Comparison of Deep Neural Network Models of Face Mask Detection in Multi-Angle Head Pose

P. Sreevani
Department of CSE
GNITS, Hyderabad
Telangana, India.

P. Sunitha Devi
Department of CSE
GNITS, Hyderabad
Telangana, India.

Abstract—**Wearing a Face Mask in public areas has become mandatory to all people in this present Covid-19 pandemic situation. As more number of people gather or visit public places like Supermarkets, Shopping malls, Office etc for their daily activities, air borne disease is more likely to spread from one person to another very fastly. So, wearing face mask not only helps oneself but also it protects others from spread of disease. Recognition of face is a popular and significant technology in recent years. Previously, detecting a face of a person and his face expression has been done. But identifying whether a person is wearing a face mask or not and that too in various angles like frontal face, side angle face is a present challenging task. The proposed system uses face mask dataset which consists of images like person with mask and person without mask. Input images are pre-processed, detection of face take place and given for deep neural networks for training, finally it classifies whether a person is without or with mask. The proposed system uses Convolutional Neural Network which is compared with different networks like MobileNetV2 and Vgg16. In real-time, proposed system is implemented by taking input from webcam which helps in detecting whether a person is wearing a face mask or not wearing a face mask.**

Keywords—Convolutional Neural Networks, Vgg16, MobileNetV2.

I. INTRODUCTION

Face-Mask is a protective covering for face. It covers nose and mouth parts of human face. Generally while inhalation, a face mask is used to filter pollution in air and prevents the entry of airborne virus, bacteria. In present pandemic situation, COVID-19 is a contagious disease which spreads from one person to another person either through physical contact or by air. So far there is no medicine or vaccine is found to this pandemic disease.

Wearing face mask helps in preventing the spread of disease up to a certain point. The Ministry of Health and Family Welfare, Government of India has issued some instructions and precautions to be taken by all Indian citizens like wearing a face mask, regular hand wash, using sanitizer, maintaining a social distance in public areas and eating healthy Vitamin food etc. Many State governments are imposing One thousand rupee fine for those people who are not wearing a mask. Recently lockdown is removed and all public services are reopened and made available to people. Now the challenging task is to identify people who are wearing a face mask and not wearing a face mask in public areas.

Face Detection is one of the trending & latest technologies used to identify human faces in the images which are digital. In Face Detection, the object-class detection takes place and the aim is to identify all different size objects present in the image and classify those objects to the respective classes. Face Detection, Face Mask Detection, etc can be done using the Image Processing Techniques. Image processing is a technique of performing some operations on an image and extracting useful information from it. The steps involved in it are importing the image, performing operations, analyzing the input image and finally in the output the resultant altered image is displayed.

Day-by-day the data transfer and data storage is rapidly increasing - managing, storing and analyzing the data takes a long time. So for this purpose latest technology is used to perform this work in less time. This data can be in different forms like text, image, video formats. Machines can perform number of tasks in less time and produce the accurate results. Machine learning algorithms help in training the machine. Technologies like machine learning and image processing helps in identifying whether a person is wearing a face mask or not.

Artificial Intelligence is a branch of Computer-Science which helps computer machines to simulate like human intelligence. Some of the applications are Speech Recognition, Natural Language Processing etc. Machine Learning (ML) is a subset of Artificial Intelligence and superset of Deep Learning as shown in Fig 1. Machine Learning uses different algorithms for data analysis which automates the analytical model building. Some of the applications of ML are Image processing, classification, prediction etc.
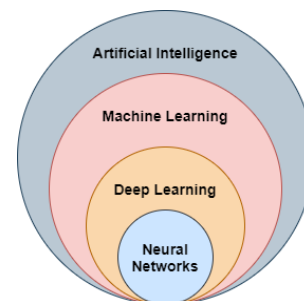

Fig.1. Artificial Intelligence

Deep Learning is a part of machine learning concept based on the artificial neural networks. The trend towards Deep Learning has increased. There are mainly three layers of neurons present in deep neural network. First layer is Input Layer, Second layer is Hidden Layer and finally the Third Layer is the Output Layer. If more number of hidden layers is used in neural network this concept is called as the Deep Neural Network. The applications of Deep Neural Network

are Voice Recognition, Image Classification, and Image Recognition etc.

The main differences between Machine Learning (ML) and Deep Learning (DL) are:

- In Deep Learning, when the raw data is given as input to the layers it learns features on its own that is machine automatically identifies important features or patterns from the training data. Whereas in Machine Learning, we have to manually identify features that the system will use as shown in Fig 2.
- ML contains one input layer, hidden layer and one output layer. Whereas DL contains one input layer, more number of hidden layers and one output layer.
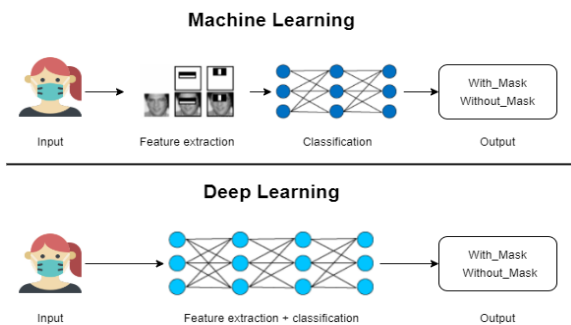

Fig.2. Machine Learning and Deep Learning

- Recognition of face is a popular and significant technology in recent years. Previously, detecting Face of a person and his face expression has done using different machine learning techniques like support vector machine.
- But the work towards identifying whether a person is wearing a face mask or not wearing a face mask is challenging task and there is no large dataset available for face masks detection.

A. *Proposed System*

- The proposed system uses face mask dataset which consists of images person with mask and person without mask.
- The input images are preprocessed and given for neural network for training, in which the features are extracted and learned during training; finally it classifies whether person is wearing a mask or not wearing a mask.
- The system uses deep neural networks like Convolutional Neural Network which is compared with different networks like Vgg16, MobileNet to know which is more computationally efficient and give accurate results.
- In real time, using these algorithms for training a system we can detect whether person is wearing a mask or not wearing a mask by taking live input from webcam.

B. *Objectives of proposed system*

The main objective of proposed system is to detect whether a person is wearing a face mask or not wearing a face mask.

- Pre-processing the images in which augmentation is performed and are resized.
- The detection of face takes place using techniques such as Haar cascade and single shot object detection.
- Training using deep neural network such as convolutional neural networks, Vgg16 and MobileNeV2.

II. LITERATURE SURVEY

Earlier, training a machine to detect and recognize a face automatically from the images was a challenging task but now a day's different machine learning and deep learning algorithms are available[17]. With the help of computer vision, neural network algorithms, python programming language and its libraries, it became easy to implement the projects like Face Detection and Recognition, etc.

Digital image is an image which consists of number of pixels. A pixel is a smallest element in an image, its value ranges from 0 to 255. The digital images are basically of two types: one is colored image and another is black & white image (gray image). In Gray scale image the pixel value is either "0" or "255" but in color image it has three layers – RGB.[3] In the color image the pixel value is between "0" and "255". As computer can understand only binary values, the image is represented as array of matrix and these values are converted to binary digits so that system can understand.

For image handling the concept of the Computer Vision is used [2]. The main aim of the computer vision is to develop the code or algorithm that allows computer to "see" i.e. image understanding, machine vision, and image analysis. It is concerned with the automatic extraction, analysis and understanding of useful information from the images. In Computer Vision, an image or a video is taken as input and is used to understand the image and its content. To solve some of its task computer vision uses image processing algorithms.

In Face Detection system, the face images are given as input to the system for training purpose [5]. Next the pre-processing is done which converts input image to gray color & are resized and facial features are extracted like eyes, nose, mouth, chin, etc and machine learning algorithms are applied on it. Now the trained system is ready to identify or detect the Faces from the images.

Rajeev Ranjan et al. [2015] proposed a fast and accurate system for face detection [10]. The easy accessibility of large amount of dataset, affordable power of computation has shown good improvement in the performance of the CNNs (Convolutional Neural Networks) on different tasks of face analysis. In this the pipeline of the deep learning for face verification and identification achieves improvement in the performance on several face datasets. In this a loss function which is called as the Crystal Loss is used for the proposed system of face verification and identification.

A system to detect and recognize face is proposed [11]. There are many applications where the face detection system is used in computer vision. Some of the examples are: automatic control system and communication application. The method of detecting the face from an image is called Face detection. More research is required in the area of face

detection, face tracking, expression recognition and pose estimation. The main challenge here was to detect the face from the input image, as all the people faces are not fixed and it changes in shape, size, color etc. Face detection becomes even more challenging task when the input image is not clear, dull lightning in the image, occluded by any other thing, not facing the camera etc. Viola Jones algorithm is used for face detection technique and for analysis of the face recognition. Some of the drawbacks of these techniques are high computation time when the input image size is having high resolution, large size and if high dimensional data is used.

For face detection and tracking system [6], the application of proposed system is to track and detect faces from the videos and in cameras which can be used for multiple purposes / activities. This is one of the deep studies of face detection using Open CV. In this work the comparison is done between different algorithms like Haar cascade, Adaboost methods. In this method of face detection technique, the Haarcascade gave the best results / accuracy rate.

In Computer Vision community Face Detection is one of the most studied topics. By the availability of datasets of the face detection much of the progress has been made. There is a gap between real world requirement and current face detection performance. To facilitate the future face detection WIDER FACE dataset has been introduced[14], which is ten times larger than the previous existing datasets. Face images present in this proposed dataset is extremely challenging because there is a large variations in pose, scale and occlusion. Furthermore, for an effective training source for detection of face - WIDER FACE dataset is used. Several detection systems provide an overview of state-of-art performance and to deal with large scale variation. Finally, failure cases need to be future investigated. Some disadvantages are faces cannot detect the small scale, occlusion and extreme poses of images from given dataset.

Menglong Zhu [2017] for mobile vision applications MobileNet is efficient model[9] . It uses depthwise separable convolutions to build light weight deep neural networks which is a streamlined architecture. In this two global hyper parameters are used which tradeoff between accuracy and latency effectively. The right sized model can be chosen with the help of these parameters based on the problem constraints. An experiment done to compare with the other ImageNet classification models.

The concept of deep learning has been originated from the field of Artificial Neural Network which refers to a class of neural networks with deep structure. The deep neural network is a powerful technology and an effective method of training tool in artificial intelligence. In this, face recognition is based on method of deep learning concept combined with the relevant theory[16]. The disadvantage is, it has to improve the image recognition rate and robustness.

Ashu Kumar [2018] proposed technique which can detect face from the video[1]. Now there is a need of intelligent systems which can automatically understand and examine the information because while doing manually it can get plainly distant. Face plays a prominent role in the social communication for conveying the feelings and the identity of a person. Human beings do not have the ability to identify different faces as accurate as machines do. So, the automatic face detection system plays an important role in Face

recognition, Human – Computer Interaction (HCI), Head-pose estimation, Facial expression recognition, etc. Face detection is a computer technology that determines the location of human face and size of the human face in a digital image. In Computer Vision literature the concept of Face detection is standout topic.

In multi-branch face detection architecture, helps in identifying the faces with small scale [13]. The feature map of neighboring are fused so that features coming from large scale can detect small scale faces. Finally, the simultaneously adopt multi scale training and testing to make the model robust towards the various scale.

Face detector uses a small subset of landmarks to estimate the facial shape and then after removing the similarity transformations a subsequent refining step is performed which estimates the facial shape of each person in high resolution. A Novel multi-view hourglass model estimates both profile facial landmarks and semi-frontal[18]. The joint training model is robust and stable under continuous variation views. By this there is a huge improvement over the state of the art results in the benchmarks for face alignment.

Fine tuning the image classification for transfer learning algorithms are proposed [7]. In this networks which are fine tuned are Alex Net, Vgg16[15]. These algorithms are used in human action recognition and object detection etc.

III. METHODOLOGY

The methodology of the Face mask detection system is shown in figure 3. In this, the first step is to collect the images dataset from kaggle containing both person wearing a face mask images and person without a face mask images followed by pre-processing, face detection from images is done and system training using neural networks and finally classifies as person with mask or without mask.

A. Dataset

- In this the face mask dataset containing 5000 images of person with mask and without mask are used. The dataset is downloaded from kaggle.
- The dataset is divided into two sets one for training and other for testing the system. For training 80% of images are given and for testing 20% of images are given.
- The same dataset is used for CNN, VGG16 and MobileNetV2.

B. Pre-processing

Images present in the dataset are pre-processed before sending to the neural network for training. As the dataset images are less so augmentation is done in which images are increased by 20,000 and are resized.
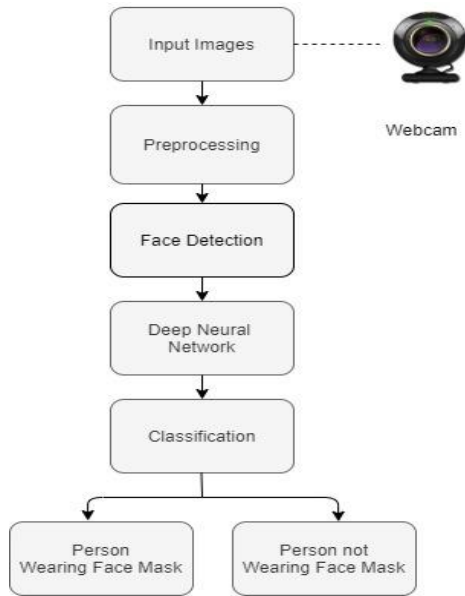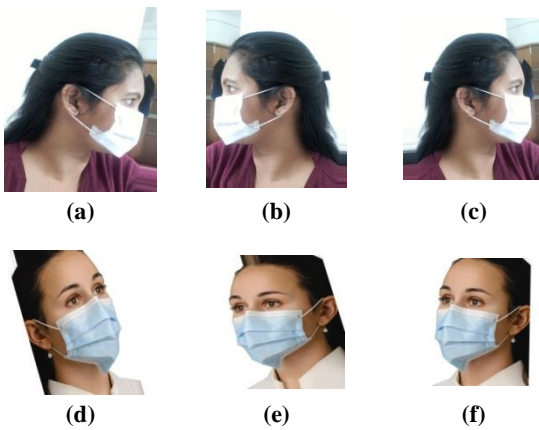
Fig.3. Process of System

Augmentation

Before training, the dataset images are augmented. Generally augmentation is the process of increasing the number of images by performing different operations like rotation, zooming, flipping, shearing, cropping the image. Below figure (a),(b),(c),(d),(e),(f) are some examples of augmented images.

| (a) | (b) | (c) |
| (d) | (e) | (f) |

C. Face Detection

For the detection of face the haarcascade and single shot object detection techniques are used. It is an effective way for detection of objects. After face detection images are given for the model to train.

D. Training Using Deep Neural Networks

After the face detection is done the training takes place using different deep neural networks like CNN, Vgg16 and MobileNetV2 model. The features are extracted and classified during training.

Convolutional Neural Networks

The CNN stands for Convolutional Neural Network. CNN is one of the Neural Networks model[12]. CNN helps system to extract features automatically. The concept of neural networks came from the neurons present in the human beings. For communication of message in humans, multiple neurons participate and the output of one neuron becomes input to the other neuron this goes on till it reaches the final destination. In a similar manner the CNN contains input layer, many convolutional layers and output layer as shown in Fig 4.
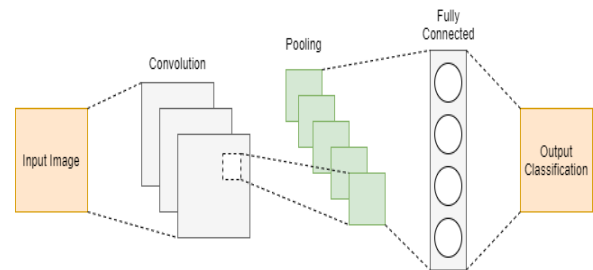

Fig.4. Convolutional Neural Network

The architecture of CNN contains many layers like Convolutional Layer, Pooling Layer, and fully connected layer. In Convolutional Layer: the input image matrix is convoluted with the help of filters (kernels) to generate the feature map. Convolution is simply the multiplication of 200 kernel with the input image having 3x3pixels. The Pooling layer helps in reducing the large size features to the smaller size features.

The pooling layer scales down the amount of information that the convolutional layer generated for each feature and it maintains the essential features. Several types of pooling methods are available like Max pooling, Min pooling, Average pooling etc. Mostly max pooling method is used for pooling purpose. It is max pooled to 2x2 pixels. By using ReLU function which flattens the layer and converts the input values to the positive numbers. It neglects the negative numbers. ReLU stands for Rectified Layer Unit. The fully connected layer is the last layer in the CNN.

Fully connected input layer flattens the previous layer generated output. Fully connected layer applies weights over the generated feature analysis on input to predict an accurate label. The fully connected output layer uses softmax for binary classification i.e with mask and without mask.

VGG16

VGG16 model is proposed by A.Zisserman and Simonyan from the Oxford University in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition"[4]. Image size of 224 x 224 RGB is given as input to the conv1 layer. The input image is passed through some convolutional layers with the filter size of 3x3. By using 2x2 pixel window, 5 max pooling is performed. The Three Dense layers follow a stack of convolutional layers with different depths. Dense layer is also called as fully connected layer.

VGG16 is a sixteen layer Convolutional Neural Network (CNN) as shown in Fig 5. The input image is passed through the VGG16 network model and the output is a classification data. VGG16 is a sequential model. A sequential model is that where the output of one layer is taken as input for the next layer. In this model a stack of sequential tasks of convolutional and max pooling layers are performed. VGG16 is used for the classification tasks and studies shows that VGG16 is giving best performance when compared with the previous models.
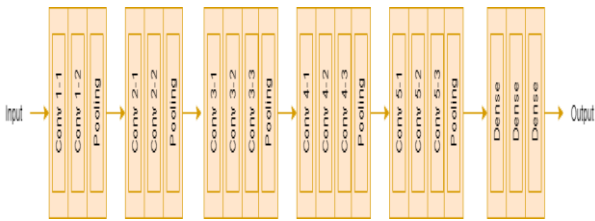

Fig.5. Vgg16

MobileNet

MobileNet is a deep convolutional neural network. It is a streamlined architecture. It consists of various versions such as MobileNetv1, MobileNetV2. In this model MobileNetV2 [8] is used. It consists of the depthwise separable convolution concept which uses the depthwise separable filters. Depthwise separable convolution filters consists of depthwise convolution filters and Point convolution filters. Single convolution is performed on each input channel in depthwise convolution filter whereas in point convolution filter combines 1*1 convolutions with the output of depthwise convolution. In MobileNetV2 the fully connected layer (final layer) uses the softmax for classification. It is an efficient model for mobile and embedded vision applications. It is faster in computing when compared to convolutional neural network and vgg16. MobileNet architecture is shown in Fig 6.
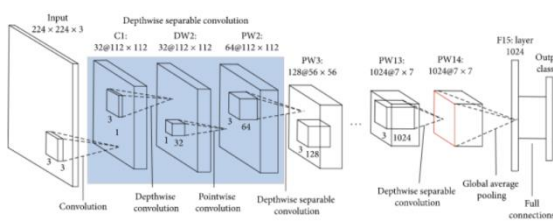

Fig.6. MobileNet

E. Classification

The system is trained and tested by using CNN, Vgg16 and MobileNetv2 model on input face mask dataset in the ratio of 80:20. After training the images are classified whether a person is wearing a mask or not. In real time, we can detect person wearing mask or not by taking input from webcam and system displays the output results with label as face with mask or without mask.

IV. RESULTS

In this system the face mask dataset comprise of images of the person with and without mask which are pre-processed and given for the neural network for the training and it classifies whether person is wearing a face mask or not.

- Comparing of three algorithms CNN, Vggg16 and MobileNetV2 takes place for face mask detection system.
- The same dataset is used for training and testing these algorithms.
- In which 80% of images are given for training and 20% for testing.
- Among 20,000 images, 16,000 images are given for training and 4000 images are given for testing.
- Convolutional Neural Network got an accuracy of 85%.
- Vgg16 got an accuracy of 89%.
- MobileNetV2 got an accuracy of 94%.
- Compared to CNN and Vgg16, the MobileNetV2 got the best accurate rate and is more computationally efficient.
- MobileNetV2 is able to detect the side faced person with mask efficiently.

A. Training Loss and Accuracy for CNN, Vgg16 and MobileNetV2

Training loss is the error on training set of data and Validation loss is error after running the validation set of data through trained network. As the number of epochs increases both the validation and training error decreases. The training loss are calculated by moving over 1 epoch on an average and validation loss is been calculated after learning phase of the same epoch. The training loss can be reduced by using the following methods:

- By increasing dataset using data augmentation.
- The training epochs has to be increased.

Train accuracy is the accuracy of a model on dataset on which it was constructed and Validation accuracy is the accuracy that we calculate on the untrained images.

Fig 7, Fig 8, Fig 9 are graphs representing the training loss and accuracy rate for CNN, Vgg16 and MobileNet. For these graphs, epochs are taken on the X-axis and loss & accuracy is taken on the Y-axis.

The training loss is shown in the red line in the below graphs and validation loss is shown in blue line. It is observed that as the number of epochs increases, both the training and validation loss are decreasing.

The purple line shows the training accuracy and grey line shows the validation accuracy. From the graph it can be seen that as the number of epochs increases both the training accuracy and the validation accuracy gradually increases.
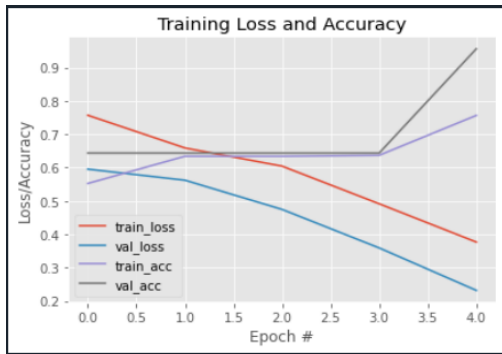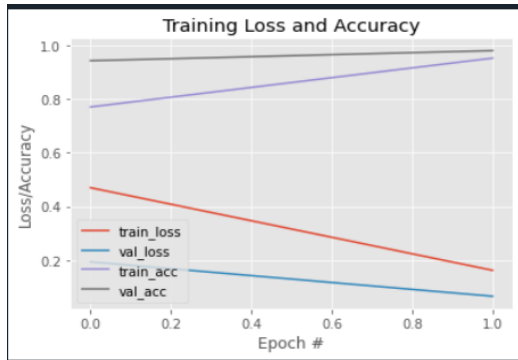
Fig.7. Training Loss and Accuracy For CNN

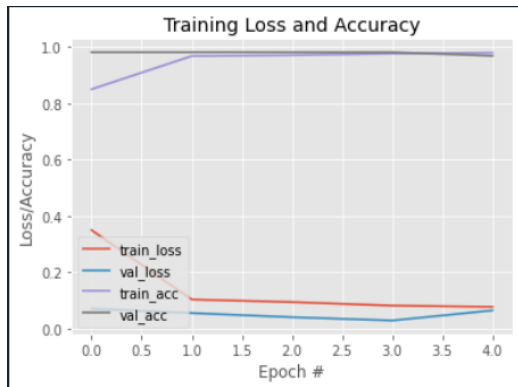Fig.8. Training Loss and Accuracy For VGG16

Fig.9. Training Loss and Accuracy For MobileNetV2

- In real-time, proposed system is implemented using MobileNetV2, Vgg16 and CNN.
- System takes input from webcam and detects whether a person is wearing a face mask or not.
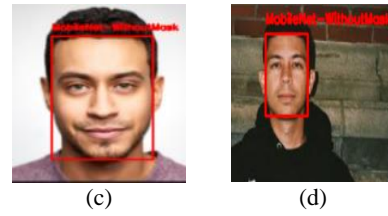- The below figures (a), (b), (c), (d) shows output detected by trained system model as person with mask and person without mask.

| (a) | (b) |

| (c) | (d) |
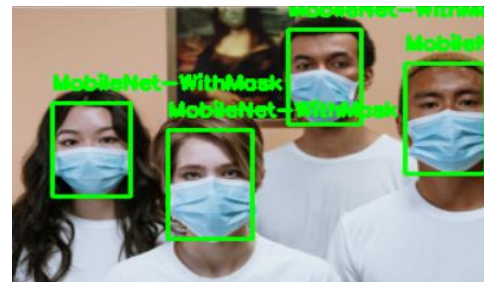
➢ System is able to detect multiple faces using MobileNetV2.
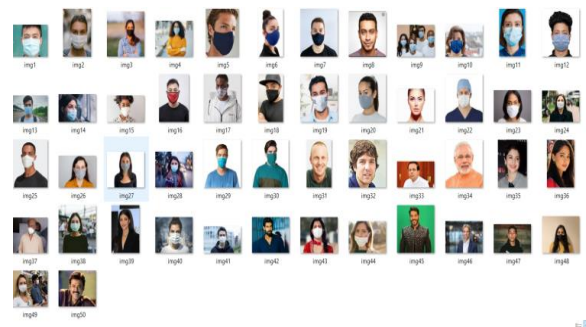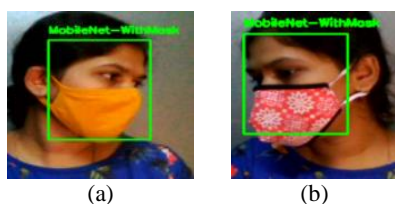
External sample images

Fig.10. Sample images

To check the performance of the trained system we have collected 100 images which are not present in the dataset. We have given these images as input to the system which is trained using CNN, Vgg16 and MobileNetV2 models as shown in Fig10.

B. Confusion Matrix

The confusion matrix is to evaluate the performance of a classification model to know the true values based on a set of test data. Generally confusion matrix is a table shown in 5.1. By using various parameters such as accuracy, precision, recall and f1-score we can find how good our model is trained and gives the results.

TABLE I.   CONFUSION MATRIX

Model	Actual Class		
Predict Class	*Class*	*With Mask*	*Without Mask*
	With Mask	TP = True Positive	FP = False Positive
	Class	FN = False Negative	TN = True Negative

- *Accuracy* is simply a ratio of correctly predicted observation values to the total observations values. The formula to calculate the Accuracy is shown below.

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \quad (1)$$

- *Precision* is the ratio of correctly predicted positive observation values to the total predicted positive observations plus wrongly predicted negative observations. The below formula is used to calculate the Precision.

$$Precision = TP/(TP+FP) \quad (2)$$

- *Recall* (Sensitivity) is the ratio of correctly predicted positive observations values to the all observations in actual class - yes. The below shown formula is used to calculate the Recall.

$$Recall = TP/(TP+FN) \quad (3)$$

- *F1 Score* is a weighted average of the Precision and Recall. Therefore, this considers both false positives and false negatives.

$$F1\text{-}Score=2*(Recall*Precision)/(Recall+Precision) \quad (4)$$

TABLE II. USING CNN

Internal Validation On 4000 Images			External Validation On 100 Images		
Class	*M*	*WM*	*Class*	*M*	*WM*
M	TP = 2090	FP = 290	*M*	TP = 34	FP = 12
WM	FN = 310	TN = 1310	*WM*	FN = 16	TN = 38
Accuracy = 0.85 Precision= 0.87 Recall= 0.8 F1-score=0.8			Accuracy = 0.72 Precision= 0.73 Recall= 0.68 F1-score =0.69		

Where M = with mask , WM = without mask

Table II shows the internal and external validation results when the system is trained on neural network models CNN. Internal validation is done on dataset consisting of 4000 images and external validation is done on 100 images which are not present in the dataset.

TABLE III. USING VGG16

Internal Validation On 4000 Images			External Validation On 100 Images		
Class	*M*	*WM*	*Class*	*M*	*WM*
M	TP = 2148	FP = 188	*M*	TP = 37	FP = 10
WM	FN = 252	TN = 1412	*WM*	FN = 13	TN = 40
Accuracy = 0.89 Precision= 0.89 Recall= 0.8 F1-score =0.8			Accuracy = 0.77 Precision= 0.78 Recall= 0.74 F1-score =0.7		

Where M = with mask , WM = without mask

Table III shows the internal and external validation results when the system is trained on neural network models VGG16. Internal validation is done on dataset consisting of 4000 images and external validation is done on 100 images which are not present in the dataset.

TABLE IV. USING MOBILENETV2

Internal Validation On 4000 Images			External Validation On 100 Images		
Class	*M*	*WM*	*Class*	*M*	*WM*
M	TP = 2272	FP = 112	*M*	TP = 42	FP = 7
WM	FN = 128	TN = 1488	*WM*	FN = 8	TN = 43
Accuracy = 0.94 Precision= 0.95 Recall= 0.9 F1-score =0.9			Accuracy = 0.85 Precision= 0.89 Recall= 0.8 F1-score =0.8		

Where M = with mask , WM = without mask

Table IV shows the internal and external validation results when the system is trained on neural network models MobileNetV2. Internal validation is done on dataset consisting of 4000 images and external validation is done on 100 images which are not present in the dataset.

From the above Tables 2, 3 and 4 results it is observed that the MobileNetV2 model is showing more efficient results when compared with CNN and VGG16.

V. CONCLUSION

The Face mask detection system is trained using Convolutional Neural Network, VGG16 and MobileNetV2 using Face Mask Dataset. Using these algorithms the system is able to detect whether a person is with mask or not. Compared to CNN and Vgg16 algorithms, the MobileNetV2 gives best accuracy rate of 94% and it is more efficient. It is able to detect the person with frontal and side face persons wearing masks. In real time, the proposed system takes input from webcam and classifies whether the person is wearing a face mask or not wearing a mask.

REFERENCES

[1] Ashu Kumar, Amandeep Kaur and Munish Kumar, "Face detection techniques: a review", Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda,2018.
[2] Hari krishnan J et al., "Vision-Face Recognition Attendance Monitoring System for Surveillance using Deep Learning Technology and Computer Vision", IEEE, 2019.
[3] Hossein Ziaei Nafchi et al., "CorrC2G: Color to Gray Conversion by Correlation" , IEE, 2017.
[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition", Conference paper at ICLR, 2015.
[5] Kriti Dang, "Review and Comparison of face detection algorithms", IEE, 2017.
[6] Kruti Goyal, Kartikey Agarwal and Rishi Kumar, "face detection and tracking", International Conference on Electronics, Communication and Aerospace Technology ,   ICECA 2017.
[7] Manali Shahal and Meenakshi Pawar, "Transfer learning for image classification", IEE, 2018.
[8] Mark Sandler et al., "MobileNet V2: Inverted Residuals and Linear Bottlenecks", IEEE, 2018.
[9] Menglong Zhu et.al., "Mobile Nets: Efficient Convolutional Neural Networks for Mobile Vision Applications", 2017.

[10] Rajeev Ranjan et al., "A Fast and Accurate System for Face Detection, Identification, and Verification", IEEE 2018.

[11] Ravi Sharan and Divya Meena, "An Approach to Face Detection and Recognition", IEEE International Conference on Recent Advances and Innovations in Engineering, December, 2016.

[12] Saad Alwabi, Tareq Abed Mohammed and Saad Al-Zawi, "Understanding of a Convolutional Neural Network", IEEE, 2017.

[13] Shi Luo , Xiongfei Li, Rui Zhu and Xiaoli Zang, "SFA: Small Faces Attention Face Detector", IEE Access , 2019.

[14] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark", In Proceedings of the IEEE conference on computer vision and pattern recognition.

[15] Srikanth Tammina, "Transfer learning using Vgg16 with deep convolutional neural networks for classifying images", International Journal of Scientific and Research Publications, Vol.9, No.10,October 2019.

[16] Xiao Han et al., "Research on Face Recognition Based on Deep Learning", IEEE,2018.

[17] Ya Wang, "Face Recognition in Real-world Surveillance Videos with Deep Learning Method", IEE, 2017.

[18] Yuxiang Zhou, Stefanos Zafeiriou, Jiankang Deng and George Trigeorgis "Joint Multi-View Face and Alignment in the Wild", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 28, NO. 7, JULY 2019.