Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.036973

ARTICLE

Deep Learning-Based Mask Identification System Using ResNet Transfer Learning Architecture

Arpit Jain[1], Nageswara Rao Moparthi[1], A. Swathi[2], Yogesh Kumar Sharma[1], Nitin Mittal[3], Ahmed Alhussen[4], Zamil S. Alzamil[5,*] and MohdAnul Haq[5]

[1]Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, 522302, India

[2]Department of Computer Science and Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, 500068, India

[3]University Centre for Research and Development, Chandigarh University, Mohali, 140413, India

[4]Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

[5]Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

*Corresponding Author: Zamil S. Alzamil. Email: z.alzamil@mu.edu.sa

Received: 18 October 2022 Accepted: 07 April 2023 Published: 19 March 2024

ABSTRACT

Recently, the coronavirus disease 2019 has shown excellent attention in the global community regarding health and the economy. World Health Organization (WHO) and many others advised controlling Corona Virus Disease in 2019. The limited treatment resources, medical resources, and unawareness of immunity is an essential horizon to unfold. Among all resources, wearing a mask is the primary non-pharmaceutical intervention to stop the spreading of the virus caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) droplets. All countries made masks mandatory to prevent infection. For such enforcement, automatic and effective face detection systems are crucial. This study presents a face mask identification approach for static photos and real-time movies that distinguishes between images with and without masks. To contribute to society, we worked on mask detection of an individual to adhere to the rule and provide awareness to the public or organization. The paper aims to get detection accuracy using transfer learning from Residual Neural Network 50 (ResNet-50) architecture and works on detection localization. The experiment is tested with other popular pre-trained models such as Deep Convolutional Neural Networks (AlexNet), Residual Neural Networks (ResNet), and Visual Geometry Group Networks (VGG-Net) advanced architecture. The proposed system generates an accuracy of 98.4% when modeled using Residual Neural Network 50 (ResNet-50). Also, the precision and recall values are proved as better when compared to the existing models. This outstanding work also can be used in video surveillance applications.

KEYWORDS

Transfer learning; depth analysis; convolutional neural networks (CNN); COVID-19

1 Introduction

The World Health Organizations (WHO) 209th report, issued on August 16, 2020, said Coronavirus Disease-19 (COVID-19). The Severe Acute Respiratory Syndrome-CovidVirus 2 (SARS-CoV2) virus has infected people worldwide. In the early days of COVID, 6 million individuals resulted in 379,941 fatalities globally [1]. The key to controlling the COVID-19 pandemic, according to the Pan American Health Organization (PAHO), is to preserve social distancing while boosting monitoring and fortifying [2] Healthcare systems. Recently, research was published on evaluating measures for University researchers are tackling the COVID-19 epidemic. According to Edinburgh University, wearing a face mask or other covering reduces the danger of Coronavirus spread. More than 90% of the distance is traveled by the exhaled breath [3]. Due to the global COVID-19 outbreak, adopting a facial mask in society is becoming more common. Because the citizens of COVID-19 refuse to protect their health by wearing air pollution masks [4], others, in contrast, are modest in their looks and hide their emotions from the general public by shifting their own faces. According to one source, using face masks can help reduce COVID-19 transmission. COVID-19 is the most recent pandemic virus to threaten human health in the twentieth century [5]; because of the rapid spread of COVID-19, WHO declared it a worldwide pandemic in 2020. The symptoms experienced by COVID-19 patients range widely from mild signs to significant sicknesses. Breathing difficulties or shortness of breath are examples of respiratory issues among them.

COVID-19 infected almost 5 million people in 188 countries in less than six months, as shown in Fig. 1: The virus spreads through close interaction in heavily crowded areas. The emergence of the Coronavirus has prompted unprecedented levels of worldwide scientific cooperation. Deep learning and machine learning, powered by computer science, will help in various ways in the fight against COVID-19. Machine learning analyses enormous amounts of information to estimate COVID-19 spread, act as a brief alerting system for potential pandemics, and classify vulnerable groups. Many nations have laws mandating people to wear face masks when in a crowd. We tend to establish such laws and regulations in response to the exponential growth of occurrences and fatalities in various fields. As a consequence, recognizing face masks is a challenging task. Due to the spread of the Coronavirus, many nations have implemented laws such as “No entry without a mask”. Face mask detection is a significant concern in COVID-19 prevention and security [6]. In the medical field, a mask reduces the risk of infection from an infected person, regardless of their symptoms. Face mask detection is everywhere, including airports, hospitals, businesses, and educational institutions. Face recognition without a mask is more ac-accessible, while face recognition with a mask is more challenging since mask face extraction of features is more complicated than ordinary face feature extraction.

The COVID-19 pandemic has grown and mutated since December 2019 (September 2, 2021), infecting more than 216,026,420 persons and resulting in 4,495,014 fatalities [7]. Researchers are rushing to understand the pathogen and provide practical countermeasures. The coronavirus cases verified by Worldometer Globally are 216,918,733 cases, 4,511,302 fatalities, and 193,849,589 recoveries [8]. Information on COVID-19, including confirmed instances, recovered cases, and deaths, was compiled by Google News COVID-19 [9] and Worldometer. COVID-19 symptoms can cause the illness, ranging from mild to severely hazardous. The incubation period for COVID-19 symptoms is 2 to 15 days after infection. Thus, anyone exhibiting COVID-19 symptoms is kept under active medical supervision under quarantine (isolation wards). The disease’s primary symptoms are shortness of breath, a sore throat, a cough, a fast pulse, chest discomfort, and fever. Typically, the virus may be transmitted from one person to another by respiratory droplets created while coughing. It can also be transmitted through contact, such as touching filthy surfaces and another person’s face [10–13].

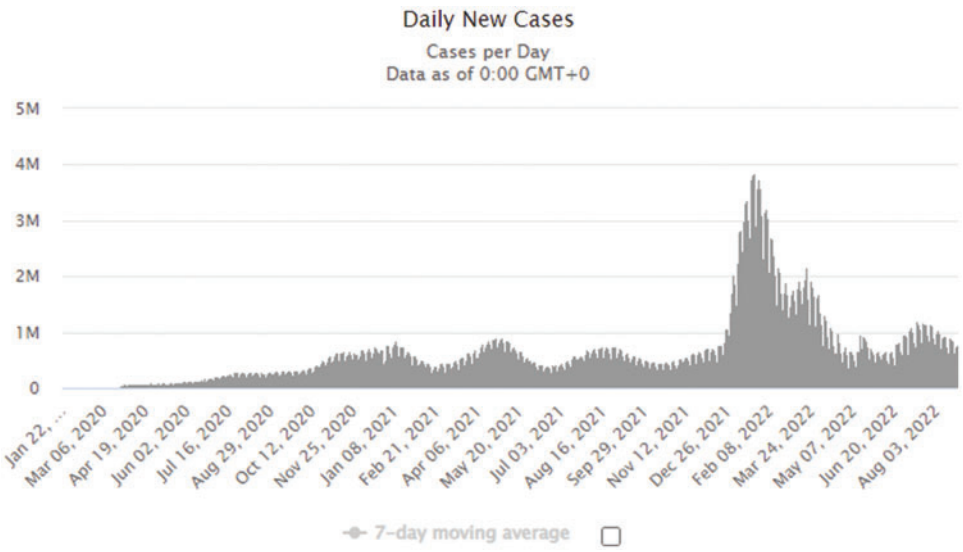


Figure 1: Analysis of COVID cases from 2020 to 2022 (Source: www.worldometers.info)

Facemask identification combining Machine Learning (ML) and Deep Learning (DL) yielded encouraging results; nonetheless, present approaches should be more resilient in real-time. Furthermore, there is a shortage of adequate resources and DL-based algorithms for face mask identification [14] that could detect a person without a face mask in real-time to prevent the spread of COVID-19. For this objective, we generated a large-scale dataset with two classes, face mask and not mask, and minimal modelling for efficient facemask recognition to establish a safe and protected environment for all individuals by limiting virus dissemination. The suggested work will be implemented in specific locations such as schools, colleges, institutions, mosques, and supermarkets. The proposed device will be installed at the main entrance, where the technology will scan each person's face. If the system finds someone without any mask, they will be denied entry. The following are the planned work's contributions which assure novelty:

- A face dataset with photos of various users generated by a generative adversarial neural network is created, and then alternative face masks are devised to be applied to these face images to produce a bespoke dataset with two classes (mask and non-mask).
- The proposed model, ResNet-50, was trained using a dataset. The generated model weights can be applied to further computer vision tasks, accelerating the learning of the new models.
- We trained multiple pre-trained Deep models such as AlexNet, VGG Net, and Inception Net to evaluate the effectiveness of the proposed lightweight DL model over the observation system. The proposed model beats previous DL-based models in terms of temporal complexity, size of the trainable model, and accuracy.
- The proposed model uses well-known deep learning techniques to build the classifier, collect images of people wearing masks, and discriminate between classes of face masks and non-facial masks. Using Open-CV, Keras, and Python, this task is carried out. Compared to other models outlined in Section 4, it takes less memory and computing time, making it simpler to deploy for monitoring.

2 Literature Survey

Researchers and scientists have demonstrated that using face masks in social places significantly inhibits the propagation rate of COVID-19, as shown in Table 1. The proposed work in [15] suggested a new approach for detecting the presence of a face mask. Their trained model can distinguish between improper, correct, and no face mask usage scenarios. Their predictive model has an accuracy of 98.70%. Reference [16] proposed suggested a face identification system based on YOLOv3. They validated their technique by training on WIDER FACE datasets [17–19]. Their model has an accuracy of 93.9%. Din et al. introduced a unique GAN-based framework that can identify and extract the face mask from a face picture and recreate the facial images using a GAN. The author in [20] presented Retina Face Mask, a face mask detector. To eliminate predictions with low confidence, they utilized a unique object-removal approach. They improved accuracy by 1.5% and recall by 5.9% and 11.0%, respectively, with state-of-the-art facial masking and detection results. On the other hand, they also investigated the suggested method's performance on lightweight neural networks. MobileNet, for example, deep learning combined with reinforcement learning, was proposed by Shashi et al. [20]. A geometric-based approach for detecting social distance and face coverings in public places. They built their model using a Raspberry Pi3.

Table 1: Survey on various mask prediction algorithms using transfer learning

Author	Year of publication	Algorithm used	Running time	Significance	Accuracy
Su et al. [7]	2021	Face mask detection using YOLOv3	13.25 flips per sec	YOLO was built using transfer learning from EfficientNet for feature extraction.	75.6
Mahurkar et al. [8]	2021	Real-time face detection using YOLOv4	50.6 flips per sec	Using a deep learning model of YOLO, with around 1000 epochs.	91.2
Jain et al. [9]	2021	AI-based mask detection	-	Used Rasberry Pi to detect the mask using multiple CNN architectures.	83.4
Ding et al. [14]	2019	Application for face mask detection	51 flops per sec	Used YOLOv5 with three different datasets to check the availability of masks.	
Draughon et al. [10]	2020	Application for personal protector detection	23 flops per sec	The application was developed to check the workers at a specific site for mask availability.	NA

(Continued)

Table 1 (continued)

Author	Year of publication	Algorithm used	Running time	Significance	Accuracy
Basha et al. [11]	2021	Deep learning application to detect the face mask	0.18 sec	Application to detect face mask using SVM classification.	Accuracy = 89.6%, Precision = 90.1% Recall = 91.1% F-measure = 87.6%
Zang et al. [16]	2017	Face mask detector application		The deep neural network was used to detect the mask using two classes.	Accuracy-83.8%, precision-89.4%, recall-91.1%

A few open-source models can handle COVID-related face-masked data sets, even though many have already been trained on benchmark datasets. Modern object detectors come in two main varieties: stage-wise detectors. However, they need to satisfy the criteria for real-time video surveillance equipment. Less computational and memory power restricts these devices. As a result, they need improved object identification models that can conduct real-time surveillance while using less memory and maintaining high accuracy. While two-stage detectors may readily give accurate answers for complicated inputs at the expense of computing time, single-stage sensors are suitable for real-time observation but need more precision. These characteristics make it necessary to create a unified model for surveillance equipment that may improve accuracy and calculation time.

Another option for the same task is to employ a transfer learning model. The contributions in light of the most recent state-of-the-art are listed below. The suggested study examines several types of research done on face mask detection about the present demand. Face masking has been the focus of many articles, although there are few observations, predictions for the future, extensive citations, current trends, etc., on this topic. The performance parameters of several algorithms are compared, and discussions on them are offered to improve the effectiveness of the review article.

Performance Analysis of the Deep Learning Classifiers Used in the Intelligent Face Mask Detection System is used to achieve the maximum accuracy, several classifiers including MobileNetV2, VGG16, and ResNet-50, 100 were compared with optimizers like Advanced Distributed Associative Memory (ADAM), etc. [21]. Utmost precision came from ADAM [22]. Using the Cascade framework to detect masked faces is based on deep learning. The CNN course structure that this algorithm is built on consists of three CNNs. The technique may be applied to Closed-Circuit Television (CCTV) footage to determine whether someone is appropriately wearing a mask so that they do not endanger others. The most effective framework may be used with alerting and precautionary frameworks shortly. This framework may be combined with a framework that facilitates social distance, making it a sound framework that can welcome emotional effects on the spread. MobileNet works like a system's backbone and may be used effectively in high- and low-calculation scenarios. Learning is used to obtain weights from a related job, face recognition, which is trained on big datasets to derive more stable features. On a public face mask dataset, the suggested approach achieves cutting-edge results. It would benefit society if face mask detection technology advanced to the point where it could tell whether someone is wearing one or not and allow them admission. A larger facemask-wearing dataset, including photographs and videos, will be gathered and labeled in the subsequent experiments to enhance the performance.

3 Proposed Model

The benchmark for object identification provided in [23] is the foundation for the suggested model. This benchmark states that all tasks involved in an object identification challenge may be grouped under the three primary components shown in Fig. 2 the Backbone, Neck, and Head. This system's backbone is a primary convolutional neural network that can extract data from pictures and turn them into feature maps. The notion of transfer learning is used in the suggested architecture's backbone to use previously learned characteristics of a potent, before a deep neural network, in collecting additional features for the network. An extensive framework-building approach using three well-known pre-ResNet50, VGG Net, and AlexNet, three trained models, is used to get the most incredible facemask detection results. ResNet50 is the best option for creating the backbone model [24–26]. The originality of our research is suggested in the Neck portion. The Neck, the intermediary part, comprises all the preparation work that must be done before classifying images. To enable our model to be used for surveillance, Neck uses various pipelines to deploy and train his devices. Phase 1: The pipeline for training is established after an objective custom dataset with ResNet50 optimization.

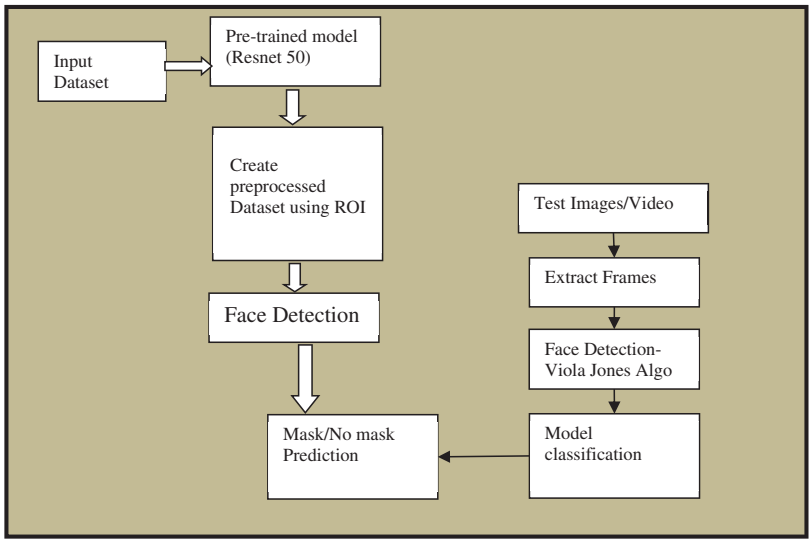


Figure 2: Proposed method for classification

The final element is an identification detector or determinant that can produce the desired results. The goal of the deep learning neural network in the proposed architecture includes achieving the learned facemask classifier after using transfer learning. Identifying faces with and without masks. The principal goal of enforcement of wearing face masks in public places will not be accomplished until the individual identity of faces is obtained, going against the mask's rules. The action may also be carried out following office or government policy. The proposed algorithm is divided into two phases. In the first phase, we train the model using two datasets for transfer learning, and in the second phase, we test the prediction model. After the first phase, the validation accuracy is still close to 90%, even after two epochs. The model quickly converges after 40 epochs. Including a few additional linked layers makes it feasible to achieve better accuracy levels. Even though Face Mask Dataset and MAFA include far-flung picture samples, ResNet-50 works reasonably well when generating features for the classifier based on the successful findings with only one hidden dense layer.

Phase-1: Algorithm

Step1: Import the necessary libraries
Step 2: Import the dataset into the working model.
Step 3: Split the data into two parts for training and validation.
Step 4: Import the pre-trained ResNet model and extract the features. ResNet with different layers in the architecture.

```
b = min(f); (b is best and P is Prob of best matching)
Update b and P:
    For i = 1:S (Data Size)
  If if < P(i)
      P(i) = f
    End
  End
    B = min(f); (B is the best match)
If B < b
    b = B;
 End
```

Step 5: Average pooling operation with 7×7 kernel.
Step 6: Linear layer with ReLu activation with 128 channels.
Step 7: Apply dropout with a threshold of 0.5
Step 8: Dense layer with the SoftMax activation for classification
Step 9: Store the weights in the model.h5 file.

Phase-2: Algorithm

Step 1: Input source image
Step 2: Resize the image.
Step 3: Split the data into training and testing.
Step 4: Detect face using Viola-Jones detection Algorithm
Step 5: Feature extraction using the trained model in Phase 1.
Step 6: Classification: Test the image for prediction as With-mask or Without_mask.

The proposed model is a compact deep neural network designed for embedded devices like mobile phones. It features quicker computation times and fewer parameters than other network models. Deep separable convolution is the central concept of ResNet, and deep convolution and argument convolution are the two main components of its construction. The method is depicted in Fig. 5 and assumes that the original input map is Conv X Conv X Maxpooling and the extracted feature map is Conv X Conv X Dense Layer. The convolution kernel's size is also assumed to be one X Conv, as shown in Fig. 3.

The resulting feature map is $12 \times 12 \times 4$ in size, whereas the given input map is $12 \times 12 \times 3$. Three 55 degrees involved in resolving are used to explore the data of three channels, and three feature maps are produced due to the depth separable convolution. The needed computation amount after the dense calculation is 12528, and four convolution kernels of 22 sizes are employed to explore three feature maps before the fusion procedure. In actuality, 43200 is the usual convolution calculation amount. The length of separable convolution may significantly lower the model's computation requirements.

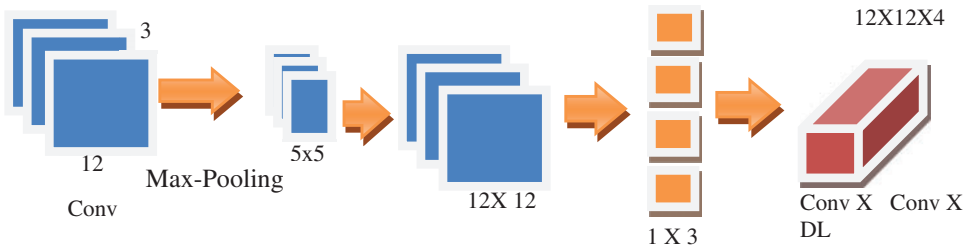


Figure 3: Depth separable training to increase the efficiency in transfer learning

3.1 Pre-Processing the Images

The proposed method fetches the images from the dataset and performs the preprocessing steps, such as: Converting RGB images to Grayscale images. Resize the input image to $160 \times 160 \times 3$. Each image is converted into a NumPy array. Align the images into vectors. Split the images into the train, test, and validation tests.

3.2 Face Detector

The input has been taken from a standard database or video, and Viola Jones's algorithm is used for face detection. The picture is transmitted to many places using convolutional layers that pull-out feature maps. Each extracted feature has a 4×4 filter to establish a low default box and forecast the bounding box offset for each box. The cascade architecture's false positive rate can be calculated using Eq. (1).

$$F_\gamma = \prod_{i=1}^{n} f_i \tag{1}$$

The detection rate is calculated by Eq. (2)

$$D = \prod_{i=1}^{n} d_i \tag{2}$$

Each bounding includes five predictions output in a box: weight, height, width, and confidence. The box's centroid is represented by the relationship between the x and y bounds for the grid cell.

3.3 Face Mask Classifier

By removing the top layers of a pre-trained CNN model trained on a sizable dataset for detection or classification issues, we may apply transfer learning as a machine learning approach (Fully connected layers). We employ the bottom layers to extract features and our newly added fully linked layers for classification. Our work inserted 64 units as fully-connected layers and Rectified Linear Unit (ReLU) with two units as an output layer using the transfer learning strategy on a pre-trained RESNET 50 on the dense layer (Softmax activation). The last two layers of our model are configured because all previous levels have already been trained.

3.4 Loss Function

The most typical loss function in classification issues is shown in Fig. 4. As the projected probability approaches the valid label, the cross-entropy loss diminishes. It evaluates how well a classification model performs when predicting an outcome with a likelihood value between 0 and 1. When there are just two classes, the categorization is binary. When there are two classes, the loss

function used is in Eq. (3).

$$Log = -\frac{1}{n}\sum_{i=1}^{n} k_i log\left(\hat{k}_i\right) \tag{3}$$

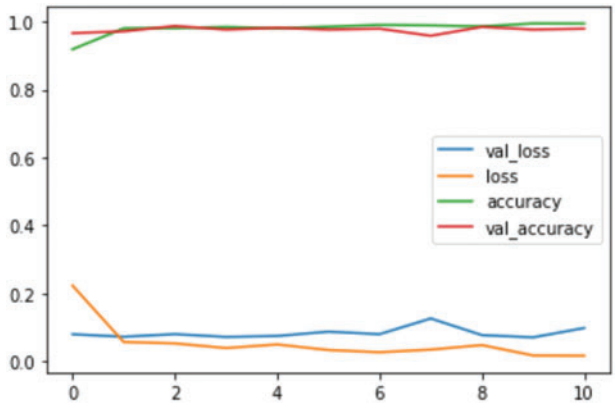


Figure 4: Loss plot of proposed method

3.5 Proposed Pre-Trained Model: Fine Tuning

The Deep neural networks are used in the proposed study to detect facemasks because of their superior performance over other classification techniques. However, deep neural network training is costly since it takes a long time and many computing resources. Deep learning-based data augmentation is used to develop the network more quickly and affordably. Transfer learning enables the neural network's taught information to be transferred to the new model as parametric weights. Even when the new model is developed on a concise dataset, it improves performance. The 13 million photos from the MAFA database were used to train several pre-trained models, including AlexNet, MobileNet, ResNet50, etc. ResNet50 is selected as a pre-trained network for facemask classification in the suggested model. Five new layers (max-pooling, flatten, relu, dropout, and softmax layers) are added to ResNet50's final layer to improve it. An average pooling layer with a pool size of 5×5, a conv layer, a dense ReLU layer with 256 neurons, a dropout of 0.5, and a decisive layer with an activation function of softmax for binary classification are among the newly added layers. These layers are depicted in Fig. 5. First, analyses and the processing of various facial photographs are conducted to address the complexity of mask detection. It has been noted that the dataset we are focusing on has two main classes: mask and non-mask. However, the mask class also has an underlying variety of occlusions not caused by surgical or cloth facemasks, such as the occlusion of ROI by a person, hand, hair, piece of food, etc. The effectiveness of facial and mask detection is discovered to be impacted by these occlusions. Finding the best compromise between precision and computation time for face identification takes time and effort. Therefore, a predictor for picture complexity is suggested here. Its goal is to divide the data into soft and hard pictures and then, using a facemask classifier, classify the data into masks and non-masks.

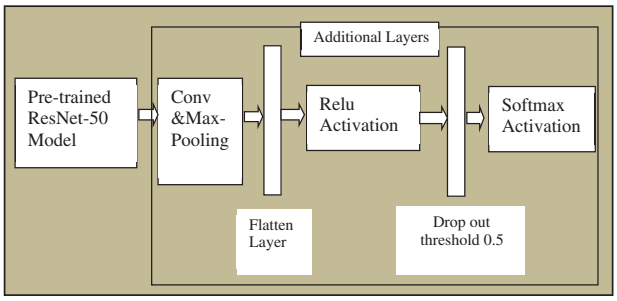


Figure 5: Finetuned model of ResNet 50

The "Semi-supervised object classification technique" put out by [13] provides the solution to this query. The proposed system is a good fit for the quasi-object classification technique since it predicts items without localizing them.

4 Results

The experimental results were evaluated on two different datasets that are available publicly. The proposed system is implemented on python 3.7, open cv, and TensorFlow libraries. The most crucial stage of any research project is data collection. To do this, we have gathered two datasets. One is the MAFA dataset, which is the publicly available dataset. Secondly, own dataset was collected from university students and employees with 150 images without a mask. For the pre-trained model of ResNet, we used the first dataset. For transfer learning, we used the second dataset. After gathering these pictures, we created several face masks and put one on each face to create a new class for face mask recognition. There are two classifications in this dataset: mask and non-mask. The mask class has 4664 pictures, whereas the non-mask type has 3076 images in the testing dataset. All photos have different channels, such as RGB, and are all 224×224 pixels. Learning, validation, and testing are the three subcategories that make up the dataset. The training uses 70% of the data in the suggested approach, while validation uses 30%.

4.1 Experimental Setup

The proposed system used the torch vision package of python for the pre-trained model. The model was trained with 43000 Masked Face Detection Benchmark Dataset (MAFA) images. At first, 25,876 photos divided into two classes—masked and unmasked- were considered. The dataset is split into 70:20:10 ratios for training, testing, and validation. The model is built and implemented using Python 3.6, and mask detection is achieved through ResNet. The learning rate was taken ass 0.0001, with a 0.5 threshold and batch size of 100. Three sets of picture samples were used to evaluate this strategy: the first set (L) comprises labeled (hard/soft) trained images, the new batch (U) includes the unlabeled training set, and the third set (T) contains labeled (hard/soft) test images. We went one step further by developing the training model for the hard/soft classifier on an expanded training set L during each iteration of the curriculum learning technique provided in [26], which runs iteratively. The learning set L is widened by transferring k samples randomly from U to L. When L tripled in size from its original size, we halted learning. L is initially filled with 450 tagged examples.

4.2 Training the Model

A supervised learning CNN model classifies the trained images into the appropriate classes by recognizing significant visual patterns once trained. The main components of the proposed model are TensorFlow and Keras. The training set comprises 70% of the dataset in this study, while the testing set comprises the remaining 30%. The procedures mentioned above are used to preprocess and enrich the supplied picture. Five Conv2D levels with ReLu activation functions and a 3×3 filter, plus five Max-Pooling layers with a 2×2 filter, make up the entire system. The fully linked layers are flattened and Dense. The output layer's activation function is softmax. 2,818,784 trainable parameters are the outcome. Another dataset, the MAFA, is also suggested by [7]. It has 160 training photos and 40 test images. The system is pre-trained with the MAFA dataset and finetuned with the classification algorithm to combat the overfitting problem brought on by a shortage of training samples. The MASKED FACE testing data is used to evaluate this covered face detection technique, and the results are promising.

The face identification model is pre-trained using 1 million iterations on a 300×300 picture size. The Single Shot Detector architecture is the foundation for the ResNet-50 architecture used by the deep Neural Network (DNN) face detector. With a multi-box, a single-shot sensor uses a single shot to recognize several items in a photograph. Consequently, it boasts a significantly faster and more precise object detection mechanism. Table 2 describes the model architecture used for the study. Most of the architecture comprises 2D convolutional layers, max-pooling layers, ReLu activation, and fully-connected layers. The suggested model has five Conv2D layers, with padding set to "same" and a stride of one. By "sliding input" over a filter or kernel at each layer, the feature map of the 2D input data is retrieved, and the following procedure is carried out using Eq. (4).

$$C = (K \times L)(P) = \left(\int_{-\infty}^{\infty} K(x)\, xL(K - c)\, \mathrm{d}x \right) \tag{4}$$

where k is the input image and, L is a kernel of the model in the convolution process, x is the input variable, c is the time variable, which gives C as the output image. For training, we considered two datasets, the Face detection dataset from Kaggle, with ideas without a mask of 3828 numbers, and photos with a mask of 3725 numbers. The discriminator receives a batch of pictures produced by the generator. The loss is determined by changing the target tags to 1 or actual. The generator does this because its goal is to deceive the discriminator. The loss modifies the generator's weights, making it more adept at producing realistic-looking pictures that fool the discriminator. Before calculating the loss, the discriminator is first given a batch of actual images, with the goal labels set to 1. T The discriminator is then given a collection of false pictures (produced by the generator), and the loss is measured with the target labels set to 0. The total loss alters the weights when the two losses are combined. The sample input images are shown in Fig. 6. The output image classification on the face mask dataset is shown in Fig. 7. The variety of live videos is offered in Fig. 8.

$$Pavg = \int_{0}^{1} P(x)\, \mathrm{d}x \tag{5}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{6}$$

where precision can be calculated using Eq. (7).

$$P = \frac{\text{True Positive}}{\text{TP} + \text{FP}} \tag{7}$$

Table 2: Hyper parameters of the proposed system

No. of epoch	Loss value	Accuracy value %	Val_loss %	Val_acc %
10	69.13	50.00	68.88	51.00
20	69.02	50.25	68.72	51.00
30	68.90	50.75	68.57	51.00
40	68.82	50.75	68.41	52.50
50	68.73	51.38	68.26	53.00
60	68.58	54.00	68.08	54.00
70	68.51	55.62	67.89	56.00
80	68.36	57.38	67.69	58.00
90	68.23	58.75	67.47	61.00
100	68.07	59.62	67.22	62.50

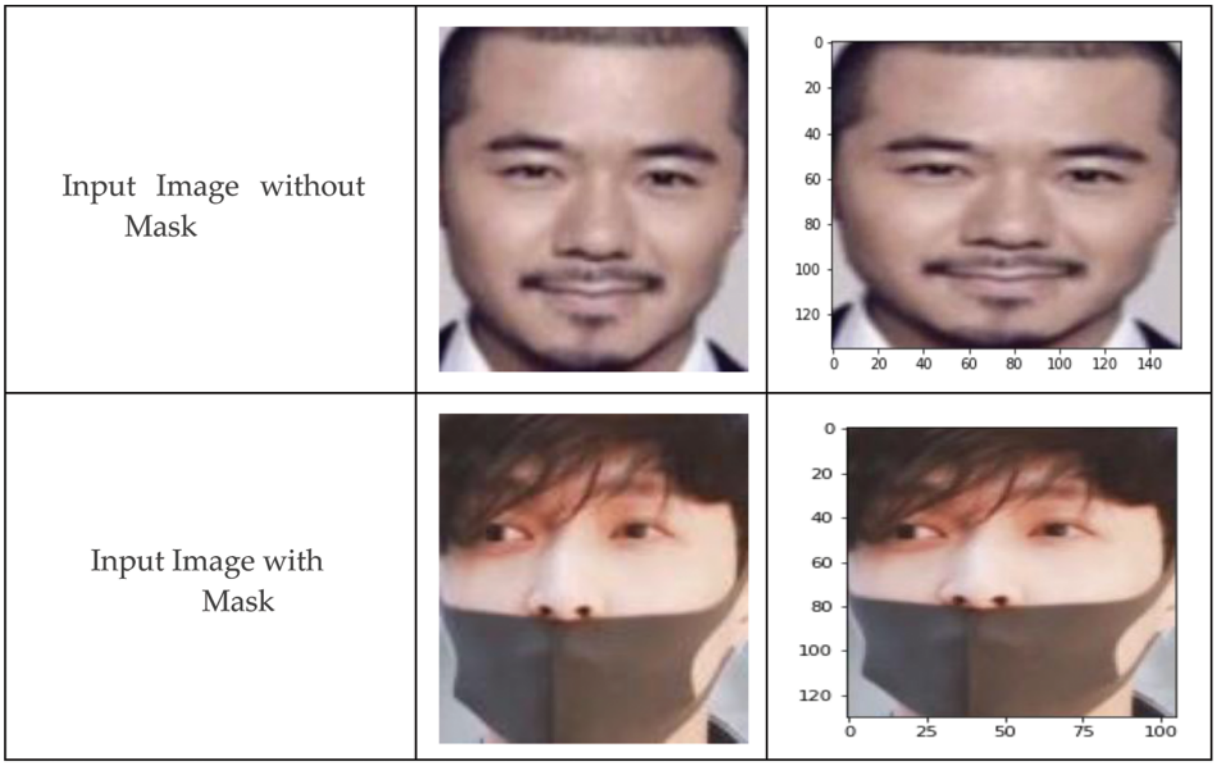


Figure 6: Sample input images for training model

Figure 7: Classification result on face mask dataset and MAFA dataset

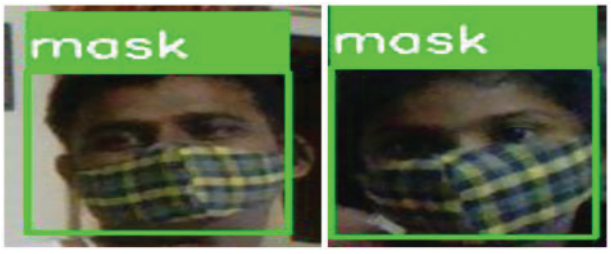


Figure 8: Detection of individual random images

And recall is calculated using Eq. (8)

$$P = \frac{\text{True Positive}}{TP + FN} \tag{8}$$

where FN is false Negative, and TP is True Positive. Accuracy is calculated using Eq. (9).

$$Acc = \frac{\text{True Positive}}{TP + FP} \tag{9}$$

F1-measure is calculated using Eq. (10)

$$F1 - Score = 2 * \frac{Precisoion * Recall}{Precision + Recall} \tag{10}$$

4.3 Evaluation Metrics

Average precision, mean average precision, and frame rate per second are used by object detection indicators (FPS). The Pavg number represents the effects of a specific object on detection, and the mAP value is the mean accuracy across all categories in the sample. The followings are the calculating Eqs. (5) and (6).

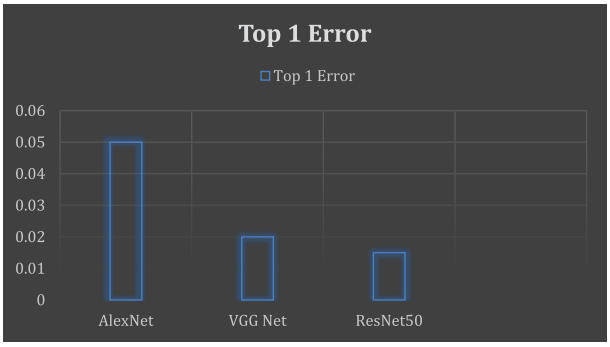
4.4 Model Evaluation

We can use transfer learning to improve pre-trained models for image classification. This section will evaluate ResNet50, AlexNet, and VGG Net-based top-1 error, which happens when the most confidently predicted class is not identical to the actual course. Secondly, the interpretation time of the CPU is the amount of time the model needs to read a picture, apply all necessary changes, and then generate the indeed predicted class to which the image belongs to forecast the class of the input image. Thirdly, CPU utilization time is the maximum number of components that may be learned that are prevalent throughout all model levels. These characteristics influence CPU utilization, model complexity, and prediction ability. Understanding the bare minimum of RAM needed for each model is made much easier with the help of this information. Additionally, it was determined by Simone Bianco et al. that the ideal number of trainable parameters is needed to accomplish a trade-off between prediction performance and memory consumption. Table 3 provides the confusion matrices for several theories under testing. Fig. 9a visual representation of the accuracy comparison of several models based on the Top-1 error. The graph shows that ResNet50 has the lowest error rate and AlexNet has the highest. The model was then contrasted according to inference time. Each model receives test pictures, and the inference times across all iterations are summed. As shown in Fig. 9b, ResNet and AlexNet infer pictures more quickly than VGG Net, as it takes longer overall. Additionally, the number of trainable parameters is determined to compare the storage utilization of various underlying models. These parameters can be found by creating a model summary for each model in Google Colab. The amount of parameters in AlexNet for our custom dataset is around 28 million, as shown in Fig. 9c. Approximately 4 million and 10 million variables are available in VGG net and ResNet 50, respectively.

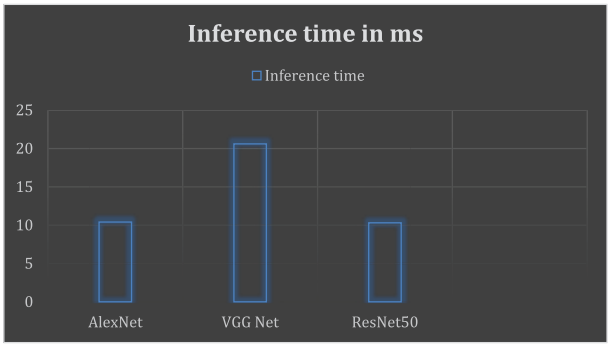
The proposed system is tested using three classifiers: decision tree, random forest classifier, Ada boost classifier, and gradient boosting classifier. The training accuracy of all classifiers is 1.0, testing accuracy of the classifiers is 0.93, 0.91, 0.97, and 0.97, respectively. The evaluation parameters for the above system are shown in Tables 4–8.

Table 3: Confusion matrix of proposed system for pre-trained models

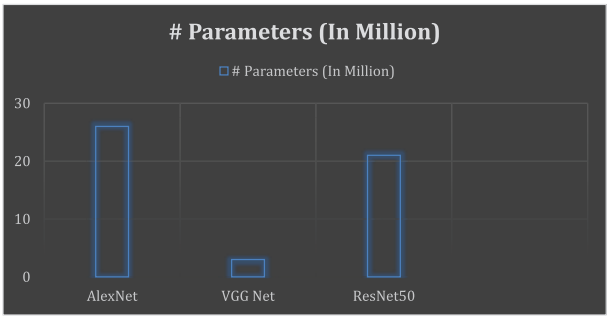
	Mask	No mask		Mask	No mask		Mask	No mask
Mask	4632 (TP)	109 (FP)	Mask	4775 (TP)	56 (FP)	Mask	4785 (TP)	78 (FP)
No mask	223 (FN)	4756 (TN)	No mask	185 (FN)	4635 (TN)	No mask	139 (FN)	4231 (TN)
Alex Net			VGG Net			ResNet50		



(a)



(b)



(c)

Figure 9: Comparison of various models using different parameters

Table 4: Evaluation parameters of decision tree classifier

Parameters	Precision	Recall	F1-score	TPR (%)
0	0.78	0.65	0.74	57
1	0.51	0.67	0.57	60
Accuracy	-	-	0.74	94
Macro-Averaged	0.65	0.66	0.63	94
Weighted-Averaged	0.68	0.65	0.67	94

Table 5: Evaluation parameters of random forest classifier

	Precision	Recall	F1-score	TPR (%)
0	0.95	0.94	0.94	56
1	0.87	0.89	0.87	42
Accuracy			0.92	86
Macro- Averaged	0.92	0.90	0.91	89
Weighted-Averaged	0.93	0.92	0.85	89

Table 6: Evaluation parameters of Ada boost classifier

	Precision	Recall	F1-score	TPR (%)
0	1.00	0.89	0.97	56
1	0.98	0.87	0.97	43
Accuracy			0.97	89
Macro-Averaged	0.85	0.99	0.97	87
Weighted-Averaged	0.86	0.97	0.97	89

Table 7: Evaluation parameters of gradient boost classifier

	Precision	Recall	F1-score	TPR (%)
0	1.00	0.97	0.97	57
1	0.97	1.0	0.98	42
Accuracy	-	-	0.99	89
Macro-Averaged	0.98	0.97	0.96	89
Weighted-Averaged	0.89	0.89	0.95	89

Table 8: Evaluation parameters of proposed work

	Precision	Recall	F1-score	TPR (%)
0	1.00	0.98	0.96	64
1	0.99	0.99	0.98	65
Accuracy	-	-	0.96	98
Macro-Averaged	0.99	0.96	0.98	98
Weighted-Averaged	0.99	0.98	0.96	98

The error rate for AlexNet is high, and VGGNet takes a while to deduce conclusions. Also, ResNet50 is the best option for face mask detection utilizing transfer learning in high accuracy, speed, and memory utilization. After several feature learning iterations, the initial output might learn offset. After adding delay to the following result, the throughput structure is achieved. The offset in the X and Y axes is represented by 2n of the initial input. The extracted feature map is a great way to determine the following translation following deformable convolution. Transfer learning improves neural network performance without requiring much data by using the dataset's training weights as the dataset's values. The system learns more robust and powerful features since there is a need for more learning sample data, and performance suffers greatly. The experimental findings demonstrate that without transfer learning, face and face masks are dropped by 12.15% and 6.25%, respectively. Fig. 10 shows the time consumption of models. The time spent during each training period for several models on the database. Since the batch size and dataset size affect the training duration, When the sample size is 3420 and the step size is 20, the training utilization time achieved in this study is shown in Fig. 10. It is evident from Fig. ResNet50's training time is the longest at 33 s, whereas our method's training time is the quickest at 10 s. The time needed for model training has significantly decreased with the addition of transfer learning.

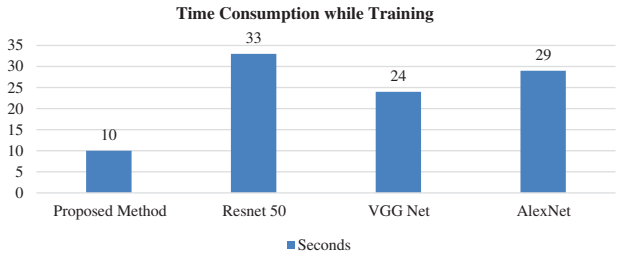


Figure 10: The consumption at each epoch while training the model

As shown visually in Fig. 10, it is further noticed that model accuracy goes on rising at numerous points and stabilizes at the 0th epoch. Compared to more contemporary strategies, the suggested model delivers excellent face and mask identification precision with reduced inference time and memory usage, according to a summary of the experimental findings. The old MAFA dataset's data imbalance issue has been addressed with much effort, and the outcome was a new, impartial dataset ideal for COVID-related mask recognition tasks. Several prominent that can be deployed in an integrated platform at public locations due to the newly developed dataset, efficient face detection technique, localizing personal identification, and prevention of overfitting to stop the spread of COVID-19 further.

4.5 Comparison of the Proposed System with State-of-Art Systems

We compared the proposed work with the existing state-of-art systems, which aim to answer the detection rate. The famous Retina Face Mask method [11] that uses the MAFA dataset performed the detection in the same environment. The comparison is specified in Table 9. The data shows the high proposed method detection rate.

Table 9: Outcome of state-of-art methods

Model	Mask detection			No mask detection		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
RetinaFaceMask model using AlexNet	84.3	81.2	87.2	82.4	82.4	81.6
RetinaFaceMask model using ResNet	91.4	92.5	86.4	89.6	87.5	92.3
RetinaFaceMask model using VGG Net	92.8	93.2	93.2	85.6	87.6	87.3
RetinaFaceMask model using proposed model	94.6	97.2	95.6	94.6	90.6	92.6

4.6 Scope of the Work and Future Scope

Although many research papers have been enrolled to show the real-time COVID-19 problem situation, deploying the systems in real-time is challenging. It is getting harder to create adaptable systems for all conditions and environments. The suggested gadget might be used to maintain a tight check on people in active regions. When estimating the project's costs, we find that most metropolitan areas already have cameras in public places. Therefore there will be little financial outlay. The single essential component of the suggested concept, a camera, exists.

Although many research papers have been enrolled to show the real-time COVID-19 problem situation, deploying the systems in real-time is challenging. It is getting harder to create adaptable systems for all conditions and environments. The suggested gadget might be used to maintain a tight check on people in active regions. When estimating the project's costs, we find that most metropolitan areas already have cameras in public places. Therefore there will be little financial outlay. The single essential component of the suggested concept, a camera, exists. Different facial photos with and without masks make up the dataset used by the face mask identification system based on the CNN model. With a dataset containing pictures relevant to that job, the same model may be used for other image-processing tasks in neuroscience.

Lastly, the research opens up exciting new avenues for study. The proposed method is not restricted to mask detection alone and can be included in other high-resolution video surveillance equipment. Second, the model may use a facemask to identify facial landmarks for biometric applications. With a 3.6% error rate, ResNet outperformed the pre-trained model.

5 Conclusion

To stop COVID-19 from spreading across the community, a deep learning-based method for spotting coverings over faces in public spaces is provided in this study. The suggested method employs

an aggregation of single- and double-stage sensors at the previous level to effectively manage occlusions in crowded environments. The ensemble technique significantly increases detection speed while also reaching high accuracy. Applying transfer learning to pre-trained models and considerable testing on an unbiased dataset produced a reliable and affordable solution. In conclusion, our CNN model obtained a top-1 failure rate of 1.8% better on the MAFA validation set, almost matching the efficiency. This difference can be the result of a streamlined training methodology. The experimental findings demonstrate that without transfer learning, face and face masks are dropped by 12.15% and 6.25%, respectively. The system's utility for the benefit of the public is increased by the identification detection of faces that further violate the mask regulations. We pre-trained the CNN with discriminative features on the original biased MAFA dataset. The free Caffe Python package was used for the pre-training.

Acknowledgement: Ahmed Alhussen would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2023-356.

Funding Statement: This work was supported by Deanship of Scientific Research at MajmaahUniversity under Project No. R-2023-356.

Author Contributions: All authors equally participated in this research paper.

Availability of Data and Materials: All data and materials used in this work is accessible and open.

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

References

[1] L. R. G. Godoy, A. E. Jones, N. T. C. L. Anderson, K. M. S. Fisher, E. A. Beeson *et al.,* "Facial protection for healthcare workers during pandemics: A scoping review," *BMJ Global Health*, vol. 5, no. 1, pp. 2353–2563, 2020.

[2] H. Goyal, K. Sidana, C. Singh, A. Jain and S. Jindal, "A real time face mask detection system using convolutional neural network," *Multimedia Tools & Application*, vol. 81, no. 11, pp. 14999–15015, 2022.

[3] T. Meenpal, A. Balakrishnan and A. Verma, "Facial mask detection using semantic segmentation," in *Int. Conf. on Advances in Science, Engineering, and Robotics Technology*, Rome, Italy, pp. 1–5, 2019.

[4] A. Kumar and A. Jain, "Image smog restoration using oblique gradient profile prior and energy minimization," *Frontiers of Computer Science*, vol. 15, no. 6, pp. 1–7, 2021.

[5] X. Su, M. Gao, J. Ren, Y. Li, M. Dong *et al.,* "Face mask detection and classification via deep transfer learning," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 4475–4494, 2022.

[6] R. R. Mahurkar and N. G. Gadge, "Real-time COVID-19 face mask detection with YOLOv4," in *Conf. on Electronics and Sustainable Communication Systems*, Dalian, China, pp. 1250–1255, 2021.

[7] S. Singh, U. Ahuja, M. Kumar, K. Kumar and M. Sachdeva, "Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment," *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 19753–19768, 2021.

[8] T. Tsai, A. Mukundan, Y. Chi, Y. Tsao, Y. Wang *et al.,* "Intelligent identification of early esophageal cancer by band-selective hyperspectral imaging," *Cancers*, vol. 14, no. 17, pp. 4292, 2022.

[9] A. Jain and A. Kumar, "Desmogging of still smoggy images using a novel channel prior," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 1161–1177, 2021.

[10] G. T. S. Draughon, P. Sun and J. P. Lynch, "Implementation of a computer vision framework for tracking and visualizing face mask usage in urban environments," in *IEEE Int. Smart Cities Conf. (ISC2)*, Piscataway, NJ, USA, pp. 1–8, 2020.

[11] C. Z. Basha, B. N. L. Pravallika and E. B. Shankar, "An efficient face mask detector with pytorch and deep learning," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 25, no. 7, pp. 1–8, 2021.

[12] Y. K. Sharma and M. D. Rokade, "Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic," *IOSR Journal of Engineering*, vol. 3, no. 1, pp. 63–67, 2019.
[13] M. Jadhav, Y. K. Sharma and G. M. Bhandari, "Currency identification and forged banknote detection using deep learning," in *Int. Conf. on Innovative Trends and Advances in Engineering and Technology*, Shegoaon, India, pp. 178–183, 2019.
[14] S. Hossain and D. Lee, "Deep learning based real time multiple object detection and tracking from aerial imagery via a flying robot with GPU based embedded devices," *Sensors*, vol. 19, no. 15, pp. 3371, 2019.
[15] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Engineering Applications of Artificial Intelligence*, vol. 51, no. 1, pp. 191–201, 2016.
[16] D. Griol, J. M. Molina and Z. Callejas, "Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances," *Neurocomputing*, vol. 326, no. 1, pp. 132–140, 2017.
[17] P. Partila, J. Tovarek, M. Voznak and J. Safarik, "Classification methods accuracy for speech emotion recognition system," *Nostradamus 2014: Prediction, Modeling and Analysis of Complex Systems Prediction*, vol. 289, no. 3, pp. 439–447, 2014.
[18] M. A. R. Khan and M. K. Jain, "Feature point detection for repacked android apps," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1359–1373, 2020.
[19] Z. Liu, M. Wu, W. Cao, J. Mao, J. Xu *et al.,* "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, no. 2, pp. 253–265, 2017.
[20] F. I. Shashi, M. S. Salman, A. Khatun, T. Sultana and T. Alam, "A study on deep reinforcement learning based traffic signal control for mitigating traffic congestion," in *IEEE 3rd Eurasia Conf. on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, Tainan, Taiwan, pp. 288–291, 2021.
[21] S. Ge, J. Li, Q. Ye and Z. Luonn, "MAFA-dataset," 2020. [Online]. Available: https://www.kaggle.com/datasets/revanthrex/mafadataset (accessed on 12/07/2022).
[22] J. Y. Zhu, T. Park, P. Isola and A. AEfros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*. Venice, Italy, pp. 2223–2232, 2017.
[23] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu *et al.,* "Generative image inpainting with contextual attention," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5505–5514, 2018.
[24] J. Xinbei, T. Gao, Z. Zhu and Y. Zhao, "Real-time face mask detection method based on YOLOv3," *Electronics*, vol. 10, no. 7, pp. 837, 2021.
[25] S. Yadav, "Deep learning based safe social distancing and face mask detection in public areas for COVID-19 safety guidelines adherence," *International Journal of Residence Applied Science Engineering and Technology*, vol. 7, no. 1, pp. 1368–1375, 2020.
[26] N. Binti, M. Ahmad, Z. Mahmoud and R. M. Mehmood, "A pursuit of sustainable privacy protection in big data environment by an optimized clustered-purpose based algorithm," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1217–1231, 2020.

Appendix A
Experiment with GAN Generated Images

The images for collecting random images dataset is gathered from internet and few from known people. This dataset consists of 150 images without mask. Since we need some preparation and processing work to fit the data into our model, we separated test and train data with images. While y train and y test sections have labels from 0 to 9 that indicate what number they truly are, x train and x test portions have gray scale image RGB values (from 0 to 255). Since we are performing an unsupervised learning job our dataset also has to be reshaped using the reshape function to make it 4-dimensional. Finally, for more effective training, we transform our NumPy array into a TensorFlow Dataset object. The 128×128 grayscale false pictures produced by our generator network are created

from random noise. It must thus take 1-dimensional arrays and produce 128×128 pixel pictures. Transposed Convolution layers are required for this operation once our 1-dimensional array has been transformed into a 2-dimensional array. A smaller array's size can be increased by using transposed Convolution layers. Leaky ReLU layers and batch normalization are also used.

Deep learning GAN-based approaches have recently merged as a promising paradigm for a variety of applications, including data augmentation, pose estimation, and image contrast enhancement. This is due to the GAN's nature of unlabeled data, ability to produce high-quality, natural, and realistic images, and the power of training data. However, because there is so much literature on the topic, we only analyze a few sample works on unwanted objects including sunglasses, microphones, hand and face masks.

Non-learning based object removal algorithms attempted to resolve the issue by eliminating unwanted things like sunglasses and random objects and synthesizing the missing material by matching comparable patches from other areas of the picture. In order to eliminate eyeglasses from face photos, the authors of [23] proposed a regularized factor to modify the route priority function. However, these techniques are limited to very small holes with little variations in colour and texture.

This study presents a dataset of high-quality, paired face photos with and without mask concurrently, which is not achievable in the actual world, to address the shortage of the masked face dataset. To do this, we suggest using the free and open-source masking application "MaskTheFace" to overlay synthetic face masks onto photographs of actual faces based on the positions of important facial landmarks. To do this, we first gathered 10249 high-quality face photographs from 12 individuals (10 for training and 2 for testing), then we used the YOLO algorithm to identify every face in the images. Through the trained model.

We modify design for our generative networks, which consists of two parts: discriminators and generators. Each generator has three initial convolutions, nine ResNET blocks with 64 channels each, two fractionally stride convolutions, and a final convolution to narrow the output channel. Additionally, each discriminator is a 70×70 Patch GAN, which penalizes pictures on a patch-by-patch basis rather than per-pixel or per-image. To quantify cycle loss, we trained the model for 50 iterations with 170 unpaired, 256×256 face photos with and without masks, using a learning rate of 0.0002 and a lambda value of 10. Following model training, we assess the model's performance using 50 masked face test photos from our own dataset.

The Fig. A1 shows the sample images of our own dataset and GAN generated datasets.

At 32nd epoch we got the accuracy of 81.58 with the value loss of 0.17. However, the work is still in progress as the results on GAN were not accurate. Soon we will come with accurate results.

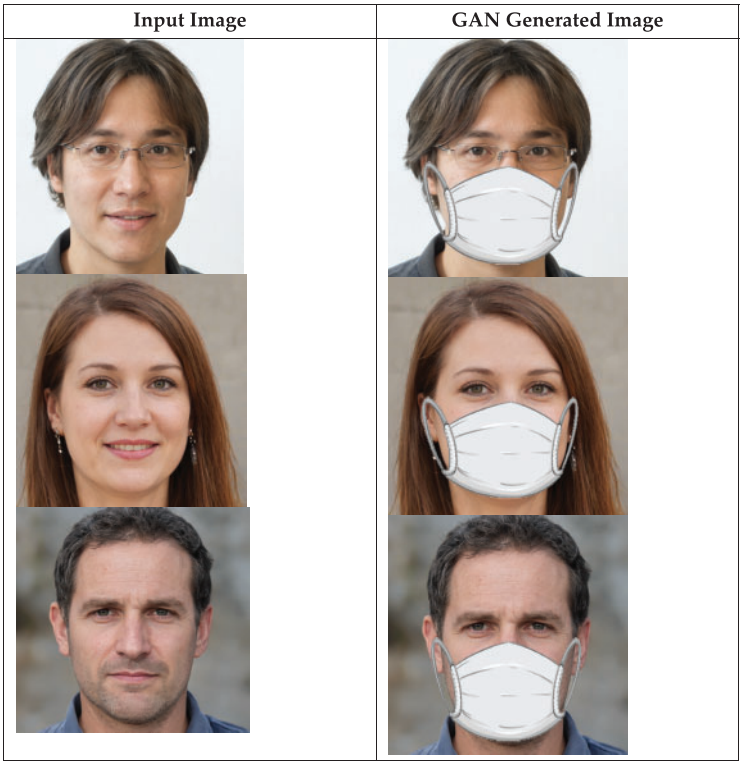


Figure A1: Sample images and masks generated through GAN