Scaling Up Face Masks Classification Using a Deep Neural Network and Classical Method Inspired Hybrid Technique

Akhil Kumar[1], Arvind Kalia[1], Kinshuk Verma[2], Akashdeep Sharma[2*], Manisha Kaushal[3], and Aayushi Kalia[3]

[1]Department of Computer Science, Himachal Pradesh University
Shimla, Himachal Pradesh 171005 India
[e-mail: akhil.hpucs@gmail.com, arvkalia@gmail.com]
[2]CSE, UIET, Panjab University
Chandigarh, UT of Chandigarh 160014 India
[e-mail: chd.kinshuk@gmail.com, akashdeep@pu.ac.in]
[3]CSED, Thapar Institute of Engineering & Technology
Patiala, Punjab 147004 India
[e-mail: manisha.kaushal@thapar.edu, aayushikalia26@gmail.com]
[*]Corresponding author: Akashdeep Sharma

Received September 11, 2021; revised April 4, 2022; revised June 5, 2022; revised July 2, 2022; accepted November 5, 2022; published November 30, 2022

Abstract

Classification of persons wearing and not wearing face masks in images has emerged as a new computer vision problem during the COVID-19 pandemic. In order to address this problem and scale up the research in this domain, in this paper a hybrid technique by employing ResNet-101 and multi-layer perceptron (MLP) classifier has been proposed. The proposed technique is tested and validated on a self-created face masks classification dataset and a standard dataset. On self-created dataset, the proposed technique achieved a classification accuracy of 97.3%. To embrace the proposed technique, six other state-of-the-art CNN feature extractors with six other classical machine learning classifiers have been tested and compared with the proposed technique. The proposed technique achieved better classification accuracy and 1-6% higher precision, recall, and F1 score as compared to other tested deep feature extractors and machine learning classifiers.

Keywords: CNNs, Face masks, Machine learning, Multi-layer perceptron, ResNet-101

1. Introduction

In these times of the COVID-19 pandemic, people across the globe are wearing face masks to protect themselves from the Coronavirus. Previous research suggests more crimes are carried out by criminals hiding their identity by wearing face masks [7, 8] and hyper-realistic face masks [9]. Artificial intelligence, machine learning, and deep learning combined with computer vision can help to develop such face recognition and identification methods that can help in identifying criminals wearing a face mask. However, there exists a bottleneck of availability of data for people wearing masks and effective face masks classifiers that can aid in developing a face recognition system for uncontrolled surveillance environments using CCTV cameras that are based on low-end computation hardware resources. However, research based on deep learning has been fascinating as deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks have applications in the fields of computer vision, machine vision, and image analysis, where they have produced results comparable to human expert performance. In recent years, several deep neural network inspired convolutional neural network (CNN) architectures such as VGG-16, Inception v3, ResNet-50, ResNet-101, ResNet-152 v2, DenseNet-121, MobileNet v2 and Xception has been proposed to perform image classification. These CNN-based architectures can be used as a method to extract features from a set of images or classify the whole set of images to their corresponding classes. The advantage of using deep neural network-based CNN architectures is their ability to self-extract the features and delivering high-quality results. However, there is still a scarcity of studies using deep learning for the identification of people wearing masks.

This work proposes a face masks classification technique by employing pre-trained ResNet-101 [1] as a feature extractor and classical multi-layer perceptron as a classifier for the classification of persons wearing face masks and not wearing face masks. The efficacy of the proposed technique was evaluated based on performance metrics: accuracy, precision, recall, and F1 score. To back and support the proposed technique various pre-trained convolution neural network architectures such as VGG-16 [2], Inception v3 [3], MobileNet v2 [4], DenseNet-121 [5], Xception [6], and ResNet-152 v2 were used for feature extraction and machine learning (ML) classifiers such as multi-layer perceptron, support vector machine, extra trees, random forest, k-nearest neighbors, gaussian naive bayes and decision trees were used for classification and evaluated based on performance metrics. Furthermore, a comparative analysis has been drawn among the proposed technique and other combinations explored to justify the validity of the proposed technique. To carry out this work, the number of layers to produce feature vectors from pre-trained networks has been suitably chosen with a trade-off of reducing training time and speeding up classifier performance. The entire work was carried out on a custom self-created dataset consisting of 23,500 images with samples for two classes namely, with masks and without masks. The proposed technique has high implications as a step towards developing a face recognition system that classifies persons wearing face masks and not wearing face masks and further extended as a robust system to detect the identity of persons by using left-over features of the face not covered by the mask.

The major contributions of this work are:

- This work presents a hybrid face masks classification technique based on ResNet-101 and multi-layer perceptron (MLP) classifier. The proposed technique achieved a classification accuracy of 97.3% on the employed dataset which was the highest across the experiments conducted. Furthermore, the proposed technique achieved a 1-6% higher precision, recall, and F1 score as compared to other tested classifiers.

- To validate the proposed technique, forty-nine combinations were experimented with by employing seven deep neural network inspired by CNN feature extractors on top of the seven classical machine learning classifiers and evaluated based on performance metrics such as accuracy, precision, recall, and F1 score. Furthermore, comparative analyses have been drawn to back and support the proposed technique.
- To embrace the proposed technique, comparisons have been drawn with face masks classification techniques proposed in recent years. The results indicate that the ResNet-101 and MLP classifier-based technique outperforms the related work in literature.
- The novelty of this work lies in the exhaustive set of experiments performed to gauge the validity of the proposed technique that can help to add new knowledge to the existing literature and further exploration of other problems by the researchers.

This paper is organized in the following sections: Section two presents the related work in relevant field; section three describes the materials and methods; section four presents the experiment design, results and comparative analysis; section five is composed of the conclusion and possibilities of future work.

2. Related Work

Face masks classification is a process to determine the presence of a mask on the face in a given image or video. In the area of face masks classification problem, most of the publications focus on face identification, face construction when wearing face masks. Recent works published in this domain are addressed using deep learning and generative adversarial networks.

In Ejaz et al. [10], to recognize the person, the authors have applied the principal component analysis (PCA) on masked and unmasked face recognition. The results of this study suggest a drop in accuracy of face recognition in a masked face. The authors Park et al. [11] proposed a method based on PCA that is used for removing glasses from a human frontal facial image. The recursive error compensation using PCA reconstruction was used to reconstruct the removed part. In Nieto-Rodríguez et al. [12], a method for detecting the presence or absence of a medical mask in the operating room was proposed. The objective of this work is to trigger alarms only for medical staff who do not wear a surgical mask in the operating room by minimizing the false positive face detections as possible without missing mask detections. The proposed method achieved 95% accuracy. Loey et al. [13] proposed a hybrid deep transfer learning model with machine learning methods for face mask detection. In the proposed model, the authors have used Resnet-50 as the feature extractor and classification process of face masks is performed using decision trees, support vector machines (SVM) and ensemble algorithms. To test the validity of the proposed model the authors have investigated three datasets namely, RMFD, SMFD and LFW. The results of the proposed model show SVM classifier achieved 99.6% testing accuracy on RMFD, 99.49% testing accuracy on SMFD and 100% testing accuracy on LFW dataset. In Qin and Li [14], a face mask-wearing classification system by embedding image super-resolution using classification network (SRCNet) was proposed. The proposed model quantified mask, no mask and incorrectly worn masks based on 2D facial pictures. Image pre-processing, face detection, crop, image super-resolution and face mask wearing conditions identification are the backbone of the proposed model. The proposed model gave an accuracy of 98.70%. In Li et al. [15], an HGL method for head pose classification with masks using color texture analysis

of pictures and line portraits was proposed. This method achieved a front accuracy of 93.64% along with a side accuracy of 87.17%. The proposed method recognizes between a face mask and not wearing a face mask. A face mask detector is proposed by Nagrath et al. [16] using SSD-MobileNet v2 for detection of people wearing or not wearing face masks on a custom dataset and achieved a detection accuracy of 92.64%. In Ud Din et al. [17], a novel GAN-based network that can automatically remove masks covering the face area was proposed. The proposed method also regenerates the image by building the missing hole. The output of this work produces a natural and realistic image of a complete face. Hussain and Balushi [18] proposed a real-time face emotion classification using deep learning. They have classified seven facial expressions. They have used VGG-16 architecture as a backend classifier. The proposed model achieved 88% accuracy on the KDEF dataset.

In Inamdar and Mehendale [19], a real-time face mask identification method using FacemaskNet deep learning network was proposed. The proposed network can classify three classes namely a person wearing a mask, improperly worn masks, or no mask detected. The authors have created a custom dataset with 10 pictures of individuals wearing a mask, 15 pictures of improperly worn masks and 10 pictures involved a person's face without a mask. The proposed work performs detection with 98.6% accuracy. Khandelwal et al. [20] proposed a deep learning model that binarizes an image as a mask is used or not. In the proposed work authors used 380 images having a mask and 460 images having no mask to train the MobileNet v2 model. The AUROC of the model was 97.6%. The limitation of this work is its inability to classify partially hidden faces. Jiang and Fan [21] proposed Retina face mask, which is a high-accuracy and efficient face mask detector. The authors have used transfer learning to extract robust characteristics trained on a large dataset of 7,959 images. The backend models used in the proposed work are ResNet and MobileNet. In Li et al. [22], authors used the YOLO v3 algorithm for face detection. The proposed method achieved 93.9% accuracy on FDDB dataset. The proposed algorithm was trained on CelebA and WIDER FACE dataset. The training dataset has more than 600,000 images.

From the existing literature, this is evident that there exists a scope of exploration and employment of state-of-the-art CNN feature extractors such as ResNet-101 and ResNet-152 combined with classical machine learning classifiers such as k-nearest neighbors and multi-layer perceptron to develop effective face masks classification techniques that can achieve better classification accuracy in a limited computation resource environment.

3. Materials and Methods

In order to propose an effective face masks classification technique, in the first step we prepared an image dataset for two classes namely, with_masks and without_masks. The images were extracted from the internet using a Python-based image crawler and a few images were also extracted from the RMFD [23] dataset. We further applied the image pre-processing technique, specifically resizing, in order to resize all the images to a standard size of 299x299 and 224x224 pixels. Moreover, data augmentation has been applied to enhance the dataset. Once the input images were prepared, these were passed to the ResNet-101 architecture for extracting deep features from each image. The feature vectors were wisely chosen and the generated features were then fed into multi-layer perceptron (MLP) classifier. The entire training of ResNet-101 architecture was carried out by mechanism of transfer learning i.e. we used a pre-trained model trained on the ImageNet dataset and re-trained the model on the self-created face masks classification dataset to extract the feature vector. To validate the effectiveness of the proposed technique, we further explored various CNN based feature

extractors such as VGG-16, Inception v3, MobileNet v2, DenseNet 121, Xception, and ResNet-152 v2 combined with classical ML classifiers such as multi-layer perceptron (MLP), support vector machine (SVM), extra trees (ET), random forest (RF), k-nearest neighbors (KNN), gaussian naive bayes (GNB) and decision trees (DT). Furthermore, comparative analysis based on classification accuracy has been drawn to support and brace the proposed technique.

3.1 Proposed Face Masks Classification Technique

The technique proposed for face masks classification in this work is a hybrid combination of ResNet-101 and multi-layer perceptron classifier. ResNet-101 is a convolutional neural network architecture that is 101 layers deep with the core idea of identity shortcut connections that skips one or more layers while learning features from the input image. To address the problem of network degradation, it uses pre-activated residual blocks with ReLU activation function. For the proposed technique, to reduce the computation complexity and training ResNet-101 from scratch, we employed the strategy of transfer learning and selected the feature representation produced by the network by the average pool layer applied after the 101st layer of the network providing a feature representation of dimension 1x1x1000. In order to reduce the computation resources dependency, we applied a dense layer of size 1024 after the average pool layer that converted the feature representation from 1x1x1000 to a 2-D feature vector. To perform feature extraction using the transfer learning mechanism, we removed the softmax layer used by ResNet-101 for classification and passed the obtained feature vector to the multi-layer perceptron classifier to perform the classification task. The advantage of using a multi-layer perceptron classifier is its ability to classify unknown patterns with other known patterns that share the same distinguishing features and its simplicity and ability to work on low computation power. It can classify noisy and incomplete inputs because of their similarity with pure and complete inputs.

In the proposed technique, the CNN-based ResNet-101 feature extractor takes input as $X \in R^{w*h*c}$ where, R is an RGB image with w: $width$, h: $height$, and c: $channel$. Each layer of the CNN feature extractor takes X and a set of parameters W as input and outputs a new image $Y = f(X, W)$. Since the ResNet-101 feature extractor is based on residual connections and identity shortcut connections; it skips one or more layers while learning features from the input image and outputs an image as shown in equation (1).

$$Y = f(X, \{W_i\}) + W_S X \qquad (1)$$

where, $f(X)$ and X have different dimensions for the input and W_S represents the 1x1 convolutions added as identity shortcuts.

For Y_i images, ResNet-101 produces a 2-D feature vector which is passed to the multi-layer perceptron (MLP) to classify the data into two classes of the dataset. The multi-layer perceptron (MLP) maps the 2-D feature vector into linearly separable hyper-plane as shown in equation (2).

$$Y = [fg_1(X) \dots \dots \dots \dots fg_k(X)] \qquad (2)$$

where, $f(.)$ is the activation function and $g_k(x)$ is the hyperplane realized by k_{th} neuron.
The problem addressed in this work is the binary classification of persons wearing face masks and not wearing face masks. The features related to the face area are present in all the images. The distinguishing feature that differentiates between the persons with face masks and without face masks is the presence of face masks on the face area. The technique proposed to solve this

problem uses ResNet-101 as a feature extractor that generates a 2-D feature vector in the form of numbers. Further, the 2-D feature vector is passed to the multi-layer perceptron (MLP) classifier for the task of classification between persons with face masks and without face masks. The multi-layer perceptron (MLP) classifier distinguishes between the obtained features by performing probability-based predictions based on the two classes i.e. with_masks and without_masks. The multi-layer perceptron (MLP) classifier employs backpropagation to learn and classify the non-linearly separable data and perform the final prediction. The overall workflow of the proposed ResNet-101 and MLP based technique is presented in **Fig. 1**. In the proposed technique, an input image of size 299x299 pixels was passed to ResNet-101 architecture which extracted refined features from the input image by applying convolution and pooling operations to generate a 2-D feature vector. Furthermore, the feature vector obtained from ResNet-101 was fed to multi-layer perceptron classifier to classify whether the input image passed was for class with_masks or without_masks.

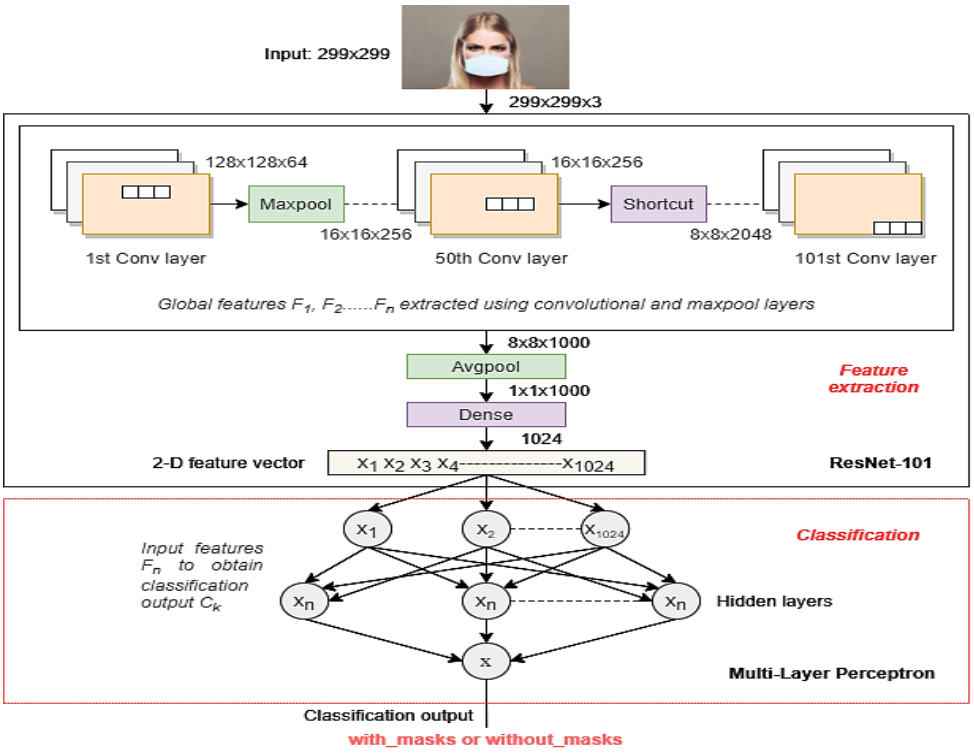


Fig. 1. Proposed face masks classification technique

In order to gauge the validity of the proposed technique tests were conducted with ResNet-101 and other state-of-the-art CNN-based feature extractors combined with multi-layer perceptron and six other classical machine learning classifiers. For each combination, performance metrics were evaluated and comparison analysis was drawn to justify the effectiveness of the proposed ResNet-101 and multi-layer perceptron-based technique.

4. Experiments and Results Analysis

4.1 Experiment Design

To embrace the proposed ResNet-101 and multi-layer perceptron-based face masks classification technique, we employed different state-of-the-art CNN architectures such as

VGG-16, Inception v3, MobileNet v2, DenseNet 121, Xception, and ResNet-152 v2 for feature extraction with the possibility of transfer learning on a limited size dataset and compared with multi-layer perceptron (MLP) and other machine learning classifiers satisfying their performance in computer vision tasks. The detailed experiment design to validate the proposed technique is presented in **Fig. 2**. As shown in **Fig. 2**, to carry out this work initially in the first step dataset was prepared. The image samples for persons with masks and without masks were scrapped from the internet and a benchmark dataset. Furthermore, image processing techniques were applied to obtain standard data. The images from the dataset were passed to seven CNN-based feature extractors which provided a 2-D feature vector. For the classification of input images, the 2-D feature vector was passed to different classical machine learning classifiers. Furthermore, the performance of each CNN feature extractor and machine learning classifier-based hybrid classifier was evaluated based on performance metrics. The description of CNN-based feature extractors employed in this work corresponding to their input size, number of CNN layers, and trainable parameters is presented in **Table 1**.

Table 1. CNN feature extractor parameters

CNN feature extractor	Input size in pixels	CNN Layers	Trainable parameters
ResNet-101	299x299	101 Conv layers	4,25,52,832
VGG-16	224x224	13 Conv layers	1,47,14,688
Inception v3	299x299	48 Conv layers	2,17,68,352
MobileNet v2	299x299	35 Conv layers	22,23,872
DenseNet-121	299x299	121 Conv layers	69,53,856
Xception	299x299	33 Conv layers	2,08,06,952
ResNet-152 v2	299x299	152 Conv layers	5,81,87,904

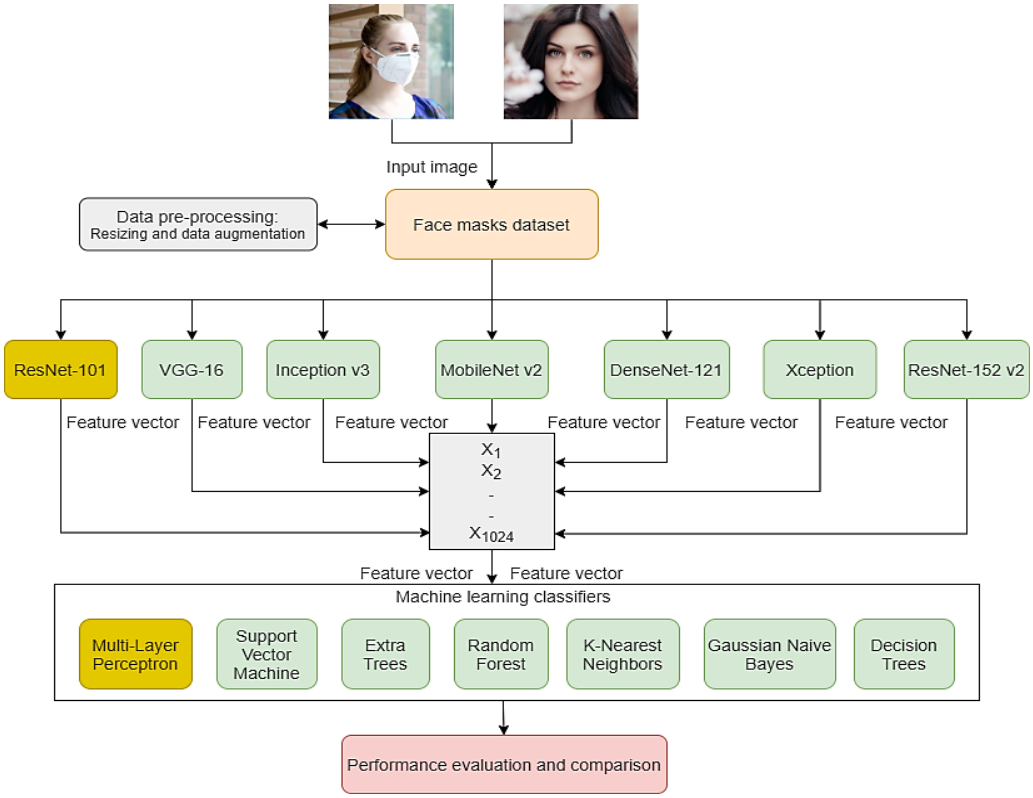


Fig. 2. Experiment design to validate the proposed technique

The proposed ResNet-101 and MLP based technique and other comparative experiments conducted in this work were developed and performed on an Intel i5-8th Generation-based system with 8GB DDR4 RAM and NVIDIA GeForce GTX 1050 GPU. For implementing CNN architectures and machine learning classifiers open-source libraries such as TensorFlow, Keras, and ScikitLearn were used and graphs were plotted using Matplotlib library.

4.2 Dataset Description

To perform classification of people wearing or not wearing face masks very few datasets are available in the public domain. In order to carry out this work we created a custom image dataset for people wearing face masks and not wearing face masks. The images for the dataset for persons not wearing face masks were collected using a Python-based image crawler-Google API [24] with 1,950 images and for persons wearing face masks 2,000 images were extracted from the RMFD [23] dataset. The collected images were divided into two classes with labels with_masks and without_masks. The dataset obtained after performing data pre-processing operations: resizing and data augmentation operations namely, flip, rotate, black and white, skew and zoom consists of 23,500 images with varying features and complexities. The class label with_masks have 12,000 images and the class label without_masks have 11,500 images. The dataset was split in the ratio of 70:30 respectively for training and test. The two classes of the dataset are illustrated in **Fig. 3**.

a). Image samples for class with_masks

b). Image samples for class without_masks

Fig. 3. Dataset description

4.3 Evaluation Criteria

To predict the performance of the proposed technique and other employed combinations, we have utilized evaluation metrics such as accuracy, precision, recall and F1 score and are shown in equations (3-6).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

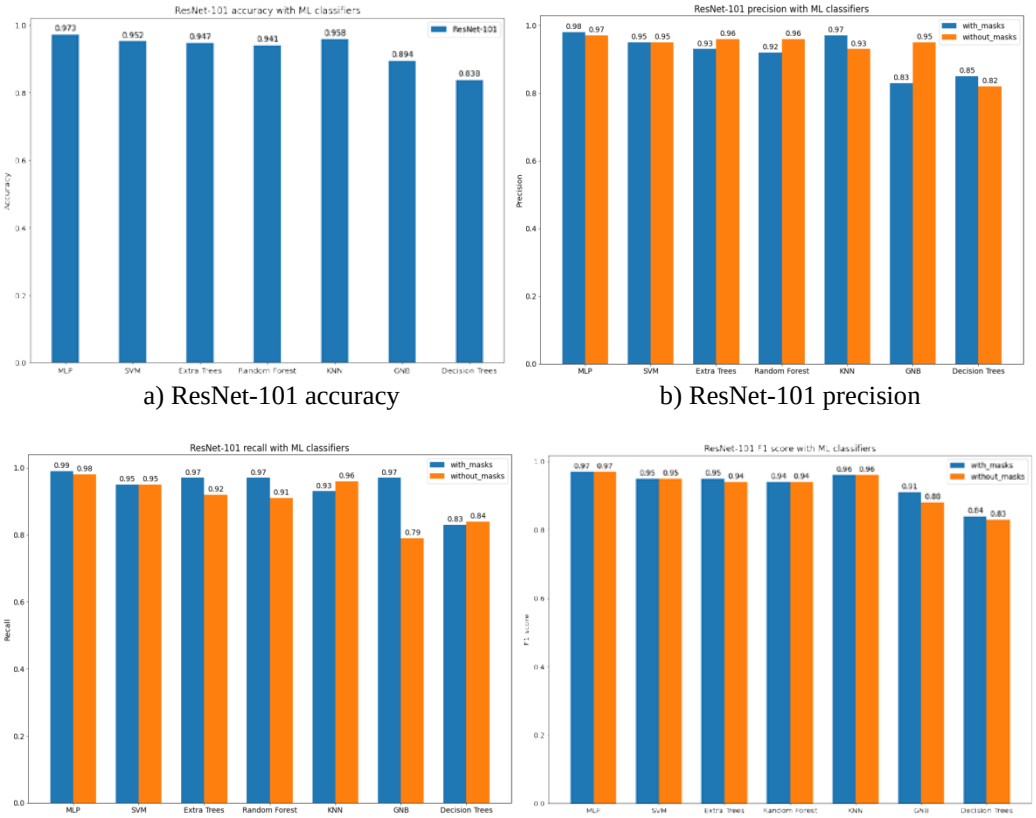
$$F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (6)$$

In equations (3-6), True positive (TP) is the number of positive instances that are correctly predicted; false negative (FN) is the number of positive instances that are incorrectly predicted. True negative (TN) is the number of negative instances that are predicted correctly; false positive (FP) is the number of negative instances that are incorrectly predicted.

4.4 Results

4.4.1 Performance of ResNet-101 Based Proposed Technique

In order to gauge the performance of the proposed technique, we trained and tested it on the employed dataset using the evaluation metrics. The ResNet-101 and multi-layer perceptron classifier (MLP) based proposed technique achieved a classification accuracy of 97.3% which was the highest among the employed classical machine learning classifiers. However, the least accuracy of 83.8% was achieved with decision trees as a classifier. The results indicate that most of the samples predicted by the proposed technique were correct. The performance of ResNet-101 corresponding to multi-layer perceptron (MLP) and other employed classifiers is presented in **Fig. 4**.



a) ResNet-101 accuracy b) ResNet-101 precision

c) ResNet-101 recall d) ResNet-101 F1 score

Fig. 4. ResNet-101 performance with machine learning classifiers

The ResNet-101 and multi-layer perceptron (MLP) combination achieved a precision of 98% for images with masks and a precision of 97% for images without masks that indicate a high number of true predictions with the proposed technique. However, the least precision for both the classes of the dataset was achieved with the decision trees classifier that indicates more false predictions with it. The proposed ResNet-101 and multi-layer perceptron (MLP) technique achieved a recall of 99% for images with face masks and a recall of 98% for images without face masks that indicate a prediction of more true positive samples. Furthermore, ResNet-101 with multi-layer perceptron (MLP) achieved an F1 score of 97% for images with masks and without masks outperforming other employed machine learning classifiers. The results indicate that the proposed ResNet-101 and MLP based technique predicted low false positives and low false negatives on the employed dataset thereby justifying the high classification accuracy for the task of face masks classification. The values of precision, recall, and F1 score remained poor with decision trees classifier when tested with ResNet-101 feature extractor.

The results as shown in **Fig. 4** indicate that the multi-layer perceptron (MLP) classifier achieved the highest accuracy, precision, recall, and F1 score among all the employed machine learning classifiers when tested with ResNet-101. To justify the effectiveness of the proposed ResNet-101 and multi-layer perceptron (MLP) based hybrid technique and advantage of multi-layer perceptron classifier, we tested other stated CNN feature extractors with multi-layer perceptron and evaluated the accuracy metric that indicates the percentage of correct predictions for the test data. The accuracy achieved by ResNet-101 and other employed CNN-based feature extractors corresponding to multi-layer perceptron (MLP) classifier is presented in **Fig. 5**.

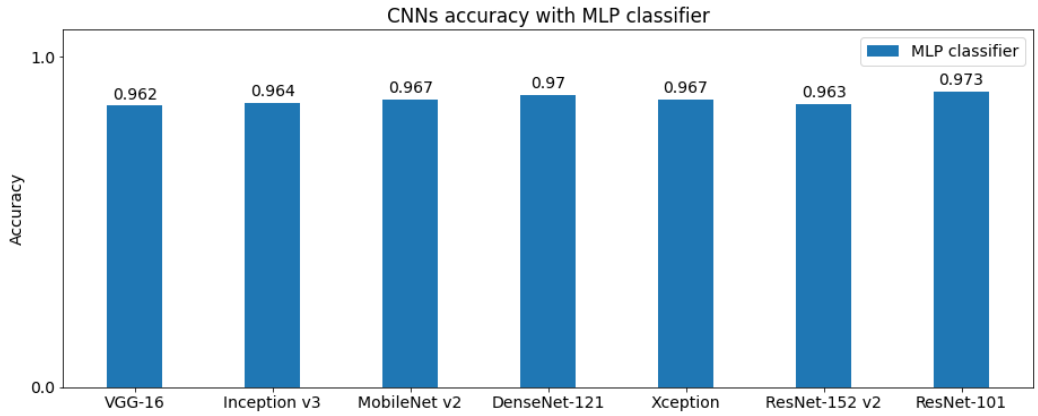


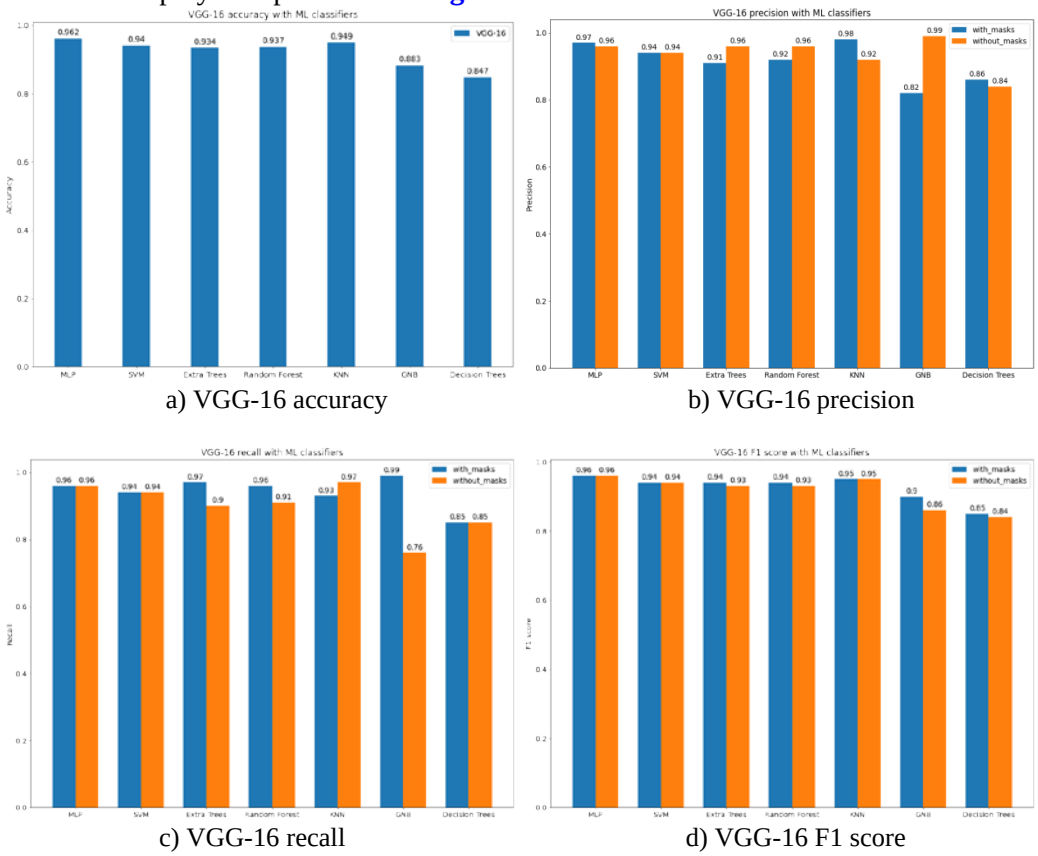
Fig. 5. CNN feature extractors accuracy with MLP classifier

The results as shown in **Fig. 5** indicate that the ResNet-101 feature extractor achieved the highest accuracy with a value of 97.3% whereas, VGG-16 achieved the lowest accuracy with a value of 96.2% among all the employed CNN feature extractors when combined with multi-layer perceptron (MLP) classifier. The results also indicate that DenseNet-121 which is 121 CNN layers deep and ResNet-152 v2 which is 152 CNN layers deep performed below par as compared to ResNet-101 having a 101 CNN layers deep architecture. ResNet-101 with multi-layer perceptron classifier achieved 1% higher classification accuracy as compared to the combination of ResNet-152 v2 and multi-layer perceptron. ResNet-101 performed better than ResNet-152 v2 due to its large size of feature extraction layers filters. Furthermore,

ResNet-101 achieved 0.3% higher classification accuracy as compared to the combination of DenseNet-121 and multi-layer perceptron classifier. The reason behind low accuracy with DenseNet-121 was its dependency on the large dataset required for feature extraction. Moreover, the ResNet-101 and multi-layer combination achieved better accuracy as compared to other tested CNN feature extractors like VGG-16, Inception v3, MobileNet v2, and Xception architecture. The reason behind achieving low accuracy with these architectures was their smaller feature extraction network and ability to generate a lesser number of trainable parameters on training data.

4.4.2 Performance Analysis of VGG-16

On testing the VGG-16 feature extractor with different ML classifiers employed in this work it achieved a significant accuracy of 96.2% with multi-layer perceptron which was highest among all the classifiers employed. The least accuracy of 84.7% was achieved with decision trees as a classifier. The performance of VGG-16 corresponding to different machine learning classifiers employed is presented in **Fig. 6**.



a) VGG-16 accuracy b) VGG-16 precision

c) VGG-16 recall d) VGG-16 F1 score

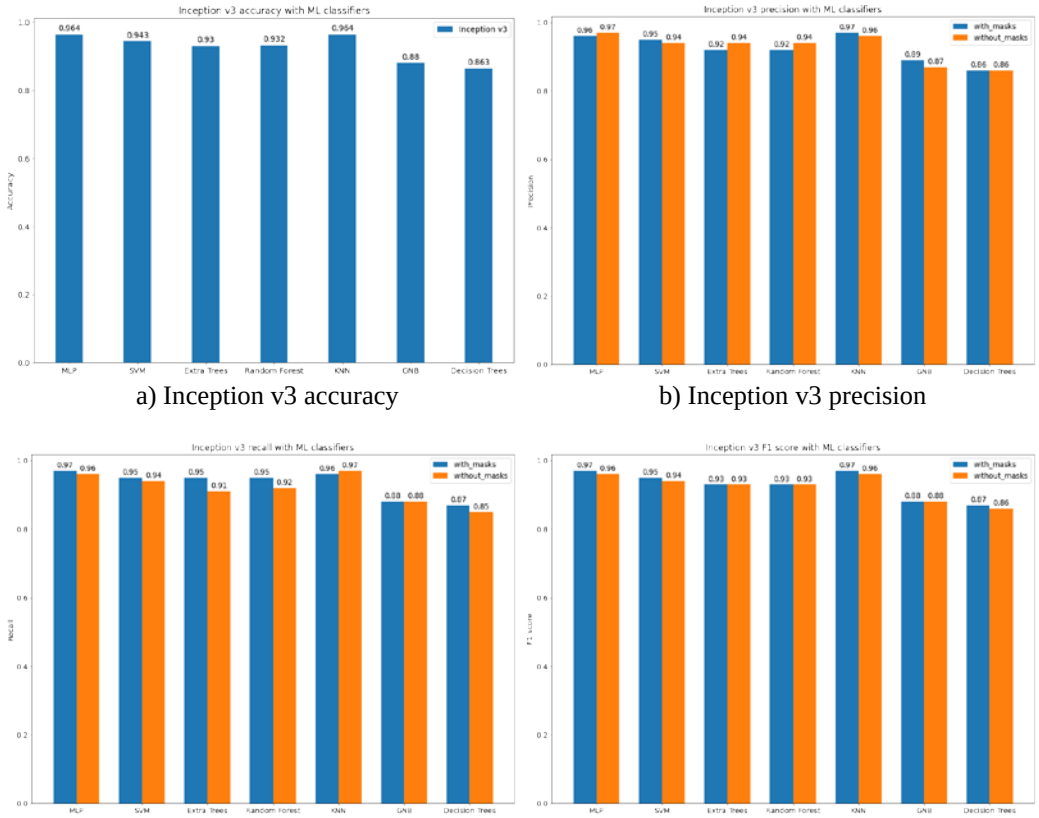
Fig. 6. VGG-16 performance with machine learning classifiers

The VGG-16 feature extractor showed varying results with different machine learning classifiers. For images with masks, it achieved the highest precision with k-nearest neighbors and for images without masks, it achieved the highest precision with gaussian naive bayes. The highest value of recall for images with masks was achieved with gaussian naive bayes and for images without masks, it was obtained with k-nearest neighbors. The highest value of F1 score for images with masks and without masks was obtained with multi-layer perceptron. The

least values for precision and F1 score were achieved with decision trees and least recall was achieved with decision trees and gaussian naive bayes.

4.4.3 Performance Analysis of Inception v3

The test results of Inception v3 with different ML classifiers showed that it achieved the highest accuracy with a value of 96.4% with multi-layer perceptron and k-nearest neighbors classifier respectively. The least accuracy of 86.3% was achieved with decision trees. The performance of Inception v3 corresponding to different machine learning classifiers employed is presented in **Fig. 7**.



a) Inception v3 accuracy b) Inception v3 precision

c) Inception v3 recall d) Inception v3 F1 score

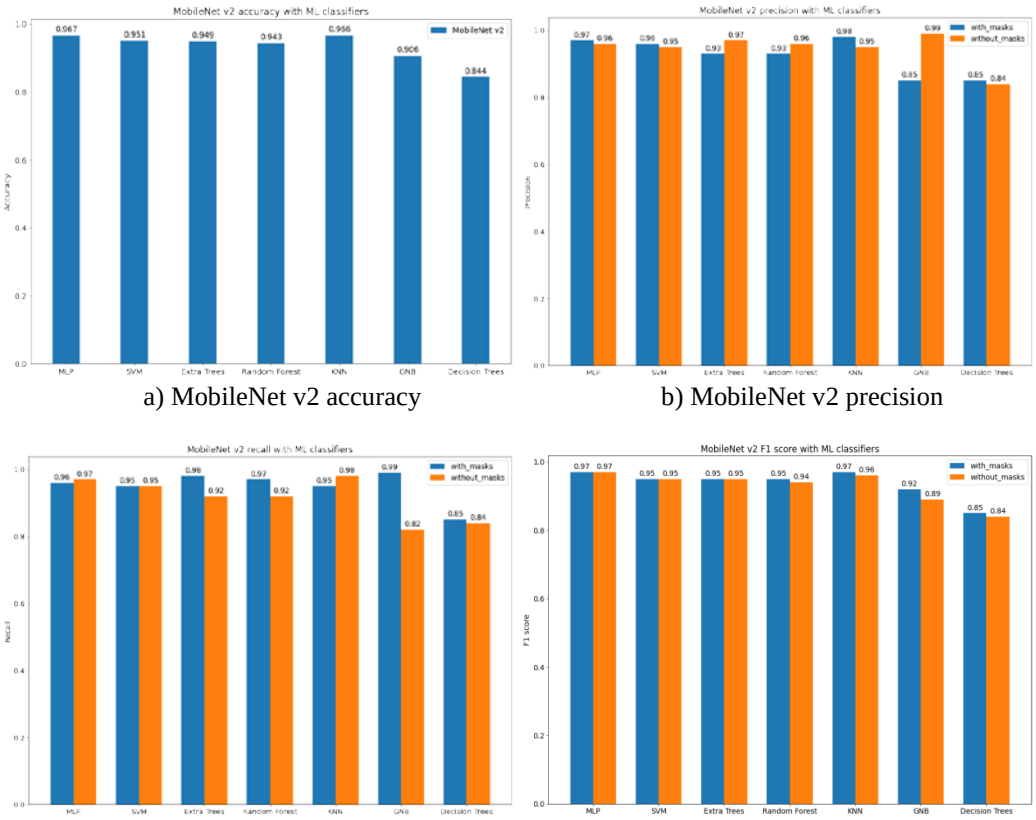
Fig. 7. Inception v3 performance with machine learning classifiers

For images with masks, Inception v3 achieved the highest precision with k-nearest neighbors whereas, for image samples without masks, it achieved the highest precision with multi-layer perceptron. The highest value of recall for images with masks was obtained with multi-layer perceptron and for samples without masks, the highest recall was achieved with k-nearest neighbors. The highest value of the F1 score for both the classes of the dataset was achieved with multi-layer perceptron and k-nearest neighbors. The least values of precision, recall, and F1 score for both the classes of the dataset were achieved with decision trees as a classifier.

4.4.4 Performance Analysis of MobileNet v2

On testing MobileNet v2 feature extractor with different ML classifiers employed in this work it achieved a significant accuracy of 96.7% with multi-layer perceptron which was highest among all the classifiers employed. The least accuracy of 84.4% was achieved with decision

trees as a classifier. The performance of MobileNet v2 corresponding to different ML classifiers employed is illustrated in **Fig. 8**.



a) MobileNet v2 accuracy b) MobileNet v2 precision

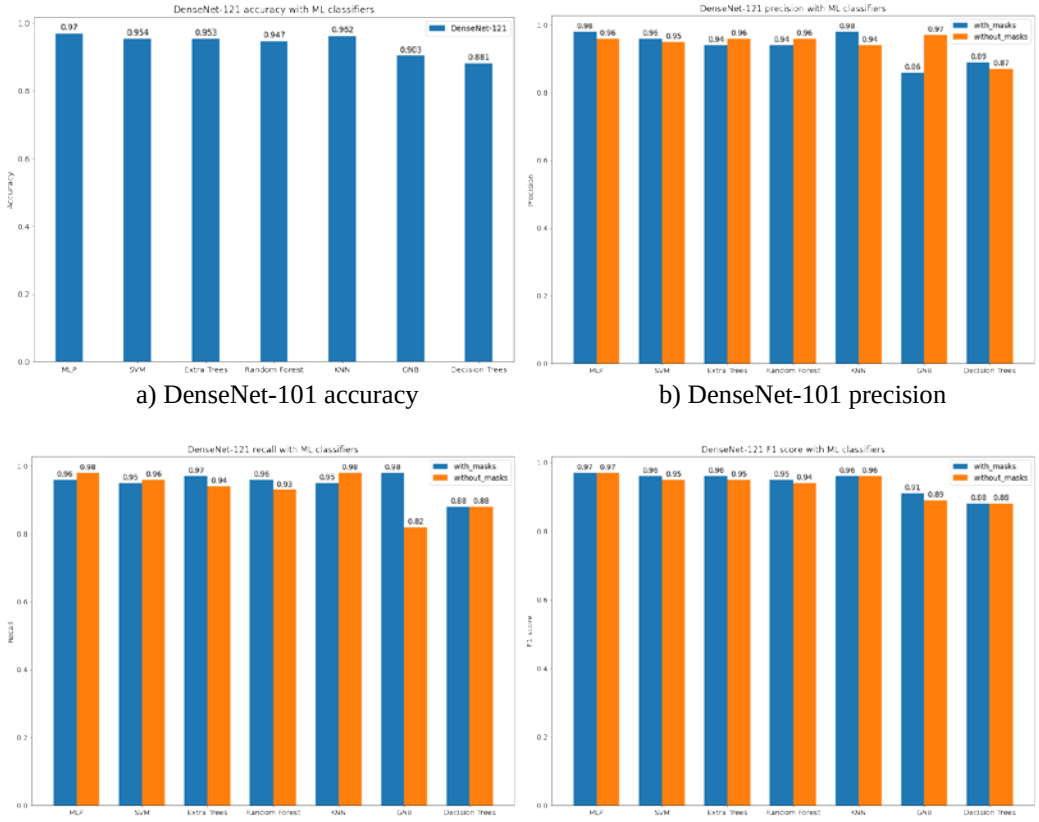
c) MobileNet v2 recall d) MobileNet v2 F1 score

Fig. 8. MobileNet v2 performance with machine learning classifiers

MobileNet v2 achieved the highest precision for images with masks with k-nearest neighbors whereas, for images without masks, it achieved the highest precision with gaussian naive bayes as a classifier. It achieved the highest recall for images with masks with gaussian naive bayes whereas, the highest recall for images without masks was achieved with k-nearest neighbors. The highest value of F1 score for images with masks was achieved with multi-layer perceptron and k-nearest neighbors whereas, for images without masks, the same has been achieved with multi-layer perceptron. The least values for precision, recall, and F1 score were achieved with decision trees as a classifier. Specifically, for images without masks, gaussian naive bayes achieved the least value of recall.

4.4.5 Performance Analysis of DenseNet-121

The test results of DenseNet-121 with different ML classifiers showed that it achieved the highest accuracy with a value of 97% with multi-layer perceptron classifier. However, the minimum accuracy of 88.1% was achieved with the decision trees classifier. The performance of DenseNet-121 corresponding to different ML classifiers employed is presented in **Fig. 9**.



a) DenseNet-101 accuracy b) DenseNet-101 precision

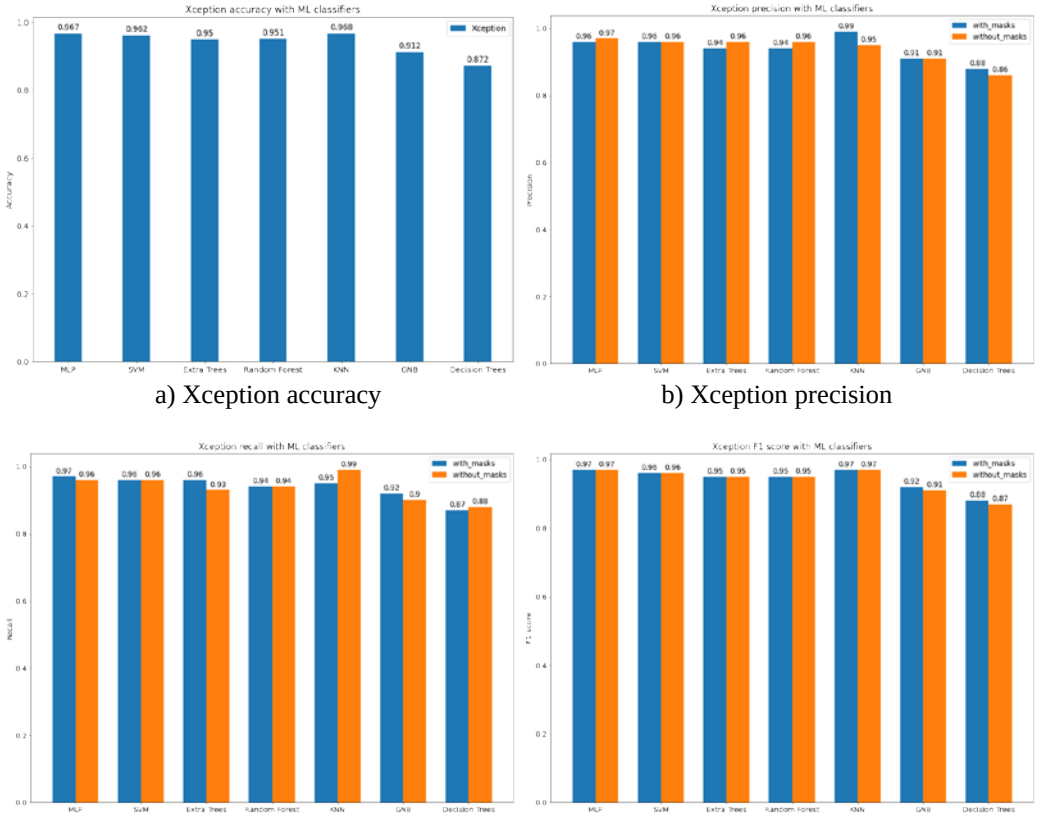
c) DenseNet-101 recall d) DenseNet-101 F1 score

Fig. 9. DenseNet-101 performance with machine learning classifiers

DenseNet-101 achieved the highest precision for images with masks with multi-layer perceptron and k-nearest neighbors whereas, for images without masks, it achieved the highest precision with gaussian naive bayes. The highest value of recall with DenseNet-101 for images with masks was achieved with gaussian naive bayes classifier whereas, for images without masks, the same has been achieved with multi-layer perceptron and k-nearest neighbors. DenseNet-121 achieved the highest F1 score for both the classes of the dataset with multi-layer perceptron as a classifier. The least values for precision, recall, and F1 score were achieved with decision trees classifier. Specifically, for images without masks, the least value of recall was achieved with gaussian naive bayes classifier.

4.4.6 Performance Analysis of Xception

On testing Xception Net feature extractor with different ML classifiers employed in this work it achieved an accuracy of 96.8% with k-nearest neighbors which was 0.1% more than the accuracy achieved with multi-layer perceptron and highest among all the classifiers employed. However, the least accuracy of 87.2% was achieved with decision trees as a classifier. The performance of Xception Net corresponding to different ML classifiers employed is illustrated in **Fig. 10**.



a) Xception accuracy b) Xception precision

c) Xception recall d) Xception F1 score

Fig. 10. Xception performance with machine learning classifiers

Xception Net achieved the highest value of precision for images with masks with k-nearest neighbors whereas, for images without masks, the same has been achieved with multi-layer perceptron. The highest value of recall for images with masks was achieved with multi-layer perceptron and for images without masks, it was achieved with k-nearest neighbors. The highest value of the F1 score was achieved with multi-layer perceptron and k-nearest neighbors for both the classes of the dataset. The least values of precision, recall, and F1 score were achieved with decision trees as a classifier.

4.4.7 Performance analysis of ResNet-152 v2

ResNet-152 v2 on tests with different ML classifiers achieved the highest accuracy of 96.3% with multi-layer perceptron whereas, the least accuracy of 85.2% was achieved with decision trees classifier. The performance of ResNet-152 v2 corresponding to different ML classifiers employed is shown in **Fig. 11**.

ResNet-152 v2 feature extractor achieved the highest value of precision for images with masks with k-nearest neighbors whereas, for images without masks, the same has been achieved with multi-layer perceptron. The highest value for recall for images with masks was achieved with multi-layer perceptron, random forest, and extra trees whereas, for images without masks, the same has been achieved with k-nearest neighbors. The top value of the F1 score for both the classes of the dataset was achieved with multi-layer perceptron and k-nearest neighbors. The least values of precision, recall, and F1 score for both the classes of the dataset were achieved with decision trees as a classifier.

a) ResNet-152 v2 accuracy b) ResNet-152 v2 precision

c) ResNet-152 v2 recall d) ResNet-152 v2 F1 score

Fig. 11. ResNet-152 v2 performance with machine learning classifiers

4.5 Summarization of Findings

Below is the summarization of findings of the experiments conducted that justify the validity of the proposed ResNet-101 and multi-layer perceptron-based hybrid face masks classification technique. Furthermore, observations are highlighted with justifications.

1. Results show that features extracted by ResNet-101 when passed to multi-layer perceptron classifier yield the best result for accuracy achieving a value of 97.3% which was highest among all the experiments performed justifying the validity of the proposed technique.
2. Results indicate that the ResNet-101 and multi-layer perceptron-based technique achieved the highest values for precision, recall, and F1 score for both the classes of the dataset as compared to other tested machine learning classifiers and hybrid classifiers. ResNet-101 and MLP classifier achieved a precision of 98% for images with masks and 97% for images without masks. The recall achieved by the proposed technique for images with masks was 99% and for images without masks, it reached 98%. Furthermore, the F1 score for images with masks and without masks with the proposed technique was 97%. The values of precision, recall, and F1 score with the proposed technique were 1-6% higher as compared to other tested combinations.
3. Results also indicate that multi-layer perceptron tops the accuracy charts when features were passed from other CNN architectures to multi-layer perceptron classifier. With multi-layer perceptron classifier, a significant accuracy was achieved

with VGG-16 - 96.2%; Inception v3 - 96.4%; MobileNet v2 - 96.7%; DenseNet-121 - 97%; Xception - 96.7%; and ResNet-152 v2 - 96.3%. This is due to the ability of multi-layer perceptron to do generalization i.e. the ability to classify an unknown pattern with other known patterns that share the same distinguishing features. The other reason behind high accuracy with multi-layer perceptron classifier was its ability of fault tolerance.

4. The CNN architectures were used for generating the feature vector and classification was performed by the machine learning (ML) classifiers. Each CNN feature extractor has a different number of layers and parameters thus, produced a 2-D feature vector consisting of varying numbers. However, for the task of classification machine learning (ML) classifiers namely, multi-layer perceptron, support vector machine, extra trees, random forest, k-nearest neighbors, gaussian naive bayes, and decision trees were used. To perform classification, machine learning (ML) classifiers were fed with the feature vector generated by the CNN architectures. Amongst all the classifiers, multi-layer perceptron performed better in terms of accuracy because of its ability to classify between non-linearly separable data. Furthermore, the multi-layer perceptron (MLP) classifier can construct its non-linear projection for sparse data and handle the noise better. The other machine learning (ML) classifier does not possess this quality thus. performed weaker as compared to the multi-layer perceptron (MLP) classifier. Therefore, the combination of ResNet-101 generating 4,25,52,832 trainable parameters and multi-layer perceptron (MLP) produced the best results**.**
5. Furthermore, the results indicate that ResNet-101 having a smaller feature extraction network and producing a lesser number of trainable parameters as compared to ResNet-152 v2 achieved better accuracy, precision, recall, and F1 score with almost all the machine learning classifiers. The extra layers in ResNet-152 v2 only contribute to computational power rather than generating the appropriate feature vector to be used for classification by a machine learning classifier.
6. Across the experiments conducted, the performance of the decision trees classifier remained below par. This is due to the reason of instability of decision trees, that is, a small change in the data can lead to a large structural change in the optimal decision tree, and in the case of classification of categorical variables with a different number of levels, information gain in decision trees bias towards attributes with more levels.

From the results, this can be summarized that ResNet-101 as backend feature extraction architecture can be combined with multi-layer perceptron to create an end-to-end product for the classification of people wearing face masks for low-end devices such as facial identification systems and surveillance systems with CCTV cameras in uncontrolled environments.

4.6 Comparison with Existing Work

In order to get more intuitive results for the proposed ResNet-101 and MLP classifier-based face masks classification technique, we tested and compared it with other related work present in the literature. The authors Loey et al. [13] proposed a ResNet-50 and SVM based technique which was tested on the RMFD dataset and achieved a classification accuracy of 99.49%. In this work, we have employed ResNet-101 which is a successor of ResNet-50, and applied a dense layer to reduce the number of trainable parameters and make the technique to get train in a lesser time, and utilize lesser computation resources such as, RAM and GPU. The comparative results of the proposed technique with ResNet-50 and SVM [13] are shown in **Table 2**.

Table 2. Comparison with related work

Work	Technique	Dataset	Dense layer	Accuracy
Loey et al. [13]	ResNet-50 + SVM	Custom (Ours)	Yes (1024)	94.7%
Proposed	ResNet-101 + MLP	Custom (Ours)	Yes (1024)	**97.3%**
Loey et al. [13]	ResNet-50 + SVM	RMFD	No	99.49%
Proposed	ResNet-101 + MLP	RMFD	No	**99.86%**

As shown in **Table 2**, the proposed ResNet-101 and MLP based technique outperform ResNet-50 and SVM based technique when tested with and without a dense layer on the dataset employed in this work and the RMFD dataset. The results achieved justify the validity of the proposed technique and dataset. Furthermore, from the results obtained this can also be concluded that ResNet-101 and MLP with a reduced number of parameters obtained after applying a dense layer perform better as compared to ResNet-50 and SVM based face masks classification technique thus, is suitable for devices with low computation resources.

As an effort to validate the proposed technique and employed dataset, we carried out two ablation studies by training and testing the proposed technique on the benchmark SMFD [23] dataset and LFW [25] dataset. The SMFD dataset is composed of 1,570 images where 785 images are for faces with simulated masks and 785 images are for faces without masks. To carry out the ablation study on the SMFD dataset, we labeled faces with simulated masks as class with_masks and faces without masks as class without_masks. The LFW dataset consists of 5,749 images of faces without masks. However, to carry out the task of classification of faces with and without masks, we extracted 1,950 images from the LFW dataset and labeled them as class without_masks, and utilized 2,000 images for faces with masks as present in the dataset employed to carry out this work and labeled as with_masks.

For the two ablation studies, we trained and tested the ResNet-101 and multi-layer perceptron based proposed technique on the SMFD dataset and custom LFW dataset and evaluated for accuracy metric. The comparative results of the proposed technique on the dataset employed in this work, the SMFD dataset, and the custom LFW dataset are presented in **Table 3**.

Table 3. Test results on SMFD and custom LFW dataset

Technique	Dataset	Dense layer	Accuracy
ResNet-101 + MLP	SMFD	Yes (1024)	95.83%
ResNet-101 + MLP	LFW (Custom)	Yes (1024)	96.14%
ResNet-101 + MLP	Custom (Ours)	Yes (1024)	97.3%

As shown in **Table 3**, the proposed ResNet-101 and multi-layer perceptron-based face masks classification technique achieved an accuracy of 95.83% on the SMFD dataset, an accuracy of 96.14% on the custom LFW dataset, and an accuracy of 97.3% on the dataset employed in this work. For the two ablation studies, we have used a dense layer of size 1024 as we have proposed our technique with the dense layer. There was a difference of 1.16-1.47% when the proposed technique was tested on the SMFD dataset, custom LFW dataset, and the dataset employed in this work. The difference is due to the reason of lesser number of images in the SMFD dataset and the custom LFW dataset. However, better results could be achieved if these datasets consisted of a large number of images. The results achieved on the SMFD dataset and custom LFW dataset justify the validity of the proposed technique and the dataset as the proposed technique has achieved close values for accuracy on exploited datasets.

5. Conclusions and Future Work

This work has proposed a hybrid face masks classification technique by combining ResNet-101 with classical multi-layer perceptron (MLP) classifier. The proposed technique achieved a classification accuracy of 97.3% on the employed self-created face masks classification dataset. To validate the effectiveness of the proposed technique, six other deep neural network-inspired CNN feature extractors namely, VGG-16, Inception v3, MobileNet v2, DenseNet-121, Xception Net, and ResNet-152 v2 were tested with multi-layer perceptron and evaluated for classification accuracy. The result of the tests highlighted that a combination of ResNet-101 and multi-layer perceptron performs best as a hybrid combination for face masks classification. Furthermore, to scale up the domain of face masks classification and embrace the proposed technique, a total of forty-nine experiments were conducted by employing seven deep neural networks inspired CNN feature extractors and seven classical machine learning classifiers and evaluated based on performance metrics. The results of the experiments showed ResNet-101 as an effective feature extractor and multi-layer perceptron as a stable classifier. The ResNet-101 feature extractor achieved 1% higher accuracy as compared to its successor and other tested CNN architectures. Furthermore, the proposed technique achieved a 1-6% higher precision, recall, and F1 score as compared to other tested combinations. The multi-layer perceptron classifier achieved accuracy comparable to its combination with ResNet-101 with CNN feature extractors namely, DenseNet-121, Xception Net, MobileNet v2, Inception v3, ResNet-152 v2, and VGG-16. To carry out this work, a dense layer of size 1024 was applied with all the exploited CNN-based feature extractors to reduce the size of the feature vector, therefore, better accuracy can be achieved by removing the dense layer and passing the full-scale feature vector to the machine learning classifiers. This was done to develop a technique under minimal dependency on high-end computation resources. Future work can be extended to the use of similar strategies for face mask detection using algorithms like YOLO, SSD, etc. The methodology can further be extended to the use of Generative Adversarial Networks for creating new proposals for face masks and faces to get more intuitive results.

Acknowledgements

The authors are thankful to All India Council of Technical Education, India for funding this work. This research work is funded under Research Promotion Scheme of AICTE, India vide file no. 8-108/FDC/RPS (POLICY-1/2019-20).

References

[1] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 770-778, 2016. Article (CrossRef Link)

[2] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv*, 2015. Article (CrossRef Link)

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR*), Las Vegas, USA, pp. 2818-2826, 2016. Article (CrossRef Link)

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 4510-4520, 2018. Article (CrossRef Link)

[5] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, "Densely connected convolutional networks," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 2261-2269, 2017. Article (CrossRef Link)
[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 1800-1807, 2017. Article (CrossRef Link)
[7] B. Hamdan, K. Mokhtar, "The detection of spoofing by 3D mask in a 2D identity recognition system," *Egyptian Informatics Journal*, vol. 19(2), pp. 75-82, 2018. Article (CrossRef Link)
[8] Q. Chen, L. Sang, "Face-mask recognition for fraud prevention using Gaussian mixture model," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 795-801, 2018. Article (CrossRef Link)
[9] J.G. Sanders, Y. Ueda, K. Minemoto, E. Noyes, S. Yoshikawa, R. Jenkins, "Hyper-realistic face masks: a new challenge in person identification," *Cognitive Research: Principles and Implications*, vol. 2, 2017. Article (CrossRef Link)
[10] M.S. Ejaz, M.R. Islam, M. Sifatullah, A. Sarker, "Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition," in *Proc. of 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, BANGLADESH, pp. 1-5, 2019. Article (CrossRef Link)
[11] J.S. Park, Y.H. Oh, S.C. Ahn, S.W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(5), pp. 805-811, 2005. Article (CrossRef Link)
[12] A. Nieto-Rodríguez, M. Mucientes, V.M. Brea, "System for medical mask detection in the operating room through facial attributes," in *Proc. of 7th Iberian Conference, IbPRIA 2015*, Santiago de Compostela, SPAIN, pp. 138-145, 2015. Article (CrossRef Link)
[13] M. Loey, G. Manogaran, M.H.N. Taha, 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: Journal of the International Measurement Confederation*, vol. 167, 108288, 2021. Article (CrossRef Link)
[14] B. Qin, D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19," *Sensors*, vol. 20(18), p. 5236, 2020. Article (CrossRef Link)
[15] S. Li, X. Ning, L. Yu, L. Zhang, X. Dong, Y. Shi, W. He, "Multi-Angle Head Pose Classification when Wearing the Mask for Face Recognition under the COVID-19 Coronavirus Epidemic," in *Proc. of 2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, Shenzhen, CHINA, pp. 1-5, 2020. Article (CrossRef Link)
[16] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, J. Hemanth, "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2," *Sustainable Cities and Society*, vol. 66, p. 102692, 2021. Article (CrossRef Link)
[17] N. Ud Din, K. Javed, S. Bae, J. Yi, "A Novel GAN-Based Network for Unmasking of Masked Face," *IEEE Access*, vol. 8, pp. 44276-44287, 2020. Article (CrossRef Link)
[18] S.A. Hussain, A. Salim Abdallah Al Balushi, "A real time face emotion classification and recognition using deep learning model," *Journal of Physics: Conference Series*, vol. 1432, p. 012087, 2020. Article (CrossRef Link)
[19] M. Inamdar, N. Mehendale, "Real-Time face mask identification using face masknet deep learning network," *SSRN Electronic Journal*, 2020. Article (CrossRef Link)
[20] P. Khandelwal, A. Khandelwal, S. Agarwal, "Using computer vision to enhance safety of workforce in manufacturing in a post covid world," *arXiv*, 2020. Article (CrossRef Link)
[21] M. Jiang, X. Fan, "Retinamask: A face mask detector," *arXiv*, 2020. Article (CrossRef Link)
[22] C. Li, R. Wang, J. Li, L. Fei, "Face detection based on YOLOv3," in *Proc. of Intelligent Computing, Communication and Devices*, SINGAPORE, pp. 277-284, 2020. Article (CrossRef Link)

[23] 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, J. Liang, "Masked face recognition dataset and application," *arXiv*, 2020. Article (CrossRef Link)
[24] Google API. [Online]. Available: https://pypi.org/project/google_images_download/
[25] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. of the 10th European Conference on Computer Vision*, Marseille, FRANCE, 2008. Article (CrossRef Link)

Akhil Kumar is pursuing Ph.D. in Computer Science from Himachal Pradesh University, Shimla, India. He has obtained B.Tech and M.Tech in Information Technology from USIT, GGS Indraprastha University, New Delhi, India. His research areas include deep learning, computer vision and object detection.

Arvind Kalia is serving as Professor in Department of Computer Science, Himachal Pradesh University, Shimla, India. He has obtained Ph.D. from Punjabi University, Patiala, Punjab, India. His research areas include software engineering, deep learning and computer vision.

Kinshuk Verma is pursuing B.E in Computer Science & Engineering from UIET, Panjab University, Chandigarh, India. His research areas include deep learning and object detection.

Akashdeep Sharma is serving as Assistant Professor in Department of Computer Science & Engineering, UIET, Panjab University, Chandigarh, India. He has obtained Ph.D. from Guru Nanak Dev University, Amritsar, Punjab, India. His research areas include video analytics, deep learning, computer vision and object detection.

Manisha Kaushal is serving as Assistant Professor in Computer Science & Engineering Department, Thapar Institute of Engineering & Technology, Patiala (Derabassi Campus), Punjab, India. She has obtained Ph.D. from I.K. GujralPunjab Technical University, Jalandhar, Punjab, India. Her research areas include image processing, machine learning and deep learning.

Aayushi Kalia is pursuing M.E in Computer Science & Engineering from Computer Science & Engineering Department, Thapar Institute of Engineering & Technology, Patiala, Punjab, India. She has obtained B.E. in Computer Science & Engineering from the same institute. Her research areas include image processing, machine learning and object detection.