Comparison of Convolutional Neural Network Architectures for Face Mask Detection

Siti Nadia Yahya[1]
Postgraduate Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Aizat Faiz Ramli[2*]
Electronics Technology Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Muhammad Noor Nordin[3]
Medical Engineering Technology Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Hafiz Basarudin[4]
Communication Technology Section
Universiti Kuala Lumpur British Malaysian Institute
Batu 8, Jalan Sungai Pusu, 53100, Selangor, Malaysia

Mohd Azlan Abu[5]
Malaysia-Japan International Institute of Technology
Universiti Teknologi Malaysia
Jalan Sultan Yahya Petra, 54100, Kuala Lumpur, Malaysia

***Abstract*—In 2020 World Health Organization (WHO) has declared that the Coronaviruses (COVID-19) pandemic is causing a worldwide health disaster. One of the most effective protections for reducing the spread of COVID-19 is by wearing a face mask in densely and close populated areas. In various countries, it has become mandatory to wear a face mask in public areas. The process of monitoring large numbers of individuals to comply with the new rule can be a challenging task. A cost-effective method to monitor a large number of individuals to comply with this new law is through computer vision and Convolution Neural Network (CNN). This paper demonstrates the application of transfer learning on pre-trained CNN architectures namely; AlexNet, GoogleNet ResNet-18, ResNet-50, ResNet-101, to classify whether or not a person in the image is wearing a facemask. The number of training images are varied in order to compare the performance of these networks. It is found that AlexNet performed the worst and requires 400 training images to achieve Specificity, Accuracy, Precision, and F-score of more than 95%. Whereas, GoogleNet and Resnet can achieve the same level of performance with 10 times fewer number of training images.**

***Keywords*—Convolution neural network; deep learning; transfer learning; computer vision; facemask detection; COVID-19**

I. Introduction

Wearing face masks in public area is becoming more common due to the prevalence of COVID-19 outbreak all over the world [1]. Before the pandemic, small minority of the population especially in east Asian countries have been wearing face masks as a prevention against common flu. COVID-19 is the most recent pandemic virus to make a huge impact on human health in the past century. The exponential rate of COVID-19 transmission has forced the World Health Organization (WHO) to declare COVID-19 a worldwide pandemic in 2020. 150,047,341 have been infected by COVID-19 as of April 2021 across 188 countries. The virus is spreading through close contact, as well as in overcrowded publica areas. In multiple countries, individuals are constrained by the law to wear face mask in public areas. The rule was implemented as a reaction to a sudden spike in cases and fatalities in a various country. To enforce the public to comply with this rules, governmental agencies such police and health agency have to allocate significant number of their workforce to continuously public areas.

This paper demonstrates the application of Convolution Neural Network CNN to automate the classification of images of an individual wearing facemask and those without facemask. The ability to automate facemask detection can significantly reduce man power requirement and governmental expenditure. This paper also presents the performance comparison of popular CNN architectures for images classifications namely AlexNet, GoogleNet, ResNet-18, ResNet-50, and ResNet-101. The results presented in this research can be used by other researchers and machine learning engineers to identify suitable CNN architectures given the number of training data set, hardware capabilities and required accuracy to automate the images classification of a person wearing face mask.

This paper is organized as follows. Section II, provides discussion on the findings by other researchers on facemask detection and classification. Section III, describes the methodology on how the five different CNN architectures; AlexNet, GoogleNet, ResNet-18, ResNet-50 and ResNet-101 are being trained. Section IV, discussed the performance evaluation metrics that were used to compare the 5 different CNN architectures. The results of the study are presented and discussed in Section V. Finally, conclusions are drawn.

[*] Corresponding Author

A. Deep Learning

Deep Learning is a subset of Machine Learning, which in turn is a subset of Artificial Intelligence. The term "Artificial Intelligence" (AI) refers to techniques that enable computers to mimic human behavior. Machine Learning is an algorithm that has been trained using data to emulate human like decision making. Deep Learning and Artificial Neural Network are a subset of Machine Learning that are inspired by human brain structure. [2] Deep learning methods take the opportunity to achieve the same findings as humans by consistently assessing data using a specified structured methodology. Deep learning does this through the use of a multi-layered structure of algorithms known as neural networks.

As shown in Fig. 1, the design of the Artificial Neural Network is based on the anatomy of the human brain. Artificial Neural Network can be trained to detect objects and categorize various sorts of data in the same way that humans do. Singular layers of neural organizations might be considered as a type of filter that capacities from coarse to fine, expanding the likelihood of recognizing and creating the right outcome. The human brain works in a similar way. While going up against with new data, the brain endeavors to contrast it with recently known objects. Deep neural networks employ the same principle. It may be use neural networks to accomplish a variety of tasks such as grouping, classification, and regression. Can be use neural networks to categorize or classify unlabeled data based on similarities between samples. In the classification process, it might prepare the network on a labeled dataset to characterize the examples in the dataset into discrete classifications.

B. Convolution Neural Networks (CNN)

As illustrated in Fig. 2 and Fig. 3, a CNN contains three layers as a convolutional layer, a pooling layer, and a fully connected layer [3]. The 'input layer' of each CNN utilized receives pictures and recompress them before passing data on to following layers for extracting features. The next layers are referred to as 'Convolution layers,' and they serve as image filters, extracting features from pictures and generating match local features during testing. Activation function 'Rectified Linear Unit' (ReLU) is employed to replace every negative integer in the pooling layer with zero. ReLU also helps the CNN maintain mathematical stability by avoiding learnt values from being stuck at zero or blowing up toward infinity. Then the data is transferred to the 'pooling layer' [4]. This layer decreases the size of large pictures yet retaining the most important information. It maintains the most value from each frame by retaining the best fits of every feature within the frame. Flatten is the way toward changing over information into a one-dimensional cluster for contribution to the next layer. The yield of the convolutional layers is flattened to make a solitary extensive component vector. It is also connected to the final classification algorithm, forming a fully connected layer. The next-to-last layer is a fully connected layer that turns the greater filtered pictures into labeled with probability for every class of every picture being categorized. To give classification output, the last layer of the CNN architecture employs a classification layer such as softmax [5].

C. AlexNet

AlexNet was created by Alex Krizhevsky and is a convolutional neural network model that has made huge contributions to Artificial Intelligence (AI), particularly the use of deep learning to machine vision [6]. The CNN model won the ImageNet Large Scale Visual Detection (ILSVRC) competition in 2012, which evaluates methods for huge object recognizing and picture classifications. AlexNet consists of 60 million parameters, three fully connected layers, 650,000 neurons and five convolutional layers [7]. Convolutions of 11x11, 5x5, and 3x3 dimensions were used, as well as max pooling, dropout, data augmentation, ReLU activations, and SGD with momentum. The initial two convolutional layers are normalization and a maximum pooling layer. Plus, the third and fourth are directly connected while the fifth is joined by a maximum pooling layer. [6] The input is given it into softmax classifier, the second of which feeds into a softmax classifier. The authors used a regularisation approach termed "dropout" with a ratio of 0.5 to avoid overfitting in the fully-connected layers. The use of Rectified Linear Unit (ReLU) on each of the first seven layers is another AlexNet model feature.

D. GoogleNet

Szegedy proposed GoogleNet, which was the champion of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014 [6]. For the major auxiliary classifiers in the network, GoogleNet features four convolutional layers, three softmax layers, seven million parameters, five fully connected layers, nine inception modules, three average pooling layers and four max-pooling layers. In the fully connected layer, dropout regularization is used, and ReLU activation is used in all of the convolutional layers. GoogleNet has 22 total layers and is significantly deeper and wider than AlexNet, but it has a much smaller number of network parameters. The DistBelief distributed machine learning system was used to train GoogLeNet architecture with a small amount of model and data parallelism. In the RGB color space, the size of the receptive field in this network is 224x224 with a zero mean. GoogleNet was developed with the intent of being able to function on a variety of devices, including those with limited computational resources, such as those with a low memory footprint [6]. GoogleNet is 22 layers deep if just layers with parameters are counted, or 27 levels if pooling is counted, and has 7 million parameters. This network has a 27MB file size and a 224-by-224 image input size. The GoogLeNet was intended to be a computational force and reckoned with higher computational proficiency than a portion of its archetypes or equivalent organizations created at that point. The main convolution layer utilizes a filter patch that is altogether huge in contrast with other patch sizes in the network. The significant objective of this layer is to quickly limit the input image while holding spatial data by utilizing enormous filter sizes. The size of the input image is diminished by a factor of four at the subsequent convolution layer and another factor of eight preceding arriving at the primary initiation module, however a more noteworthy number of highlight maps are produced. The GoogLeNet architecture is comprised of nine inception modules. Furthermore, some inception modules include two max-pooling layers.

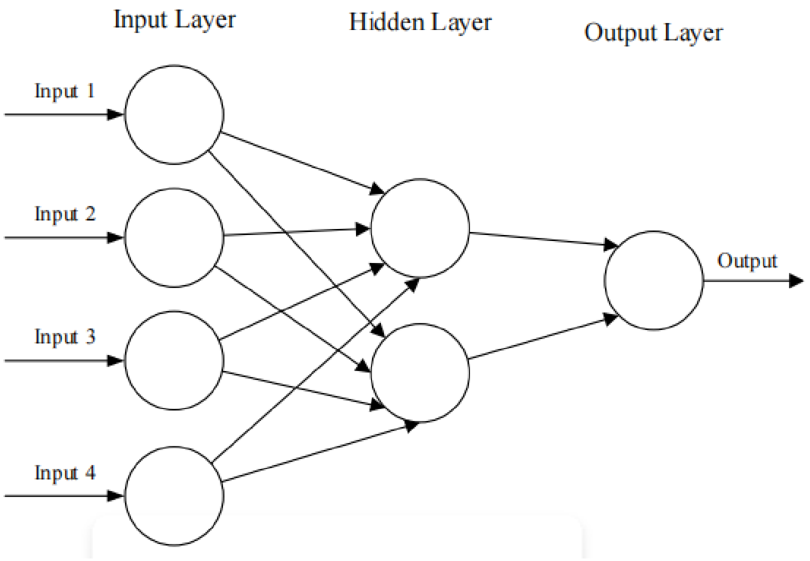


Fig. 1. Artificial Neural Network Connection [8].

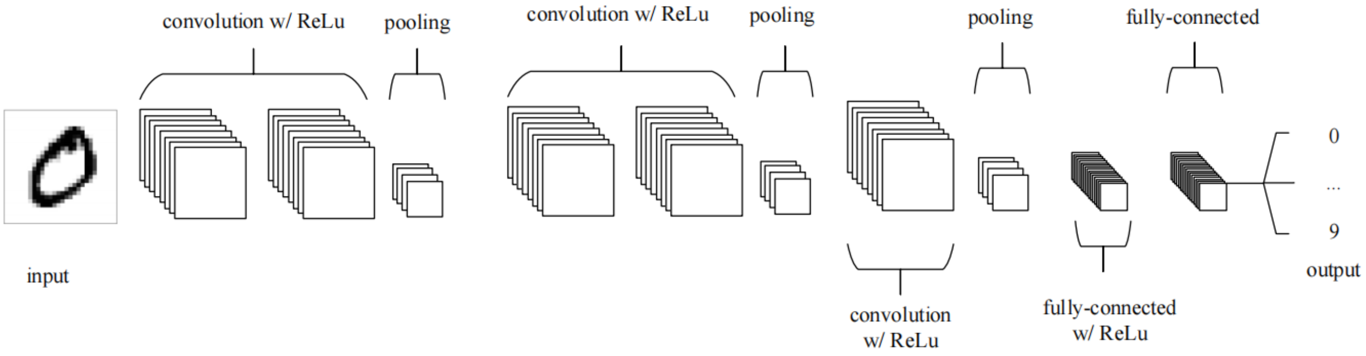


Fig. 2. Convolutional Neural Networks [8].

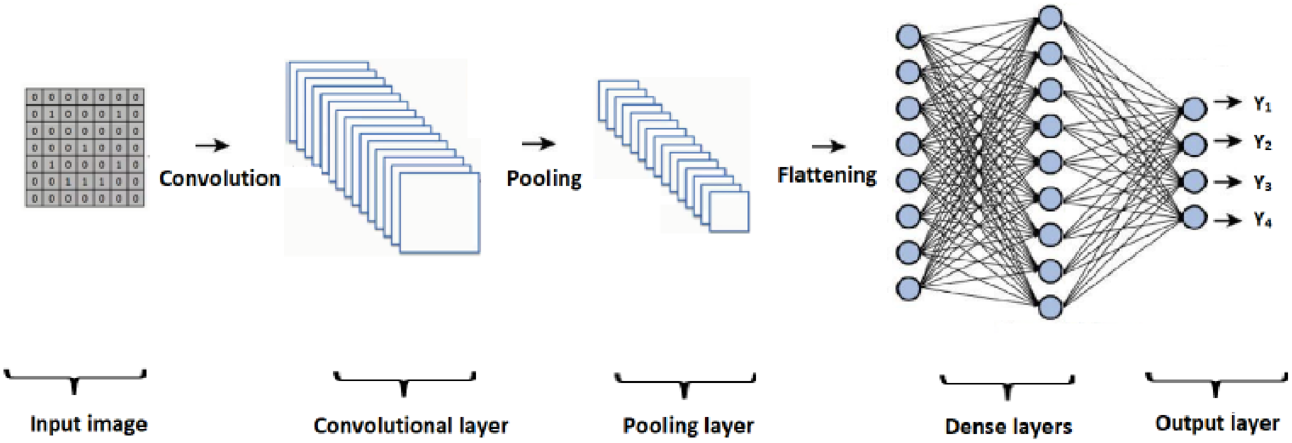


Fig. 3. Representation of a Convolutional Neural Network's Architecture [6].

Just before the linear layer, a dropout layer is being utilized. The dropout layer is a regularization procedure utilized during preparing to keep the network from overfitting. The linear layer is comprised of 1000 hidden units that compare to the 1000 classes in the Imagenet dataset. The last layer is the softmax layer, which utilizes the softmax function, an actuation function used to assess the likelihood circulation of an assortment of the number contained inside an information vector. A softmax actuation function is a vector wherein the assortment of qualities addresses the likelihood of a class or occasion event. The vector's qualities all amount to one.

E. ResNet (Residual Network)

The ResNet feature is based on deep architectures that have demonstrated good convergence and accuracy, were created by He et al. [6] They won the ImageNet Large Scale Visual Recognition Challenge classification competitions in 2015. A residual neural network (ResNet) is a form of Artificial Neural Network (ANN) that is based on pyramidal cell frameworks in the cerebral cortex. Skip connections, or shortcuts, are used by residual neural networks to skip over some layers. ResNet was created using numerous stacked residual units and a variety of layer counts such as 18, 34, 50, 101, 152, and 1202. The number of operations, on the other hand, might vary depending on the architecture. The residual units for all of the preceding are made up of convolutional, pooling, and layering operations. The ResNet 18 network provides an excellent balance of depth and performance, and it is made up of a fully connected layer with a softmax, five convolutional layers, and one average pooling layer. ResNet-18 comprises 11.7 million parameters and an 18 depth layer with a size of 44MB.

Resnet-50 has 49 convolutional layers, 25.6 million parameters, 50 depth layer with a size of 96MB and a fully connected layer at the end of the network. [9] ResNet-101 is a deep convolutional neural network with 101 layers. This architecture has 44.6 million parameters and 101 depth layers, and it is 167MB in size. It will load a pre-trained rendition of the network that has been trained on over 1,000,000 images from the ImageNet information base. The network was pre-trained to recognize images into 1000 distinctive item classes. ResNet is a reliable architecture for identifying a wide range of classes, having won the 2016 ImageNet competition.

A deep residual network (ResNet) is made up of modules, which are entities with identical loops layered on top of each other. Each module is made up of multiple convolutional layers that are used to become familiar with the features of the input space. After the second convolutional layer, a dropout layer was added. Each module delivers more generalized output with greater regularization with the inclusion of the Dropout layer. Many architectures in the literature use dropout, and it is often used on layers with a large number of parameters to minimize feature adaptation and overfitting. Dropouts outperform in generalization. As a default, Softmax is utilized after fully connected layers.

II. Related Work

This section discusses similar research that has been conducted relating to face mask detection.

A. Facial Mask Detection using Semantic Segmentation

The objective of the paper presented by Meenpal et al., 2019 [10] was to develop a binary face classifier that can identify each face in the frame regardless of alignments, including a strategy for generating accurate face segmentation masks of any arbitrary size input image. The approach begins with an RGB image of any size and utilizes Predefined Training Weights of VGG – 16 Architecture for feature extraction. For segmented face masks, experiments on the Multi Parsing Human Dataset revealed a mean pixel-level efficiency of 93.884%.

B. Real-Time Face Mask Identification using Facemasknet Deep Learning Network

Inamdar & Mehendale, 2020 develop a deep learning architecture called Facemasknet [11] COVID-19 face mask classification. The proposed architecture provides three characterizations which are people wearing a mask, erroneously worn masks, and no mask detected. Utilizing a deep learning technique called Facemasknet, they got a precision of 98.6%.

C. Covid-19 Facemask Detection with Deep Learning and Computer Vision

Vinitha & Velantina, 2020 developed a real-time face detection from a live feed via their webcam [12]. The research was conducted using OpenCV framework and A.I framework such as Python, Tensor Flow and Keras. Their aim is to employ deep learning and computer vision to determine if the person in the picture or video feed is wearing a mask.

D. Deep Learning based Safe Social Distancing and Face Mask Detection in Public Areas for COVID-19 Safety Guidelines Adherence

Yadav, 2020 presents a technique for forestalling the transmission of the virus by observing individuals progressively to check whether they are utilizing safe social distance and wearing face mask in public area [13]. The method used for this research includes Raspberry pi4, OpenCV framework, MobileNetV2 and TensorFlow. The detection of face mask wearing achieves an accuracy of 91.2%.

E. A Hybrid Deep Transfer Learning Model with Machine Learning Methods for Face Mask Detection in the Era of the COVID-19 Pandemic

Loey et al., 2021 presents a hybrid model for face mask identification based on deep and conventional machine learning. There are two elements to the technique that follows [1]. The first element will use Resnet-50 to extract features. The ensemble method, decision trees, and Support Vector Machine (SVM) are used in the second element to categorise face masks.

F. Validating the Correct Wearing of Protection Mask by Taking a Selfie: Design of a Mobile Application 'CheckYourMask' to Limit the Spread of COVID-19

The COVID-19 contagiousness is considered to be high in comparison to the flu. Hammoudi et al., 2020 presents a mobile application design that allows anybody with a smartphone the ability to snap a picture to verify that his or her protective mask is properly positioned on his or her face [14]. The technique used in this research included Android, OpenCV, and Haar-like. True Detection (TD) accuracy for the face is 99.92% and the nose is 100%.

G. Identifying Facemask-Wearing Condition using Image Super-Resolution with Classification Network to Prevent COVID-19

Qin & Li, 2020 proposed another facemask wearing condition identification technique as a cooperation with picture super-resolution with classification algorithm (SRCNet) to quantify three different classification issues using unconstrained 2D facial image photos [15]. The suggested technique included four major steps which are pictures pre-processing, image recognition and cropping, picture super-resolution and recognition of wearing a face mask circumstances. The proposed technique reported a 98.7% accuracy rate.

H. An Application of Mask Detector for Prevent Covid-19 in Public Services Area

Henderi et al., 2020 created systems for real-time monitoring of those who do not wear a face masks in public areas [16]. The authors utilize images and video input from a camera and connects it to a Speed Maix Bit CPU to process data and show it onto the LCD display. The materials used to develop the system are MicroPython, Sipeed Maix Bit, MaixPy and Python 3.

I. An Automated System to Limit COVID-19 using Facial Mask Detection in Smart City Network

Rahman et al., 2020 propose a framework that limits COVID-19 spread in an active city network where all open spots are monitored by Closed-Circuit Television (CCTV) cameras by recognizing people who are not wearing any facial masks [4]. When an individual without a face mask is detected, the city network alerts the necessary authorities. The materials used for this system is CCTV and GPS. The system achieved an accuracy rate of 98.7% of facemask detection.

J. Retinamask: A Face Mask Detector

Jiang & Fan, 2020 presented RetinaFaceMask, a high-accuracy and effective face mask detector [17]. The presented RetinaFaceMask detector is a one-stage detector comprised of a feature pyramid network that combines high-level semantic information with numerous feature vectors and a novel context attentive modules focused on identifying face masks. Furthermore, they provide a cross-class object removal approach for rejecting predictions with low confidence and a high intersection of a union. The framework used in this research were MobileNet, ResNet, and PyTorch.

K. Performance Evaluation of Intelligent Face Mask Detection System with Various Deep Learning Classifiers Keywords

The research by [18] focuses on the use of deep learning algorithms to identify persons who do and do not wear a face mask. The framework has been trained to decide if an individual is wearing a face mask or not. At the point when the algorithm perceives an individual without a mask, an alarm will be set off to caution the people around or the relevant authorities, so that appropriate action may be taken against such offenders. Like most establishments, associations, businesses, shopping centers, and hospital must resume normal operations before the epidemic is removed, to incorporate a face mask recognition strategy with the current information system at the passage and leave entryways is emphatically encouraged.

III. METHODOLOGY

The main objectives of this research are to demonstrates the application of using existing CNN architectures; ResNet-101, ResNet-50, ResNet-18, GoogleNet and AlexNet in classifying images of an individual wearing and not wearing facemask. The first stage is to conduct transfer learning on the networks using image datasets. The most essential criterion for evaluating the performance is to see if the prediction accuracy varies among all CNN architectures used in this research. This study is divided into two phases which are training and testing of the face mask detector.

In the training phase, the dataset is loaded for the model to be trained, and the model is serialized. The training datasets consist of images of faces with and without facemask. Fig. 4, 5, 6, 7 and 8 shows example of training images of faces at various angles without facemask. To gauge and compare the performance of ResNet-101, ResNet-50, ResNet-18, GoogleNet and AlexNet, the dataset were varied as follows:

- 20 datasets (10 images with facemask, 10 without facemask)
- 40 datasets (20 images with facemask, 20 without facemask)
- 60 datasets (30 images with facemask, 30 without facemask)
- 80 datasets (40 images with facemask, 40 without facemask)
- 100 datasets (50 images with facemask, 50 without facemask)
- 200 datasets (100 images with facemask, 100 without facemask)
- 300 datasets (150 images with facemask, 150 without facemask)
- 400 datasets (200 images with facemask, 200 without facemask)

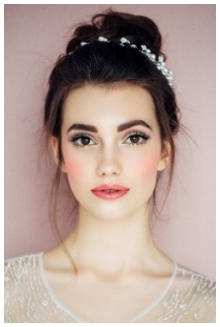


Fig. 4. Example on Front Facing Image [19].

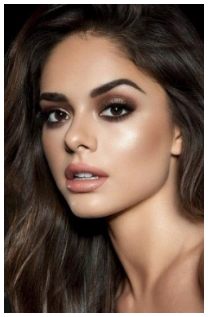
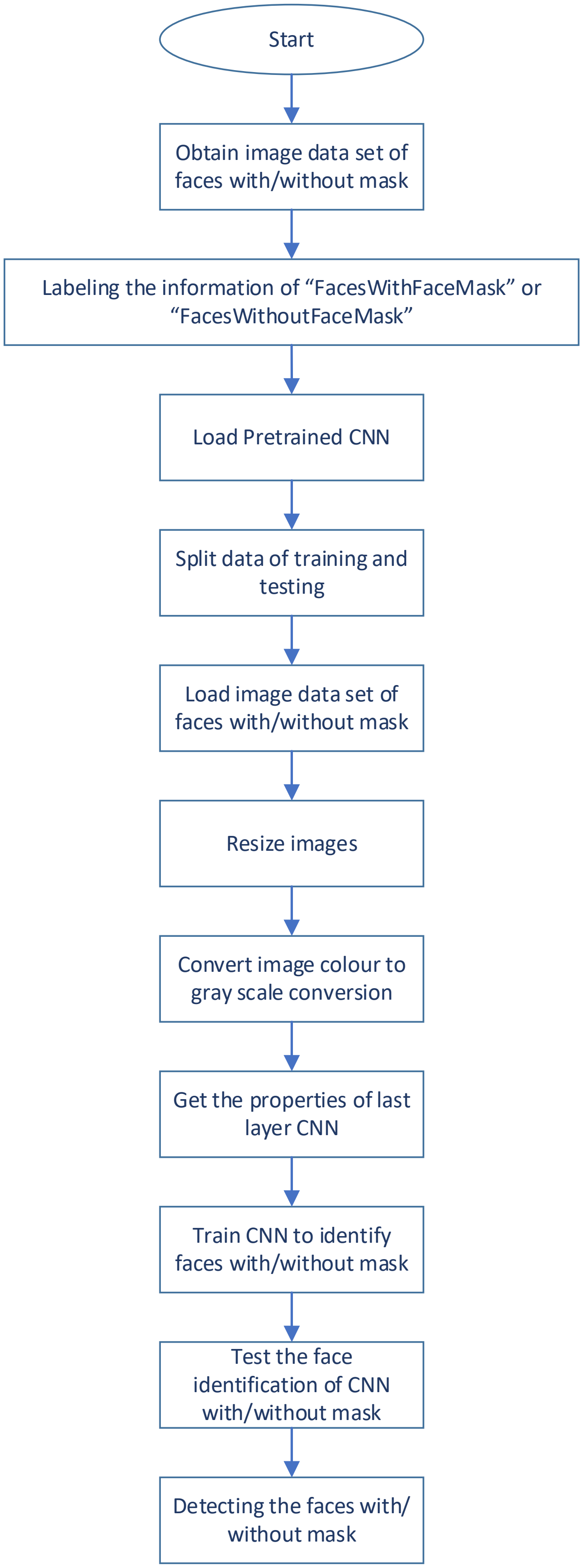


Fig. 5. Example on Left Facing Image [19].

Fig. 6. Example on Right Facing Image [19].

Fig. 7. Example on Bottom Facing Image [19].

Fig. 8. Example on Top Facing Image [19].



Start
Obtain image data set of faces with/without mask
Labeling the information of “FacesWithFaceMask” or “FacesWithoutFaceMask”
Load Pretrained CNN
Split data of training and testing
Load image data set of faces with/without mask
Resize images
Convert image colour to gray scale conversion
Get the properties of last layer CNN
Train CNN to identify faces with/without mask
Test the face identification of CNN with/without mask
Detecting the faces with/ without mask

Fig. 9. Process Flow for the Transfer Learning.

The limitation of this research each of the training images used consists of only one face and Matlab programming was used to perform the transfer learning. For Alexnet, all the training images were resized to 227 x 227 x 3 pixels. Whilst for the ResNet-18, ResNet-50, ResNet-101 and GoogleNet the training images were resized to 224x224x3 pixels. After the model was trained, the images were stacked as input to recognize whether an individual was wearing a face mask or not.

In the testing phase, the dataset consists of a total of 200 images which were not part of the training data set, 100 images of faces without face masks and 100 images of faces with face masks. The testing datasets consist of faces at various angles and were fed into a trained CNN architectures to classify if the face detected in the image is wearing a face mask or not. After that detection of face mask takes place, the result will state either the face on the images uploaded wearing a face mask or not wearing a face mask and appeared on the screen display. The screen will mention 'FacesWithFaceMask' for the result of images of people wearing a face mask and 'FacesWithoutFaceMask' for the images of people not wearing a face mask.

Fig. 9 summarizes the flowchart for the transfer learning of CNN for face mask classification. First, the data set of human faces wearing and not wearing a face mask were compiled. The data sets were labelled with their respective categories, "FacesWithFaceMask" and "FacesWithoutFaceMask." A specific CNN architecture is then loaded. The training datasets are then randomly divided into training and testing with a ratio of 70:30. All the images are resized according to the requirements of specific CNN architectures and converted into a grayscale. The features of the final layer CNN are obtained as every CNN network has a different final layer. The CNN network is then train to recognize faces with and without a face mask using the training dataset. Using an image which are not part of the training dataset, the performance of the various trained CNN networks in classifying face with and without facemask are evaluated.

IV. Performance Metrics

Several performance measures are used in this research to compare the performance of the various CNN networks in classifying images of a person with and without facemask.

The sensitivity performance, also known as recall, refers to the accuracy of true positives and how many positive class samples were appropriately labelled [20]. Sensitivity can be calculated using (1), where True Positive (TP) is the number of events that are both positive and accurately identified, which means the number of images with a person are wearing a face mask and are correctly identified as such. While, False Negative FN is defined as the number of positive events that are incorrectly classified as negative. In this research FN, is the number of images of a person wearing facemask but has been incorrectly classified by CNN as not wearing a facemask.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP + FN}} \tag{1}$$

Specificity is defined as the probability distribution of true negatives with a secondary class, which generally corresponds to the possibility that the negative label was correct, and is presented in (2). True Negatives (TN), also known as negative cases that are classified as negative, showing that individuals are not wearing a face mask but are labeled as wearing a face mask. False Positive (FP) is the individuals who are not wearing a face mask but are erroneously categorized as wearing a face mask.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN + FP}} \tag{2}$$

Accuracy is the most often used parameter for assessing classification performance. This measure computes the proportion of properly identified samples and is represented by (3) [20].

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

Precision is calculated by using (4), which divides the total number of true positives plus false positives by the number of true positives. This statistic evaluates the algorithm's predicting capabilities and is concerned with accuracy. Precision refers to the model's "accuracy" in terms of both the number of positive predictions and the number of positive predictions that occur. [20].

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

The harmonic mean accuracy and recall are used to generate the F-score, as shown in (5). It is related with the examination of positive classes. A perfect score for this parameter means that the model leads the positive class [20].

$$\text{F-score} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \tag{5}$$

V. Result and Discussion

Fig. 10 and 11 illustrates the correct classification of trained CNN networks in identifying 'FacesWithFaceMask' and 'FacesWithoutFaceMask'. Results shown to Fig. 12, 13, 14, 15 and 16, all of the networks performed differently in terms of its statistical significance.

The performance results of AlexNet shows that the Sensitivity is highest at 20 and 40 training images, as shown in Fig. 12. However, for Specificity, Accuracy, Precision, and F-score, the results are its lowest at 20 training images. The performance of AlexNet to classify images of a person wearing facemask can be improved by providing it with more training data set. AlexNet has the lowest accuracy and performed the worst when compare to other CNN architectures. Similar findings was also reported by Neha Sharma et al. [21]. To achieve an acceptable level of performances, AlexNet requires 200 training images to achieve average performance (sensitivity, accuracy, specificity, precession and F-score) of more than 95%. More training data can result in longer training processing time which can undesirable for low powered machines. AlexNet has relatively poorest performance compared to other CNN architectures is because the network has far fewer layers (AlexNet consists of only 8 layers).

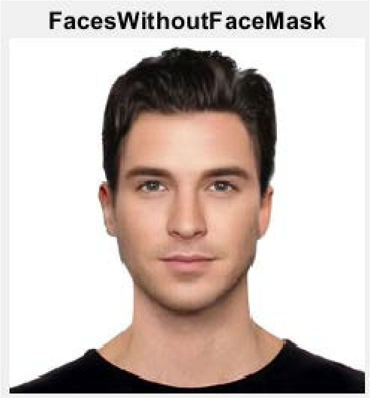


Fig. 10. Result will show “FacesWithoutFaceMask’ for the Image of People Not Wearing a Face Mask.

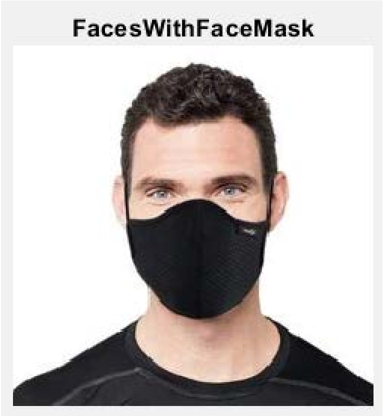


Fig. 11. Result will Show “FacesWithFaceMask’ for the Image of People Wearing a Face Mask.

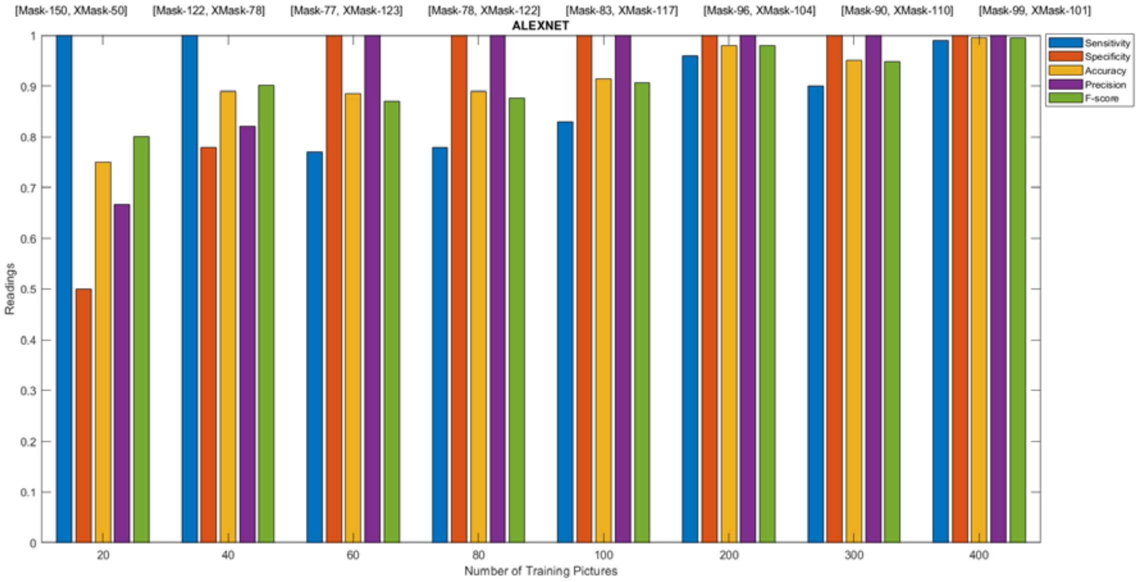


Fig. 12. Performance Results (%) for AlexNet.

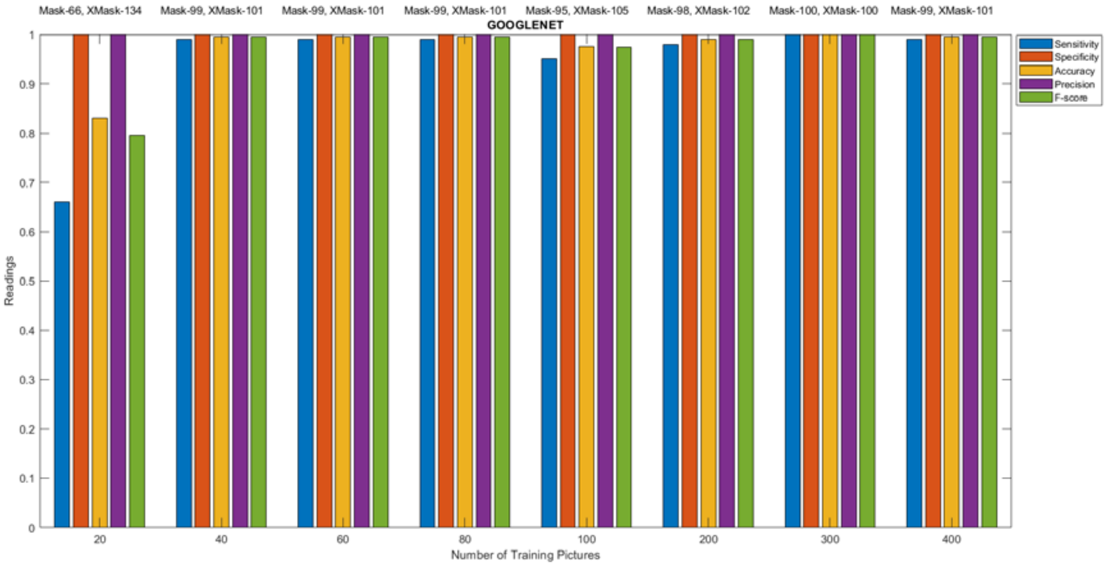


Fig. 13. Performance Results (%) for GoogleNet.

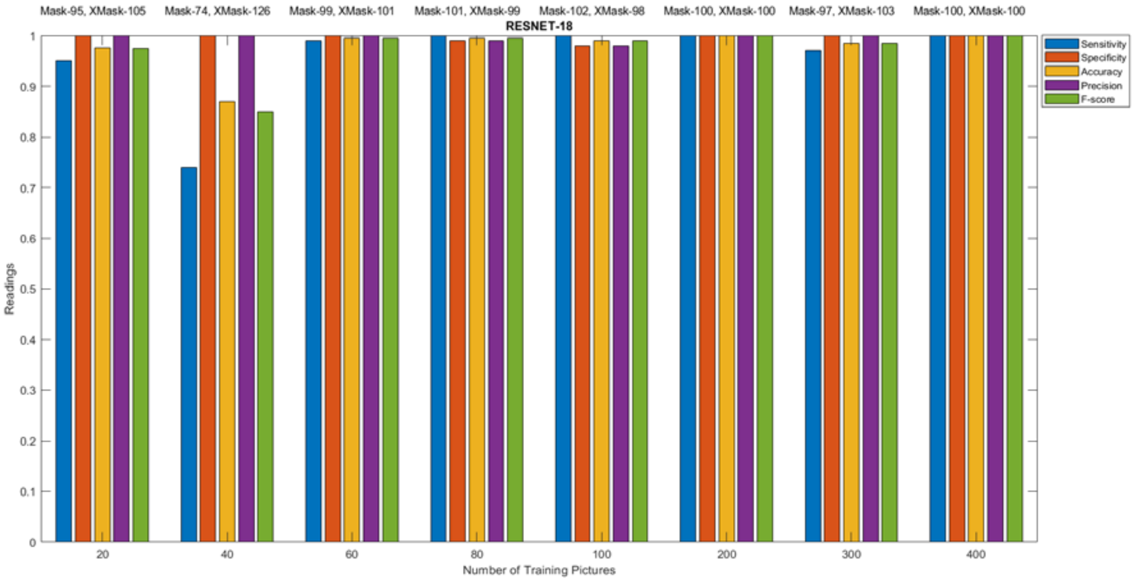


Fig. 14. Performance Results (%) for ResNet-18.

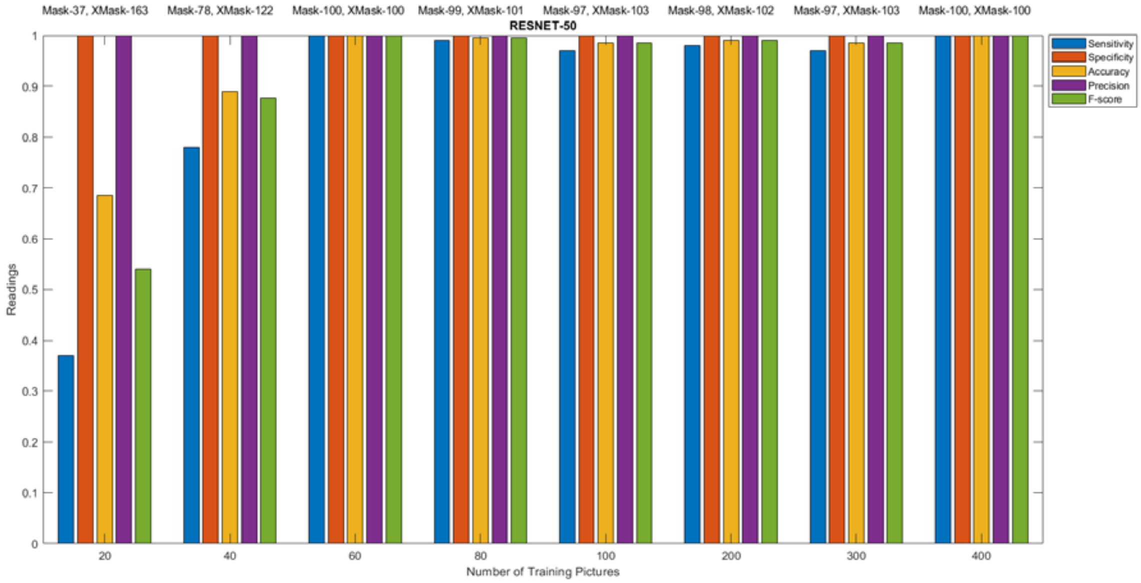


Fig. 15. Performance Results (%) for ResNet-50.

In comparison, GoogleNet which consists of 22 layers only needs 40 images of training data to achieve Accuracy, Sensitivity, Specificity, Precision and F-score of more than 95%.

Fig. 14 and 15 shows that the percentage of performance results for ResNet-18 and ResNet-50 for Sensitivity, Specificity, Accuracy, Precision, and F-score are unstable at 20 to 40 training images. However, as the number of training images increases to 60, the results stabilize with an average performance evaluation of 97%. According to Fig. 16, the best performed network based on the performance evaluation metrics is ResNet-101. ResNet-101 only needs 40 images of training data to achieve Accuracy, Sensitivity, Specificity, Precision and F-score of more than 98%. ResNet-101 achieved the highest average performance as the network has far more layers compared to the CNN architectures presented in this paper. As the name implies, ResNet-101 consists of 101 layers. However, the complexity of the network means it requires more processing power.

The Specificity and Precision performance of GoogleNet, ResNet-50, and ResNet-101 are consistently more than 98% from the start. However this is not the case for AlexNet and ResNet-18 which indicates that these architectures struggle to classify faces without facemask with low number of training data. Low Precision value means that the networks made several incorrectly classification on faces without facemask as wearing facemask. Overall, as the number of training images increases for AlexNet, GoogleNet, ResNet-18, ResNet-50 and ResNet-101, the performance in terms Specificity, Accuracy, Precision, and F-score also improve.

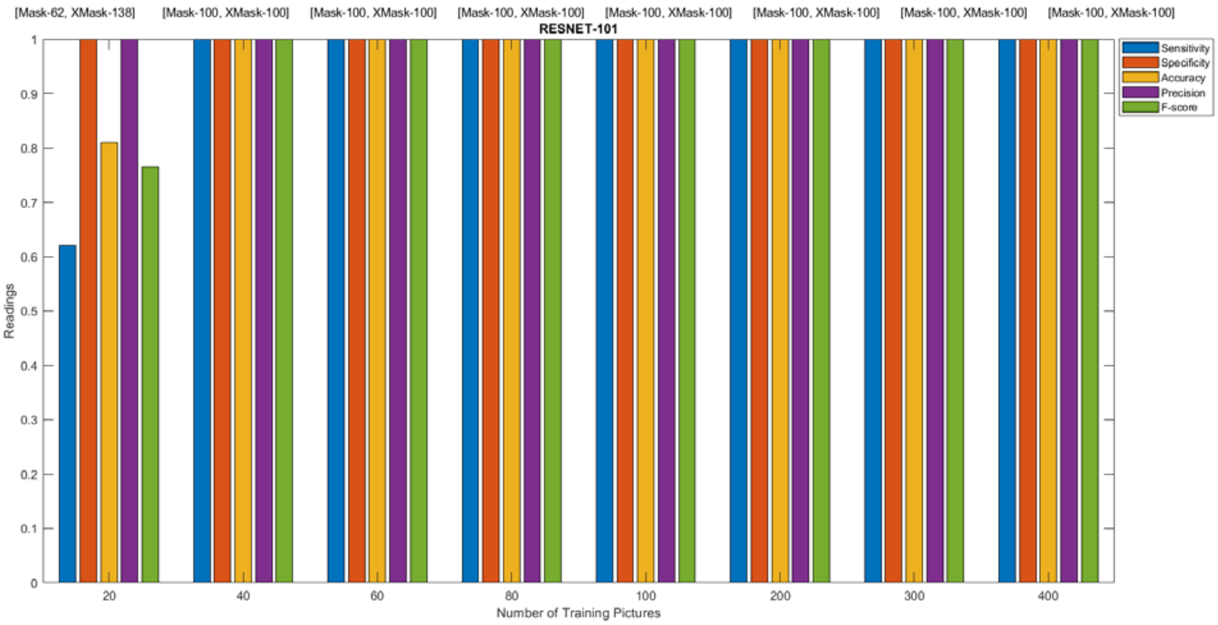


Fig. 16. Performance Results (%) for ResNet-101.

VI. Conclusion

This papers successfully demonstrates application of transfer learning in popular CNN architectures to classify images of a person wearing or not wearing a facemask. In this study, five network architectures were compared using performance metrics; AlexNet, ResNet-18, ResNet-50, GoogleNet, and ResNet-101. It is found that on average AlexNet performed the worst and requires far greater training data to achieve accuracy of more than 95%. ResNet-101 achieved the highest accuracy and requires smallest number of training data to far more 95% accuracy. However the complexity of ResNet-101 means that it requires far greater processing power. GoogleNet strikes the balance of not being overly complex whilst achieving an acceptable level of performance with relatively small number of training data images. Further research can be conducted to gauge the performance and computational resources requirement of various CNN architectures to determine whether a person is wearing a face mask correctly or incorrectly.

References

[1] 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,” Meas. J. Int. Meas. Confed., vol. 167, no. May 2020, p. 108288, 2021, doi: 10.1016/j.measurement.2020.108288.

[2] H. Kim, “5 . 1 Artificial Intelligence , Machine Learning , and Deep,” pp. 151–193, 2020.

[3] G. Wu et al., “Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks,” Remote Sens., vol. 10, no. 3, pp. 1–18, 2018, doi: 10.3390/rs10030407.

[4] M. M. Rahman, M. M. H. Manik, M. M. Islam, S. Mahmud, and J. H. Kim, “An automated system to limit COVID-19 using facial mask detection in smart city network,” IEMTRONICS 2020 - Int. IOT, Electron. Mechatronics Conf. Proc., 2020, doi: 10.1109/IEMTRONICS51293.2020.9216386.

[5] S. Datta, “A review on convolutional neural networks,” Lect. Notes Electr. Eng., vol. 662, no. 03, pp. 445–452, 2020, doi: 10.1007/978-981-15-4932-8_50.

[6] V. Maeda-Gutiérrez et al., “Comparison of convolutional neural network architectures for classification of tomato plant diseases,” Appl. Sci., vol. 10, no. 4, 2020, doi: 10.3390/app10041245.

[7] Y. Anavi, I. Kogan, E. Gelbart, O. Geva, and H. Greenspan, “A comparative study for chest radiograph image retrieval using binary texture and deep learning classification,” Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2015-Novem, pp. 2940–2943, 2015, doi: 10.1109/EMBC.2015.7319008.

[8] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” no. December, 2015, [Online]. Available: http://arxiv.org/abs/1511.08458.

[9] M. D. Putro, D. L. Nguyen, and K. H. Jo, “Real-Time Multi-view Face Mask Detector on Edge Device for Supporting Service Robots in the COVID-19 Pandemic,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12672 LNAI. pp. 507–517, 2021, doi: 10.1007/978-3-030-73280-6_40.

[10] T. Meenpal, A. Balakrishnan, and A. Verma, “Facial Mask Detection using Semantic Segmentation,” 2019 4th Int. Conf. Comput. Commun. Secur. ICCCS 2019, pp. 1–5, 2019, doi: 10.1109/CCCS.2019.8888092.

[11] M. Inamdar and N. Mehendale, “Real-Time Face Mask Identification Using Facemasknet Deep Learning Network,” SSRN Electron. J., 2020, doi: 10.2139/ssrn.3663305.

[12] V. Vinitha and V. Velantina, “Covid-19 Facemask Detection With Deep Learning and Computer Vision,” Int. Res. J. Eng. Technol., vol. 07, no. 08, pp. 3127–3132, 2020.

[13] 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, 2020, doi: 10.22214/ijraset.2020.30560.

[14] K. Hammoudi, A. Cabani, H. Benhabiles, and M. Melkemi, “Validating the correct wearing of protection mask by taking a selfie: Design of a mobile application ‘CheckYourMask’ to limit the spread of COVID-19,” C. - Comput. Model. Eng. Sci., vol. 124, no. 3, pp. 1049–1059, 2020, doi: 10.32604/cmes.2020.011663.

[15] B. Qin and D. Li, “Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19,”

Sensors (Switzerland), vol. 20, no. 18, pp. 1–23, 2020, doi: 10.3390/s20185236.

[16] Henderi, A. S. Rafika, H. L. H. Spits Warnar, and M. A. Saputra, “An Application of Mask Detector for Prevent Covid-19 in Public Services Area,” J. Phys. Conf. Ser., vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012063.

[17] M. Jiang and X. Fan, “Retinamask: A Face Mask Detector,” arXiv, 2020.

[18] C. Jagadeeswari and M. U. Theja, “Performance Evaluation of Intelligent Face Mask Detection System with various Deep Learning Classifiers Keywords :,” Int. J. Adv. Sci. Technol., vol. 29, no. 11, pp. 3074–3082, 2020.

[19] iStock, “Stock Images, Royalty-Free Pictures, Illustrations & Videos - iStock.” 2020, [Online]. Available: https://www.istockphoto.com/.

[20] M. Rahimzadeh and A. Attar, “A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2,” Informatics Med. Unlocked, vol. 19, p. 100360, 2020, doi: 10.1016/j.imu.2020.100360.

[21] N. Sharma, V. Jain, and A. Mishra, “An Analysis of Convolutional Neural Networks for Image Classification,” Procedia Comput. Sci., vol. 132, no. Iccids, pp. 377–384, 2018, doi: 10.1016/j.procs.2018.05.198.