All Sciences Proceedings

1st International Conference on Scientific and Innovative Studies

April 18-20, 2023 : Konya, Turkey

ICSIS

All Sciences Proceedings
http://as-proceeding.com/

https://as-proceeding.com/index.php/icsis

Multimodal CNN-Based System For Mask And Maskless Face Detection

Saed ALQARALEH

Computer Engineering Department, Hasan Kalyoncu University, Turkey

saed.alqaraleh@hku.edu.tr

Abstract **–** Face masks existed for much longer before the pandemic corresponding to COVID-19, wherever staff in several sectors, such as medical, chemical, and nuclear, needed to wear masks throughout duties. Following the pandemic caused by the COVID-19 virus, most countries requested publicly covering the nose and mouth as vital life to keep the communities safe. However, 24/7 human superintendence is almost impossible.

In this paper, an efficient and automatic multimodal face mask detection was developed. The model was engineered based on intensive investigations, where first, the performance of two well-known deep learning models, particularly MobileNetV2 and VGG19, was investigated. Next, the performance was further improved using the late fusion principle. Four datasets consisting of roughly 6K, 12K, 4k, and 4k images, respectively, are used to confirm the results robustness of the developed model. Overall, the results of the experimental works showed that fusion leads to a more stable and outperforming model compared to five base CNN models, i.e., MobileNetV2, VGG19, and three sequent models.

Keywords – Convolution Neural Network, COVID-19, Deep Learning, Image Classification, Mask Detection, Multimodal Classification System

I. INTRODUCTION

Image classification typically works by assigning categories or labels to associate input images. Manually classifying is a straightforward task for our brain; however, it is feverish once we process an enormous range of images. Hence, an automated, efficient classifier is preferred. Following the pandemic caused by the COVID-19 virus, most countries require covering the nose and mouth in public places as a necessary live to keep the communities safe. Thus, as physical 24/7 oversight is nearly impossible, developing an alternative way to monitor people who seem to need to comply with the principles of wearing masks wherever it is mandatory. This downside may be solved by taking advantage of deep learning and machine learning and using some image process algorithms, and the system can be easily used worldwide. In addition, such a system can be designed by taking advantage of the spectacular performance achieved by general deep learning and Convolutional Neural Networks (CNN) in most computer-based systems.

Furthermore, the performance can be further improved using the multimodal principle, where the system uses more than one classifier, providing an output with a richer performance than individual modalities.

The main contributions of this paper are:

i. Investigating the performance of two deep learning models, particularly Convolutional neural networks(MobileNetV2, VGG19).

ii. Budling is an efficient multimodal face mask detection based on the late fusion of MobileNetV2 and VGG19.

iii. To ensure the results robustness of the developed model, four datasets consisting of 6K, 12K, 4k, and 4k images, respectively, have been used to compare the performance with the following base models: MobileNetV2, VGG19, and three sequential models.

The topic of improving classification systems in general and building mask detection systems, in particular, has recently attracted the researcher's attention. Some recent mask detection studies are summarized below. [1] introduced a DNN network model named SSDMNV2 that uses Single Shot Multibox Detector as a face detector and the MobileNetV2 to recognize face masks in real-time without consuming many resources. This model was trained and tested using a dataset combined of the Kaggle dataset that has 678 images of people wearing medical masks and an XML file containing descriptions of these images, and the "Prajna Bhandary" dataset that contains Artificial Intelligence created 1376 pictures; 690 images are for people wearing masks and 686 without masks. As a result of using the SSDMNV2 model with the MobilenetV2 classifier, the system was able to achieve a 92.64% accuracy.

In [2], a new model with three CNN network stages was developed. A Proposal Network (P-Net) obtains the candidate windows associated with their bounding box regression vectors and is employed in stage one. Stage 2 uses a Refine Network (R-Net) to accomplish activity mistreatment bounding box regression. The non-maximum suppression (NMS) candidate merges and rejects several incorrect candidates. Finally, stage three uses an output network (O-Net) that describes the face in additional detail. Specifically, this network can output the positions of 5 facial landmarks. Three datasets were employed in this to validate the work of [2], which was ready to achieve an associate accuracy of 97.14%.

Another face mask detection system called FMD-Yolo is proposed in [3]. This model uses Res2Net and deep residual networks in feature extraction. Next, the extracted features are fusion using the En-PAN network. Also, the Matrix NMS method was adapted at the inference stage to improve detection efficiency. Results indicated that the proposed model achieved a precision AP50 of 92.0% and 88.4% for the two used datasets.

II. THE DEVELOPED MASK DETECTION SYSTEMS

In general, image classification (shown in Figure 1) and the proposed mask detection systems consisted of the following main steps.

A. Training and Testing Data Preparation:

In this work, four mask detection image datasets that are publicly available were used. These datasets are Dataset 1 [4]: The total number of images in the dataset is 5988; it has three classes, 2994 images where their class is "with a mask," a class without a mask has 2994 images, and the remaining images represent class for masks that are worn incorrectly. In this paper, we utilized only the first two categories. This dataset was created by Vijay Kumar and cleared and equally distributed across each class. The dataset contains images with a unified dimension in which each image is 128 x 128.

Dataset 2 [5]: This dataset was created by scraping images from Google and Jessica Li's CelebFace dataset. The two classes' dataset consists of almost 12K images, where the first class consists of images for people "with masks" and the other "without masks." The images' dimensions are not fixed, so the images are of different sizes, where the smallest size is 25x25, and the largest one is 563x563.

The RMFD dataset, i.e., "Dataset 3" [6]: is created by combing a Kaggle dataset and photos gathered via the Bing search API. This dataset has 4079 photos split into "with mask" and "without mask" classes. The photos' sizes vary, with the lowest being 26x37 and the biggest 4734x5412.

Face Mask Detection (FMD) dataset, i.e., "Dataset 4" [7]. This dataset represents the natural world, as all its images are of real people with masks (unlike the other datasets that have some images created by AI), and it has 3833 images, where 1915 images are with masks and 1918 without masks. The images in this dataset are of different dimensions and similar to "Dataset 3", the smallest size is 26x37, and the largest is 4734x5412.

B. Feature extraction:

Mainly, this step works on detecting and extracting distinguishing features that represent the input image. In this work, the array resulting from the pre-processing step will be normalized to the range [-1 to 1]. Then, a vector of 49152 features is extracted to represent the image.

C. System classifier:

Here, we work on selecting the classification algorithm, an essential part of an efficient classification system. This work investigates the performance of five CNN-based mask detection models, and their details are summarized below.

"VGG19": VGG is a Convolutional Neural network that stands for Visual Geometry group proposed by Simonyan and Zisserman in 2014 at the University of Oxford. VGG19 consisted of 19 layers, where 16 convolution layers are used for feature extraction, and three fully connected layers are used for classification. Note that 5 MaxPooling layers follow each convolution layer in VGG19. Mainly, VGG19 was pre-trained using the ImageNet database, which has 14 million images that belong to 1000 classes. In this paper, we adapted the mask detection model of [8], based on VGG19, and uses a fixed size of RGB input images (224*224).

"MobileNet-v2": The model was introduced in early 2018 by google. It is consisted of 53 layers and was also trained initially on the ImageNet database. MobileNetV2 is an improved version of MobileNetV1 that was developed to offer the possibility of developing mobile-oriented deep learning models with low computational power. It is worth mentioning that MobileNetV2 is a 35% faster and more effective extracting tool compared to V1 while still able to achieve the same accuracy. In this paper, we adapted the mask detection model of [9] based on the pre-trained MobileNetV2, and the input images are resized into 224x224.

"SEQUENTIAL1": This model was created using Keras and consists of two convolution layers (Conv2D), where the first one has 200 filters of (3x3) and the second one with 100 filters of (3x3). Both convolution layers are followed by ReLU, MaxPooling, flatten, and Dropout layers. Then, two dense layers were used, where the first has 50 nodes(neuron) and ReLU activation, and the other has two nodes and SoftMax activation [10].

"SEQUENTIAL2": This model consists of four convolution layers (Conv2D). The first layer has 64 filters of (3x3), and the second layer has 256 filters of (3x3). These two layers are followed by the ReLU activation layer and the MaxPooling2D layer (2x2). Then, the third convolution layer has 128 filters of (3x3), followed by the ReLU activation layer and Dropout layer. The last convolution with 32 kernel/filter of (3x3) is followed by the ReLU activation layer and MaxPooling2D layer of (2x2). Next, another Dropout and flatten layer are added to this model, followed by three Dense layers, 100 nodes and ReLU activation, 16 nodes and ReLU activation, and two nodes with SoftMax activation, respectively [11].

"SEQUENTIAL3": This model was created sequentially like the previous two models. This model has three convolution layers, where all of these three layers have 32 filters of (3x3) followed by the ReLU layer and MaxPooling layer of (2x2). The model also has flatten layer and a dense layer of 100 nodes and ReLU activation; next is the Dropout layer, and the last layer is another dense layer with two nodes and SoftMax activation [12].

In this work, one further step is performed to achieve our goals, i.e., the fusion mechanism was adopted, which is the joining of information from two or more modalities to perform a more comprehensive prediction. Fusion can generally be implemented using early, Intermediate/Joint, and Late/Decision Fusions. Based on our preliminary investigation, this paper shows that late fusion schemes tend to perform better than others adopted. Figure 2 shows the proposed late fusion mask detection system

D. Class assignment (Classification output):

Here, each input image is assigned to one of the predefined two classes, i.e., a person wearing a face mask, whereas the second class is a person who is not.

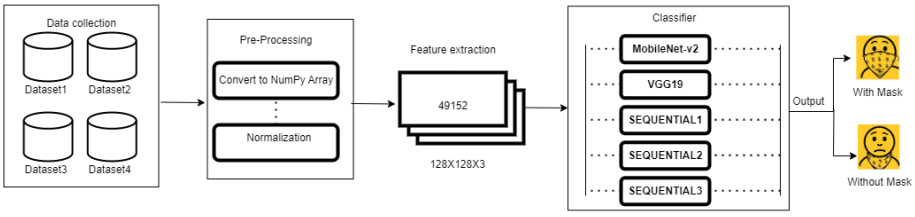

Fig. 1. Main steps of the image classification system.

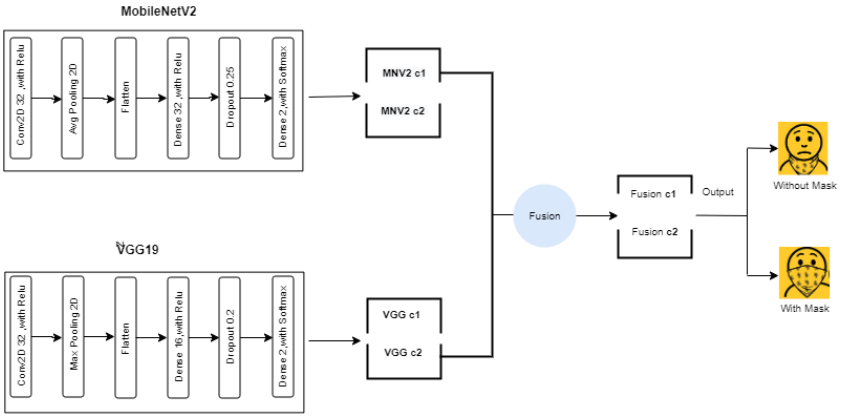

Fig. 2. The structure of the proposed late fusion classifier. Where VGG c1 and VGG c2 represent the membership of the input sample to the two predefined classes, i.e., with Mask and Without Mask classes using VGG19; similarly, MobileNetV2 produces MNV2 c1 and MNV2 c2.

III. EXPERIMENT AND RESULTS

In this section, we first compare the performance of two well-known CNN models, i.e., MobileNetV2 and VGG19. Then we worked on improving the performance by applying the late fusion and to produce our enhanced model. Next its performance was compared with five CNN base models. In addition, the accuracy, precision, recall, and F1 score evaluation metrics were used to guarantee the results' reliability. Note that the 3-fold cross-validation method is also used.

Experiment 1. Performance of MobileNetV2 and VGG19

In this experiment, the performance of MobileNetV2 and VGG19 was investigated using two datasets. The first consisted of 12K and the second 4k images. Results are shown in Figure 3, and both models achieved a pretty good performance using the first dataset. However, MobileNetV2 was able to outperform VGG19 when it comes to the second dataset, which has low-quality images.

Experiment 2. Performance of the Developed Late Fusion Model vs Some State of Art CNN Classification Models

This experiment compares the performance of the developed model vs five mentioned base CNN models, i.e., MobileNetV2, VGG19, and three Sequential models (SEQUENTIAL1, SEQUENTIAL2, SEQUENTIAL3) using the four mentioned datasets. Results are shown in Figure 4, and the stability of the performance over time and using the four datasets is shown in Figure 5. Based on these results, the following can be concluded:

a) The developed model outperformed all five investigated models. In addition, the base version of MobileV2 and VGG19 achieved second and third-best performance, respectively.

b) Based on the stability of the performance over time and using the four data sets, as shown in Figure 4, although all other models, including MobileV2 and VGG19, significantly suffered from performance degrading when it came to processing some challenging actual life samples when our model was able to keep its performance successfully and process this dataset while achieving high performance.

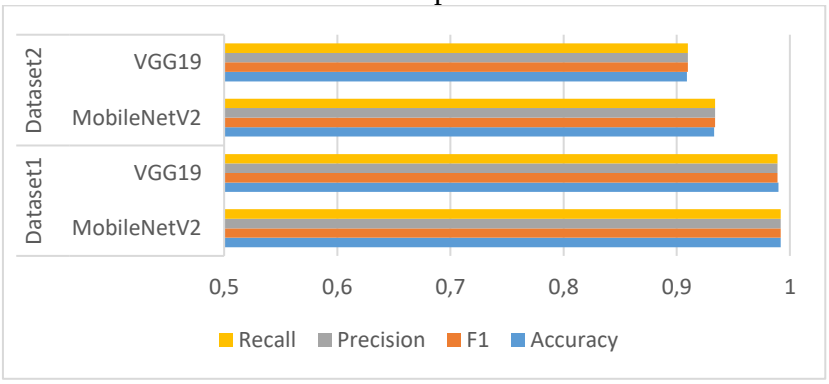

Fig. 3. The Accuracy, F1, Precision, and Recall for MobileNetV2 and VGG19 using the first and second datasets.

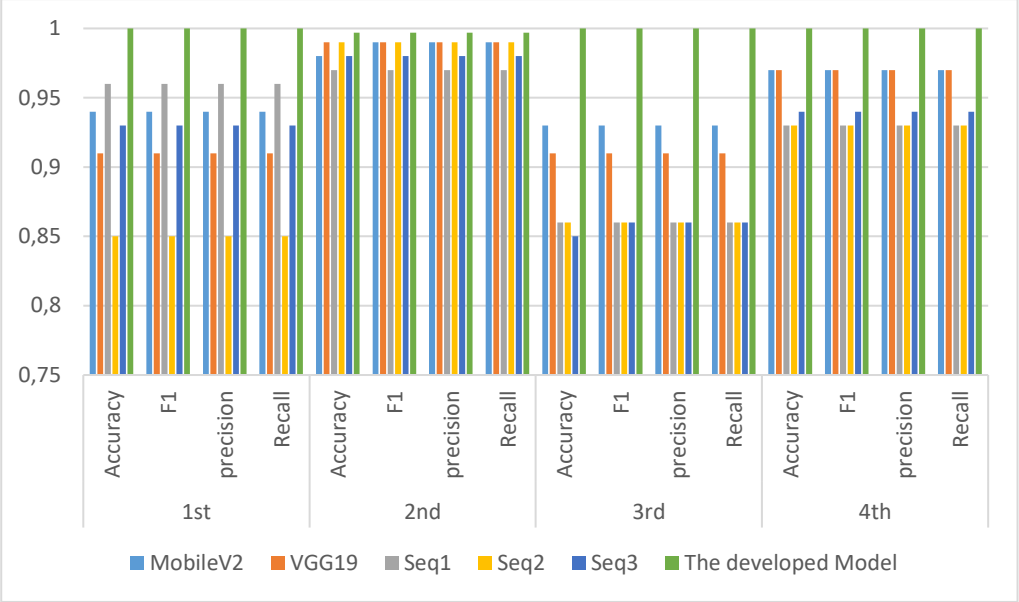

Fig. 5. The Accuracy, F1, Precision, and recall for the developed model, the base version of MobileNetV2, VGG19, and three Sequential models using the four used datasets.

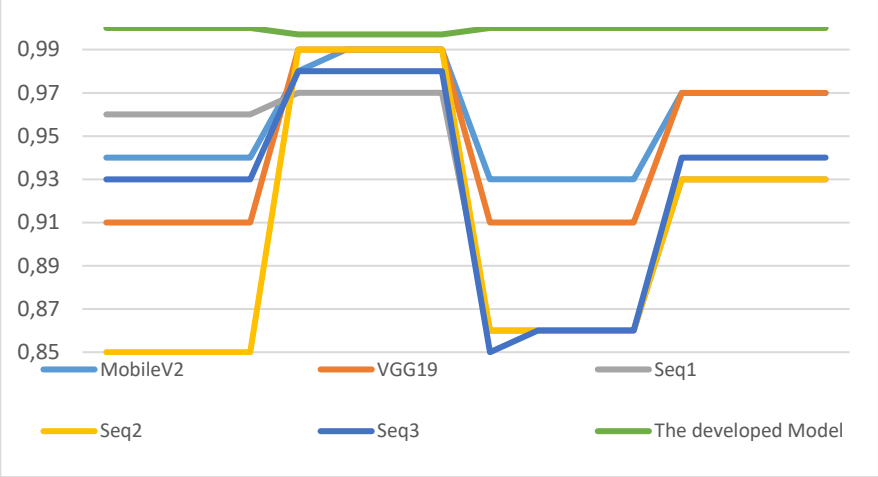

Fig. 6. The performance stability over time and using the four data sets for the developed model and the base version of MobileNetV2, VGG19, and three Sequential models.

IV. CONCLUSION

Following the pandemic of COVID-19 virus, we have been required to cover our noses and mouth in public places to a critical degree to hold the communities safe. The 24/7 human manual control and supervision is complex and sometimes near impossible. As a solution, an automatic tracking machine that can report individuals who are not complying with the regulations of wearing a mask is preferred.

This paper introduced a fusion and CNN-based automatic mask detection system. Its performance was compared with five deep learning models, along with MobileNetV2 and VGG19, while it was used for mask detection. Overall, the developed model could efficiently classify the most complicated samples and show performance stability over time and multiple datasets.

Investigating more advanced feature extraction and encoding methodologies can be one path for future work. Another path is investigating the opportunity of taking advantage of the images associated with textual content whenever available.

REFERENCES

[1] Nagrath, P., Jain, R., Madan, A., Arora, R., Kataria, P., & Hemanth, J. (2021). SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. Sustainable cities and society, 66, 102692.

[2] Özkaya, U., Öztürk, Ş., & Barstugan, M. (2020). Coronavirus (COVID-19) classification using deep features fusion and ranking technique. Big Data Analytics and Artificial Intelligence Against COVID-19:

Innovation Vision and Approach, 281-295.Wu, P., Li, H., Zeng, N., & Li, F. (2022). FMD-Yolo: An efficient face mask detection method for COVID-19 prevention and control in public. Image and vision computing, 117, 104341.

[3] VIJAY KUMAR, (2021) Face Mask Detection, [Online]. Available: https://www.kaggle.com/datasets/vijaykumar1799/face-mask-detection.

[4] ASHISH JANGRA, (2021) Face Mask Detection~12K Images Dataset, [Online]. Available: https://www.kaggle.com/datasets/ashishjangra27 /face-mask-12k-images-dataset.

[5] Real World Fasked Face Recognition Dataset (RMFRD), (2021) [Online]. Available: https://www.kaggle.com/datasets/ muhammeddalkran/masked-facerecognition.

[6] Balaji S, (2021) Face-Mask-Detection, [Online]. Available: https://github.com/balajisrinivas/Face-Mask-Detection/tree/master /dataset/ with_mask.

[7] Teboulbi, S., Messaoud, S., Hajjaji, M. A., & Mtibaa, A. (2022). Face Mask Classification Based on Deep Learning Framework. In Advanced Practical Approaches to Web Mining Techniques and Application (pp. 175-188). IGI Global.

[8] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[9] Perceptron, (2021) face-mask-detection-keras, [Online]. Available: https://github.com/aieml /face-mask-detection-keras.

[10] Real-time Face Mask Detection with OpenCV, [Online]. Available: https://projectgurukul.org/face-mask-detection/.

[11] Karan Malik, (2021) FaceMaskDetector, [Online]. Available: https://github.com/Karan-Malik/FaceMaskDetector.