboard. https://covid19.who.int/WHO-COVID-19-global-table-data.csv.

[2] J. B. Soriano, S. Murthy, J. C. Marshall, P. Relan and J. V. Diaz, "A clinical case definition of post-COVID-19 condition by a Delphi consensus," *The Lancet Infectious Disease*, vol. 1, no. 1, pp. 1–5, 2021.

[3] M. Liao, H. Liu, X. Wang, X. Hu, Y. Huang *et al.,* "A technical review of face mask wearing in preventing respiratory COVID-19 transmission," *Current Opinion in Colloid & Interface Science*, vol. 52, no. 1, pp. 1–19, 2021.

[4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard *et al.,* "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[5] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolution neural networks," in *Proc. NIPS 2012*, Nevada, USA, pp. 1–12, 2012.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," 1, arXiv.1409.4842, 1–12, 2014.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, pp. 1–14, 2015.

[8] K. Weiss, T. M. Khoshgoftaar and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 9, pp. 1–40, 2016.

[9] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 10, pp. 137–154, 2004.

[10] B. Yang, J. Yan, Z. Lei and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. IJCB*, Clearwater, USA, pp. 1–8, 2014.

[11] X. X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. CVPR*, Providence, USA, pp. 2879–2886, 2012.

[12] S. Yang, P. Luo, C. C. Loy and X. Tang, "Wider face: A face detection benchmark," in *Proc. CVPR*, Lasvegas, USA, pp. 5525–5533, 2016.

[13] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 1, no. 1, pp. 1–16, 2021.

[14] L. Rokach, "Ensemble-based classifier," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

[15] A. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. CVPR*, Columbus, Ohio, USA, pp. 806–813, 2014.