Application of machine learning and deep learning in mask detection

Zihan Wang 1*†, Xuefei Xu 2*†

1 School of Mathematics(BS), University of Massachusetts Amherst, Amherst, Massachusetts, 01002, the United States
2 School of Automation Engineering, Northeast Electric Power University, Jilin, Jilin 132012, China
* Corresponding author: zihanwang@umass.edu, 2846374452@qq.com
†These authors contributed equally.

Abstract. The world witnessed the outbreak of coronavirus disease (COVID-19) at the end of 2019. Even in the 2021, this disease still caused difficulties and inconvenience to millions of people. COVID-19 pandemic has made it a necessity for almost everyone to wear a facial mask as a protective measure. There are several methods to detect masked face, including deep learning and machine learning, which constitutes an integral part on people’s normal life. In order to find a faster and more convenient way to detect whether people wearing masks, this paper compared advantages and disadvantages between deep learning and machine learning based on existing literature, and proposed the most suitable method for detecting masks under different practical conditions. Therefore, we can monitor whether people wear masks correctly and give them timely guidance.

1. Introduction
At the end of 2019, a new kind of virus had spread around the world. The Corona virus-19, which is also known as COVID-19, is the virus that will be introduced in this passage. COVID-19 is a contagious virus that can spread easily through the air. As we know, the COVID-19 often cause irreparable damage to the body and may cause the loss of property, and social economy. Under this situation, scientists had found that wearing a mask is effective against spreading the virus. Hence, wearing a mask is necessary to protect the population from the virus. However, some people reject to wear a mask. In this case, the mask-wearing detection system is necessary to control disease. The samples in the dataset could be classified into three groups, the first group is the person who wears the masks correctly, the second group is the person who wears the mask but not wearing correctly, and the third group is the person who is not wearing the mask. And we will mostly concentrate on the detection of the population from the previous article to know how to find the person who is not wearing the mask. We now have a proven technique of AI technology to detect whether a person wears a mask, which is what we want to discuss in this review paper. In this paper, we will illustrate these two kinds of techniques to achieve the goal. The first one is Machine Learning, and the second one is Deep Learning.

Machine learning and deep learning are based on three different types. They are supervised learning which machine learned from labeled dataset, unsupervised learning which learned from unlabeled dataset, and semi-supervised learning which learned from both labeled and unlabeled
dataset. Now we will talk about Machine Learning first. Machine learning is a kind of study using computer data and improves algorithm in the model automatically. This means Machine Learning needs data to train the algorithm itself. The training data will help the machine learn from the feature and improve the algorithm accuracy. Moreover, validation data is also important, this data is used to fix the feature collected from the training data and to make the algorithm more accurate. We need the test data to test whether the feature obtains from the training data is accurate or not. Then we will introduce Machine Learning in face mask detection. Usually, there are two steps to detect masks through Machine Learning. First it will learn from extracting feature from the data to find the features from three groups and collect them as the input information for later use. The second step is to use the algorithm through the existed Machine Learning model to detect whether the person wears mask using the features collected from the first step. There are several machine learning algorithms to detect the masked face. The first is decision tree, this algorithm will use a mapping relationship between object attributes and its values. Decision tree uses nodes to test and the child nodes are arranged to correct branch. The second one is SVM (Support-Vector Machine). SVM is a classical supervised machine learning working for face mask detection. This algorithm will build a model that assigns new examples to one or another category according to the training set, then it will set up a non-probabilistic binary linear classifier. The third one, the KNN (k-nearest neighbors), is a kind of classification algorithm first developed in statistic field, and used to classify dataset in the machine learning. This algorithm will output the class membership from its neighbor. The object will assign the nearest neighbor the number of the group. The regression is also important in machine learning and we will also introduce some regression algorithms. The first one, Linear Regression, is a method to complete regression from statistics and mathematics analysis that determines the independent quantitively relationship between two or more variables. It will estimate the values of their coefficients used in the representation from the dataset. The second one, logistic regression, is kind of logistic model that used to model the probability of events that could use for image. It evaluates the performance of the model to find whether it is fitting well.

Deep learning is based on artificial neural networks and it has the same or similar concepts as it in artificial neural networks. In deep learning, the lower layer has a symbol that could be identified by those symbols, and the higher layer would have more specific digits, letters, or images. In this paper, we introduced three kinds of deep learning algorithms. They are ResNet-34, YOLOv5, and Faster-RCNN. For Faster-RCNN, when the images were put into the Faster-RCNN, it will form a convolutional feature map. However, it will use another feature map, Region Proposal Network, to feature and it would be much faster than the original RCNN and fast CNN because it will use the shared map. In this case, Faster-RCNN successfully integrates the region of interest generation into the neural network. For YOLO (You Only Look Once), it is a neural network to do the detection without specific features and will also base on CNN. However, it is different from Faster-RCNN. Faster-RCNN will collect the feature and generate an independent neural network and classify where YOLO will classify the feature during generalization. Hence, YOLO will always use for monitoring. ResNet-34 is a kind of artificial neural network that builds on constructs known from pyramidal cells. The ResNet-34 Model will use two layers and is based on Residual to feature the data. This will make the error become lower.

Under the circumstances of virus outbreaks, the masked face is the most efficient and useful way to protect people from virus, especially in the crowded space. The mask face detection will help to protect people from danger. Considering about the environment all over the world, the region required to wear masks has a lower rate of cases and deaths which also shows that wearing mask is necessary to protect people from the virus.

2. Method

Based on existing literature, machine learning and deep learning were used for mask recognition.
2.1 Machine learning for mask recognition

When using machine learning for face mask detection, facial features were extracted first, then classical machine learning algorithms were utilized to train model and recognize masked face. Two methods, deep learning and machine learning method, could be used for feature extraction.

2.1.1 Deep learning for feature extraction

Neural network can automatically extract image features [1]. \( x_{i,j} \) is used to represent the image element of the \( i \) row and \( j \) column, \( \omega_{m,n} \) represents the weight of the \( m \) row and \( n \) column, \( \alpha_b \) represents the offset term of filter, \( a_{i,j} \) represents the element of the \( i \) row and \( j \) column of the Feature Map, \( f \) represents activation function. The convolution is calculated by the following formula:

\[
a_{i,j} = f \left( \sum_{m=0}^{2} \sum_{n=0}^{2} \omega_{m,n} x_{i+m,j+n} + \alpha_b \right)
\]

(1)

Image size, step length and Feature Map size after convolution are satisfy with the following relationship:

\[
W_2 = \left( W_1 - F + 2P \right) / S + 1
\]

\[
H_2 = \left( H_1 - F + 2P \right) / S + 1
\]

(2)

where \( W_2 \) is the width of Feature Map after convolution; \( W_1 \) is the width of the image before convolution; \( F \) is the width of filter; \( P \) is Zero Padding; \( S \) is stride length; \( H_2 \) is the height of Feature Map after convolution; \( H_1 \) is the width of the image before convolution.

Pooling part can reduce the size of a large image, reducing pixel information while preserving important information. When all feature maps are pooled, the large graphs corresponding to a series of inputs become a small one. At the full connection layer, the convolution results are arranged into a row of vector \( x_i \), which are fully connected with the two output neurons. Each connection has a weight \( \omega_i \), and two neurons are calculated respectively according to \( y = \sum \omega_i * x_i \).

Finally, back propagation is performed in the process of comparing predicted value with the actual and then returning to modify the network parameters. At first, the parameters of the convolution kernel are randomly initialized, and then values of the convolution kernel are adjusted under the guidance of the error through back propagation algorithm, so as to minimize the error between the predicted value and the actual value.

The features in deep learning method are obtained automatically rather than manually designed, which lacks the ability to maintain excellent generalization performance and robustness.

2.1.2 Machine learning for feature extraction

The basement of machine learning is feature extraction, which can increase operation efficiency by removing irrelevant and redundant data. For many models without regularization, feature selection and feature extraction are necessary in pattern recognition. He Y u-min [2] extracted HSV and HOG features to detect masked face. HSV features were extracted through calcHist() function in OpenCV. HOC features were generated from hog() function in the skimage.feature module.

Traditional machine learning technology exists great limitations in processing original natural data. To accurately extract typical features, skilled engineers and experienced domain experts are required to design feature extractors and transform original data into appropriate intermediate vectors forms.

2.1.3 Classical machine learning algorithm for detecting face mask

Machine learning is the core of artificial intelligence, which aims to improve algorithms performance in empirical learning and enhance optimizing computer programs performance. Elliot Mbunge [1] used three methods to detect
COVID-19 face masks, including decision trees, support vector machine (SVM), and ensemble. Ensemble methods used K-Nearest Neighbors Algorithm, Linear Regression and Logistic Regression.

(1) Decision trees
Decision tree is a tree structure representing a mapping relationship between object attributes and object values. Each internal node represents a test on an attribute. Each branch indicates a test output and each leaf node stands for a category. These categories are determined in advance and a classifier can be obtained through learning, which can bring correct classification to new objects.

Entropy measures the uncertainty of classification. Joint entropy expression for two variables X and Y is shown as follow:

\[ H(X,Y) = - \sum_{x \in X} \sum_{y \in Y} p(x_i, y_j) \log p(x_i, y_j) \]  \hspace{1cm} (3)

where \( p(x_i, y_j) \) is the probability with values \( x_i \) and \( y_j \).

Conditional entropy is similar to conditional probability which measures the uncertainty of \( X \).

The expression is as follows:

\[ H(X|Y) = - \sum_{x \in X} \sum_{y \in Y} p(x_i, y_j) \log p(x_i | y_j) = \sum_{j=1}^{n} p(y_j) H(X|y_j) \]  \hspace{1cm} (4)

(2) SVM
SVM is a two-class classifier. A hyperplane is confirmed by placing two types of data in a farther place from the hyperplane.

Suppose a training data set is given in a feature space:

\[ T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\} \]  \hspace{1cm} (5)

where \( x_i \in \mathbb{R}^n \), \( y_i \in \{1, -1\}, i = 1, 2, \cdots, N \), \( x_i \) is the eigenvector of \( i \), and \( y_i \) is the class marker.

For a given dataset \( T \) and hyperplane \( \omega \cdot x + b = 0 \), the geometric interval of the hyperplane with respect to the sample points \( (x_i, y_i) \) is defined as:

\[ y_i = y_i \left( \frac{\omega}{||\omega||} \cdot x_i + b \right) \]  \hspace{1cm} (6)

\[ \omega = \frac{\omega}{||\omega||} \]  \hspace{1cm} (7)

\[ b = \frac{b}{||\omega||} \]  \hspace{1cm} (8)

Therefore, the maximum segmented hyperplane problem of SVM model can be expressed as the following constrained optimization problem:

\[ \min_{\alpha, b} \frac{1}{2} ||\alpha||^2 \]  \hspace{1cm} (9)

\[ y_i (\omega \cdot x_i + b) \geq 1, i = 1, 2, \cdots, N \]

The dual problem of convex quadratic programming with inequality constraints can be obtained by Lagrange multiplier method. The original objective function with constraints is transformed into a newly constructed Lagrange objective function without constraints:

\[ L(\omega, b, \alpha) = \frac{1}{2} ||\omega||^2 - \sum_{i=1}^{N} \alpha_i (y_i (\omega \cdot x_i + b) - 1) \]  \hspace{1cm} (10)

where \( \alpha_i \) is the Lagrange multiplier.

The optimization process satisfies KKT condition:
In order to obtain the specific form solving the duality problem, the partial derivative of $L(\omega, b, \alpha)$ with respect to $\omega$ and $b$ is 0, then:

$$\omega = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Substitute the above two equations into the Lagrange objective function, then eliminate $\omega$ and $b$, the equation is designed as follow:

$$\min_{\omega, b} L(\omega, b, \alpha) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i$$

Duality problem in SVM is to find the maximum of $\min_{\omega, b} L(\omega, b, \alpha)$ to $\alpha$.

(3) KNN

K-Nearest Neighbors algorithm is a classification algorithm, which applied in the fields of character recognition, text classification, image recognition and so on. When a given sample is most similar to K sample in the dataset, if the majority of the K samples belong to one category, we can confirm that this sample belongs to this kind of category.

Euclidean distance is used in KNN algorithm, and the calculation formula of Euclidean distance in multi-dimensional space is as follows:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

(4) Linear regression

Linear regression is a statistical analysis method that uses regression analysis in mathematical statistics to determine interdependent quantitative relationship between two or more variables. In short, it aims to select a linear function to fit known data well and predict the unknown data accurately.

Suppose the following functions:

$$h(x) = \theta_0 + \theta_1 x$$

where $x$ is an input function, $h(x)$ is an output function, and $\theta$ is the parameter of the function, which needs to be learned from samples.

Assuming features $x$ has $m$ numbers, the above formula can be changed to:

$$h(x) = \sum_{i=0}^{m} \theta_i x_i = \theta^T x$$

In order to express the degree of closeness, the sum of errors of all samples in the training data can be used as the loss function:

$$J(\theta) = j_1(\theta) + j_2(\theta) + \cdots + j_m(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)}\right)^2$$

where $j(\theta)$ is the error of a single sample.

(5) Logistic regression

Logistic regression solves the optimal model parameters through optimization method. The
The advantages of logistic regression are fast and suitable for binary classification problems, making algorithm constructions simple and easy to understand.

The form of sigmoid function is:

$$g(z) = \frac{1}{1 + e^{-z}}$$  \hspace{1cm} (18)

The form of linear boundary is designed as follow:

$$z = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{i=0}^{n} \theta_i x_i$$  \hspace{1cm} (19)

Training data is $$x = [x_0, x_1, \cdots, x_n]^T$$, the best parameter is $$\theta = [\theta_0, \theta_1, \cdots, \theta_n]$$. The prediction function is constructed as follow:

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$  \hspace{1cm} (20)

6) Comparison of classical machine learning algorithm for detecting masked face

Considering all data sets in the experiment, the time consumed by SVM is less than that of decision tree classifier. The model performance in SVM classifier was better than that in decision tree classifier. When comparing ensembled algorithm with SVM, ensembled algorithm took more time than decision tree and SVM. Reference [2] also showed that using SVM to detect masked face has high accuracy, which indicated that utilizing SVM met the requirements in mask detection. To sum up, SVM was the most suitable algorithm to recognize masked face in practice.

2.2 Deep learning for mask recognition

Since the technologies of deep learning in image recognition gradually matured, neural network algorithm was widely used for mask detection. There are several algorithms commonly used in mask recognition, including RCNN, Faster-RCNN, Yolo and SDD. Faster-RCNN, YOLOv5 and Rsenet-34 are going to serve as typical examples to compare the advantages and disadvantages in using deep learning to detect masks.

2.2.1 Faster-RCNN. Faster-RCNN can be mainly divided into four parts, including Feature extraction, RPN, ROI Pooling and Classification and regression [3].

1) Feature extraction

Faster-RCNN is a target detection method, which first uses in a set of basic convolutional layers to extract image feature maps. The convolution layer includes a series of convolution and pooling operations, which are used to extract image features. Generally, existing classical network models ZF or VGG16 are directly used in this process. The weight parameters of the convolution layer are shared for RPN and Fast RCNN, which are also the key to speed up the training process and improve the real-time performance of the model.

2) RPN

RPN network is used to generate the regional candidate box Proposal. The Proposal is to find out possible position of the target in advance, and ensure a high recall rate when selecting fewer Windows (thousands or even hundreds) by using texture, edge, color and other information in the image. The Region Proposal method has higher quality than traditional sliding window. Commonly used Region Proposal methods include Selective Search (SS) and Edge Boxes (EB). Based on the multi-scale Anchor introduced by network model, Softmax defines anchors that are foreground or background. Anchors regression prediction is performed by Bounding Box Regression to obtain the exact position of the Proposal for subsequent goal identification and detection.

Anchor is a relatively important concept in RPN network. In traditional detection methods, multi-scale sampling of images or filters (sliding Windows) are required by establishing an image pyramid in order to obtain multi-scale detection frame. RPN network uses a 3×3 convolution kernel to slide on the last feature graph (conv5-3) and maps the corresponding position of the center of
convolution kernel back to the input image, generating a total of 9 anchors of 3 scales and 3 aspect ratios. The sliding window approach ensures that all feature spaces are associated, resulting in a multi-scale and multi-aspect ratio anchors on the original image.

RPN loss function is designed as follow:

\[ L(p_i, \{t_i\}) = \frac{1}{N_{cls}} \sum_{i} L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_{i} p_i^* L_{reg}(t_i, t_i^*) \]  

(21)

where \( p_i \) represents the probability that the \( i \)-th anchor predicts a true label; \( p_i^* \) equals 1 for a positive sample, 0 for a negative sample; \( t_i \) represents the boundary box regression parameter predicting the \( i \)-th anchor; \( t_i^* \) represents the real boundary box corresponding to the \( i \)-th anchor; \( N_{cls} \) indicates the number of all samples in a mini-batch 256; \( N_{reg} \) represents the number of anchor locations, which is about 2400.

In practice, \( \frac{1}{N_{cls}} \) is usually used instead of \( \frac{1}{N_{reg}} \).

(3) ROI Pooling

By integrating feature maps information of convolutional layer and candidate box Proposal, the coordinates of Proposal are mapped in the input image to the feature map of the last layer (Conv5-3), then pool the corresponding areas in the feature map. The pooled results of the fixed-size output are connected to the full connection layer behind. In short, ROI Pooling layer collects input feature maps and proposals, then extracts these elements after synthesizing information, sending them to subsequent full connection layer to determine target category. The labeled data sets were put into Faster-RCNN network for training.

(4) Classification and regression

Obtaining the feature map selected in the suggestion box from the global feature map, input the feature map into the ROI Pooling layer to obtain 7*7 scaled feature map, then go through two fully connected layers, associate two fully connected layers in parallel to obtain ROI feature vector, and go through one fully connected layer respectively. Finally, the probability prediction and regression parameter prediction of classification are obtained. The full connection layer is followed by two sub-connection layers, which are classification layer and regression layer. The classification layer is used to determine the category of Proposal, and the regression layer predicts the exact position of Proposal through bounding box regression.

Pan Liu [4] used Faster-RCNN to detect masks. The data sets were trained 100 times, and the average detection accuracy of the network reaches 97.57%, which showed that the model gained high accuracy and could be applied to various detection systems. However, some problems were still existed in this model. When the detected person was sheltered or standing in a long distance, the model’s detection ability would decline. Therefore, it is necessary to strengthen the detection ability with small targets and covered objective.

2.2.2 YOLOv5

YOLOv5 is a single-stage target detection algorithm, which can be divided into four parts, including: input, benchmark network, Neck network and Head output layer [5-7]. Based on YOLOv4 [8-9], this algorithm adds some new ideas to improve its speed and accuracy greatly. The main improvement ideas are as follows:

(1) Input

In the training stage, some improvement ideas are proposed, including Mosaic data enhancement, adaptive anchor frame calculation and adaptive image scaling. The input side represents the input picture. The input image size of the network is 608*608. This stage usually includes an image pre-processing stage, that is, the input image is scaled to the input size of the network and normalized. In the network training stage, YOLOv5 uses Mosaic data enhancement operation to improve training speed and network accuracy. Adaptive anchor frame calculation and adaptive image scaling are
proposed in this part.

Mosaic data enhancement method used in YOLOv5 in the training model stage is improved on the basis of CutMix data enhancement method. CutMix uses just two images, while Mosaic uses four images that are randomly scaled, cropped, and arranged. This enhancement method combines several images into one image, which can not only enrich data set but also greatly improve network training speed, reduce memory requirement of the model.

In network training stage, the model outputs corresponding prediction frame on the basis of the initial anchor frame, calculates the gap between it and GT frame, performs reverse update operation to update parameters of whole network. In YOLOv3 and YOLOv4 detection algorithms, when training different data sets, the initial anchor point box is obtained by running a separate program. YOLOv5 embedded this function into training code. Data sets adaptively are calculated in the best anchor box and users can turn the function off or on according to their own needs.

(2) Benchmark network

The benchmark network usually contains the network of some excellent classifiers. This module is used to extract feature representation. In YOLOv5, some new ideas are integrated from other detection algorithms, including Focus structure and CSP structure.

The main idea of Focus structure is slicing input image. The original input image size is 608*608*3. After Slice and Concat operation, a feature map of 304*304*12 becomes an output. Then through a Conv layer with 32 channels, a 304*304*32 size feature map is obtained. For CSP structure, in YOLOv4 network, CSP structure is designed only in the backbone network, referring to the design idea of CSPNet. Two kinds of CSP structures are designed in YOLOv5, including CSP1_X and CSP2_X structure. CSP1_X structure is applied in Backbone network, and CSP2_X structure is applied in Neck network.

(3) Neck network

Neck network is usually located in the middle of the reference network and the head network, which can further improve the diversity and robustness of features. YOLOv5 also uses SPP and FPN+PAN modules, but the implementation details are different.

YOLOv5's Neck network still uses FPN+PAN structure, but some improvements have made on it. The Neck structure of YOLOv4 adopts ordinary convolution operations. In YOLOv5 Neck network, CSP2 structure designed by CSPnet is used to strengthen the ability of network feature fusion. By comparing YOLOv4 with YOLOv5, we can find that: 1) YOLOv5 not only replaces part of CBL module with CSP2_1 structure, but also removes the lower CBL module. 2) YOLOv5 not only replaces CBL module operated after Concat with CSP2_1 module, but also changes another CBL module position. 3) Original CBL module is replaced by CSP2_1 module in YOLOv5.

(4) Head output layer

Head is used to complete target detection results. For different detection algorithms, the number of output branches are different, usually including a classification branch and a regression branch. YOLOv5 uses GIOU_Loss to replace Smooth L1 Loss function to further improve the detection accuracy of the algorithm.

The loss function of target detection task generally includes classification loss function and regression loss function. In YOLOv5, GIOU_Loss is used as loss function of Bounding box. IOU_Loss is intersection between prediction frames and GT frames, or the union between prediction frames and GT frames. IOU_Loss mainly considers the overlap between detection frame and GT frame. On the basis of IOU, GIOU_Loss adds measure method of intersection scale to solve problems when boundary boxes do not coincide, and alleviate problems in simple IOU_Loss.

Xiao [10] used improved YOLOv5 model to study mask wearing recognition. Xiao adjusted the input size of original model and initial candidate box, then improved the convolution and loss function to make YOLOv5 more suitable for mask wearing identification. The model used accuracy, recall, average precision mean and harmonic mean as evaluation indexes. The larger the value, the better the mask recognition effect. After 1000 epochs, the model reached convergence state. The accuracy rate was stable above 95%, the recall rate was stable at 100%, the average accuracy was stable around 1.0,
and the harmonic mean was stable above 0.9. All statistics indicated that improved YOLOv5 could be effectively applied in mask recognition system.

2.2.3 Resnet-34. Resnet is a deep residual network. The network is deeper, the more information can be obtained, and the characteristics are richer. However, experiments show that with the deepening of the network, the optimization effect became worse, and the accuracy was declined. But Resnet can increase the network depth as much as possible to obtain the best effect under the condition of ensuring test accuracy. The Figure 1 is the structure diagram of Resnet-34:

![Resnet-34 structure diagram](image)

The main branch of Resnet-34 residual structure is composed of two 3*3 convolution layers. The connecting line on the right side of residual structure is shortcut branch. In order to add output matrix on main branch to the output matrix on shortcut branch, two output characteristic matrices are required to have the same shape. The dotted line residual structure can reduce the dimension by 1x1 convolution kernel on the shortcut branch.

Residual network is made up of a series of residual blocks. A residual unit can be expressed as follow:

\[
y_i = h(x_i) + F(x_i, W_i)
\]

\[
x_{i+1} = f(y_i)
\]

where \(x_i\) and \(x_{i+1}\) respectively represent the input and output of the \(l\)th residual unit. Each residual unit generally contains a multi-layer structure. \(F(x_i, W_i)\) is the ReLU activation function, which
represents learned residual, while \( h(x_i) \) represents identity mapping. Based on the above formula, we find that the learning characteristics from shallow \( l \) to deep \( L \) are:

\[
\frac{\partial \text{loss}}{\partial x_i} = \frac{\partial \text{loss}}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_i} = \frac{\partial \text{loss}}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=1}^{l-1} F(x, W_i)\right)
\]

(24)

where \( \frac{\partial \text{loss}}{\partial x_L} \) represents the loss function reaches the gradient of \( L \); \( 1 \) indicates that the short-circuit mechanism can propagate the gradient without loss; while the other residual gradient \( \frac{\partial}{\partial x_L} \sum_{i=1}^{l-1} F(x, W_i) \) needs to pass through the layer with weights, which cannot directly transmit. The residual gradient will not always equal to \(-1\), and the existence of \( 1 \) will not cause the gradient to disappear. So residual learning will be easier.

The construction of residual network is divided into two steps: using VGG formula to build plain VGG network, and then inserting identity mapping between convolution networks of plain VGG. Direct mapping is the best choice in residual network. Moving the activation function to the residual part can improve the model accuracy.

Liu [11] studied the preprocessing method of input pictures, finding optimal learning rate and appropriate size of batch data based on ResNet-34, then the model was trained with mask detection dataset. Compared with other target detection algorithms, the accuracy of Resnet-34 had improved to 97.25%, and the detection speed was greatly accelerated. This algorithm could also detect small targets and the faces wearing masks on the side. ResNet-34 had high anti-interference ability when detecting shielding masks, special-shaped masks and pattern masks, which possessed great capability in accurately identifying different types of masks, which could avoid false detection dramatically.

2.2.4 Comparison of deep learning algorithm for detecting masked face. At present, deep learning methods in the field of target detection are mainly divided into two categories: two-stage and one-stage target detection algorithms. In the former, a series of candidate boxes are generated by the algorithm as samples, and then the samples are classified by the convolutional neural network. The latter does not generate candidate boxes and directly transforms the problem of target frame positioning into regression problem. Just because of the difference between the two methods, there exist differences in performance. The former is superior in detection accuracy and positioning accuracy, while the latter is superior in algorithm speed.

Faster-RCNN achieves high-precision detection performance by adding RPN in two stages. Compared with other first-stage detection networks, the two-stage network is more accurate and can easily solve multi-scale and small-target problems. Besides, Faster-RCNN performs well on multiple data sets and is easy to migrate. Changing the target class in the data set can also change the test model. The model can be improved in many places and the network algorithm can be optimized. However, Faster-RCNN still contains a lot of shortages. No matter using VGGNet or ResNet, the extracted feature maps are all single layer and the resolution is relatively small. Therefore, Faster-RCNN has limitations in solving multi-scale and small-target problems. In order to avoid overlapping candidate boxes, NMS is used for post-processing based on classification precision. In fact, this method will cause missed detection when targets are occluded. The original Faster-RCNN uses the full connection layer, which occupies a large part of the number of parameters. The speed of the two stages is obviously slower than that of the single stage, and the actual application is not real-time. Therefore, the network consumes a long time and cannot obtain a high accuracy in detecting small targets and covered objective.

YOLOv5 uses the pytorch framework which is very user-friendly and can easily train data sets. Besides, the code is easy to read and integrate a large number of computer vision technologies, which is very conducive to learning and reference. YOLOv5 is not only easy to configure the environment,
but also trains model and produces real-time results very fast, which can effectively infer the input of single image, batch image, video and even webcam port directly. However, YOLOv5 filters real target frames with extreme size and aspect ratio, which plays an important role in real detection tasks and significant detection problems. Therefore, YOLOv5 is facilitate to lose a lot of important information which leads to bad detection results in small targets and the subsequent detection and regression cannot meet the requirements.

ResNet-34 uses the method of adding residual networks to the network to solve the problem that the stacking effect of deeper networks becomes worse when the network depth increases to a certain extent. When the network depth reaches a certain level, the error increases, the effect deteriorates, and the gradient disappears more obviously. During backward propagation, the gradient cannot feed back to the front network layer, and the front network parameters cannot be updated, leading to the deterioration of training. In the residual network, an identity mapping is added to skip local or multi-layer operations. Meanwhile, in the backward propagation process, the gradient of the network at the next layer is directly transmitted to the one at the next layer, so as to solve the problem of the gradient disappearing in the deep network. Therefore, ResNet-34 makes the feedback propagation algorithm smoothly and the structure simply. The increase in identity mapping does not degrade network performance. However, ResNet-34 requires long training time. When increasing the number of layers, although ResNet-34 can learn features better, it will also increase the amount of calculation and be prone to over-fitting, gradient disappearance, and gradient explosion.

Above all, ResNet-34 has high accuracy in detecting small target but spend a lot of time to complete the process. Faster-RCNN has high prediction accuracy but loses its advantage when the detected targets are sheltered or have a long distance away from the testing machine. YOLOv5 is easy to use and do not take much time to complete detection, but it is not suitable for small target detection and is facilitate to lose significant information.

2.3 Comparison between deep learning and machine learning algorithm for detecting masked face

Machine learning is a subset of artificial intelligence that uses statistical techniques to provide the ability to learn data from a computer without complex programming. In short, machine learning enables computers to act and learn like humans to improve their ability to study in an independent manner by providing them with actual interactions and observations information. Unlike task-specific algorithms, deep learning is a subset of machine learning based on learning data. The inspiration comes from the function and structure which are known as artificial neural networks. Deep learning displays the world as simpler concepts and hierarchies through learning.

The most important difference between traditional machine learning and deep learning is data scaling performance. When the number of data is small, deep learning algorithms do not work well because they need big data to perfectly identify and understand. However, machine learning algorithms can work in less data, using small sample to train model and possessing high accuracy. Feature extraction aims to reduce data complexity and make the model work more accurately and efficiently. The whole process is expensive and difficult, requiring a lot of time and expertise. However, in traditional machine learning, features of all applications are identified by experts that features such as shapes, pixel values, textures, directions and positions need to be manually coded according to data types and domains. The performance of machine learning algorithms depends on the accuracy that features identified and extracted. Deep learning algorithms recognize these advanced features from the sample data, thus reducing the workload of developing a new feature extractor for each problem. In addition, interpretation is one of the factors that must be considered whether deep learning can be applied to the industry. Deep learning can find nodes that activate deep neural networks, but cannot identify those neurons model and their cooperating functions. However, machine learning algorithms provide us with a clear set of rules which is easy to explain the logic behind it.

All in all, when choosing models for mask recognition, we should fully consider the size of sample data, types of features and extraction methods, as well as the difficulty of model interpretation and innovation, so as to select the most suitable model under different situations.
3. Conclusion
Under the circumstances that the COVID-19 suffered all over the world, protection of infection from COVID-19 is necessary and sufficient. The masked face will help the people and the governments to protect themselves from the virus and successful AI models will help us in detecting the face mask in the real world to save people from the global crisis caused by the COVID-19. We have learned about machine learning and deep learning that significantly help in the detection and listed the models’ type and algorithm to do the comparison and contrast between all the models which will play a role in future research from the model selection.

Reference
[1] E. Mbunge, S. Simelane, S.G. Fashoto, et al. (2021) Application of deep learning and machine learning models to detect COVID-19 face masks – A review. Sustainable Operations and Computers, 2: 235-245.
[2] Y.M. He, Z.H. Wang, S.Y. Guo, et al. (2021) Face mask detection algorithm based on HSV+HOG features and SVM. Journal of Measurement Science and Instrumentation, 12.
[3] Z.C. Li, H.C. Li, W.S. Hu, et al. (2021) Masked face detection model based on multi-scale attention-driven Faster R-CNN. Journal of Southwest Jiaotong University.
[4] P. Liu, H.Y. Wen. (2021) Research on mask wearing detection based on Faster RCNN. Computer Knowledge and Technology 17: 14-16.
[5] F. Wang. (2020) Improved artificial intelligence detection and recognition algorithm for wearing masks and helmets in YOLOv5. Construction and Budget, 11: 67-69.
[6] S.L. Tan, X.B. Bie, G.L. Lu, et al. (2021) Real-time detection for mask-wearing of personnel based on YOLOv5 network model. Laser Journal 42: 147-150.
[7] C. Peng, Q.H. Zhang, Z.H. Tang, et al. (2021) A face mask wearing detection method based on YOLOv5 enhancement model. Computer Engineering.
[8] T.H. Cheng, Y.C. Cui, X.M. Wu. (2021) Research on wearing mask detection in public places based on YOLOv4 convolutional neural network. Modern Information Technology, 4: 9-12.
[9] J.L. Guan, X. Zhi. (2020) Mask wearing detection based on YOLOv4 convolutional neural network. Modern Information Technology, 4: 9-12.
[10] B.J. Xiao, L.J. Wan, J.Q. Chen. (2021) Research on mask wearing recognition using YOLOV5 model. Journal of Fujian Computer, 37: 35-37.
[11] G.M. Liu, J.L. Jiang, B. Zha, et al. (2021) Mask wearing detection based on deep neural network. Journal of An qing Normal University (Natural Science Edition), 27: 54-58.