IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 18, 2021, accepted July 22, 2021, date of publication July 26, 2021, date of current version August 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100070

Comprehensive Review on Facemask Detection Techniques in the Context of Covid-19

AFSANA NOWRIN[1], SHARMIN AFROZ[1], MD. SAZZADUR RAHMAN[2], IMTIAZ MAHMUD[3], AND YOU-ZE CHO[3], (Senior Member, IEEE)

[1]Department of Information and Communication Technology (ICT), Bangladesh University of Professionals, Dhaka 1216, Bangladesh
[2]Institute of Information Technology, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh
[3]School of Electronics and Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding authors: You-Ze Cho (yzcho@ee.knu.ac.kr) and Md. Sazzadur Rahman (sazzad@juniv.edu)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) by the Ministry of Education under Grant NRF-2018R1A6A1A03025109, and in part by the National Research Foundation of Korea (NRF) Grant by the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2019R1A2C1006249.

ABSTRACT The outbreak of Coronavirus Disease 2019 (Covid-19) had an enormous impact on humanity. Till May 2021, almost 172 million people have been affected globally due to the contagious spread of Covid-19. Although the distribution of vaccines has been started, the worldwide mass distribution is yet to happen. According to the World Health Organization (WHO), wearing a facemask can reduce the contagious spread of Covid-19 significantly. The governments of different countries have recommended implementing the ''no mask, no service'' method to impede the spread of Covid-19. However, even the improper wearing of a facemask can obstruct the goal and lead to the spread of the virus. Therefore, to ensure public safety, a system for monitoring facemasks on faces, commonly known as a facemask detection algorithm, is essential for overcoming this crisis. The facemask detection algorithms are part of the object detection algorithms which are used to detect objects in an image. Among the various object detection algorithms, deep learning showed tremendous performance in facemask detection for its excellent feature extraction capability than the traditional machine learning algorithms. However, there remains a lot of scope for future research to build an efficient facemask detection system. Therefore, this study aims to draw attention to the researchers by providing a narrative and meta-analytic review on all the published works related to facemask detection in the context of Covid-19. Because facemask detection algorithms are run mainly by adopting object detection algorithms, this paper also explores the progress of object detection algorithms over the last few decades. A comprehensive analysis of different datasets used in facemask detection techniques by many studies has been explored. The performance comparison among these algorithms is discussed in narrative and meta-analytic approaches. Finally, this study concludes with a discussion of some of the major challenges and future scope in the related field.

INDEX TERMS Covid-19, convolutional neural network, deep neural network, facemask detection, Covid-19 health, object detection, machine learning.

I. INTRODUCTION

The novel coronavirus pandemic, also known as Covid-19 has led the world to a crisis affecting more than 172 million people and causing the death of approximately 3.7 million based on the global report provided by the WHO on 2nd June, 2021 [1]. Covid-19 is a family of coronaviruses whose previous outbreaks were included Severe Acute Respiratory Syndrome (SARS-CoV) in 2003 [2] and Middle East Respiratory Syndrome (MERS- CoV) in 2012 [3]. The SARS-CoV-2, also a replaceable term for Covid-19, is a novel strain of coronavirus first detected in the city of Wuhan, in the province of Hubei, China [4]. The virus spreads contagiously from person to person, as well as by human contact and contaminated surfaces. Covid-19 was recognized as a pandemic by the WHO in March 2020 [5]. During the Covid-19 outbreak, the highest rate of confirmed cases and deaths has been found in America, followed by Europe. Since this is a novel virus that has not been found before, researchers are still working on effective vaccines to eradicate Covid-19.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

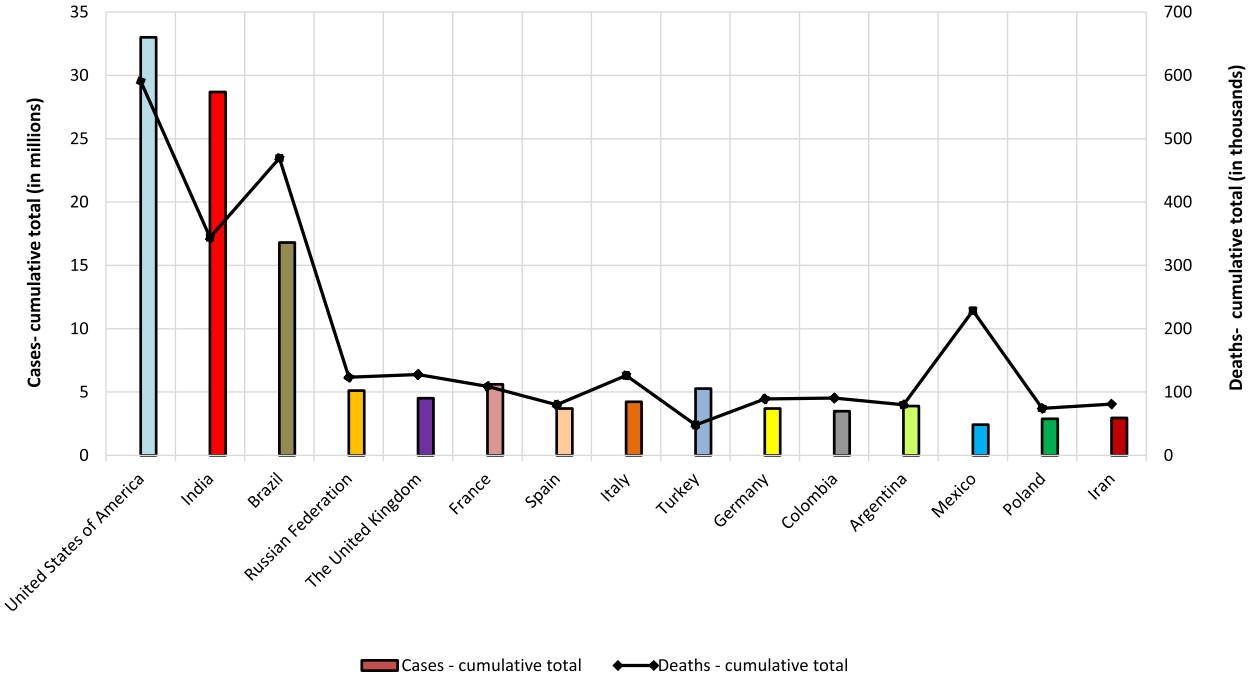


FIGURE 1. Trend of cumulative total cases and cumulative total deaths in different countries worldwide due to Covid-19. The straight line and columns indicate the cumulative death cases and cumulative Covid-19 affected cases from January 3, 2020 to June 2, 2021, respectively.

In the last year, Covid-19 devastated humanity without any consideration of age, gender, and region. The virus stopped the world for a while. Not only physical sufferings but also Covid-19 resulted in a financial crisis in many developed and developing countries. The number of confirmed cases and deaths raised to the midpoint of the year is beyond description. To date (2^{nd} June, 2021), the globally affected count and death count is 172.2 million and 3.7 million, respectively. Fig. 1 presents the statistics according to the WHO, where the most affected first-world countries have been taken as a sample for visualizing the acuteness and contagiousness of this virus. The straight line represents the cumulative total deaths, and the columns represent the cumulative confirmed cases across a particular country. The acuteness of Covid-19 last year escalated gently from the graphical representation. The number of cases in America is nearly 68 million, and the number of deaths rose to 1.7 million (June 2, 2021). As a continent, America alone has reported almost 48.3% of the world's deaths from Covid-19. Other first-world countries in Europe and other regions, such as Brazil, Germany, United Kingdom, Italy, France, and India, have lost people to Covid-19. The curve of cases is also significant. The ratio of confirmed and recovered cases was very low. The whole world stopped due to one fatal virus because of the unavailability of vaccines and lack of awareness. Fig. 2 presents a pie chart to show the deaths in different regions around the world, where America and Europe reports the highest death cases.

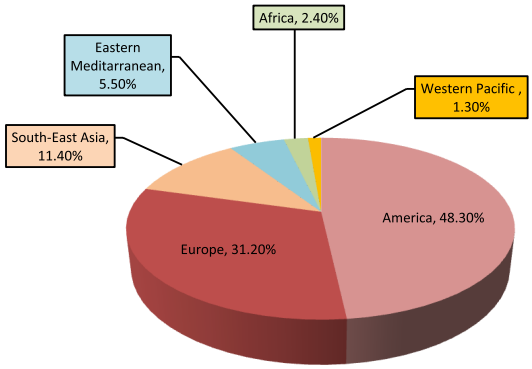


FIGURE 2. Cumulative death cases reported by the world health organization (WHO) from different regions globally are shown in this pie chart. The reported number of deaths is taken within a time limit from January 3, 2020, to June 02, 2021.

Within a short period of time, the virus has spread contagiously than any others. There are many reasons for this outrageous transmission, such as, lack of awareness, not maintaining social distancing, not wearing facemasks in public gatherings, and so on. Recent studies have shown that wearing a facemask can prevent the contagious transmission of coronavirus as well as any respiratory diseases considerably [6]–[8]. Although vaccines of coronavirus have been invented and their mass distribution has started in early December 2020, they only reduce the complications and morbidity of Covid-19 rather than eradicating the virus. Hence, one of the most efficient and safest ways to protect an

individual from this virus is to wear a facemask [9]–[11]. The WHO strongly recommends wearing a facemask in public gatherings and outside because it blocks the transmission of the virus through the nasal or oral cavity [12]. Facemasks (i.e., cotton fabrics, surgical, N-95) provide from 50% to 95% protection against the Covid-19 virus [13]. In the mentioned scenario, to be safe from getting affected by Covid-19, it is the best practice to always wear a facemask. Governments from different countries have made wearing a mask mandatory. ''No mask, no service'' tags have been initialized to spread awareness. In this scenario, ensuring whether a person in a public gathering or an organization is wearing a mask or not has been an immense area of research. Conventional procedures, referred to as human force or guards, are not always feasible to monitor if a person is wearing a facemask. Therefore, facemask detection using machine learning or deep learning is necessary. In contrast, ensuring the proper wearing of a facemask is essential. Otherwise, although the mask will be detected, the contagious transmission of the virus will not be taken into account. Consequently, the purpose of wearing a mask, which is to protect against Covid-19, will be negated. Recently, many studies have been conducted to determine if a person is wearing a facemask in public and maintaining precautions.

A facemask detection algorithm is a type of object detection algorithm that detects and localizes facemasks in an image or video stream with bounding boxes. Object detection is a combination of image classification and image localization. The image classification produces the class of an object. For example, the classification of facemasks will classify an image as ''a masked face'' class or a ''no-masked face class.'' The image localization decides where the facemask is located and draws bounding boxes around its extent. Currently, some facemask detection algorithms focus only on image classification and others on image localization. Moreover, facemask detection algorithms have been adopted and trained by existing object detection algorithms, such as, You Only Look Once (YOLO), Single-Shot Detector (SSD), and convolutional neural networks (CNNs) (Briefly discussed in II). Most facemask detection algorithms proposed after the post-Covid-19 world are mainly deep learning (DL)-based algorithms, which is a sub-set of machine learning (ML) algorithms [14]–[18]. A DL-based networks include deep neural network (DNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). A neural network that uses multiple hidden layers is called DNN and the process can be referred to as DL-based algorithm. The DL-based algorithms have superior feature extraction capabilities than traditional ML-based algorithms. The features which refer to the edges, corners, and textures of an image, need to be extracted precisely using an algorithm during the learning period. DNN extracts those high-level features from an image automatically, whereas ML-based algorithms require human supervision and are mostly handcrafted. This is why most of the facemask detection algorithms, especially the feature extraction part of it, have been conducted using DL-based algorithms. The CNN, which is a class of DNN, has been widely used in facemask classification and detection techniques. As the name suggests, it has a convolutional layer and a pooling layer. The convolutional property of CNN extracts useful features using filters from an image, and the pooling layer reduces the dimensionality. A typical CNN has five to seven layers. However, there are other types of CNN, e.g., MobileNet [19] (28 layers), MobilNetV2 [20] (53 layers), ResNet-50 [21] (50 layers), VGGNet [22] (16 layers), GoogLeNet [23] (19 layers), and so on.

The facemask detection algorithms perform well when the model is trained with DL-based algorithms along with significant amount of data. Previously, a facemask detection model based on the Viola-Jones detector has been introduced to detect masked faces in a medical operating room [24]. This method used colour filter to differentiate between masked face and bare face by portioning faces into two halves, i.e., upper and lower half. Colour filter could successfully detect facemasks but the feature extraction capability is not as effective as DNNs since it has hand-coded features. Because of being lightweight classifiers of CNN family, MobileNetV1 and MobileNetV2 are being immensely used in facemask detection achieving state-of-the-art performance. CNN is a type of DNN where convolution and pooling operation is performed. Also, MobileNetV1 and MobileNetV2 has been solely used as feature extractor along with classifier [14], [25], [26]. In some cases, it has been used with different detectors as well, e.g., YOLO and SSD [18]. However, it requires a bit high computational power and memory. Feature Pyramid Network (FPN) is used in the detection phase of facemask detection process [27], [28].

Another method of successful detection of facemask is using Residual Neural Network (ResNet) which is a very deep CNN. After the outbreak of Covid-19, ResNet and ResNet-50 both are seen being used in facemask detection algorithms. It is observed that DL-based algorithms are best suited in case of object detection earlier [29]. Since many facemask detection algorithms have been adopted from object detection algorithms, like in object detection algorithms, DL-based algorithms have superiority over ML-based algorithms in case of facemask detection too. For classification algorithm, the use of ML-based classifiers, e.g., support vector machines (SVM), decision tree (DT) provide precise performances [30], [31]. To amplify the model's accuracy, data augmentation strategy and use of image-resolution networks have also been initiated in current state-of-the-art of facemask detection [14], [30], so that the resolution does not degrade poorly after the face gets cropped from an image.

Although a significant amount of research has been conducted on facemask detection techniques, a proper review paper that systematically provides both the shortcomings of current research and future scope for guiding future research is still missing. Even, a systematic review article related to diagnosis and prognosis of the Covid-19 using chest radiographs (CXR) and computed tomography (CT) imaging by adopting ML algorithms has been presented in [32].

The authors considered all the published papers and preprints from 1 January 2020 to 3 October 2020 related to using ML to diagnose Covid-19 by using CXR and CT-scans, presented the main weaknesses of the considered literature, and provided several recommendations for future research. However, there are no such review articles regarding facemask detection techniques and their potential in the context of Covid-19. To the best of the authors' knowledge, there is only one review paper on facemask detection techniques in [33], which has significant shortcomings. First, they reviewed only a handful of studies that focused mainly on CNN-based algorithms. Second, they only discussed those works without providing systematic information, such as their shortcomings, considered scopes, and considered datasets. Third, a proper outline for future research scope was absent. On the other hand, some review works related to object detection algorithms summarize the object detection algorithms practiced for decades. These reviews are precise for object-detection algorithms, but it is unclear how facemask detection algorithms work in those detection models. Therefore, a proper comprehensive review related to facemask detection covering all the existing facemask detection algorithms in the context of Covid-19 along with the existing dataset for doing so, has not been presented yet.

This paper focuses on studies conducted in the detection of facemasks and serves as a narrative analysis regarding them in the context of Covid-19. In doing so, one can understand the facemask detection algorithm along with the advantages and disadvantages of the existing algorithms. Researchers can choose the best algorithm among the existing facemask detection techniques according to their application. This study also highlights the scopes and underrated issues in this field and introduces an accurate description of the datasets related to the field. Because there has not been any review covering all of the existing algorithms related to facemask detection, this study incorporates all the works regarding facemask detection algorithms done in the pre-Covid-19 and post-Covid-19 world. The major contributions of the study are discussed below,

- This paper provides a precise concept of the major object detection techniques using traditional ML-based and DL-based algorithms so that it is easier to choose the best-suited algorithm among them and to apply it in case of detecting facemask.
- The study explores the existing literatures related to facemask detection algorithms in the context of the pre-Covid-19 and post-Covid-19 world.
- A description of different types of convenient as well as effective datasets have been reported, which have already been deployed by researchers for implementing facemask detection algorithms.
- A comparison of the performance and features of the existing facemask detection literature have been highlighted so that researchers can find gaps and conduct further research to overcome those gaps.
- Finally, the work discusses the difficulties related to implementing facemask detection studies and future research ideas that will pave the way for further developments in the related fields.

The remainder of the paper has been organized as follows. Section II presents a detailed conceptual discussion regarding the state-of-the-art object detection using conventional ML and DL algorithms. Section III reports the existing remarkable approaches of facemask detection that have been done in the pre and post-Covid-19 world thus far. Section IV is organized by mentioning effective datasets and their descriptions for detecting facemasks and comparing those datasets used in different techniques. Section V summarizes the performance of different approaches mentioned in section III. Section VI reports the difficulties implementing these facemask detection algorithms and future directions in the context of post-Covid world. Finally, Section VII concludes the study.

II. OBJECT DETECTION ALGORITHMS

Various types of object detection algorithms and their performance will be described in this section. As well, the classifications and applications will be presented here. In the next section, we will discuss the facemask detection algorithms in detail.

Object detection is a process that can detect an object from images or video frames using computer vision. It has brought a revolution in computer technology as well as image processing. Machines can easily detect multiple objects in images using an object detection algorithm. The machine can identify which is a human face, a cat, or anything else within nanoseconds. The applications of object detection are increasing continually. As a result, researchers have been researching in this field for almost 20 years. New algorithms are being invented that outperform the previous ones. Object detection combines two major learning algorithms: object localization and image classification. Image classification allocates a class label to an image, whereas image localization finds the location of an object in an image with the help of a bounding box. Face detection is a kind of object detection with high priority from researchers because of its numerous applications. Initially, traditional object detection methods are used for identifying multiple faces in an image, but there are some unsolved issues, such as, it does not consider any occlusion. For this reason, deep learning-based object detection has been developed. This process concentrates on contextual information learning, complex feature learning, and handling the occlusion. It decreases the processing time and increases the accuracy. Recently, a well-researched domain of object detection is facemask detection due to the Covid-19 pandemic. Therefore, it follows the methods adopted from object detection algorithms. Indeed, the facemask detection process is very difficult because feature extraction from a masked face is more complicated than without a masked face, and facial features (e.g., nose, mouth, and chin) are invisible. Most facemask detection algorithms maintain two elementary

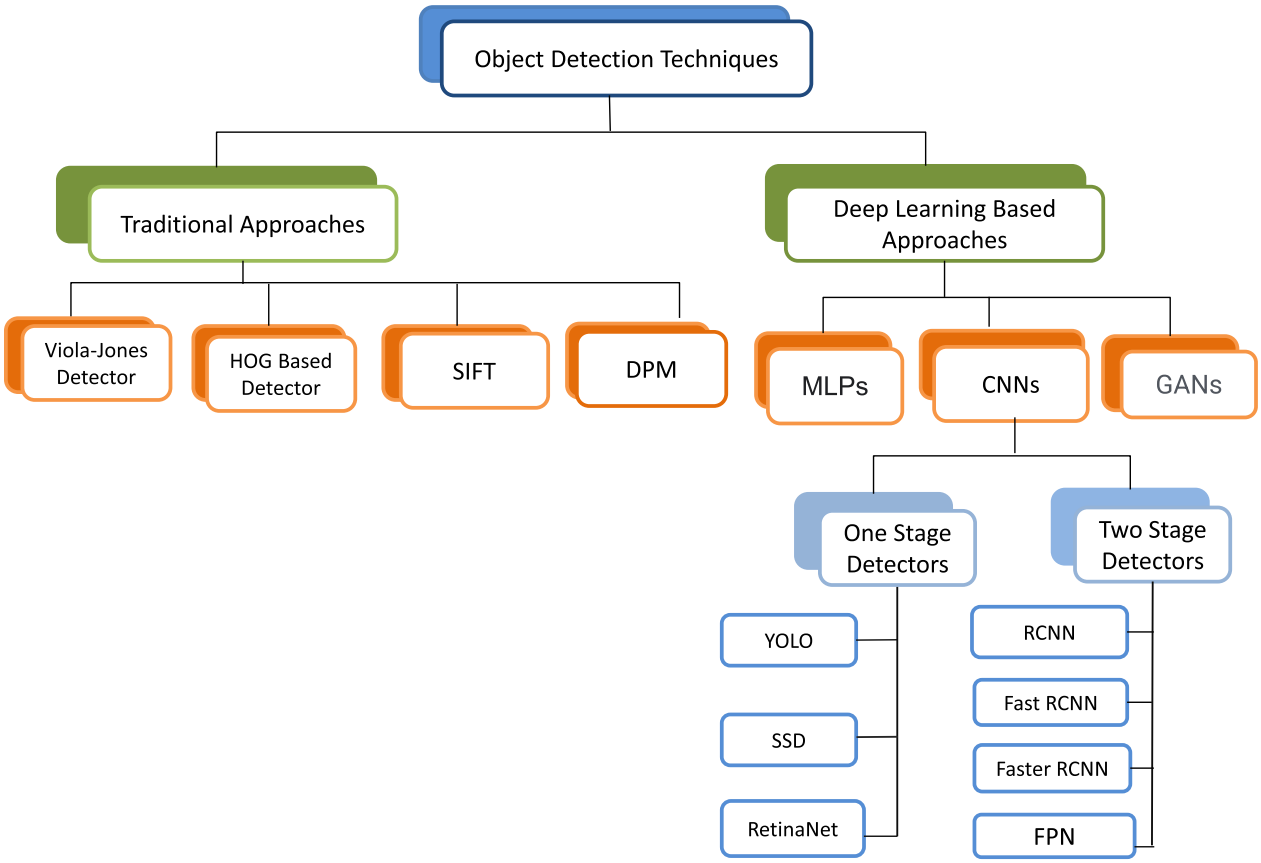


FIGURE 3. A hierarchical representation of object detection algorithms based on the discussion in Section II.

stages: i) face identification and ii) feature extraction. Since CNN-based algorithms are efficient for face detection, they also provide a better result in facemask detection.

Object detection has two major groups of parent algorithms. One is traditional object detection algorithms, and the other is deep learning-based object detection algorithms. Fig. 3 presents the hierarchical representation of object detection algorithms.

A. TRADITIONAL OBJECT DETECTION

1) VIOLA–JONES OBJECT DETECTION

In 2001, Viola and Jones proposed a real-time object detection called the Viola-Jones object detection framework [34], [35]. This process can determine competitive object detection rates in real-time to solve various detection problems. Moreover, it was used primarily in face detection. Although this algorithm was very slow in training, it could detect faces in real-time with impressive speed [36]. The method divided the image into many smaller sub-regions and required checking many different positions and scales because an image had many faces of different sizes. This algorithm has four main steps to perform:

i. **Selecting Haar-like Features**: Vila *et al.* adapted and developed the idea of Haar wavelets from the so-called Haar-like features [34]. Any kind of human face that shares some universal properties like the eye region is gloomier than its neighbor pixels, and the nose region is shinier than the eye region. A simple way is needed to find out which region is light or dark. Therefore, the sum-up of the pixel values of both regions is counted and compared. If one side is lighter than the other, it may be the edge of the eyebrows. Sometimes, the middle part may be shinier than the surrounding boxes, which can be considered a nose.

ii. **Creating An Integral Image**: In the previous step, all the pixel values of a particular feature must be counted; this calculation is quite complicated. The integral of images plays an important role in performing these intensive calculations quickly to understand whether a feature of several features fits the criteria [37]. In an integral image, the value of each point is the summation of all pixels above and to the left, including the target pixel [38]. These integral images save considerable time calculating the summation of all the pixels in a rectangle because only four edges of the rectangle need to be counted.

iii. **Running Adaptive Boosting (AdaBoost) Training**: The number of features present in the 24×24 detector window is nearly 1,60,000, but only a few of these features are essential in identifying a face. Therefore, scientists use the AdaBoost algorithm to identify the best

features of the 1,60,000 features. In the Viola-Jones algorithm, each Haar-like feature represents a weak entity. To determine the type and size of a feature in the final classifier, AdaBoost checks the performance of all classifiers given to it. Those classifiers evaluate all sub-regions of all the images used for training. The classifiers that performed well gain higher importance. The final result is a strong classifier, also called a boosted classifier. The algorithm sets a minimum threshold to determine whether something can be classified as a useful feature or not [39].

iv. **Cascading Classifiers**: Still now, AdaBoost is a time consuming process to calculate these features for each region. The main job of the cascade is to reject non-faces quickly and avoid wasting valuable time and computations. Consequently, it can achieve the necessary speed for real-time face detection. Multiple stages are followed to identify a face. If all classifiers approve the image, it is classified as a human face and presented to the user as a detection. If there is a negative evaluation in the first stage, the image will immediately be rejected because there is no human face. If it passes the first stage but fails the second stage, it is also canceled.

2) HISTOGRAM OF ORIENTED GRADIENTS

A feature descriptor named the Histogram of Oriented Gradients (HOG) is often used to extract features from image data [40]. Dalal and Triggs [41] proposed this descriptor in 2005. The HOG is generally used in computer vision, pattern recognition [42], and image processing to detect and recognize visual objects (e.g., faces, hands). These descriptors are considered powerful for identifying faces with occlusions, posture, and movement changes because they are extracted in a regular grid. The structure or the shape of an object is considered by the HOG descriptor [43]. This method is used to detect the edge direction and determine the gradient (magnitude) and orientation (direction) of the edges. It measures the existing edges of the image and produces individual histograms for each of these edge regions.

3) SCALE INVARIANT FEATURE TRANSFORM

David *et al.* proposed an image descriptor called Scale Invariant Feature Transform (SIFT) in 1999, which can detect image-based matching and recognition [44]. The descriptor can relate to point matching among different views of a 3-D scene and view-based object recognition [45]. In the image domain, the translation, rotation, and scaling transformations can be achieved by SIFT. It can also be used in an uninterrupted and medium perspective conversion and strengthens the illumination variation. This algorithm can be decomposed into four steps

i. Detection of a feature point (also called a keypoint)
ii. Localization of the keypoint
iii. Orientation of an assignment
iv. The generation of feature descriptor

4) DEFORMABLE PART-BASED MODEL

Deformable Part-Based Model (DPM) is an extension of the HOG detector and it can win the detection challenges of VOC-07, -08, and -09. This was first proposed by Felzenszwalb *et al.* [46], after which Girshick *et al.* had made a series of improvements. A root-filter and several part-filters are included in a typical DPM detector. Its main function works manually. Hence, all the configurations of part filters can be learned automatically by a weakly supervised learning method. The DPM follows the rule of ''divide and conquer''. To train the model, the training data can be simply considered as the learning of a proper way. For the ensemble of detection on different object parts, the inference can be considered.

B. DEEP LEARNING BASED OBJECT DETECTION

Previously, shallow learning-based methods were used for face detection. It had been facing challenges in some issues, such as pose variation, facial disguises, lighting of a scene, the complexity of the image background, and changes in facial expression. The methods based on shallow learning use some basic features of images and depend on artificial experience to extract the sample features. The deep learning-based methods can solve those challenges and extract more complicated face features.

A subsection of an artificial neural network (ANN) is a deep neural network (DNN) that consists of the input and output layers. In between, there are some hidden layers. The hidden layers are used to perform the nonlinear transformation of the inputs entered into the network. Hidden layers permit the function of a neural network to be broken down into specific transformations of the data. The DNN is one of the feed-forward networks. In this network, there is no data flow from the input layer to the output layer with looping back. There are many algorithms under deep learning such as, CNNs, Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), Multilayer Perceptrons (MLPs), Self-Organizing Maps (SOMs), etc.

1) CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is one of the DL-based algorithms is that can recognize and classify features in images for computer vision. This is one of the multi-layer neural networks. This network aims to research visual inputs and perform tasks. The CNN can perform image classification, segmentation, and object detection. In contrast to classical image detection, CNN defines image properties itself and takes raw pixel data from the image, trains the model, then automatically extracts the features for more advanced classification. A convoluted layer can apply multiple filters, and each time, a filter is applied across the input. This means that any convoluted level can generate more data. A convoluted level usually follows a pooling level to reduce the dimensionality and noise from the output. It uses predictions from levels to generate a final output representing a vector of probability scores to characterize the probability so that a particular feature belongs to

a particular class. The CNN-based modern object detection algorithms can be divided into two categories: Multi-Stage Detectors and Single-Stage Detectors.

2) TWO STAGE DETECTORS

In a multi-stage detector, the detection process is divided into several stages. A two-stage detector such as RCNN [47], first determines and processes a set of regions of interest (RoIs) using a selective search. Subsequently, the CNN property vectors are extracted individually from each region. Other algorithms based on regional proposal networks, such as Fast RCNN [48] and Faster RCNN [49], have attained greater accuracy and better results than other single-phase identifiers.

- RCNN: Since 2012, many face-related applications, especially in computer vision and pattern recognition, have been using neural networks (e.g,. CNNs) [50]. Girshick *et al.* [47] proposed a region-based CNN (RCNN) for object detection, which brings extraordinary accuracy over CNNs on classification tasks to solve object detection. Its advancement is significant to transfer the supervised pre-trained image representation for image classification to object detection. The RCNN requires a forward pass through the convolutional neural network for each object proposal to extract the features, leading to a heavy computational burden. It requires a multi-stage pipeline, expensive time, and space for training. Two approaches have been recommended to solve this problem: the Fast RCNN and the Faster RCNN. The main drawbacks of RCNN are as follows:
  i. At first, this algorithm does not know how many objects will be in the picture. As the input is of variable length, this makes it difficult to use the CNN.
  ii. There is a dilemma about identifying objects in the image. It is possible to arbitrarily select a few regions and categorize them, but there is a risk of missing important regions. It can check every possible area in the figure, which will take too long to run.
- SPPNet: The SPPNet employs spatial pyramid pooling to overcome the constant size limitations of the network. In this network, an SPP layer is connected above the last compromise level. The SPP pools the layer properties and generates the output of a certain length, which is fed into the fully connected layers [51]. In contrast, the information aggregation function at the deeper level of the network is used at the beginning to avoid the need for cropping or warping. The accuracy of SPPNet is (VOC07 mAP = 59.2%).
- Fast RCNN: One of the object detection model is Fast RCNN that enhances its predecessor over RCNN [48] in various ways. Instead of retrieving CNN properties individually in each RoI, Fast RCNN quickly integrates these into a single forward pass of the image. In Fig. 4(b), this model works jointly to train the framework which increases the computational accuracy. Compared to a previous study, Fast RCNN has made various discoveries to improve the speed of training and test, and increase the detection accuracy. On VOC 2007 test set, Fast RCNN increased the mAP from 58.5% (RCNN) to 66.9.0% (Fast RCNN) [52]. The improvements of Fast RCNN in contrast with RCNN/SPPNet are as given follows:
  i. It has better detection quality (mAP) compare to RCNN, SPPNet;
  ii. It uses single-stage for training with a multi-task loss;
  iii. Fast RCNN training can update all network layers, and it is higher than SPPNet, which only updates the fully connected layer;
  iv. It does not require disk storage for feature caching.
- Faster RCNN: An advanced algorithm called faster RCNN was proposed [49], eliminating a selective search and permits learning of region proposals directly. The Faster RCNN source obtains an image and enters it into a CNN layer called the RPN [53]. It counts a greater number of potential regions than the original algorithm and uses a well-organized DL method to estimate which regions may be of most interest. The predicted region proposals are then reshaped using an (RoI) pooling layer. This layer is used to classify the images within each region and predict the offset values for the bounding boxes. The Faster RCNN can automatically learn a feature representation from data and compare it with other neural-based methods. Fig. 4(c) shows the architecture of Faster RCNN, which is complex because it has several moving parts.
- FPN: The FPN is a type of feature extractor proposed by Lin *et al.* [28]. It is used for feature fusion and detection. It receives a single-scale image as an input, and the outputs are feature maps at multiple levels. The size of the output is as same as the input. Previously, the detectors mostly used top-level feature detection or independent detection in different detection processes, but FPN consists of a bottom-up feature and a top-down feature. The network can merge the feature maps of the same size and detect integrated multi-layer feature layers.

Fig. 4 presents the mechanism of RCNN, Fast RCNN, and Faster RCNN and compares them with each other. The RCNN normally uses an SVM classifier, whereas the Fast RCNN and Faster RCNN use the softmax classifier. The RCNN and Fast RCNN focus mainly on the region proposal, but the Faster RCNN does not consider the region proposal.

3) ONE STAGE DETECTORS

A single-stage detector detects a direct sample by looking at a dense sample of potential locations. These algorithms bypass the stage of the region proposition consumed in multi-stage

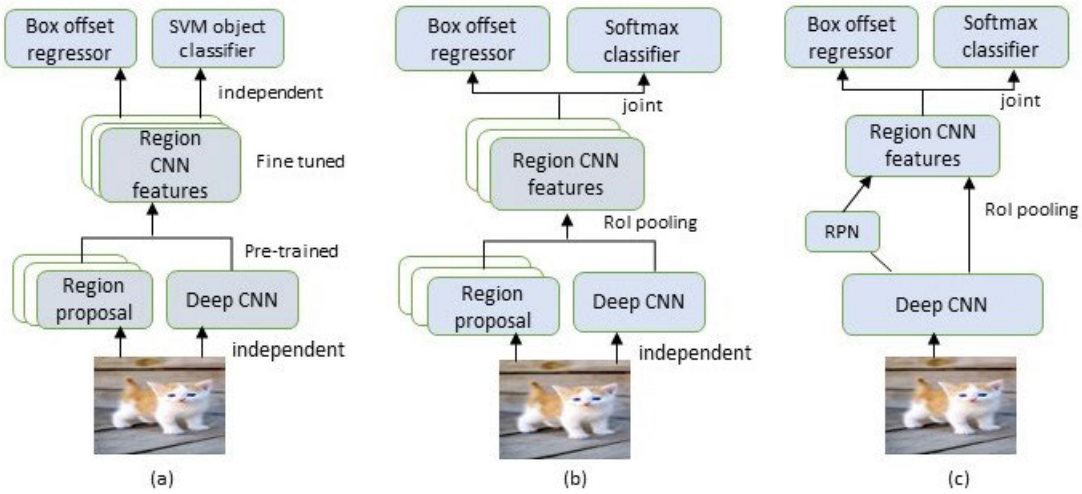


FIGURE 4. Comparison among two stage detectors (a) RCNN, (b) Fast RCNN, and (c) Faster RCNN. The architecture of Faster RCNN is more complicated.

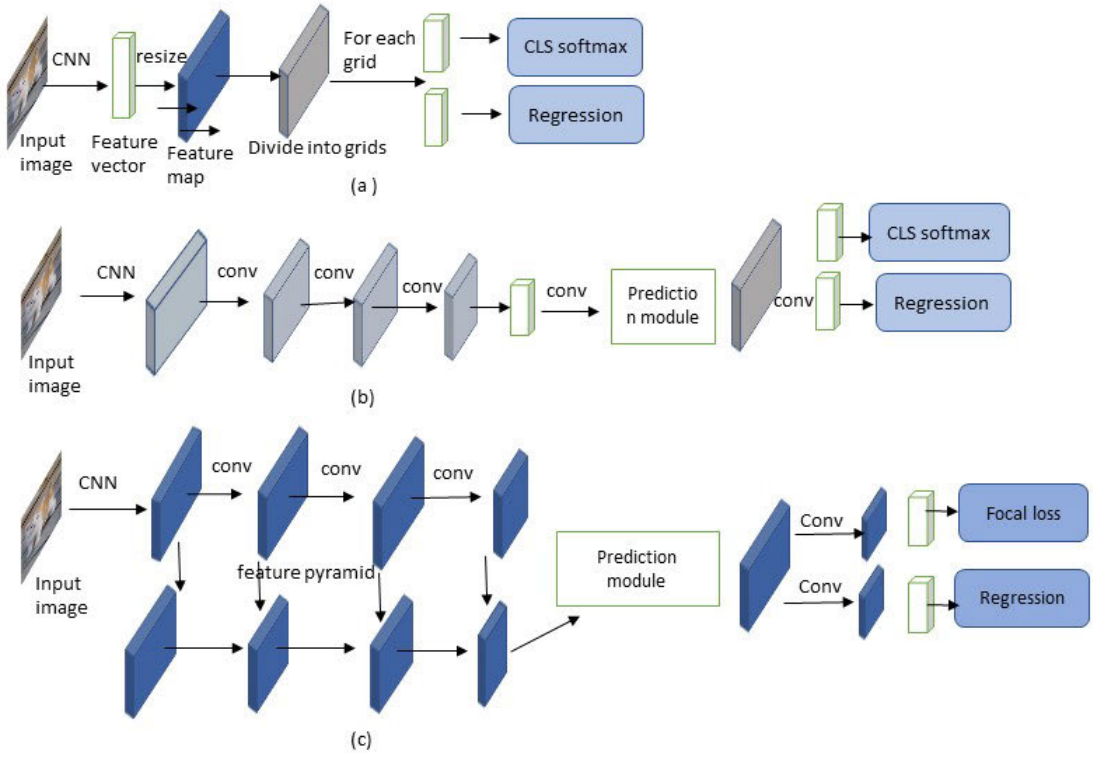


FIGURE 5. The architecture of various type of one stage detectors (a) YOLO, (b) SSD, and (c) RetinaNet.

identifiers. This type of detector is generally considered to be faster at the expense of some loss of accuracy. One of the most popular single-stage algorithms, YOLO [54] achieves close to real-time performance. A SSD [55] is conducted for object detection, which provides outstanding results. RetinaNet is based on Feature Pyramid networks [28] and uses focal loss.

- YOLO: In the DL era, YOLO was the first one-stage detector proposed by Redmon *et al.* [54]. The primary purpose of YOLO is the detection of full images in real-time and webcam. Fig. 5(a) shows that YOLO splits the image into many regions and simultaneously predicts the bounding boxes and probabilities of each region. An old version of YOLO runs at 155 fps with VOC 2007 mAP = 52.7%, and an updated version runs at 45 fps with VOC 2007 mAP = 63.4% and VOC 2012 mAP = 57.9% [29]. Subsequently, Joseph *et al.* made a series of improvements based on YOLO, and its v2 and v3 editions have been proposed [56], [57].
- SSD: Liu *et al.* proposed SSD in 2015, which contributes to the field of multi-reference and multi-resolution

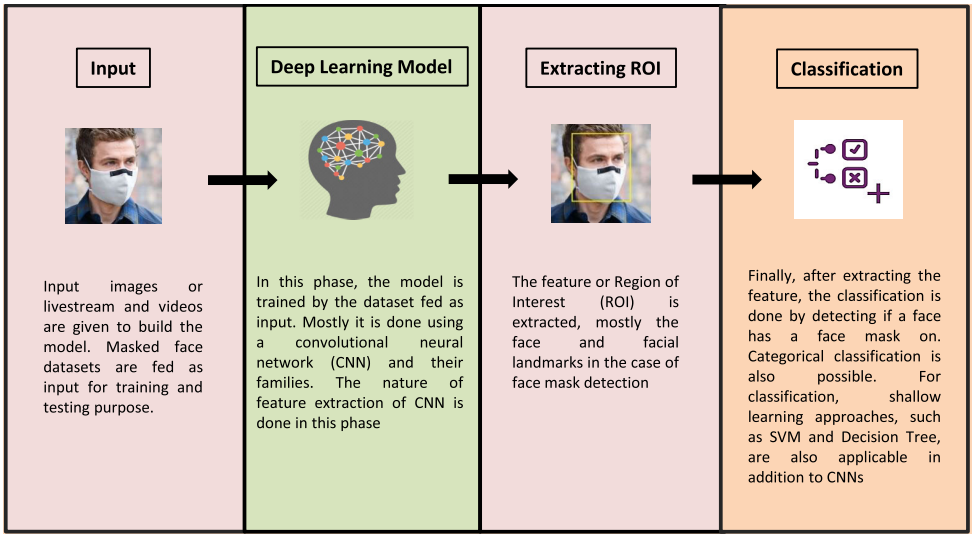


FIGURE 6. General infrastructure for how facemask detection algorithms work for almost all the techniques stated in Section III.

detection techniques. An SDD precisely predicts the category scores and box offsets for a constant set of default bounding boxes of different scales at each location in several feature maps with different scales. It runs at 59 fps with VOC 2007 mAP = 76.8%, VOC 2012 mAP = 74.9% [29]. Fig. 5(b) presents the architecture of SSD.

- RetinaNet: Lin *et al.* proposed RetinaNet in 2018 [58] with focal loss as the classification loss function. He claimed that the intense foreground-background class produces an imbalance during the training of dense detectors. A new loss function called ''focal loss'' was invented in RetinaNet by resizing the standard cross-entropy loss to overcome this imbalance. Detectors will set more focus on hard, misclassified examples during training. Focal loss supports the one-stage detectors to obtain more accuracy over two-stage detectors while maintaining a very fast detection speed.

Fig. 5 shows the basic structure of YOLO, SSD, and RetinaNet. The main difference among them is YOLO and SDD use softmax, but RetinaNet focuses on focal loss. One-stage detectors are used for regression tasks.

III. EXISTING ALGORITHMS FOR FACEMASK DETECTION

In this section, the existing algorithms for facemask detection are reviewed with their features, performances, and shortcomings. Most studies used the CNN architecture because CNN gives excellent performance in extracting features than any other algorithms. Others used shallow ML with/without DL models. Hence, all existing facemask detection algorithms in two groups named are summarized as follows: 1) CNN Based Approaches, and 2) Hybrid Approaches. The posterior class contains algorithms that either used a combination of classical ML-based and DL-based approaches or only shallow ML-based approaches. The reason behind this classification is that most of the studies related to masked face detection or classification algorithms used the DL-based approach, and there is only one study that has detected facemask using solely traditional ML-based approach [24]. Fig. 6 gives a general flowchart that represents how most of the facemask detection model operates. Fig. 7 presents a chronological diagram based on all the facemask detection approaches of different times. In Fig. 11, a hierarchy of the state-of-the-art of facemask detection algorithms is shown based on the algorithms' types and the number of their used classifiers is summarized by the following section.

A. CNN BASED APPROACHES

This sub-section discusses about the algorithms that have been done mostly using CNN. Algorithms inclusive of this approach do not require handcrafted features as ML-based approaches. The CNN and the Multi-Layer Perceptron (MLP) are subsets of DNN. Because MLP is barely used in facemask classification or detection algorithms of facemasks, this sub-section is named after CNN. It has the extraordinary capability of feature extraction by its convolutional and pooling function. Some popular CNN methods, such as MobileNet, ResNet, VGG- 16, NASNet-Mobile, and DenseNet, have been used immensely in existing facemask detection algorithms (See Appendix). The algorithms that used CNN to detect facemasks are discussed in this sub-section.

Shimming *et al.* [16] designed an algorithm that aimed to detect masked faces, where the ''mask'' refers to any occlusion on the face along with physical facemasks. Faces occluded with diverse subjects, such as scarves, hair, hands, and niqab, are considered to be masks covering the face. The types of masks are described in Section IV. A complete dataset called MAFA or Masked Faces, was introduced in their study to detect occluded faces [16]. MAFA contains more than 35,000 masked face images, ensuring at least one part of the face is covered. The dataset also contains faces

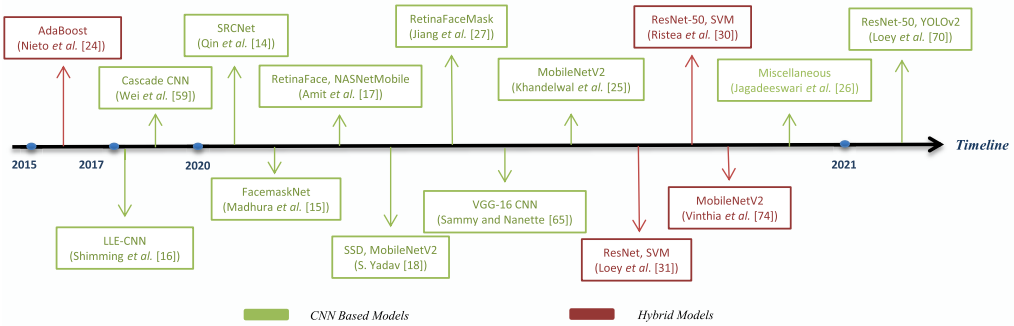


FIGURE 7. Chronological diagram of facemask detection algorithms of a time interval from 2015 to 2021. Almost all of the studies were conducted in 2020, after the post-Covid-19 world in the case of necessity.

with various degrees and orientations. They mentioned six attributes of MAFA. They divided their model into three major parts: (a) a proposal module, (b) an embedded module, and (c) a verification module. The first module combines two CNN and extracts the features from face images. The second module is dedicated to finding the missing facial landmarks that occurred by occlusion. In this phase, the Locally Linear Embedded (LLE) algorithm is used. In the last module, the classification and regression tasks are performed using unified CNN to determine if it is a face or not and scales the position of missing facial cues. They compared their model with six state-of-the-art face detectors and achieved the best performance. Their diverse dataset (MAFA) is one of the main attractions and utilities of their research. Although this mask/occlusion detection algorithm achieves state-of-the-art performance compared to six different algorithms, but the model has some drawbacks to mention in the context of facemask detection. First, detection with a side face orientation affected the performance significantly using this model. Second, the dataset contains occluded faces more than masked faces. So, training with this dataset will not always be feasible for solely mask detection. They calculated the individual precision and took the average precision among the various attributes to measure the performance. The reported average precision was 74.6%.

Wei *et al*. [59] proposed a cascaded CNN architecture for masked face identification, which was comprised of three complete CNN. They also proposed a dataset and called it the ''MASKED FACE Dataset.'' Their dataset contained only 200 images, which are not sufficient to feed into any DL-based framework. DL-based frameworks require a large number of datasets for training to achieve higher performance. To overcome this problem, they used a pre-trained model with the WiderFace dataset [60], and they then tuned it with their dataset. The first, second, and third CNN architectures have the number of five, seven, and seven layers, respectively. The first layer is a very shallow CNN that scales the input image. The classification ability of the CNNs ranges from low to high. Each of the first scales the image and then evaluates the result according to a pre-set threshold. If the evaluated value or probability is less than the given threshold, it is considered a false detection and is discarded. An attribute called the non-maximum suppression is added after the three CNNs to unify the overlapped candidate windows. Here, the use of three cascaded CNNs has an advantage and disadvantage. The three-stage CNN makes the prediction stronger because each false detection is eliminated and is not calculated. The demerit is that it also increases computational complexity. Finally, according to the last CNN, the detection of the masked face is evaluated. The context of this masked face identification was dedicated to preventing terrorist attacks rather than Covid-19. Hence, the dataset they used has a majority of scarves rather than masks. PASCAL VOC was used to evaluate the detection performance. Their model had an accuracy of 86.6% and a recall of 87.8% on their MASKED FACE test dataset.

Qin *et al*. [14] proposed a mask-wearing condition identification algorithm using two networks called image super-resolution (SR) and classification network. They combined and named these two as SRCNet with three classification categories, such as no facemask-wearing (NFW), Incorrect Facemask-Wearing (IFW), and Correct Facemask-Wearing (CFW). The input of the research was 2D images from the real world, so pre-processing of images is necessary to eliminate unwanted information. The dataset used for the experiment was obtained from the literature [61] and applied to MATLAB for proper pre- processing. Among the pictures, only the face was detected using a cascaded neural network [62], which is a robust face detector, and the facial images were then cropped. The cropped images were fed into SR network because the cropping of images causes low resolution. The condition of an image going into the SR network is that if the cropped face image size is less than 150, it is sent to the SR network to achieve a higher resolution. Otherwise, it is sent directly to the CNN for classification. The CNN algorithm used here was adopted from MobileNetV2 [20]. The classifier predicted the image condition and classified them into three classes: NFW, IFW, and CFW. In Fig. 9, the process of classification using SRCNet is given by a pictorial representation. The SRCNet algorithm was well organized and showed better performance. Nevertheless, there were several limitations. The dataset was relatively small and lower in attributes. The detection speed was slower than the other existing algorithms. To evaluate performance,

they used the accuracy, and their algorithm showed 98.70% accuracy.

Madhura *et al*. [15] proposed a model called Facemasknet to identify if a person is wearing a facemask properly or not, which summed up to a three-class classification: no mask, improperly worn mask, and with a mask. A customized dataset including 35 images was used to train the model. Those 35 images were a combination of the masked and unmasked faces. Before training, the input data were pre-processed and resized in the desired value. After pre-processing, the input image or live streams were passed through the Facemasknet model, which detected the face and then extracted the Region of Interest (RoI). They stated their works have two detectors. First, the face is detected. The RoI is extracted, and the Facemasknet model is then applied to those cropped images or live streams for classification. The green and yellow bounding boxes prominently refer to the face and facemask in an image, respectively. It contains very small and region-biased data. Fig. 8 presents the training and detection stage of the model. Facemasknet model reported an accuracy of 98.6%.

Amit *et al*. [17] developed a two-stage-based detector using two pre-trained CNN models. The first stage of the detector completes the face detection in an image, and the next stage classifies the detected images into a mask and no-mask class, as reported elsewhere [15]. The difference between the two studies was that they used two CNNs for face and mask detection. They used various datasets and combined them to produce versatile and geologically bias-free data. In the first stage, the input RGB image was passed through a face detector that could detect faces, even though any scenario of two overlapping faces occurs. The RoI was extracted and passed to the next stage detector, which classified the faces retrieved from the first phase as masked face or unmasked face. The development of a face detector requires a considerable amount of completely trained datasets and a lengthy processing time. Therefore, the authors chose the RetinaFace model [63] as the first stage detector and NASNetMobile [64] as the classifier model for comparison. Between the mentioned phase, there is another intermediary phase that collects the detected faces from stage 1 and batches them, enlarges the bounding box of faces according to height and width, and resizes them according to the requirement to pass through the next stage for classification. The algorithm would allow a further extension, where live video streams could be used as an input. There were some drawbacks of this model. It used two different detectors that created complexity. In addition, the video frame rate was comparatively low. Their model reported a precision, recall, and F1-score for the facemask classifier of 98.28%, 100%, and 99.13%, respectively.

Shashi Yadav [18] proposed a system combining DL and geometric learning, which detects if a person is maintaining social distance and wearing a mask. social distance and wearing a mask. The study was developed in raspberry pi4 for deployment in public live-stream footage using RTSP. The RGB frames were converted to gray-scale for

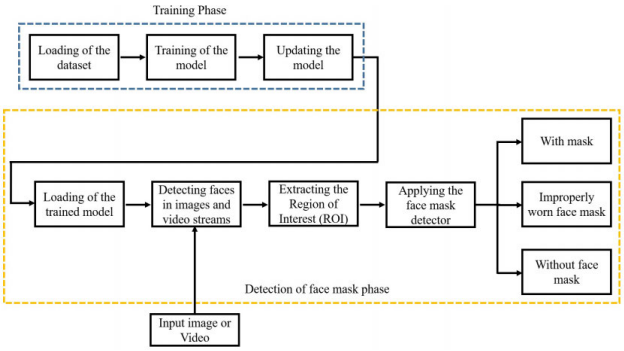


FIGURE 8. Training and detection stage of Facemasknet model introduced in [15].

a lower computational overhead. They used a pre-trained SSD [55] framework with MobileNetV2 [20], which could extract multiple faces from one image by bounding boxes. The study contributes three important criteria in the post-Covid world: (i) detecting people using MobileNetV2 carried out by OpenCV and TensorFlow, (ii) measuring the Euclidean distances between detected people, and (iii) detected whether people have masks on or not. For (iii), they further trained masked and non-masked face images to classify mask and no mask class successfully. The images fed into the classification network were cropped into faces according to the requirements. If a person is not wearing a mask and/or maintaining social distance, the alarm will be turned on to the authority. The detection of social distancing is considered a performance factor. The model achieved a precision of 91.7% with a confidence score of 0.7.

Jiang *et al*. [27] constructed a significant face detector model called RetinaFaceMask, which has the infrastructure of a backbone network, a neck, and head networks. The backbone network refers to the feature extraction segment in DL. RetinaFacemask adopts ResNet as a preliminary backbone, and MobileNet as a backbone for comparison. As a biological neck lies between the back and head, the neck of this framework also implies the strategy. The neck comprises a Feature Pyramid Network (FPN) [28] built inside the CNN for high-level precision. The head refers to the classifier or the detector, where a context attention module has been intrduced to increase the detection performance. In this algorithm, transfer learning (TL) is applied because of the limited dataset. It is a form of DL, where knowledge of a model is transferred from one framework to another for scaling purposes. The RetinaFaceMask model is comprised of a very strong network, which sometimes results in a high-computation overhead.

Sammy and Nanette [65] proposed a simplistic real-time mask detection framework using VGG-16 CNN. Their dataset was labeled and contained 25,000 images to train the model. Although they did not mention the source of their dataset, the number was quite acceptable. The images were pre-processed to avoid unnecessary information like other models, and the segmentation and extraction of the mask-covered area from the face then take place. The classification

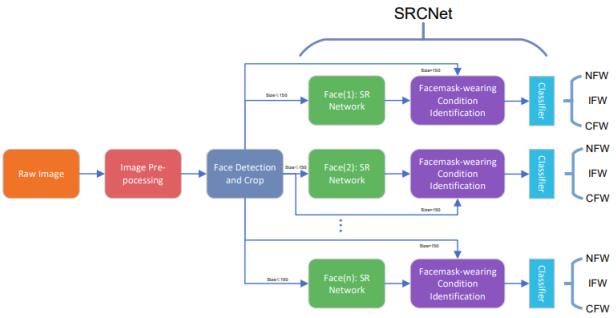


FIGURE 9. Image super-resolution and classification network (SRCNet) proposed by Qin *et al* [14].

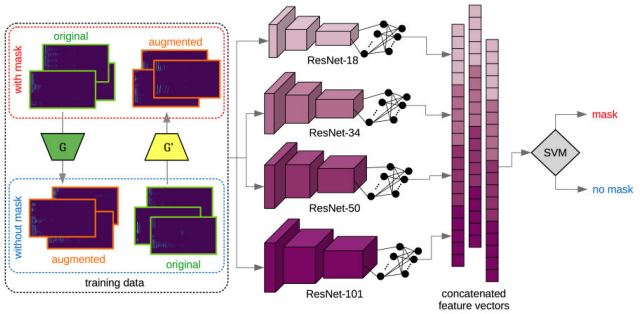


FIGURE 10. Facemask detection from speech by combining SVM and ResNet by cycle-consistent GAN [30].

was done using the VGG-16 CNN model. In the training phase, they used the ADAM optimizer for optimizing the parameters [66]. ADAM is a derived from the term estimation of adaptive moment. Comparisons with other state-of-the-art models are not described here. Their algorithm achieved 96% accuracy in detecting facemasks.

Khandelwal *et al*. [25] proposed a model and deployed it in a real-life application that detects whether a mask is used or not in an image. The work was separated into two steps. One was face detection in the image, and the other was mask classification. They used a pre-trained CNN-based MobileNetV2 model for face detection. A face size less than a certain number of pixels cannot be detected using this model. After detecting faces, the data was prepared to feed into the mask detection stage by cropping and labeling faces using semi-supervised learning [67]. MobileNetV2 was used to build the model. Before feeding into the network, the images were resized in accordance with their requirement. They also used an augmentation strategy to bring diversity to the data. They took a validation set of 840 images combined with a mask and no mask among 4,225 annotated images. This work achieved high performance and was already implemented, but the model had two major drawbacks. First, classification or detection of partially overlapped faces cannot be done using this method. Secondly, this model cannot detect faces where the height of the camera exceeds 10 feet. Their model achieved an Area Under Region of Convergence (AUROC) of 97.6%.

Jagadeeswari *et al* [26] proposed a system that classifies facemasks and used images and video streams as input. Their dataset contained 1300 images. Like other training approaches, they trained their model with MobileNetV2, ResNet50, and VGG-16 CNN. Faces in the images were detected, and the RoI was extracted just like other models. One of the innovative things regarding their work was that they compared three different models by training them in the same training and test sets and set the same epochs. Another is that they used three optimizers (ADAM [66], SGD [68], ADAGRAD [69]) to amplify the performance of their model. In the DNN, an optimizer is a computer algorithm that tunes parameters and attributes in the model to reduce the training loss effectively. This comparison showed how effectively a DL model could work. MobileNetV2 performed best among the three DL models, and in the case of optimizers, ADAM showed the best performance.

Loey *et al*. [70] introduced a novel deep learning-based facemask detection model by implementing YOLOv2 [71] and ResNet-50 [21] together. They paid attention to medical or surgical facemasks. The YOLOv2, an updated version of YOLO, is a feature extracting and classification algorithm, and ResNet-50 is mainly a deep transfer learning-based residual network for feature extraction. They combined facemask Dataset (FMD) [72] and Medical Mask Dataset (MMD) [61] to test and train their model. A data augmentation strategy was used to manipulate the data. They used an estimation of the anchor boxes to improve the model. Two optimizers (SGDM [68] and ADAM [66]) were used to compare their performances. The ADAM optimizer reaches greater accuracy than SGDM. The mean [73] Intersection over Union (IoU) is used to estimate the anchor boxes. This model could not identify masked faces from videos. The average precision reached 81% for their algorithm.

B. HYBRID APPROACHES

Face detection using solely shallow ML is not a feasible solution. Therefore, most of the authors choose to train their model using a DNN, mostly CNN, along with the classical ML approaches (e.g., SVM and DT Algorithm) as classification algorithm. Mostly SVM are used as a classifier along with DL-based algorithms. Owing to the limited number of studies in facemask detection, this sub-section is called the Hybrid Approach. Here, either DL-based and shallow ML-based algorithms are combined, or only shallow ML-based algorithms are used.

Ristea *et al*. [30] developed a method for facemask detection from speech. This model consisted of two parts, i) training Generative Adversarial Networks (GANs) with cycle -consistency loss to transform the unpaired utterances in between two classes (with mask and without mask), and ii) Assigning opposite labels to each transformed pronunciation, producing new training accents using cycle-consistent GANs. The initial and transformed accents were converted to spectra that were used as input to ResNet networks with different depths. The networks were grouped by classifying

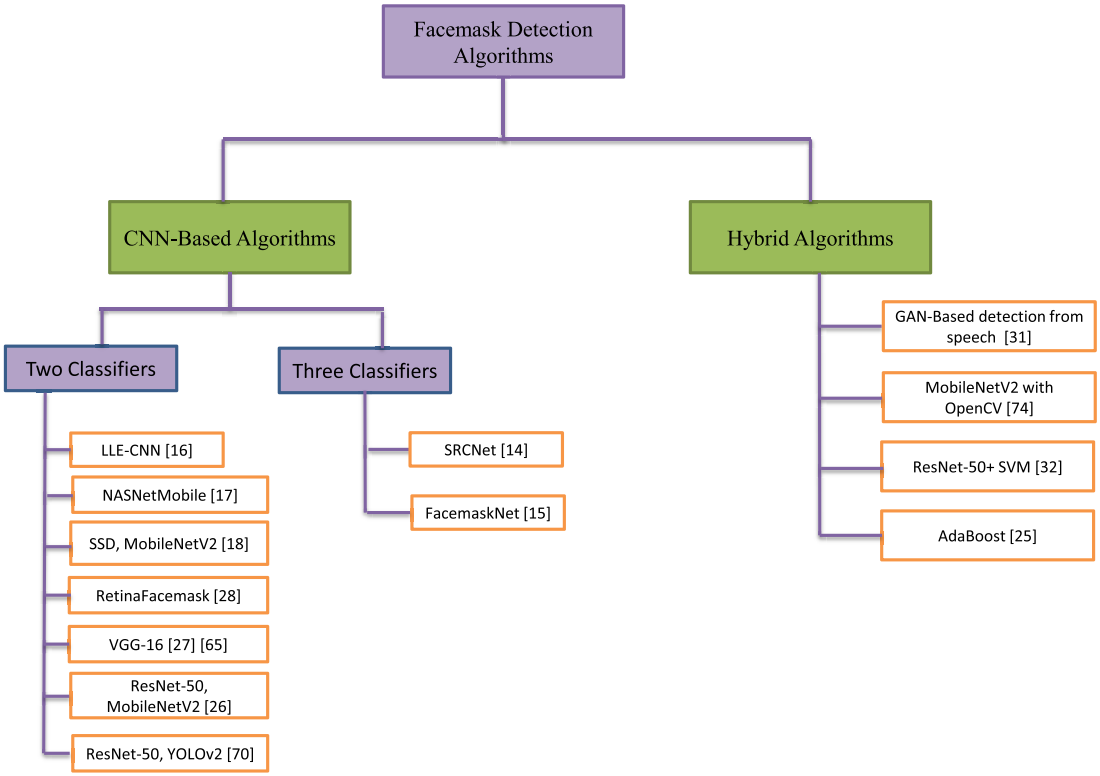


FIGURE 11. A hierarchical representation of current state-of-the-art of all the existing facemask detection algorithms summarized in Section III.

the SVMs. In this process, augmented spectrograms were used to train the model. Training spectrograms were also converted from one class to another class using G and G'. The ResNet was used; the start layer was 18, and the end layer was 101. All the outputs of ResNet were mixed in concatenated feature vectors and were considered as input for the SVM, which predicts the final result. The datasets were provided by the ComParE organizers having 36554 samples. Among them, 10,895 samples were used for training, 14,647 samples for development, and the remaining samples for testing. They reported that their data augmentation method yielded better results than other baseline methods. Fig. 10 presents the infrastructure of the model. The model required high processing time. Therefore, the ratio of the consumption of time and accuracy is the main drawback of this model. They compared their result with and without the augmentation of data and achieved mixed results and an accuracy of 74.6%.

Loey *et al*. [31] proposed a hybrid infrastructure where the feature extraction was done using ResNet-50, and the classification was done using three classifiers (SVM, decision tree, and ensemble classifier). The dataset was divided into three parts: 70% for training, 10% for validation, and 20% for testing. The model was trained, tested, and validated in Real-World Masked Face Dataset (RMFD), and Simulated Masked Face Dataset (SMFD), but tested in Labeled faces in the wild (LFW). After the feature extraction was done using ResNet-50, ML algorithms were used for classification. The decision tree, SVM, and ensemble classifiers were used for the comparison. Among them, SVM provided better performance with low time consumption than the others. In the testing phase, the accuracy was varied according to datasets and classifiers. This model used ML-based algorithm along with DL-based algorithm. They combined three datasets. Although this model performed very well in three of the classifiers, the classical ML classifiers consumed more time with higher accuracy. The ML classifier can be replaced with MobileNetV2 because it is a very lightweight but efficient classifier. The model was tested in three datasets. The model using the RMFD, SMFD, and LFW datasets showed 99.64%, 99.49%, and 100% accuracy in the testing phase, respectively.

Vinitha and Venlantina [74] proposed a model based on computer vision and a DL approach. This model was capable of real-time facemask detection from surveillance cameras and images. They used TensorFlow, OpenCV, Keras, and MobileNetV2 architectures in their model, which was trained on a large dataset. In this model, most of the images were added by OpenCV. Pre-processing of the input images was achieved by resizing them, and they were applied to color filters (RGB) over the channels. The images were scaled using the standard mean of Pytorch build in weights 4.5. Finally, it was converted to tensors (Similar to NumPy array). The model was tested with real-time images and real-time video streams. They did not report their used dataset and performance but they mentioned of using a large dataset.

FIGURE 12. Samples of popular datasets used for the detection of facemasks by different authors; (a) Samples of Maskedface-Net Dataset [75], (b) Samples of Medical Mask Dataset (MMD) [61], (c) Samples of Real-world masked face recognition dataset (RMFRD) dataset [76], (d) Samples of Facemask Detection Dataset (FMDD) [72], and (e) samples of Simulated masked face recognition dataset (SMFRD) dataset inspired by Labeled Faces in the Wild (LFW) [77], [78].

They also reported their performance achieved state-of-the-art results.

Nieto-Rodr′ıguez *et al*. [24] introduced a real-time facemask detection system that triggers an alarm when healthcare staff do not wear surgical masks in the medical or operating room. They used two detectors and two color filters for each detector. One of them was a face detector, and another was a medical mask detector. The face detection was done using the traditional Viola-Jones face detection algorithm. They used a variant AdaBoost called LogitBoost [79] for detecting facemasks. One of the challenges was that any clothing near the mask area could give a false mask detection result because it is based on color filtering. This problem was solved by synthetic rotation [80]. Also, they overcame the problem of missing facial attributes while a person bends or leans for some reason. This was solved by increasing the frame rate to detect the person in the frontal position. The detection model had some shortcomings. First, this model was only trained by surgical masks. One kind of conventional object detection algorithm is AdaBoost. It performs well but not as much as DNNs. Moreover, the use of a color filter for object detection limits the accuracy of a model compared to all existing DL algorithms. Their reported recall was above 95%, and the false-positive rate was below 5%.

Table 1 summarizes all the methods of the literature in facemask detection for convenience with a different timeline based on the different approaches discussed in Section III. From this table, it is seen that in the year of 2020 and after, the highest number of researches have been done related to facemask detection.

IV. DATASET

This section states a comprehensive description of some popular datasets used in facemask detection techniques mentioned in Section III. This section also focuses on the characteristics, compositions, quantities and limitations of those datasets.

One of the major requirements for implementing the DL algorithms is that the number of training sets should be plentiful. After the burst outbreak of Covid-19, voluminous datasets have been required to perform mask detection tasks using DL-based algorithms. Although there is a significant number of face datasets for face detection and recognition, the masked face data is comparatively small.

TABLE 1. Summary of used methodology in different facemask detection algorithms.

	Ref.	Year	Summary of Used Method
CNN Based Approaches	[16]	2017	A novel dataset and face occlusion detection technique using LLE-CNN and its three separate modules
	[59]		Three CNN classifiers were combined to detect a masked face and named as cascade framework of CNN
	[14]	2020	Image Super-Resolution network (SR) and a classification network were combined to identify facemask-wearing conditions. Images with low resolution improved with an SR network, and MobileNetV2 was used for the classification network
	[15]		An 8 layer CNN was used to detect faces, extract RoI and detect facemasks. The detector was applied both on input images and video streams.
	[17]		A two-stage detection algorithm comprised of face detection and facemask detection. It first detected a face (RetinaFace) and then detected a facemask (NASNetMobile)
	[18]		A real-time social distance maintaining and a facemask detector was deployed in raspberry pi4 using a combination of SSD and MobileNetV2. Violating instructions resulted in alarm
	[27]		A facemask detector consisted of CNNs, FPN, and a context attention module. The models used were MobileNet and ResNet
	[65]		A simplistic CNN (VGG-16) model was used to classify the facemask deployed in raspberry pi4
	[25]		This work detected social distancing and masks on the face and was deployed in a commercial area using MobileNetV2
	[26]		A comparison of MobileNetV2, ResNet-50 and VGG-16 CNN classifier on same dataset with three different optimizers
	[70]	2021	A facemask classification algorithm with the help of YOLOv2 and ResNet-50 along with a novel dataset
Hybrid Approaches	[24]	2015	This method used colour filters for classification of face and facemask in an operating room with the help of skin texture in HSV color space
	[30]	2020	Detects facemask from speech using data augmentation technique with multiple ResNet that have been trained by GANs
	[31]		Facemask classification algorithm whose feature extraction was done by ResNet-50 and classification is done using SVM, DT and ensemble classifiers
	[74]		Simplistic real-time facemask detection algorithm that is done by MobileNetV2 and PyTorch

As soon as the mask detection models gained attention after the post-Covid-19 world, the number of masked face datasets increased. The datasets that are explicitly being used in the facemask detection process are described here. Table 2 summarizes the datasets described in this section along with their sources, characteristics, compositions and limitations. Hence, it is easier to interpret the shortcomings and utility of a particular dataset. Fig. 12 refers to a pictorial representation of the mostly used datasets for facemask detection, which have been frequently used in the detection process of facemask. In the scarcity of diverse datasets initially, most of the researchers chose to build and train a model by using combined datasets rather than using a single one. One major reason behind this is, some datasets lead to bias decisions in terms of facial orientation, texture, or skin color. Some models are trained with a very limited customized dataset of the authors and tested on those as well. Table 3 provides an overview of the different datasets they used in their existing facemask detection procedures along with the models' performances.

A. MASKED FACE DETECTION DATASET (MFDD)

This dataset is one of the most widely used datasets that contain 24,771 masked face images [77]. It will be helpful to train models for masked face detection classification. The dataset is a mixture of some sample projects done in China [84]. Previously, they used MAFA [16], and some samples were taken from the Internet. The number of images in the dataset is appreciable and should give good accuracy in masked face detection. A drawback of this dataset is that it is biased toward Chinese faces.

B. REAL-WORLD MASKED FACE RECOGNITION DATASET/REAL-WORLD MASKED FACE DATASET (RMFRD/RMFD)

The dataset includes 5,000 pictures of 525 people wearing facemasks and 90,000 images of the same 525 people without masks [76]. The dataset contains 1000 classes. The source of this dataset was mostly Internet images of different people. The dataset is complete, diverse, and extensive. Because the dataset has samples of the same people wearing and not wearing facemasks, this can be useful for masked face recognition purposes. The dataset is also biased toward Chinese images.

C. MASKED FACES (MAFA)

This dataset is where the mask is a facemask and any occlusion found in the mouth area. The construction of their dataset

TABLE 2. General overview of the major datasets applicable for facemask detection process.

Ref.	Dataset	No. of Images	Composition	Characteristic	Limitations
[77]	MFDD	24,471	Solely masked face images	Public dataset	Biased to Chinese Face and not so sufficient in number
[76]	RMFRD	95,000	Masked and unmasked face images of same subject	Effective in accuracy since the dataset is very large	Biased toward Asian face images
[16]	MAFA	30,811	Contains masked face and any sort of occlusion on face	Categorical classification is easily deployable since mask-type is specified	Mostly preferable for occlusion detection rather than physical mask detection
[77]	SMFRD	5,00,000	Masked and unmasked face image of same person	Diverse and versatile	Training phase takes more memory and time
[75]	MaskedFace-Net	1,37,016	Contains improperly worn masked face data along with masked faces	Benchmark dataset, categorical classification is easier than with other datasets.	Biased toward surgical masks
[78]	LFW	13,233	Composed of celebrity images of different orientation	Benchmark dataset for face recognition	Does not contain any masked face image
[72]	FMDD	853	Contains only masked face images	Publicly available	Very limited in number

Legends: Ref.– Reference; MFDD– Masked Face Detection Dataset; RMFRD– Real-world masked face recognition dataset; MAFA– Masked Faces; SMFRD– Simulated masked face recognition dataset; LFW– Labeled Faces in the Wild; FMDD– Facemask Detection Dataset.

contains four types of masks. The type ''Simple'' contains a plain and pure mask with any color; type ''Complex'' contains a facemask with textures, logos, or designs; the type ''Human Body'' contains occlusion by hand or hair; the type ''Hybrid'' contains a combination of any of the two mentioned above. The dataset consists of 30,811 internet images and 35,806 masked face images [16]. The dataset has a diverse orientation of faces rather than the front face, which is one uniqueness of this dataset.

D. SIMULATED MASKED FACE RECOGNITION DATASET (SMFRD)

Because the number of masked face images was much less than any face image, Wang *et al.* [77] proposed a dataset, where a facemask is worn on an individual face with the help of software. The initiative was taken to increase the diversity in the masked face dataset. The images in the dataset were taken from both LFW [78] and Webface [85] dataset, which mostly contains celebrity faces. The dataset has 5,00,000 face images with 10,000 subjects.

E. MASKEDFACE-NET

Proper wearing of a mask is necessary. Along with mask detection on the face, the detection of correctly worn masks is also essential. Owing to insufficient data of incorrectly worn masks, the MaskedFace-Net [75] was developed to fill the gap. The database contains 1,37,016 images of properly and improperly worn masks based on the Flickr Faces HQ (FFHQ) [81] dataset. The dataset contains properly and improperly worn masked faces along with no masked faces. The improperly worn masked face has three classifications: (i) Uncovered chin, (ii) Uncovered nose, and (iii) Uncovered nose and mouth, which covers 80%, 10%, and 10% of the data, respectively. This is the largest manual masked face dataset thus far.

F. LABELED FACES IN THE WILD (LFW)

This dataset contains more than 13,000 images of faces of 5,749 people from the Internet [78]. The noticeable criterion of this dataset is the dataset is labeled with the people's names, which can also be used for face detection and recognition. The dataset comprised solely barefaced images with different popular people and was used to produce the SMFRD dataset.

G. LARXEL'S FACEMASK DETECTION DATASET (FMDD)

This dataset contained 853 images belonging to three classes [72]. The dataset might be smaller in size, but a significant number of works have been carried out using this dataset.

Moreover, there are many more effective and insightful dataset for facemask detection including masked and unmasked faces, namely Medical Mask Detection (MMD) [61], Facemask Detection (FMD) [82], Flickr Faces

TABLE 3. Dataset used by existing facemask detection literatures and their performances.

Lit.	Dataset	Description	Total No. of Images	Performance in mentioned literature
Shimming *et al.* [16]	MAFA [16]	Dataset introduced by author himself	35,806	Average precision 74.6%
Wei *et al.* [59]	1.WiderFace (Pre-trained) [60] 2.MASKED FACES [59]	Trained with WiderFace and fine tuned with the latter	200	Accuracy is 86.6% and recall is 87.8%
Qin *et al.* [14]	MMD [61]	Medical Mask Dataset	3,835	Accuracy is 98.70%
Madhura *et al.* [15]	Customized	Dataset introduced by author himself	35	Accuracy is 98.6%
Amit *et al.* [17]	1.RMFRD [77] 2.FMDD [72] 3.FFHQ [81]	Combined three dataset to avoid bias and scarcity	7,855	Precision, recall, and F1-score for the facemask classifier of 98.28%, 100%, and 99.13%, respectively
Shashi Yadav [18]	Customized dataset	Author claimed to use customized dataset	3,165	Precision of 91.7% with a confidence score of 0.7
Jiyang *et al.* [27]	FMD [82]	Facemask Dataset combined by WiderFace [60] and MAFA	7,959	-
Sammy and Nanette [65]	Not mentioned		25,000	96% accuracy
Khandelwal *et al.* [25]	Customized	Customized images of factory workers	4,225	97.6% accuracy
Jagadeeswari *et al.* [26]	Not mentioned		1,300	Accuracy is 99.8%
Loey *et al.* [70]	1.FMDD [72] 2.MMD [61]	Combination of facemask Dataset and Medical Mask Dataset	1,415	Average Precision is 81%
Ristea *et al.* [30]	Mask Augsburg Speech Corpus (MASC)	Voice recordings of 32 German speakers. Each samples contain one second, and sampling rate is 16Khz	36,554 (audio data)	74.6% accuracy
Nieto-Rodr´ıguez *et al.* [24]	1.BAO [83] 2.LFW [78]	For training phase LFW, For testing BAO	99 images with 496 faces	Recall above 95%, False positive rate below 5%
Loey *et al.* [31]	1.RMFRD [77] 2.SMFD [76] 3.LFW [78]	LFW used for testing accuracy only	Not specified	RMFD, SMFD, and LFW datasets showed 99.64%, 99.49% and 100% testing accuracy respectively
Vinthia *et al.* [74]	Not specified	Not specified	Not specified	Achieves state-of-the-art performance

Legends: Lit.– Literature; MAFA– Masked Faces; MMD– Medical Mask Dataset; RMFRD– Real-world Masked Face Recognition Dataset; FMDD– Facemask Detection Dataset; FFHQ– Flickr Faces HQ; FMD– Facemask Detection; MASC– Mask Augsburg Speech Corpus; SMFD– Simulated Masked Face Dataset ; LFW– Labeled Faces in the Wild; Khz– Kilohertz; no.– number.

HQ (FFHQ) [81], and Widerface [60]. Since, the dataset for facemask detection have not been adequate and versatile yet, the best practice is to use different datasets to increase the diversity and length and to avoid bias. In Table 3, it is seen that many works have been done by combining more than one datasets, even by creating images by their own. Descriptions of these datasets have also been given in the table. Overall, Table 3 provides an overview of the used

datasets in existing facemasks detection procedures along with their performances for comparison.

V. PERFORMANCE ANALYSIS

The following section is divided into two sub-sections. The first sub-section defines the primary performance metrics used the most in the literatures. The latter gives a comprehensive discussion on the performances, shortcomings, and possible ways of improvement for these algorithms.

A. MAJOR PERFORMANCE METRICS USED IN FACEMASK DETECTION

Different authors chose different performance metrics to evaluate the performance of their models. Although most of the studies were evaluated using multiple metrics, some used single metrics to evaluate their performance too. This sub-section acknowledges how they differ from one another.

According to different evaluation metrics used by different authors, a column graph is shown in Fig. 13 to visualize the most used and least used performance metrics. According to Fig. 13, because the accuracy is the most used performance metric to evaluate most of the models', a column graph has been added where the accuracy of different algorithms along with their used datasets have been presented for visualization in Fig. 14.

Now, we will discuss some of the popular performance metrics used to evaluate the facemask detection algorithms as follows:

1) TRUE POSITIVE, TP

In the case of mask detection in an image, a TP value also refers to a *detection* value, where there is a facemask detected on the model, and the facemask exists on an image.

2) FALSE POSITIVE, FP

This is the case when a false detection occurs by the model. That is, the model is detecting a mask, but there is no mask.

3) TRUE NEGATIVE, TN

A true negative value is more likely to describe the non-object regions and detect them as a non-object region. This value is barely used in the detection of masks as it does not affect the performance.

4) FALSE NEGATIVE, FN

A false negative refers to when the model misses a facemask on an image to detect.

5) ACCURACY

Accuracy in a facemask detection model is defined as the ratio of correctly predicted objects over all possible predictions. Eq.1. defines the accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

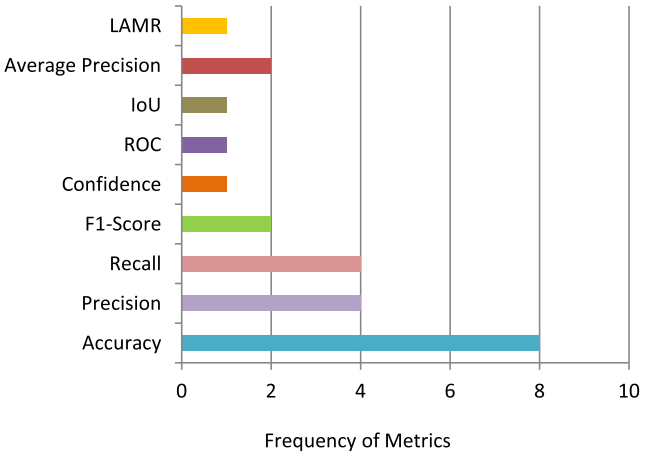


FIGURE 13. Frequency of using a particular performance metrics reported in all of the existing facemask detection algorithms described in Section III. The horizontal axis shows the frequency, and the vertical axis shows different performance metrics. Legends: LAMR– Log Average Miss Rate; IoU– Intersection over Union; ROC– Region of Convergence.

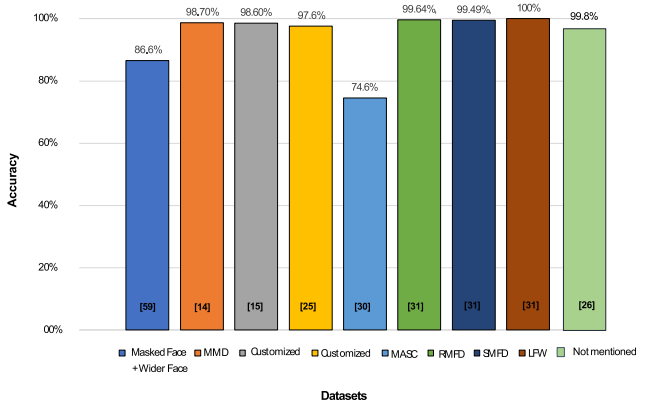


FIGURE 14. Performance measure of accuracy against different facemask detection algorithms with their corresponding datasets according to Table 3.

6) PRECISION

Precision describes the probability of predicting a bounding box on a mask that matches the ground truth box. This is a measure of facemasks that the model correctly detects among all the actual facemasks in the image. Mathematically, it is the ratio of TP and all positives defined in Eq. 2.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

7) RECALL

The recall is the ability of a detector to find and detect all possible facemasks or all possible true values correctly. Mathematically, it is the ratio of TP and the sum of TP and FN, which is defined in Eq. 3.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

8) F1-SCORE

This is defined as the harmonic mean of precision and recall because it uses both of them. A mathematical equation of

TABLE 4. Reported performance of existing facemask detection methods summarized in Section III.

Ref.	Year	Performance Metric	Reported Performance
[24]	2015	Recall, False Positive	Rec. was above 95%, FP was below 5%
[16]	2017	Average Precision	AP was 76.4%
[59]		Accuracy, Recall, IoU	Acc. was 86.6%, Rec. was 87.8%
[14]	2020	Accuracy	Acc. was 98.70%
[15]		Accuracy	Max Acc. 98.7%, Min Acc. 74.97%
[17]		Precision, Recall, F1-Score	Prec. was 98.28, Rec. was 100 and Conf. was 99.13
[18]		Confidence, Precision, Recall	Prec. score of 91.7% with Conf. score 0.7
[27]		Precision, Recall	-
[65]		Accuracy	Acc. was 96%
[25]		AUROC	97.6%
[30]		Accuracy	74.6 %
[31]		Recall, Precision, Testing Accuracy, F1- Score	Acc. was 99.64%, 99.49%, and 100% for three different dataset
[74]		Accuracy	Not Specified
[26]		Train and Test Loss, Accuracy	-
[70]	2021	Average Precision, Log Average Miss Rate	AP was 81%, LAMR was 0.4

Legends: Ref.– Reference; Rec.– Recall; FP– False Positive; AP– Average Precision; Acc.– Accuracy; Prec.– Precision; Conf.– Confidence; LAMR– Log Average Miss Rate; AUROC– area under region of convergence.

recall is expressed in Eq. 4.

$$\text{F1-Score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{4}$$

9) CONFIDENCE

Contextually, it is a measure of how much an algorithm or a model is confident to be in either a ''mask'' class or a ''no mask'' class.

10) INTERSECTION OVER UNION, IoU

This is a measure of a threshold used if the model classifies a detection class as a True Positive or a False Positive. Mathematically, it is the ratio of the area of overlap and the area of union between bounding boxes, as expressed in Eq. 5.

$$\text{IoU} = \frac{\text{Area\_of\_overlapping}}{\text{Area\_of\_Union}} \tag{5}$$

B. PERFORMANCE ANALYSIS OF DIFFERENT LITERATURE

The reported performances stated by different authors for detecting facemasks are discussed in this sub-section. Table 4 lists the reported performance stated by the existing facemask detection algorithms summarized in Section III along with their performance metrics they used. An overview of the existing literature following their performance and shortcomings according to different timeline is also summarized in this sub-section.

Before 2020, three studies related to facemask detection were conducted (Fig. 7). In [24], a facemask classification algorithm is introduced, which was trained and tested by AdaBoost, a conventional ML-based algorithm. In the real-time detection of facemasks, conventional ML algorithms would not be sufficient for robust performance. The performance metrics are also inadequate for justification. A DL-based algorithm could effectively increase the performance of a model. The algorithm might be feasible in an operating room but not for state-of-the-art facemask detection. The performance of the algorithm does not achieve state-of-the-art performance. Later in 2017, the authors proposed a novel dataset with an occlusion detection algorithm [16] on human faces. They divided their model with various attributes. The model could detect the degree of face, mask type, and occlusion type. The model achieved a very high performance compared to the other state-of-the-art performances. Among six other models, their models performed the best. Still, the overall AP was relatively lower because of the diversely oriented face dataset. Furthermore, they considered 12 attributes in case of detecting an occluded face. In 2017, another cascaded CNN-based model was reported, where the dataset was mainly made of scarves worn to cover faces in terrorist attacks. Their overall masked face detection accuracy was unsatisfactory. The model showed medium performance.

After 2020, several studies have been done providing tremendous performances. For example, SRCNet [14], Facemasknet [15], and RetinaFacemask [27] achieved very high performance with additional networks. The SRCNet model

used an image SR network that improves the accuracy of the model because low-resolution objects are challenging to detect. Also, RetinaFacemask achieved state-of-the-art results and high performance because they used a TL approach and a backbone, head, and neck network. The same author proposed two methods [31], [70]. In the first work, the model achieved very high accuracy; they performed only classification but no detection. Later they proposed an algorithm with proven performance comparing with two optimizers (ADAM and SGDM).

Two comparison studies were done separately in [26] and in [17] based on the same training and test set. The multi-stage CNN network proposed in [17] achieved high accuracy and could be improved if the two stages were combined. A comparison with three popular CNN models showed that RetinaFace and NASNetMobile were good face and mask identifiers, respectively. Similarly, in [26] they compared three CNN classifiers by training them on the same dataset. A previous study [17] proposed an unbiased dataset, whereas the number of datasets in the Facemasknet model was very small. The dataset could be adopted from the previous one to achieve high performance for Facemasknet.

The models proposed in [18], [25] used social distance monitoring. Both of them proposed a model where it first detects a person, identifies if they are maintaining social distancing, and then detects masks. Between them, S. Yadav measured the performance for both the social distance monitoring algorithm and facemask detection algorithm. If the facemask detection could be evaluated separately, the actual performance could be measured more precisely. Both models achieved high performance. In [65], they used VGG-16 CNN and achieved high performance.

Table 5 summarizes the discussions, merits, and shortcomings in the existing literature. Moreover, Table 6 summarizes the technical or experimental differences among different approaches in facemask detection. The number of classifiers used, different parameters, i.e., learning rate (LR), epochs, and optimizers (OPT) have been mentioned here. Also, the number of convolutional layers (if any) used by different models has been stated. Furthermore, the train and test ratio (in percentage or number of images) of the datasets mentioned in those algorithms have been listed. Note that, in Table 6, we only listed those approaches whose technical details (mentioned above) were mentioned in their corresponding literature. Besides the listed approaches in Table 6, the algorithm used in [31] has stated their software and hardware specifications in their literature; and the train, validation, and test ratio as 70:10:20. Since ResNet-50 was used as a feature extractor in [31], it can be assumed that it has 50 convolutional layers [21]. Moreover, in [24], as the used ML-based classification algorithm was AdaBoost, it does not own any convolution layer.

According to the discussions stated above, some general conclusive comments can be listed as follows:

- The performance of facemask detection algorithms conducted after 2020 showed a significant performance compared to the detection models before 2020.
- The characteristics of the dataset have a considerable impact on the performance of any model.
- Among the CNN-based approaches, the most widely used models are MobileNetV2, ResNet, and VGG-16 CNN (See Table 5). However, MobileNet, ResNet, and their families show better performance than VGG-16 CNN. Being a lightweight classifier, MobileNetV2 performed the best.
- Although accuracy has been used in many existing approaches to evaluate the performance of a model, it is not always a good evaluation metric because it considers all the negative sample values. The precision might be a better evaluation measure in this case (See Eq. 1, 2).
- The use of multiple CNNs increase the complexity of a detection model.
- The use of different optimizers, data augmentation strategies, and image resolution networks improve the classification performance of a model significantly.

VI. DIFFICULTIES IN FACEMASK DETECTION AND FUTURE DIRECTIONS

This section presents some major challenges and difficulties related to facemask detection of existing literature to provide a path for researchers for determining where the attention should be given.

A. INADEQUACY OF BENCHMARK DATASET

Although there are effective and efficient datasets for facemask detection, there are still some limitations. For example, there are datasets with masked and non-masked faces but no improperly worn mask images. Even if they do, the number is very limited. If a dataset has a proper or improper masked face, they are limited to cases of bare faces. On the other hand, the dataset that contains all three is lacking in number. Some datasets are biased toward a particular region, hence testing for neutral data would not be effective. Consequently, the model cannot be trained and appropriately classified. Therefore, a proper dataset is needed that is vast in number, diverse in orientation, and unbiased to a particular region. Furthermore, a dataset must include different ages of people with different structures and textures with various facemasks. A benchmark dataset should contain faces of different degrees so that side faces or angled faces can also be identified for classification or detection of facemasks. Moreover, because Covid-19 is a recent phenomenon, there is a less number of datasets of masked faces for proper training of a model, this often leads to difficulties in the training and implementation of the facemask detection algorithms. In this case, the simulated masked face dataset-making approach can help because the number of face datasets is larger than the number of masked face datasets.

TABLE 5. An overview of facemask detection algorithms reviewed in Section III related to used models, datasets, performances, and scopes for further research.

Method		Ref.	Used Dataset	Classification Type	Merits	Shortcomings
CNN Based Approach	SRCNet	[14]	MMD	Categorical	1. Achieves higher accuracy using a lightweight classifier 2. SR network increases the resolution of image	1. Relatively small dataset 2. Lack of identification on video streams 3. Detection speed is comparatively high
	Facemasknet	[15]	Customized	Categorical	Provides real-time mask wearing condition detection	Contains very small and region biased data
	LLE-CNN	[16]	MAFA	Binary	1. Robust classification of occlusion and its type 2. Diverse and large dataset	1. Side face orientation affects the model's performance 2. Detects any occlusion regardless of facemasks
	NASNetMobile	[17]	RMFRD, FMDD, FFHQ	Binary	1. Biased free and large dataset 2.Comparison of various models	1. Lower video frame rate 2. Complexity of two different detectors
	SSD, MobileNetV2	[18]	Not Mentioned	Binary	1. Real-time implementation with alarm system 2. Social distancing detection 3. Lightweight hardware and classifier	Performance is evaluated with the detection of social distancing.
	RetinaFaceMask	[27]	FMD	Binary	ResNet outperforms in face and mask detection.	ResNet requires high computation
	VGG-16 CNN	[65]	Not Mentioned	Binary	Real-time implementation with alarm systems.	No comparison with other state-of-the art models
	MobileNetV2	[25]	Customized	Binary	Achieves high performance	1. Unable to classify partially overlapped faces 2. Constraint of equipment
	ResNet-50, YOLOv2	[70]	MMD, FMDD	Binary	1. Optimizers are used to increase performance 2. Data augmentation increases quality of dataset	1. Relatively small dataset 2. Lack of video streams
	Cascaded CNN	[59]	MASKED FACES Dataset	Binary	Use of three CNN avoids false detection as much as possible	Use of three cascaded CNN is might bring complexity and easily replaceable to a robust classifier
	ResNet-50, MobileNet, VGG-16	[26]	Customized	Binary	1. Optimizers used in there stimulates each model's performance 2. MobileNetV2 performs the best	1. Relatively small dataset 2. Performance was evaluated based on accuracy only
Hybrid Approach	ResNet, SVM	[30]	MASC	Binary	1. Comparison with multiple ResNets 2. Use of data augmentation has superiority in performance	Processing time is high
	ResNet, SVM	[31]	RMFD, SMFD, LFW	Binary	1. Combines shallow ML and deep learning 2. Compares classification with three individual classifier 3. Comparatively large Dataset	Classification is done by shallow ML approach
	MobileNetV2	[74]	Not mentioned	Binary	Simplicity in real-time deployment	No mentioned dataset and performance to compare
	AdaBoost	[24]	BAO, LFW	Binary	Includes detection of rotated or leaning face	1. Only trained to detect surgical masks in operation room 2. Use of color filter does not provide high-performance always

Legends: Ref.– Reference; SRCNet– Super Resolution Network; LLE-CNN– Locally Linear Embedded Convolutional Neural Network; YOLOv2– You Only Look Once (version 2); CNN– Convolutional Neural Network; SVM– Support Vector Machine; AdaBoost– adaptive programming boost; MMD– Medical Mask Dataset; MAFA– Masked Faces; RMFRD– Real-world Masked Face Recognition Dataset; FMDD– Face Mask Detection Dataset; FFHQ– Flickr Faces HQ; FMD– Facemask Detection; MASC– Mask Augsburg Speech Corpus; LFW– Labeled Faces in the Wild; ML– Machine Learning.

B. VERSATILE TYPES OF MASK

In the existing dataset, the most common types of mask were plain and medical or surgical masks. Covid-19 could not stop people from being commercial and come up with different categories of masks. Later, different types of designed masks entered the market, which has a significantly smaller contribution to the datasets. In addition, facemasks with skin color are sometimes challenging to detect if the model is not well trained. Most studies were done on the classification of masked faces, non-masked faces, and improper wearing of

TABLE 6. Technical differences among different approaches for facemask detection algorithms mentioned in Section III.

Ref.	Model Used	No. of classifiers	No. of layers	No. of Images	Parameter			Train test validation ratio/number
					epochs	LR	OPT	
[16]	LLE-CNN	2	-	35,806	-	-	-	25,876 and 4,935
[59]	Cascaded (3) CNN	2	5,7,7	200	-	-	-	160 and 40
[14]	SRCNet	3	53	3,835	50	10^{-4}	ADAM	90:10
[15]	FacemaskNet	3	8	35	20	$1 \times e^{-4}$	-	-
[17]	NASNetMobile	2	6	7,855	50	0.001	-	80:10:10 (Train, validation, test)
[18]	MobileNetV2+SSD	2	53	3,165	20	$1 \times e^{-4}$	ADAM	80:20
[27]	RetinaFaceMask	2	6	7,959	250	10^{-4}	SGD	4,906, 1,226,1,839 (Train, validation, test)
[65]	VGG-16 CNN	2	16	25,000	100	0.001	ADAM	80:20
[74]	MobileNetV2	2	53	4,225	40	0.1	ADAM	80:20
[25]	MobileNetV2	2	53	1300	20	-	ADAM	90:10
[70]	YOLOv2+ ResNet50	2	24	1,415	60	0.001	SGDM, ADAM	70:20:10
[31]	ResNet+SVM	2	18-101	1,415	100	10^{-4}	ADAM	-

Legends: Ref.– Reference; No.– number; LR– learning rate; OPT– Optimizer; ADAM– adaptive moment estimation; SGDM–stochastic gradient descent with momentum.

facemasks. More classifications could be developed as the type of masks, e.g., N-95, surgical or fabric masks. Because N-95 masks are the most secure among them, the categorical classification of masks would highlight the need for organizations that require higher safety. More datasets will be needed to classify the mask type. In the case of classifying the mask type, it is a difficulty that needs to be considered.

C. MASK WEARING CONDITION DETECTION

In detecting facemasks, the classification as ''mask'' and ''no mask'' is inadequate. If the detector can identify a mask on a face and cannot detect if it is correctly worn or not, although the facemask detection process might be completed, the result would be of no use. This is because the motive of detecting facemasks is to bound the people to wear facemasks so that the contagious spread of coronavirus is reduced. And wearing a facemask requires wearing it properly, i.e., the facemask should cover the entire face and its orientation should be proper too. However, till now, only a few studies have detected the mask-wearing condition. There is a massive gap in this research section. Moreover, a robust detector should perform the categorical classification of masks, such as mask-wearing condition classification and mask type classification.

D. PROCESSING SPEED AND EQUIPMENT

To deploy in real-time, the most important factor to consider is how fast the model can be trained, tested, and provide an accurate result. Although the existing algorithms give excellent performance, most did not state the processing time. In real-time detection techniques, good cameras are needed to monitor areas so the model can be well trained with good quality pictures. Some good cameras that capture a high-resolution image or video frame might be expensive. Computing resources should be less complex and effective so that they could be deployed in public areas. In accordance with less expensive and lightweight equipment, maintaining good performance is a significantly difficult task for the facemask detection techniques.

E. VARIATION OF IMAGE RESOLUTION

Image resolution is one of the main factors in facemask detection. Most existing algorithms work with 224 $\times$ 224 pixels to 240 $\times$ 240 pixels resolution. They take these resolution images as input. In the pre-processing phase of the image, the images are resized to prepare them for the next step. When a single masked face image or multi- faces in an image is detected as a human face, it is cropped to feed into a mask detector. This results in a decrease in image resolution, resulting in lower quality when the face landmarks are cropped. If these low-resolution images are used to generate high-resolution images, these high-resolution images will improve the classification performance. Therefore, different types of image resolution affect the overall performance. Sometimes the accuracy of a model also depends on the image resolution. Hence, maintaining the resolution while classifying the facemask is an issue that needs to be considered in this field.

F. INCREASING PERFORMANCE ON DIFFERENT FACE ORIENTATION

A masked face data is not enough if it is taken only from the front face degree. It is difficult to detect a face and facemask if the side profile of a person is shown. A model should be trained with faces with various degrees and orientations so that the facemask detector can detect face with the least facial landmarks. More images with side profiles of faces need to be inserted into the dataset. If another face slightly covers a face, face detection becomes quite difficult. Therefore, overlapping of two faces in facemask detection is a challenging issue.

G. MASKED FACE RECONSTRUCTION

An inevitable problem in wearing a facemask is that it covers half of the human face, which is a constraint in face detection in various systems and organizations. Because face detection and recognition have tremendous applications, face recognition is difficult when a person wears a mask on their face. The most important regions of the face are covered with masks, which is not feasible for face recognition. Furthermore, a masked face compromises security because a system cannot recognize a person or authorize an individual. Thus, recognizing masked faces by reconstructing the whole face from the available region could be another important research sector. During this Covid-19 pandemic, safety and security are both issues for further research. Wearing facemasks is mandatory for safety, but it could compromise security. Some

studies on masked face recognition or reconstruction have also been done using computer vision and DL, but the number is minimal. Therefore, masked face reconstruction is a very challenging as well as difficult issue in case of masked face detection.

H. REAL-TIME FACEMASK DETECTION

Real-time facemask detection should be able to verify if a person is wearing a facemask from a video stream. But, real-time facemask detector with fast inference and high accuracy is a challenge in this field. Though the described methods are helpful for facemask detection, they are not enough to successfully implement a real-time facemask detection system. Moreover, there are several challenging issues which should be solved for the implementation of a real-time facemask detection system. We have listed some of the key challenges below:

- Lack of adequate facemask dataset: To train our machine for mask detection in real-time, we need a rich dataset both in images and in video streams.
- Memory issue: In the case of real-time implementation, we have to collaborate with webcam camera, CCTV and other tools. Advanced and light-weight machines and camera sensors is also required for successfully detecting facemasks in real-time.
- Different types of mask: People wear different types of facemasks with different colors and sizes. Some wear double masks for additional safety. Therefore, the training algorithms in real-time should be robust enough to detect all these variations.

VII. CONCLUSION

This paper presented a narrative and meta-analytic review covering all the existing facemask detection algorithms, considering the context of Covid-19. The procedure of the existing algorithms, their considerations, effectiveness, evaluation process, and outcomes were presented. Moreover, the datasets used in those algorithms were discussed briefly. The shortcomings of the existing algorithms were reviewed, and the future challenges were outlined. Although a significant amount of research has been focused on developing an efficient facemask detection algorithm, they mainly concentrated on the same set of problems neglecting some other significant issues. This paper highlighted those shortcomings, such as, maintaining image-resolution during detection process, scarcity of rich dataset, categorical classifications, and others. Also it specified the future scopes which includes diversity in datasets and facemask types, different facemask wearing conditions, reconstruction of the masked face, and so on. This comprehensive review will pave the way for the research community to understand the current facemask detection algorithms. By analyzing the shortcomings and future challenges in this field, researchers will develop novel approaches to fill those gaps.

In future work, we tend to gather upcoming facemask detection algorithms and apply those algorithms on a benchmark dataset to compare their performances fairly. By doing so, it will be easier to deduce where those algorithms differ in terms of running time, space and accuracy based on a common benchmark dataset.

NOMENCLATURE

AdaBoost	adaptive boosting/adaptive programming boost
AI	artificial intelligence
ANN	artificial neural network
AUROC	area under region of convergence
CFW	correct facemask-wearing
CNN	convolutional neural network
Covid-19	coronavirus disease-2019
CT	computed tomography
CXR	chest radiographs
DL	deep learning
DNN	deep neural network
DT	decision tree
Eq	eqnarray
FN	false negative
FP	false positive
GAN	Generative Adversarial Network
HSV	hue,saturation,value
IFW	incorrect facemask-wearing
IoU	intersection over union
LAMR	log average miss-rate
MASC	mask Augsburg speech corpus
MERS-CoV	middle east respiratory syndrome
ML	machine learning
MLP	multi-layer perceptron
NFW	no facemask-wearing
PPE	personal protective equipment
RCNN	region based convolutional neural network
ResNet	Residual Neural Network
RNN	Recurrent Neural Network
RoI	region of interest
RPN	region proposal network
RTSP	real-time streaming protocol
SARS-CoV	severe acute respiratory syndrome
SGDM	stochastic gradient descent with momentum
SVM	support vector machine
TL	transfer learning
TN	true negative
TP	true positive
VGA	video graphics array
WHO	world health organization

APPENDIX.

- MobileNetV2: MobileNetV2 is one kind of convolutional neural network consisting of a fully convolution layer with 32 filters. It performs well on mobile devices.
- ResNet: Residual neural network (ResNet) is a special type of ANN having 100 to 1000 layers according to application.

- VGG-16: VGG-16 is a convolutional neural network architecture that contains 16 layers deep
- DenseNet: DenseNet is a type of CNN that conducts dense connections between layers through Dense Blocks.
- NASNetMobile: NasNetMobile is another kind of convolutional neural network that can be on more than a million images. This network can classify images into 1000 object categories.
- EPOCHS: An epochs defines training the network with all the training data for only one cycle.
- Learning Rate: Learning rate is a parameter which is used in the training of neural network containing a small positive value,range [0.0 - 1.0].
- Optimizer: Optimizer is one kind of algorithm used to change the attributes of neural network such as weights and learning rate in order to reduce the losses.

REFERENCES

[1] *WHO Coronavirus (COVID-19) Dashboard*. Accessed: Jun. 2, 2021. [Online]. Available: https://covid19.who.int/
[2] S. K. P. Lau, P. C. Y. Woo, K. S. M. Li, Y. Huang, H.-W. Tsoi, B. H. L. Wong, S. S. Y. Wong, S.-Y. Leung, K.-H. Chan, and K.-Y. Yuen, "Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 39, pp. 14040–14045, Sep. 2005.
[3] A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, and R. A. Fouchier, "Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia," *New England J. Med.*, vol. 367, no. 19, pp. 1814–1820, 2012.
[4] T. Velavan and C. Meyer, "The COVID-19 epidemic," *Tropical Med. Int. Health*, vol. 25, no. 3, p. 278, 2020.
[5] *Declaration of COVID-19 to be a Pandemic*. Accessed: Jan. 25, 2021. [Online]. Available: https://en.wikipedia.org/wiki/COVID-19_pandemic
[6] B. J. Cowling, K.-H. Chan, V. J. Fang, C. K. Y. Cheng, R. O. P. Fung, W. Wai, J. Sin, W. H. Seto, R. Yung, D. W. Chu, and B. C. Chiu, "Facemasks and hand hygiene to prevent influenza transmission in households: A cluster randomized trial," *Ann. Internal Med.*, vol. 151, no. 7, pp. 437–446, 2009.
[7] S. M. Tracht, S. Y. Del Valle, and J. M. Hyman, "Mathematical modeling of the effectiveness of facemasks in reducing the spread of novel influenza A (H1N1)," *PLoS ONE*, vol. 5, no. 2, 2010, Art. no. e9018.
[8] T. Jefferson, R. Foxlee, C. D. Mar, L. Dooley, E. Ferroni, B. Hewak, A. Prabhala, S. Nair, and A. Rivetti, "Physical interventions to interrupt or reduce the spread of respiratory viruses: Systematic review," *BMJ*, vol. 336, no. 7635, pp. 77–80, Jan. 2008.
[9] *Coronavirus Disease (COVID-19): Vaccines*. Accessed: Jun. 2, 2021. [Online]. Available: https://www.who.int/news-room/q-a-detail/coronavirus-disease-(covid-19)-vaccines
[10] N. Leung, D. Chu, E. Shiu, K.-H. Chan, J. Mcdevitt, B. Hau, H.-L. Yen, Y. Li, D. Ip, J. S. Peiris, W.-H. Seto, G. Leung, D. Milton, and B. Cowling, "Respiratory virus shedding in exhaled breath and efficacy of face masks," *Nature Med.*, vol. 26, pp. 676–680, May 2020.
[11] J. Howard *et al.*, "An evidence review of face masks against COVID-19," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 4, 2021. [Online]. Available: https://www.pnas.org/content/118/4/e2014564118 and https://www.pnas.org/content/118/4/e2014564118.full.pdf
[12] *Coronavirus Disease (COVID-19) Advice for the Public*. Accessed: Jun. 16, 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public
[13] (2020). *How Mask Antiviral Coatings May Limit COVID-19 Transmission*. Accessed: Jun. 17, 2021. [Online]. Available: https://www.optometrytimes.com/view/how-mask-antiviral-coatings-may-limit-covid-19-transmission
[14] B. Qin and D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19," *Sensors*, vol. 20, no. 18, p. 5236, Sep. 2020.
[15] M. Inamdar and N. Mehendale, "Real-time face mask identification using facemasknet deep learning network," India, Jul. 2020. [Online]. Available: https://ssrn.com/abstract=3663305
[16] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2682–2690.
[17] A. Chavda, J. Dsouza, S. Badgujar, and A. Damani, "Multi-stage CNN architecture for face mask detection," 2020, *arXiv:2009.07627*. [Online]. Available: http://arxiv.org/abs/2009.07627
[18] S. Yadav, "Deep learning based safe social distancing and face mask detection in public areas for COVID-19 safety guidelines adherence," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 7, pp. 1368–1375, Jul. 2020.
[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[24] A. Nieto-Rodríguez, M. Mucientes, and V. M. Brea, "System for medical mask detection in the operating room through facial attributes," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.* Spain: Springer, 2015, pp. 138–145.
[25] P. Khandelwal, A. Khandelwal, S. Agarwal, D. Thomas, N. Xavier, and A. Raghuraman, "Using computer vision to enhance safety of workforce in manufacturing in a post COVID world," 2020, *arXiv:2005.05287*. [Online]. Available: http://arxiv.org/abs/2005.05287
[26] M. T. C. Jagadeeswari, "Performance evaluation of intelligent face mask detection system with various deep learning classifiers," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 11, pp. 3083–3087, May 2020. [Online]. Available: http://sersc.org/journals/index.php/IJAST/article/view/23805
[27] M. Jiang, X. Fan, and H. Yan, "RetinaMask: A face mask detector," 2020, *arXiv:2005.03950*. [Online]. Available: http://arxiv.org/abs/2005.03950
[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[29] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055
[30] N.-C. Ristea and R. T. Ionescu, "Are you wearing a mask? Improving mask detection from speech using augmentation by cycle-consistent GANs," 2020, *arXiv:2006.10147*. [Online]. Available: http://arxiv.org/abs/2006.10147
[31] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, Jan. 2021, Art. no. 108288.
[32] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, and J. R. Weir-McCall, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Mach. Intell.*, vol. 3, no. 3, pp. 199–217, 2021.
[33] J. B. K. Adithya, "A review on face mask detection using convolutional neural network," *Int. Res. J. Eng. Technol.*, vol. 7, pp. 1302–1304, Apr. 2020.
[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. 1.
[35] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[36] A. D. Egorov, "Algorithm for optimization of Viola–Jones object detection framework parameters," *J. Phys., Conf.*, vol. 945, Jan. 2018, Art. no. 012032.

[37] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 555–562.

[38] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[39] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, 1999.

[40] L. Cerna, G. Cámara-Chávez, and D. Menotti, "Face detection: Histogram of oriented gradients and bag of feature method," in *Proc. Int. Conf. Image Process., Comput. Vis., Pattern Recognit. (IPCV)*, 2013, p. 1.

[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[42] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognit. Lett.*, vol. 32, no. 12, pp. 1598–1603, 2011.

[43] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[44] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[46] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[47] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[48] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497

[50] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face R-CNN," 2017, *arXiv:1706.01061*. [Online]. Available: http://arxiv.org/abs/1706.01061

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[52] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[53] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 650–657.

[54] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* USA: Springer, 2016, pp. 21–37.

[56] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[57] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[59] W. Bu, J. Xiao, C. Zhou, M. Yang, and C. Peng, "A cascade framework for masked face detection," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst. (CIS), IEEE Conf. Robot., Autom. Mechatronics (RAM)*, Nov. 2017, pp. 458–462.

[60] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[61] *Medical Mask Dataset*. Accessed: Jun. 17, 2021. [Online]. Available: https://www.kaggle.com/shreyashwaghe/medical-mask-dataset

[62] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[63] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.

[64] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[65] S. V. Militante and N. V. Dionisio, "Real-time facemask recognition with alarm system using deep learning," in *Proc. 11th IEEE Control Syst. Graduate Res. Colloq. (ICSGRC)*, Aug. 2020, pp. 106–110.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, pp. 1–15, Dec. 2015.

[67] X. J. Zhu. (2005). *Semi-Supervised Learning Literature Survey*. [Online]. Available: http://digital.library.wisc.edu/1793/60444

[68] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

[69] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 1–39, 2011. [Online]. Available: https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf

[70] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection," *Sustain. Cities Soc.*, vol. 65, Feb. 2021, Art. no. 102600. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670720308179

[71] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6517–6525.

[72] *Larxel's Face Mask Detection Dataset*. Accessed: Jun. 17, 2021. [Online]. Available: https://www.kaggle.com/andrewmvd/face-mask-detection

[73] M. Barthakur and K. K. Sarma, "Semantic segmentation using K-means clustering and deep learning in satellite image," in *Proc. 2nd Int. Conf. Innov. Electron., Signal Process. Commun. (IESC)*, Mar. 2019, pp. 192–196.

[74] V. V. V. Vinitha, "COVID-19 facemask detection with deep learning and computer vision," *Int. Res. J. Eng. Technol.*, vol. 7, no. 8, pp. 3127–3132, 2020.

[75] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "MaskedFace-Net—A dataset of correctly/incorrectly masked face images in the context of COVID-19," *Smart Health*, vol. 19, Mar. 2021, Art. no. 100144. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352648320300362

[76] *Real World Masked Face Recognition Dataset*. Accessed: Jun. 17, 2021. [Online]. Available: https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset

[77] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," 2020, *arXiv:2003.09093*. [Online]. Available: http://arxiv.org/abs/2003.09093

[78] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.* Marseille, France: Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Oct. 2008, pp. 1–15. [Online]. Available: https://hal.inria.fr/inria-00321923

[79] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2000.

[80] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *Proc. Joint Pattern Recognit. Symp.* USA: Springer, 2003, pp. 297–304.

[81] *Flickr-Faces-HQ Dataset (FFHQ)*. Accessed: Jul. 22, 2021. [Online]. Available: https://github.com/NVlabs/ffhq-datase

[82] D. Chiang. (2020). *Detect Faces and Determine Whether People Are Wearing Mask*. Accessed: Jun. 17, 2021. [Online]. Available: https://github.com/AIZOOTech/FaceMaskDetection

[83] R. Frischholz. *Face Detection & Recognition Homepage*. Accessed: Jul. 22, 2021. [Online]. Available: https://facedetection.com/datasets/

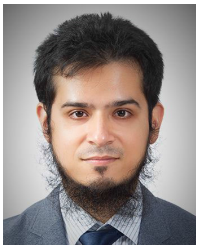
[84] *Open Source Face Mask Detection Data, Model, Code, Online Web Experience, All Open Source*. Accessed: Jun. 17, 2021. [Online]. Available: https://zhuanlan.zhihu.com/p/107719641?utm_source=com.yinxiang

[85] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

AFSANA NOWRIN was born in Bangladesh, in 1997. She received the B.Sc. degree in information and communication engineering (ICE) from the Bangladesh University of Professionals (BUP), in 2019, where she is currently pursuing the M.Sc. degree in ICE. She was a former general member of IEEE, from October 2017 to 2019. She is conducting her research based on object detection and neural networks. Her research interests include big data, the IoT, machine learning, and neural networks, along with WSNs and their routing protocols.

SHARMIN AFROZ was born in Dhaka, Bangladesh, in 1998. She received the B.Sc. degree in information and communication engineering (ICE) from the Bangladesh University of Professionals (BUP), Dhaka, in 2019, where she is currently pursuing the master's degree in ICE. She is currently working on facemask prediction, masked face reconstruction, and AI for masked face recognition. Her research interests include the application of big data, AI, and ML.

MD. SAZZADUR RAHMAN received the B.Sc. and M.S. degrees in applied physics, electronics, and communication engineering from the University of Dhaka, Dhaka, Bangladesh, in 2005 and 2006, respectively, and the Ph.D. degree in material science from Kyushu University, Japan, in 2015. He has worked as a Faculty Member with the Faculty of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, from May 2009 to November 2018. Since November 2018, he has been serving with the Institute of Information Technology, Jahangirnagar University, Dhaka, where he is currently working as an Associate Professor. His research interests include material science for computer application, surface science, ubiquitous computing, WSNs, machine learning, and the IoT.



IMTIAZ MAHMUD received the B.Sc. degree in telecommunication and electronic engineering from Hajee Mohammad Danesh Science & Technology University, in 2011, and the M.S. and Ph.D. degrees in electronics engineering from Kyungpook National University, in 2015 and 2019, respectively. He is currently working as a Postdoctoral Research Fellow with the Telecommunication and Network Laboratory, Kyungpook National University. His current research interests include artificial intelligence, multipath-TCP, TCP congestion control algorithms, and UAVs path planning and network constraints.

YOU-ZE CHO (Senior Member, IEEE) received the B.S. degree in electronics engineering from Seoul National University, South Korea, in 1982, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, in 1985 and 1988, respectively. Since 1989, he has been with Kyungpook National University (KNU), South Korea, where he is currently a Professor with the School of Electronics Engineering. He was a visiting researcher in Toronto, Canada. From March 2002 to February 2003, he worked at the National Institute of Standards and Technology, USA, as a Guest Researcher. In 2017, he served as the President of Korean Institute of Communications and Information Sciences (KICS). His research interests include mobility management and traffic engineering for wireless and mobile networks and transport protocols for the future internet.