electronics

MDPI

Article

An Integrated Approach for Monitoring Social Distancing and Face Mask Detection Using Stacked ResNet-50 and YOLOv5

Inderpreet Singh Walia [1], Deepika Kumar [1,*], Kaushal Sharma [1], Jude D. Hemanth [2] and Daniela Elena Popescu [3]

[1] Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi 110063, India; inderpreet.cse2@bvp.edu.in (I.S.W.); kaushalsharma.cse2@bvp.edu.in (K.S.)
[2] Department of Electronics & Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore 641114, India; judehemanth@karunya.edu
[3] Faculty of Electrical Engineering and Information Technology, University of Oradea, 410087 Oradea, Romania; depopescu@uoradea.ro
* Correspondence: deepika.kumar@bharatividyapeeth.edu

Abstract: SARS-CoV-19 is one of the deadliest pandemics the world has witnessed, taking around 5,049,374 lives till now across worldwide and 459,873 in India. To limit its spread numerous countries have issued many safety measures. Though vaccines are available now, still face mask detection and maintain social distance are the key aspects to prevent from this pandemic. Therefore, authors have proposed a real-time surveillance system that would take the input video feed and check whether the people detected in the video are wearing a mask, this research further monitors the humans for social distancing norms. The proposed methodology involves taking input from a CCTV feed and detecting humans in the frame, using YOLOv5. These detected faces are then processed using Stacked ResNet-50 for classification whether the person is wearing a mask or not, meanwhile, DBSCAN has been used to detect proximities within the persons detected.

Keywords: COVID-19; face mask; social-distancing; YOLOv5

Citation: Walia, I.S.; Kumar, D.; Sharma, K.; Hemanth, J.D.; Popescu, D.E. An Integrated Approach for Monitoring Social Distancing and Face Mask Detection Using Stacked ResNet-50 and YOLOv5. *Electronics* **2021**, *10*, 2996. https://doi.org/10.3390/electronics10232996

Academic Editors: Inés Sittón, Sara Rodriguez, Lilia Muñoz and Xianzhi Wang

Received: 8 October 2021
Accepted: 25 November 2021
Published: 1 December 2021

1. Introduction

COVID-19 (coronavirus disease) is an infectious virus that caused by extreme acute respiratory syndrome (SARS-CoV-2). SARS-CoV-2 is a coronavirus that triggers a respiratory tract infection. It was first discovered in December 2019 in Wuhan, Hubei, China, and has since caused an ongoing pandemic with multiple deaths all over the world. There have been 135,098,227 confirmed cases of COVID-19 and 2,922,345 deaths due to COVID-19 to date, reported to WHO [1]. Even in the most symptomatic cases, the incubation period would last anywhere from two to fourteen days [2]. The disease's symptoms range from a mild cold to extreme respiratory disease and death [3]. Coughing, sneezing, and talking are the most common methods for the virus to surge and by touching a contaminated surface and people may become sick [4]. Although chest CT imaging can help treat infection in people with a high risk of infection based on symptoms and risk factors, it is not recommended for routine screening [5].

Since the infection is spread primarily by air droplets near the infected person, it is critical to maintain social distance between people and wear a face mask, which substantially reduces the risk of disease [6,7]. In order to ensure that people wear masks and maintain social distance in public areas, tighter regulations must be implemented [8]. The current work performed in this area shows various computer vision algorithms that have been used for COVID-19 detection [9]. Some of the researchers included IoT-based systems aiming to help organizations follow COVID-19 safety rules [10]. Research shows that face mask sampling defects hyper-realistic mask detection and immobilization mask [11–13]. COVID-19 has been studied using the biological understanding of SARS/MERS, which

may or may not be accurate [14,15]. Deep learning network architectures have been developed for large-scale screening of COVID-19 affected persons depending on their respiratory pattern [16,17].

The overhead perception allows for a broader field of view and eliminates occlusion issues, making it ideal for social distance monitoring and calculation. This work aims to present face mask detection and deep learning social distance surveillance consecutively to detect that the social distancing norms are followed in the public area to minimize the rise of COVID-19 cases. This might help overcome hardware requirements, installation costs, human resources. Transfer learning is also used to improve the efficiency of the detection model. This is the first time an overhead view perspective has been used to measure social distance with transfer learning, to the authors' knowledge. The following is an overview of the research contributions:

1 A robust dataset consisting of 1916 masked and 1930 unmasked images has been developed from various sources, and then data augmentation is applied.
2 Five types of models have been trained on the dataset for face mask detection and their comparative analysis has been presented.
3 The human faces have been extracted using DSFD and then have been fed into the proposed model (Stacked ResNet-50) to be classified as masked or unmasked.
4 The integrated system used YOLOv5 to detect humans in a particular frame, which is then clustered using DBSCAN for social distancing monitoring. The Euclidean distance between each detected bounding box has been computed. A social distance violation threshold has been predefined which ensures whether social distancing rules are being followed.
5 The proposed technique was systematically analysed using different algorithms with the same set of parameters.

The rest of the research paper is organized under the following headings: Section 2 gives a summary of relevant research in this area, and Section 3 illustrates the methodology, including all the dataset and data preprocessing techniques. Section 4 discusses the architecture of the models used and how the research has been accomplished. The results and henceforth analysis have been presented in Section 5, which is followed by a conclusion and future scope in Section 6.

2. Related Work

Educating the workforce concerning new safety procedures, at the workplace, which helps lessen the probability of virus transmission has been the subject of countless research. With the expeditious development of machine learning methods, the combined problem of face mask detection and social distancing still has not been well addressed yet.

A sturdy social distancing evaluation algorithm, using a combination of modern-day deep learning and classic projective geometry techniques has been used to create a safe environment in a factory [18]. In another research LLE-CNNs have been used for masked face detection to draw out facial regions from the input image [19]. A researcher has come out with a GAN-based network which removes masks covering the face area and regenerates the image by building the missing hole, giving a complete face image [20] while [21] proposed a methodology to analyze chest X-Ray images of infected patients. The use of mobile robots with commodity sensors, such as an RGB-D camera and a 2-D lidar to perform collision detection, has been suggested as a tool for automatically detecting the distance between humans in crowded areas with no restriction on navigation [22]. To track social distancing in real time, the authors used YOLOv3 and a deep sort tracking scheme with balanced mAP and FPS ranking. The results concluded that a hybrid configuration of CCTV and robot, outperformed configurations, and achieved satisfactory results [23]. Test results on several COVID-19 diagnostic techniques based on deep learning methods show that models that do not consider defensive models against adversarial fluctuations remain vulnerable to adversarial attacks, according to study [24,25]. The method was used to predict continuous emotions from audio-visual data and encode shape features using

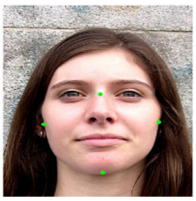
binary image masks computed from facial landmark locations [26,27]. In an active AI-based surveillance system, a pre-trained deep convolutional neural network (CNN) was used to identify individuals with bounding boxes in a monocular camera frame. The technique was capable of identifying distances between people and alerting them to prevent the deadly disease from spreading [28]. The authors used the YOLOv3 algorithm to generate an autonomous drone system for their study. The drone camera keeps track of people's social distance as well as whether or not they are wearing masks in public [29]. It was an automated system where artificial intelligence, facial detection algorithms, drones, GPS, high-end camera, radio control modules have been used for detecting whether the faces being observed are masked and socially distanced [30]. A deep learning method called Facemask net has been proposed which was working with both still images and live video stream [31]. The face mask identifier provides fast outcomes in this procedure, and can also be used in CCTV footage to evaluate if a person is correctly wearing a mask so that he does not pose a risk to others [32,33]. Another research used a transformed deep neural network to extract complex and high features from input diffused frames and tested it on the replay attack data to detect face spoofing attacks [34].

A system has been designed that automatically tracks the public places in real-time using Raspberry pi4 which captures the real-time videos of different public places to keep track of whether the people are wearing masks and abiding by the social distancing norms or not [35,36]. An integrated system was proposed where visual descriptors were used to measure the distance between the people and the methodology was able to achieve an accuracy of 93.3% [37]. Another deep learning solution was proposed that alerts the person as soon as they violate the social distancing norms via CCTV and PoseNet [38,39]. The research concluded that using a hybrid deep transfer learning model based on ResNet-50 and machine learning algorithms such as SVM, decision trees, etc. The SVM reached the best accuracy in the classification of mask/no-mask images [40].

3. Materials and Methods

A. Data Collection

The processing of CCTV feeds for mask classification is required for this study. As a result, training has been performed with low-resolution facial portraits. The images have been taken from various sources such as the Real-World Masked Face Dataset, RMFD [41]. In addition, more images have been scraped from Google Images. The facial images have been interpolated with the mask images to create the Face-mask dataset. This has been accomplished with the help of OpenCV and face-detection from dual shot face detector (DSFD). Four types of masks have been interpolated onto the faces to generate an even distribution of various types of masked images using facial points recognition. Figure 1a depicts the original unmasked image and the 4 types of masked images which have been generated is depicted in Figure 1b–e. These are the most common type of mask that are used and will make our dataset more distinctive and will result in low bias and low variance for accurate predictions.



(**a**)

(**b**)

Figure 1. *Cont*.

(**c**)

(**d**)

(**e**)

Figure 1. (**a**–**e**). Facial point detection and interpolated face masks.

B. Data Augmentation

The dataset acquired has low-resolution images containing 1916 masked images and 1930 non-masked images. Data augmentation techniques have been applied for processing images such as blurring and rotation to achieve the required results. The dataset was generalized as a result of this augmentation method, and the likelihood of overfitting was reduced. The data has been augmented for both classes (masked/unmasked images). Gaussian blur, average blur and motion blur have been used for data augmentation. The masked and unmasked image folders have been processed with these blurs techniques are discussed below:

- Gaussian blur: here a Gaussian filter is used (Equation (1)), which removes noise and reduces details.

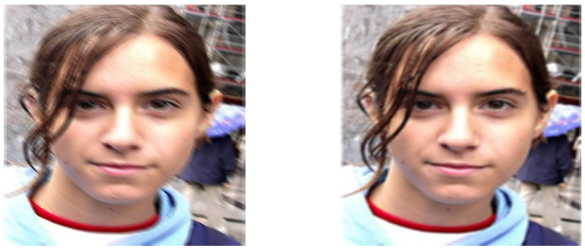
$$F_0(p,q) = Z\epsilon^{(-(p-\mu_p)^2/\sigma_p^2 + -(q-\mu_q)^2/\sigma_q^2)} \quad (1)$$

  where p and q are the inputs and μ_p and μ_q are the means while σ_p^2 and σ_q^2 are the variances of variables p and q, Z denotes the amplitude.
- Average blur: The kernel taken for this has a range of (3, 7). The image is processed with a filter box during this operation. The mean of all the pixels in the kernel region is used to replace the image's core component.
- In motion blur out of the 4 types of blurs namely vertical, horizontal, main diagonal, and anti-diagonal one was chosen at random and applied to the image. It has a kernel range of (3, 8). Figure 2 shows the sample images of original and non-masked images used in research. The dataset of 528 random images have been created using web scraping and various other sources containing masked and non-masked classes.

(**a**)

(**b**)

Figure 2. *Cont*.



(**c**) (**d**)

Figure 2. Sample of data augmentation process. (**a**) Original; (**b**) Gaussian; (**c**) motion; (**d**) average.

4. Proposed Methodology

A large amount of research has been conducted on detecting face masks and tracking social distancing violations, but none of it has succeeded in an integrated system for both. Here, the authors have proposed an integrated approach for detecting face masks on humans and monitoring social distancing violations. Face mask recognition and social distancing norm anomalies are two major concerns for reducing COVID-19 spread. The proposed research has been accomplished in two parts, first, classification of face mask using Stacked ResNet-50 and monitoring social distancing using DBSCAN clustering, both of them are covered in Section 4A and B respectively. Figure 3 shows the flowchart for the deployment phase. The initial step involves extracting a single frame from a CCTV video. It has been then processed with YOLOv5 to detect humans in it and draw bounding boxes on each detection. These boxes have then been passed for facial detection. The detected faces have then been fed into the Stacked ResNet-50 for faced mask detection and classified as masked (green color) or non-masked (red color). Meanwhile, the centroid point of the YOLOv5 human detections have been calculated from the bounding boxes. These points have been clustered using DBSCAN, those forming clusters being in a particular proximity (eps = 150) have been added to violated set and others in non-violation set. Results of both, the face mask detections and the social distancing ovulations have been displayed on the processed frame with appropriate color.

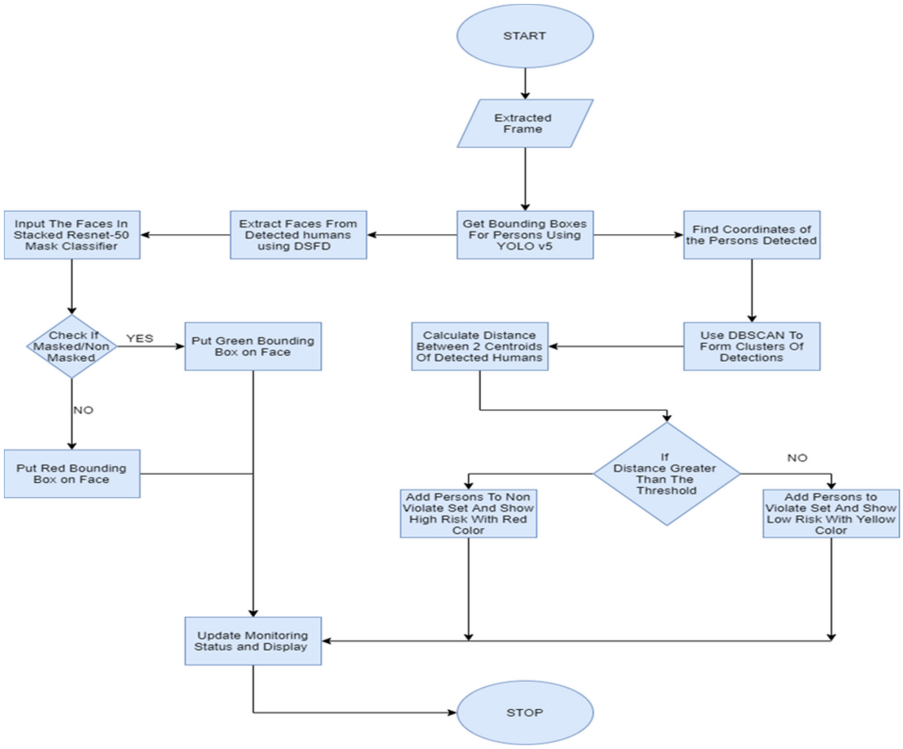


Figure 3. Flowchart for the proposed methodology.

A. Face Mask Classifier Using Stacked ResNet-50

The face mask classification task is a trivial image classification process that includes data curation, training, and inference. The conventional ResNet-50 uses a neural network consisting of 50 convolution layers that are pre-trained on a dataset known as Imagenet-21k to classify objects in different classes. The proposed Stacked ResNet-50 uses pre-trained ImageNet weights and transfer learning-based methodology. The different sizes of kernels are used in average pooling to reduce the variance and extract the low-level features from the images. The data augmentation has been used in order to boost the robustness of the Stacked ResNet-50 model. The experimentation has been performed with different hyper-parameter tuning techniques by varying the learning rate, number of epochs, activation functions and hidden layers in Stacked ResNet-50, which is explained below.

Figure 4 shows the architecture of a Stacked ResNet-50 model which has been used for the accomplishment of work. The first layer of the 8 layers network is a Convolution2D layer with kernel size 3×3, 24. After each convolution layer in the model, the ReLU activation function has been used. The first convolution's output is transmitted to a 7×7 average pooling sheet, which then is convoluted with a 3×3, 12 kernel. The output of the 2nd convolution is passed to an average pooling layer of 5×5. The output is then flattened and passed through a dense layer of 128 neurons with ReLU activation function. A dropout layer of 0.50 is used after the dense layer. The output after dropout is fed into a dense layer of one neuron. The last layer output is fed into a neural network 8 layers deep which finally gives the binary classification.

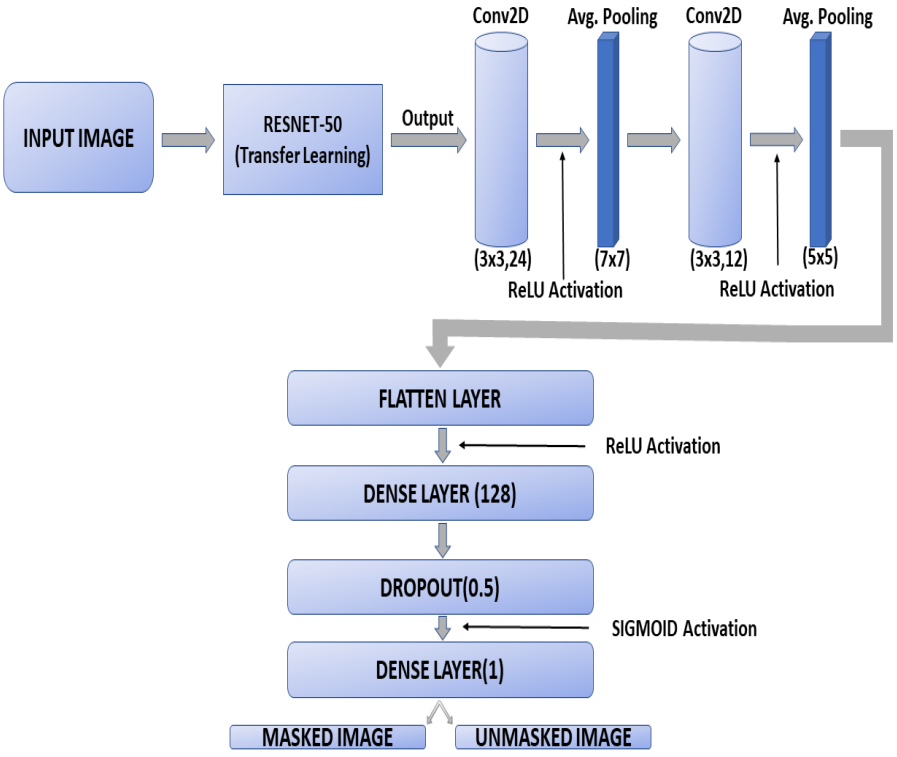


Figure 4. The model architecture of proposed Stacked Resnet-50.

The sigmoid function is used as a binary function to give binary results. The output of the layer is converted into a 1D feature vector using the method of flattening the dense layer to perform the classification process. The layer of classification consists of fully connected layers with a dropout layer of 0.50 and dense layers of size 1 respectively. The function performs the task of classification. Adam optimizer has been used with a learning rate of 0.0001.

B. Social Distance Monitoring using DBSCAN and YOLOv5

This part of the research uses the SOTA model for person detection, the YOLOv5. After detecting the humans in a frame, bounding boxes have been drawn in to find out the centroid points of each detection from the corner coordinates of the boxes. The clustering of detected humans in a frame is performed using DBSCAN, it is a clustering technique that uses density-based spatial clustering of applications and is robust to outliers thus, providing a more accurate clustering. The violations have been marked with a warning of high risk and non-violation by low risk. Other statistics have been collected for crowd analysis. The eps used in DBSCAN illustrates the maximum distance between two samples for one to be considered as in the neighborhood of the other has been set as the threshold distance that is the minimum distance expected to be maintained for social distancing. The threshold value has been calculated by taking an array consisting of different values of eps, and it has been concluded that eps of 150 forms cluster accurately. The minimum number of humans to violate social distancing should be two as per COVID-19 norms and accordingly the minimum point parameter has been set to 2 during experimentation.

a. YOLOv5

The YOLOv5 is an enhanced version of YOLOv4 with the same head. It uses a model neck to generate feature pyramids and feature pyramids help the model to generalize well on object scaling. The model used CSPResNext50, CSPDarknet53, and EfficientNet-B3 as a backbone for the YOLOv5 object detector. DenseNet is used in both the CSPResNext50 and the CSPDarknet53. The mosaic data augmentation which tiles four images together resulted in finding smaller objects in YOLOv5 with better accuracy. The YOLOv5 was trained on the video using the pre-trained weights of YOLOv5 and the scale factor of 0.00392 with the spatial size of the convolutional neural network of 416 × 416. The algorithm detects persons from the video and coordinates their position with 4 coordinates. After that, the data loader performs three types of augmentation scaling, color space changes, and mosaic augmentation. YOLOv5 produces three predictions for each spatial location in an image at different scales, which addresses the problem of identifying small objects accurately. Each prediction is monitored by computing objects, boundary box regressor, and classification scores. The overall loss function of YOLOv5 consist of localization loss, cross-entropy, and confidence loss for classification score, defined below, in Equation (2).

$$\begin{gathered}\alpha_{coord}\sum_{m=0}^{P^2}\sum_{n=0}^{A}1_{m,n}^{obj}\left((s_x-\hat{s_x})^2+(s_y-\hat{s_y})^2+(s_w-\hat{s_w})^2+(s_h-\hat{s_h})^2\right)+\\ \sum_{m=0}^{P^2}\sum_{n=0}^{A}1_{m,n}^{obj}\left(-log(\beta(s_0))+\sum_{k=1}^{C}ACD(\hat{y_k},\beta(\mu_k))\right)+\alpha_{noobj}\sum_{m=0}^{P^2}\sum_{n=0}^{A}1_{m,n}^{noobj}\left(-log(1-\beta(s_0))\right)\end{gathered} \tag{2}$$

where α_{coord} indicates the weight of the coordinate error, p^2 indicates the number of grids in the image, and A is the number of total generated bounding boxes per grid. $1_{m,n}^{obj} = 1$ describes that the object in the nth bounding box in grid m, otherwise it is 0.

b. DBSCAN

DBSCAN (density-based spatial clustering of applications with noise) groups points that are adjacent to each other using distance measurements such as Euclidean distance, Manhattan distance, etc. It has two main factors, ε which defines the size and borders of each neighbourhood and minimum points. The neighbourhood of a point x is defined mathematically as follows:

$$N_\varepsilon(x) = B_d(x,\varepsilon) = \{y|\delta(x,y) \le \varepsilon\} \tag{3}$$

5. Result and Analysis

The research has been completed in two phases, the first one being the face mask classifications and later the social distance monitoring. The subsections below discuss the results and analysis of the proposed methodology in detail.

A. Face mask Classifier using Stacked ResNet-50

Initially, five models i.e., Model (1)—CNN, Model (2)—MobileNetV3, Model (3)—InceptionV3, Model (4)—ResNet-50, Model (5)—Stacked ResNet-50 have been trained on

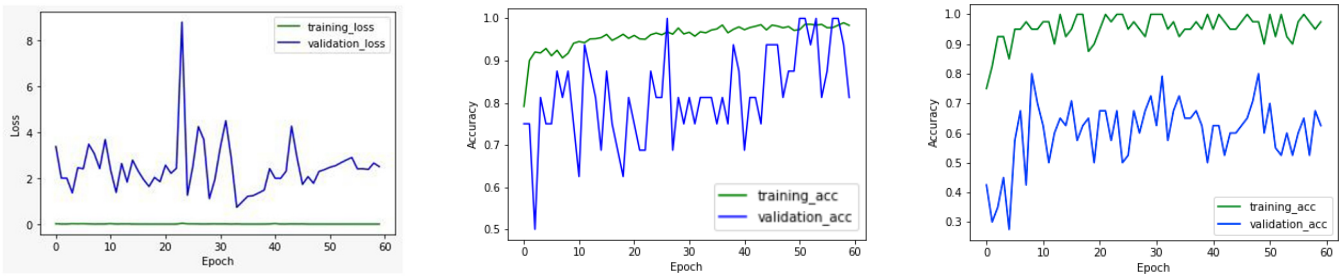
a dataset of 1916 masked images and 1930 non-masked images. A comparison of these models has been presented in Table 1 while their training and validation curves have been illustrated in Figure 5, whereas their training history for model accuracy and model loss has been represented in Figures 5a–e and 6a–e respectively.

Table 1. Comparative analysis of various models.

	Training					Testing				
Model	**CNN**	**MobileNet V3**	**Inception V3**	**Resnet-50**	**Stacked ResNet-50**	**CNN**	**MobileNet V3**	**Inception V3**	**ResNet-50**	**Stacked ResNet-50**
Accuracy	0.92	0.94	0.95	0.96	0.96	0.75	0.83	0.83	0.77	0.87
Precision	0.57	0.59	0.67	0.70	0.73	0.51	0.56	0.66	0.54	0.71
Recall	0.67	0.94	0.94	0.90	0.93	0.65	0.93	0.93	0.67	0.92
F1 Score	0.61	0.7	0.77	0.77	0.81	0.52	0.59	0.65	0.59	0.79
Loss	0.13	0.11	0.11	0.18	0.12	0.18	0.13	0.13	0.19	0.14
Specificity	0.53	0.82	0.92	0.78	0.83	0.48	0.8	0.91	0.79	0.81

The purpose of the proposed methodology is to detect whether people are wearing masks or not by using Stacked ResNet-50. Augmentation modules have been used to reduce the variance and prevent overfitting by forming new examples in the training dataset. The model consists of 8 layers and is trained using the ReLU activation function, which introduces linearity. The experiment was conducted five times, each time for two hours. The models were trained on 60 epochs with a learning rate of 0.0001 and the batch size was set to 64. The Adam optimizer has been used for optimizing the learning and reducing losses.

The described model has been trained and 1916 masked images and 1930 non-masked images, validated on 120 images and tested on 528 masked images. Since it is not recommended to test the model from the training dataset, we have tested the models on another dataset containing 528 images using web scraping and other sources. During the training of the model, the images were resized and augmented using different properties. The images were horizontally flipped and set the range of its brightness from [0.5, 1.25], zoom range from [0.8, 1] and setting the rotation to 15 degrees.



(**a**) (**b**) (**c**)

Figure 5. *Cont.*

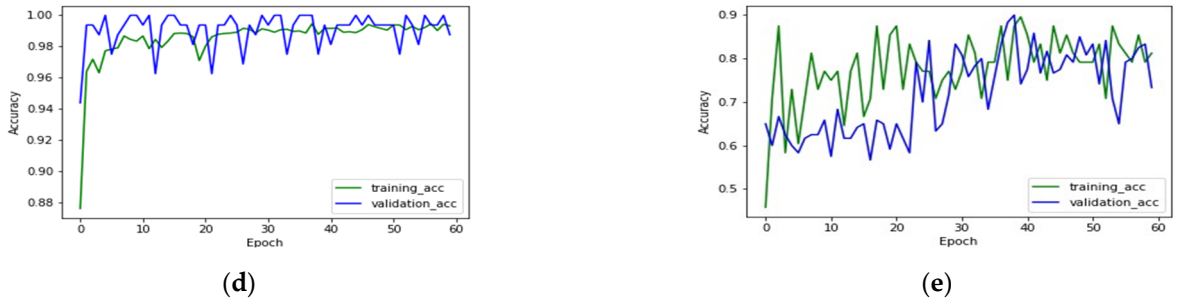


Figure 5. (**a**–**e**): Training accuracy (in green) and validation accuracy (in blue) during the training of all models for the dataset.

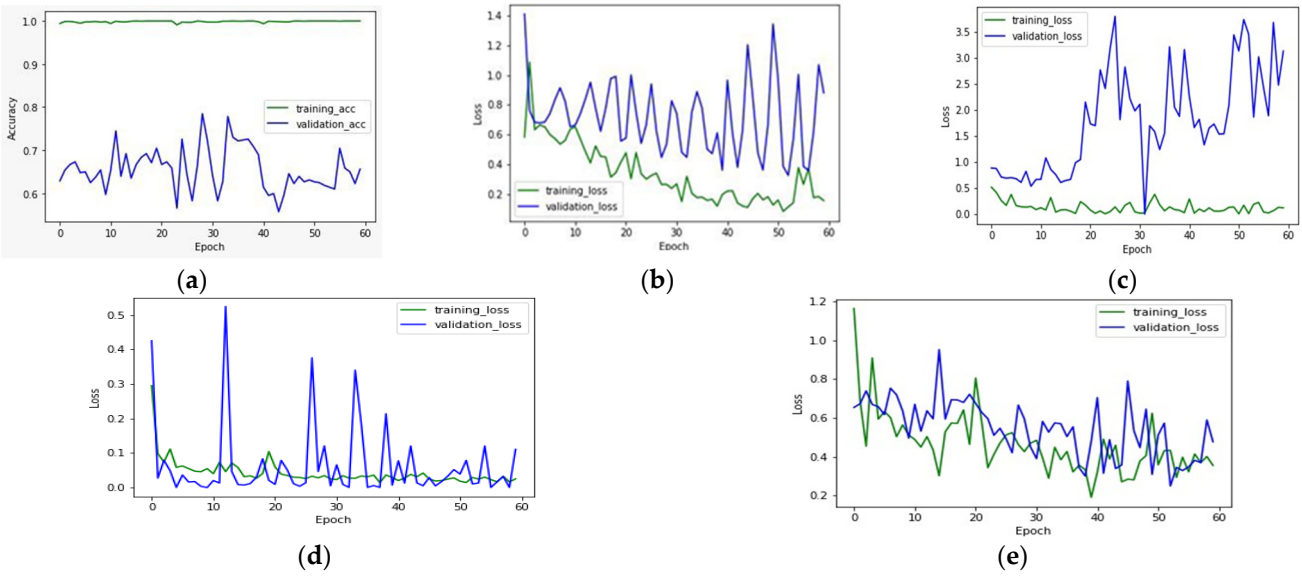
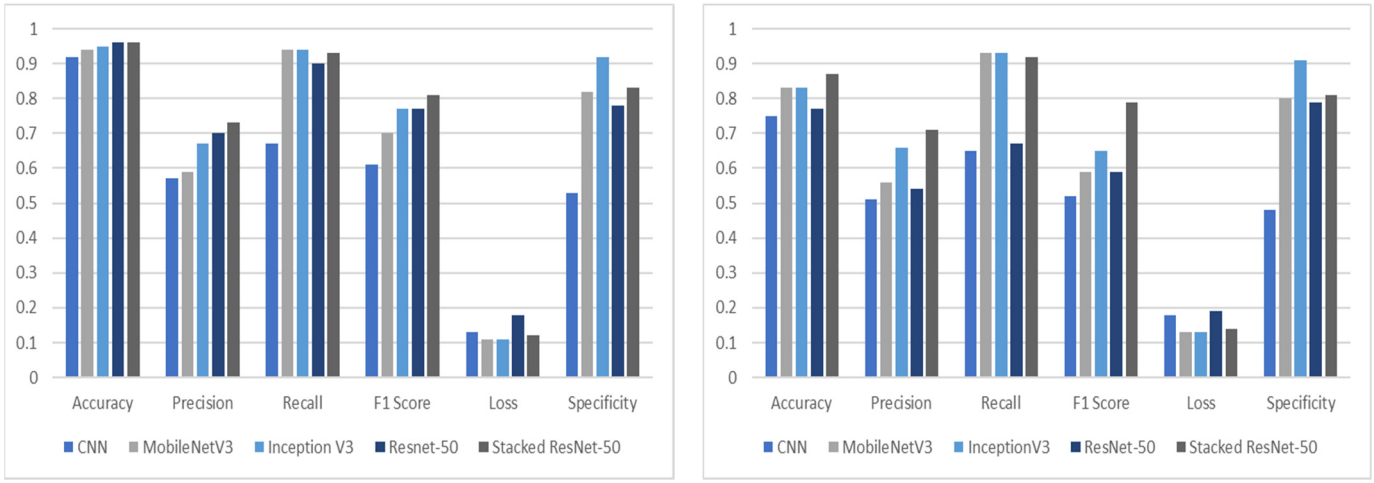


Figure 6. (**a**–**e**): Training loss (in green) and validation loss (in blue) during the training of all models for the dataset.

The final results were obtained with a training accuracy of 0.85, precision of 0.97 recall of 0.80 and having an F1 score of 0.876. The analysis of the different models on training and testing data, which indicated that the CNN was compelled to overfit due to the algorithm's need for integrated transfer learning. However, the models that used augmentation and transfer learning showed high accuracy and were more robust to the testing data. It can be substantiated that CNN without data augmentation has shown high variance and high bias on the training data with an accuracy of 92%, precision of 57% and a recall score of 67%. The MobileNetV3 and InceptionV3 with data augmentation showed an accuracy of 94%, precision of 57% and recall of 94%. Accuracy on the testing set came out to be 77% only and very low precision. On the other hand, we can observe that Stacked ResNet-50 has outperformed all the other models with an accuracy of 96%, precision of 73% and recall of 93% respectively. On the other hand, we can observe that Stacked ResNet-50 has outperformed all the other models with an accuracy of 96%, precision of 73% and recall of 93% respectively. Figure 7a,b shows the achieved percentage in various performance metrics during the training and testing phase of the five models, respectively. From the plots it can be seen that, the Stacked ResNet-50 the has low false positive and low false negative rate and hence will perform better when it comes to real life application, since high false positive rates are much a concern than high false negative rates.

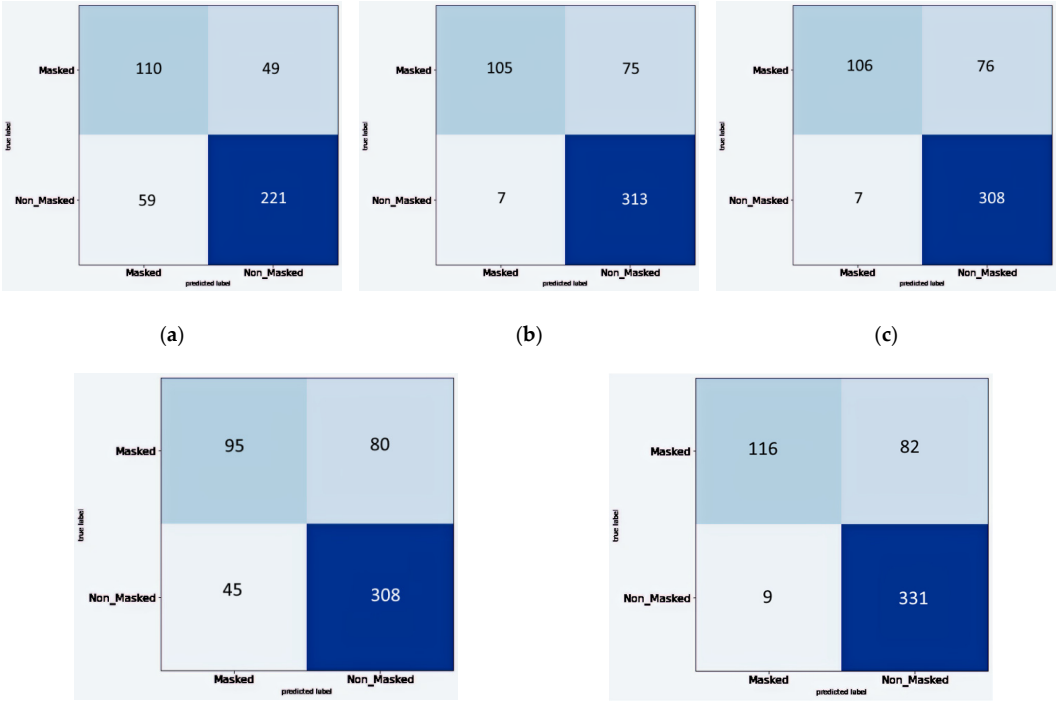


(**a**) (**b**)

Figure 7. (**a**,**b**): Comparison graph of various models in training and testing phase.

Table 1 concludes that the results of the training and testing phase with the Stacked ResNet-50 model performed best amongst the others. All models have been compared and analyzed using accuracy, precision, recall, loss, F1 score, loss, specificity as evaluation criteria. Table 1 shows the detail analysis of all models. The proposed methodology has achieved maximum accuracy 96 during training and 87 in testing phase.

The described model used data augmentation and gave desired results by reducing the false negatives. The confusion matrix for the proposed model has been shown in Figure 8a–e. It has been discovered that Stacked ResNet-50 has used a tremendous amount of data augmentation. The confusion matrix results in a recall of 92% and a precision of 71% which prevents the model from overfitting.



(**a**) (**b**) (**c**)

(**d**) (**e**)

Figure 8. (**a**–**e**): Confusion Matrix for CNN, MobileNetV3, InceptionV3, ResNet-50 and Stacked ResNet-50.

Figure 9 shows the output of the model when executed on a test image, the masked persons can be detected in the green bounding box and the unmasked faces are detected in the red bounding box. The model performed well on the side faces also, this has been achieved using our own dataset which was robust enough to deal with all the cases.

Figure 9. Mask/unmasked classification (Stacked ResNet-50).

B. Social Distance Monitoring using DBSCAN and YOLOV5

Figure 10 demonstrates the results of YOLOv5 with 100 epochs. The batch size used for YOLOv5 is 32 and the model was trained on an Nvidia GTX 1060. These plots further show how the size of the bounding box has dwindled after every iteration, which has reduced as the model detects more precisely after every iteration. Objectness shows the probability that an object exists in an image and is used as a loss function. The classification decreases as the Objectness of the model decreases with each iteration because of the decimated size of the bounding box after each iteration, which further signifies the model classifies accurately. The precision of the model, of which the true positives and the false positives are derived by the IOU (intersection over union). The mAP is the area under the precision—recall curve with the IOU of 0.5. The precision and recall of the object detection model are 0.6 and 0.98.

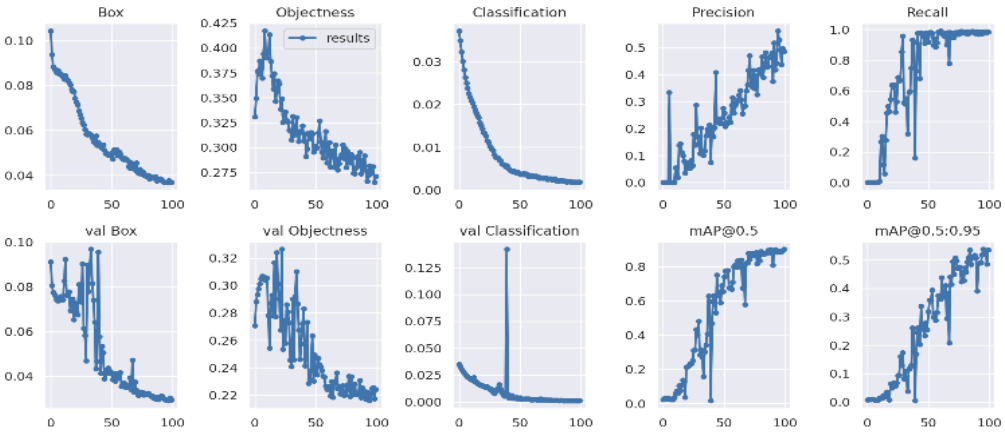


Figure 10. Analysis of YOLOv5.

An input video of length 60 s, 60 fps, and 480 p has been taken from the internet; each frame has been extracted and processed to the methodology discussed in the above sections. The persons detected have been cropped, these detections have been passed to a face detector which extracted the faces of the person. These facial images have been fed into the Stacked ResNet-50 model which classified them into masked or unmasked classes. Figure 11 shows the output frame when the video is fed into the YOLOv5 network for both face mask detection and social distance monitoring.

Figure 11. Sample outputs of the proposed framework for face mask.

For the current instance in the video, the model calculates the number of persons wearing a mask or not at a particular time and additionally calculates the social distancing in the nearby region with the help of DBSCAN clustering.

The green color in the processed frame denotes the people wearing masks and following social distancing while the red color shows the warning of social distancing not being followed. Total number of persons detected in the frame where 29, out of which 12 flouted social distancing norms and 17 where maintain appropriate distance. There were 12 people who were wearing a mask, while for other 17 no faces were detected for classification. The integrated system showed a performance of 32 fps on the input video.

C. Comparison with State of art methods

An integrated approach for recognizing human face masks and monitoring violations of social distancing have been proposed. Two significant problems for limiting COVID-19 spread are face mask detection and social distancing norm have been discussed. The proposed study is divided into two parts: the first is the categorization of face masks using Stacked ResNet-50, and the second is the monitoring of social separation using DBSCAN clustering. Table 2 discusses the comparison of various state of art methods.

Table 2. Comparison of various state of art methods on the basis of accuracy (AC) and average precision (AP).

Source	Authors	Methodology	Social Distance Monitoring	Resul t(%)
[42]	S Singh et al.	Faster R-CNN + Yolov3	No	AP-62(FasterCNN) 55(YOLOv3)
[43]	Loey, Mohamed, et al.	YOLOv2 + ResNet50	No	AP-81 *
[19]	Ge, Shiming, et al.	LLE-CNNs	No	AP-76.1 *
[44]	Ejaz, Md Sabbir, et al.	PCA	No	AC-70 *
[45]	Venkateswarlu et al.	ResNet-50	No	AC-84.1 *
[46]	Yu Jimin et al.	YOLOv4	No	AC-98.3 *
[47]	S. Sethi et al.	ResNet-50	No	AP-98.2 *
Proposed Methdology		Stacked Resnet-50	YES	87

* Experimental results may vary when different datasets are used. One of the main reasons for different results is the bias-variance trade-off. Variance is measured on how algorithm performed during data training process. We have tested proposed methodology on different datasets (we have prepared our testing dataset by collecting masked/unmasked images using image scraping) that is the reason of larger variance in the proposed model. This also shows that proposed model is robust to a new dataset and testing on the different types of datasets from training will always result in lesser accuracy. In most of the papers mentioned above, the author has used the train–test split to test the model on the same training data which has resulted in higher accuracy. Different hyperparameters usage also results in different results.

6. Conclusions and Future Scope

The deadly pandemic has taken 5,049,374 lives in the world and 459,873 in India as of now, some countries have been facing third wave of the pandemic. The authors have proposed an integrated approach for detecting face masks on humans and monitoring social distancing violations. Face mask recognition and social distancing norm anomalies are two major concerns for reducing COVID-19 spread. The proposed research has been accomplished in two parts, first, classification of face mask using Stacked ResNet-50 and monitoring social distancing using DBSCAN clustering. A single frame from a CCTV video is extracted as the first step. It was then processed through YOLOv5 to see whether there were any humans present and create bounding boxes around each one. After that, the boxes have been sent to be scanned for faces. The faces were then fed into the Stacked ResNet-50 for facial mask detection, where they were classified as masked (green colour) or non-masked (red colour) (red color). In our comparative analysis for face mask detection, the proposed Stacked ResNet-50 Model outperformed CNN, InceptionV3, MobileNetV3 and the pretrained ResNet-50 models, and it was used in our proposed Face Mask Detection technique. Using binary cross-entropy, the results showed a training accuracy of 96 percent and a training loss of 12 percent when trained on a set of 3846 images that included masked and unmasked images. On 528 testing images, the testing accuracy was 84 percent, and the testing loss was 14%. The model uses the Adam optimizer, which has a learning rate of 0.0001, as well as ReLU and Sigmoid Activation functions. The facial features have been extracted using the DSFD (dual shot face detector). The humans have been detected from real-time video stream of a minute long, having a frame rate of 60 fps and a resolution of 320 $\times$ 240 pixels, using YOLOv5 and social distancing has been monitored using DBSCAN. The YOLOv5 has a precision and recall of 0.6 and 0.98 respectively with a mAP (mean average precision) of 0.98. The authors have been able to achieve a frame rate of 32 fps on the input video for the integrated analysis of face masks and social distancing of detected humans.

This methodology can be improved further by using different state of art algorithms for face detection. Other transfer learning algorithms and techniques such as hyperparameter tuning can be used to improve the face mask classification. Using a more diverse dataset can result in a less robust model to the testing images. Different state of art algorithms can be used with different data augmentation techniques and FPN (feature pyramid network) with different sliding windows can achieve better results. The research accomplished can be easily incorporated with the CCTVs in open spaces thus preventing the spread and saving lives. This system can be made more efficient by providing real-time CCTV facial data for training of the mask classifier. It can be further used with hardware to warn authorities as soon as crowding increases and take necessary measures. Since the world has been witnessing a pandemic in the past also, this approach can be used for dealing with forthcoming pandemics as these two measures can lower the spread to a great extent.

Author Contributions: Conceptualization, I.S.W., D.K. and K.S.; methodology, D.K. and K.S.; software, I.S.W.; validation, K.S.; formal analysis, D.K.; investigation, J.D.H.; resources, J.D.H. and D.E.P.; data curation, I.S.W. and K.S.; writing—I.S.W. and K.S.; writing—review and editing, D.K. and J.D.H.; visualization, D.K. and J.D.H.; supervision, D.K., J.D.H. and D.E.P.; project administration, D.K., J.D.H. and D.E.P.; funding acquisition, J.D.H. and D.E.P. All authors have read and agreed to the published version of the manuscript.

Funding: There is no funding received for this research.

Conflicts of Interest: All authors declare that they have no conflicts of interest.

References

1. WHO Corona-Viruses (COVID-19). 2020. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed on 4 November 2021).
2. Kooraki, S.; Hosseiny, M.; Myers, L.; Gholamrezanezhad, A. Coronavirus (COVID-19) Outbreak: What the Department of Radiology Should Know. *J. Am. Coll. Radiol.* **2020**, *17*, 447–451. [CrossRef]

3. Siedner, M.J.; Gandhi, R.T.; Kim, A.Y. Desperate Times Call for Temperate Measures: Practicing Infectious Diseases During a Novel Pandemic. *J. Infect. Dis.* **2020**, *222*, 1084–1085. [CrossRef] [PubMed]
4. Dhand, R.; Li, J. Coughs and sneezes: Their role in transmission of respiratory viral infections, including SARS-CoV-2. *Am. J. Respir. Crit. Care Med.* **2020**, *202*, 651–659. [CrossRef]
5. Majidi, H.; Niksolat, F. Chest CT in patients suspected of COVID-19 infection: A reliable alternative for RT-PCR. *Am. J. Emerg. Med.* **2020**, *38*, 2730–2732. [CrossRef] [PubMed]
6. Palanisamy, V.; Thirunavukarasu, R. Implications of big data analytics in developing healthcare frameworks—A review. *J. King Saud Univ.-Comput. Inf. Sci.* **2019**, *31*, 415–425. [CrossRef]
7. Chu, D.K.; Akl, E.A.; Duda, S.; Solo, K.; Yaacoub, S.; Schünemann, H.J.; Reinap, M. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: A systematic review and meta-analysis. *Lancet* **2020**, *395*, 1973–1987. [CrossRef]
8. Gilani, S.; Roditi, R.; Naraghi, M. COVID-19 and anosmia in Tehran, Iran. *Med. Hypotheses* **2020**, *141*, 109757. [CrossRef] [PubMed]
9. Elston, D.M. The coronavirus (COVID-19) epidemic and patient safety. *J. Am. Acad. Dermatol.* **2020**, *82*, 819–820. [CrossRef]
10. Ferguson, N.M.; Cummings, D.A.; Cauchemez, S.; Fraser, C.; Riley, S.; Meeyai, A.; Iamsirithaworn, S.; Burke, D. Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia. *Nature* **2005**, *437*, 209–214. [CrossRef]
11. Wen, L.; Li, X.; Gao, L. A transfer convolutional neural network for fault diagnosis based on ResNet-50. *Neural Comput. Appl.* **2020**, *32*, 6111–6124. [CrossRef]
12. Sanders, J.G.; Jenkins, R. Individual differences in hyper-realistic mask detection. *Cogn. Res. Princ. Implic.* **2018**, *3*, 24. [CrossRef]
13. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
14. Chinazzi, M.; Davis, J.T.; Ajelli, M.; Gioannini, C.; Litvinova, M.; Merler, S.; Piontti, Y.; Pastore, A.; Mu, K.; Rossi, L.; et al. The Effect of Travel Restrictions on the Spread of the 2019 Novel Coronavirus (Covid-19) Outbreak. *Science* **2020**, *368*, 395–400. [CrossRef] [PubMed]
15. Peeri, N.C.; Shrestha, N.; Rahman, M.S.; Zaki, R.; Tan, Z.; Bibi, S.; Baghbanzadeh, M.; Aghamohammadi, N.; Zhang, W.; Haque, U. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: What lessons have we learned? *Int. J. Epidemiol.* **2020**, *49*, 717–726. [CrossRef]
16. Jain, R.; Gupta, M.; Taneja, S.; Hemanth, D.J. Deep learning based detection and analysis of COVID-19 on chest X-ray images. *Appl. Intell.* **2021**, *51*, 1690–1700. [CrossRef]
17. Ozturk, T.; Talo, M.; Yildirim, E.A.; Baloglu, U.B.; Yildirim, O.; Acharya, U.R. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **2020**, *121*, 103792. [CrossRef] [PubMed]
18. Kumar, D.; Batra, U. Classification of Invasive Ductal Carcinoma from histopathology breast cancer images using Stacked Generalized Ensemble. *J. Intell. Fuzzy Syst.* **2021**, *40*, 4919–4934. [CrossRef]
19. Qin, B.; Li, D. Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19. *Sensors* **2018**, *20*, 5236. [CrossRef]
20. Din, N.U.; Javed, K.; Bae, S.; Yi., J. A Novel GAN-Based Network for Unmasking of Masked Face. *IEEE Access* **2020**, *8*, 44276–44287. [CrossRef]
21. Loey, M.; Smarandache, F.; Khalifa, N.E.M. Within the Lack of Chest COVID-19 X-ray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning. *Symmetry* **2020**, *12*, 651. [CrossRef]
22. Sathyamoorthy, A.J.; Patel, U.; Savle, Y.A.; Paul, M.; Manocha, D. COVID-robot: Monitoring social distancing constraints in crowded scenarios. *arXiv* **2020**, arXiv:2008.06585.
23. Punn, N.S.; Sonbhadra, S.K.; Agarwal, S.; Rai, G. Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques. *arXiv* **2020**, arXiv:2005.01385.
24. Nguyen, C.T.; Saputra, Y.M.; Van Huynh, N.; Nguyen, N.T.; Khoa, T.V.; Tuan, B.M.; Ottersten, B. A comprehensive survey of enabling and emerging technologies for social distancing—Part I: Fundamentals and enabling technologies. *IEEE Access* **2020**, *8*, 153479–153507. [CrossRef] [PubMed]
25. Rahman, A.; Hossain, M.S.; Alrajeh, N.A.; Alsolami, F. Adversarial Examples—Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices. *IEEE Internet Things J.* **2021**, *8*, 9603–9610. [CrossRef]
26. Militante, S.V.; Dionisio, N.V. Real-Time Facemask Recognition with Alarm System using Deep Learning. In Proceedings of the 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, 8 August 2020; pp. 106–110.
27. Jaiswal, S.; Valstar, M. Deep learning the dynamic appearance and shape of facial action units. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–8.
28. Yang, D.; Yurtsever, E.; Renganathan, V.; Redmill, K.; Özgüner, Ü. A Vision-Based Social Distancing and Critical Density Detection System for COVID-19. *Sensors* **2021**, *21*, 4608. [CrossRef] [PubMed]
29. Ramadass, L.; Arunachalam, S.; Sagayasree, Z. Applying deep learning algorithm to maintain social distance in public place through drone technology. *Int. J. Pervasive Comput. Commun.* **2020**, *16*, 223–234. [CrossRef]
30. Hossain, M.S.; Muhammad, G.; Guizani, N. Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat COVID-I9 Like Pandemics. *IEEE Netw.* **2020**, *34*, 126–132. [CrossRef]

31. Inamdar, M.; Ninad, M. Real-Time Face Mask Identification Using Face Mask Net Deep Learning Network. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3663305 (accessed on 4 November 2021). [CrossRef]
32. Lucena, O.; Junior, A.; Moia, V.; Souza, R.; Valle, E.; Lotufo, R. Transfer Learning Using Convolutional Neural Networks for Face Anti-spoofing. In Proceedings of the Lecture Notes in Computer Science, Montreal, QC, Canada, 5–7 July 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 27–34.
33. Alotaibi, A.; Mahmood, A. Enhancing computer vision to detect face spoofing attack utilizing a single frame from a replay video attack using deep learning. In Proceedings of the 2016 International Conference on Optoelectronics and Image Processing (ICOIP), Warsaw, Poland, 10–12 June 2016; pp. 1–5.
34. Dzisi, E.K.J.; Dei, O.A. Adherence to social distancing and wearing of masks within public transportation during the COVID 19 pandemic. *Transp. Res. Interdiscip. Perspect.* **2020**, *7*, 100191. [CrossRef]
35. Ahmed, I.; Ahmad, M.; Rodrigues, J.J.; Jeon, G.; Din, S. A deep learning-based social distance monitoring framework for COVID-19. *Sustain. Cities Soc.* **2021**, *65*, 102571. [CrossRef]
36. Walia, I.; Srivastava, M.; Kumar, D.; Rani, M.; Muthreja, P.; Mohadikar, G. Pneumonia Detection using Depth-Wise Convolutional Neural Network (DW-CNN). *EAI Endorsed Trans. Pervasive Health Technol.* **2020**, *6*. [CrossRef]
37. Sener, F.; Ikizler-Cinbis, N. Two-person interaction recognition via spatial multiple instance embedding. *J. Vis. Commun. Image Represent.* **2015**, *32*, 63–73. [CrossRef]
38. Ghorai, A.; Gawde, S.; Kalbande, D. Digital Solution for Enforcing Social Distancing. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC), New Delhi, India, 20–22 February 2020.
39. Kumar, D.; Jain, N.; Khurana, A.; Mittal, S.; Satapathy, S.C.; Senkerik, R.; Hemanth, J.D. Automatic Detection of White Blood Cancer from Bone Marrow Microscopic Images Using Convolutional Neural Networks. *IEEE Access* **2020**, *8*, 142521–142531. [CrossRef]
40. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement* **2021**, *167*, 108288. [CrossRef] [PubMed]
41. Wang, Z.; Wang, G.; Huang, B.; Xiong, Z.; Hong, Q.; Wu, H.; Liang, J. Masked face recognition dataset and application. *arXiv* **2020**, arXiv:2003.09093.
42. Singh, S.; Ahuja, U.; Kumar, M.; Kumar, K.; Sachdeva, M. Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. *Multimed. Tools Appl.* **2021**, *80*, 19753–19768. [CrossRef]
43. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustain. Cities Soc.* **2021**, *65*, 102600. [CrossRef]
44. Ejaz, S.; Islam, R.; Ejaz, M.S.; Islam, M.R.; Sifatullah, M.; Sarker, A. Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition. In Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 3–5 May 2019.
45. Venkateswarlu, I.B.; Kakarla, J.; Prakash, S. Face mask detection using MobileNet and Global Pooling Block. In Proceedings of the 2020 IEEE 4th Conference on Information & Communication Technology (CICT), Chennai, India, 3–5 December 2020; pp. 1–5.
46. Yu, J.; Zhang, W. Face Mask Wearing Detection Algorithm Based on Improved YOLO-v4. *Sensors* **2021**, *21*, 3263. [CrossRef]
47. Sethi, S.; Kathuria, M.; Kaushik, T. Face Mask Detection using Deep Learning: An Approach to Reduce Risk of Coronavirus Spread. *J. Biomed. Inform.* **2021**, *120*, 103848. [CrossRef]