Detection of Face-Mask Using Deep Learning in ATM Station

Abdul Azadh Abdul Saleem
MSc in Data Analytics
National College of Ireland
Dublin
x18203621@student.ncirl.ie

Arunkumar Sudhakaran Nair
MSc in Data Analytics
National College of Ireland
Dublin
x19153317@student.ncirl.ie

Evlin Bridgit Emmanuel
MSc in Data Analytics
National College of Ireland
Dublin
x19153813@student.ncirl.ie

***Abstract*—Automatic Teller Machines (ATMs) have proved to be one of the paramount discoveries which were made in the Banking Sector. It lowered the time limit for carrying out a transaction and also reduced the workload of the banking personnel. After the occurrence of COVID-19 pandemic in late 2019, the normal lifestyle of the people has been impacted to a greater extent worldwide. Wearing a facemask has been made mandatory in most of the public places to control the further transmission of the disease. However, when it comes to an ATM transaction, a person carrying out a transaction with their mask on is always risky and the chances for ATM fraud is increased rapidly. This project is focused on an approach designed to detect whether the person carrying out the transaction is wearing a facemask or not and allows them to carry out further transaction only after removing the facemask. The facemask mask detection model is developed with ResNet50 model as its base layer and the image sliced from the video stream is classified to the maximum accuracy of 97% detecting the face mask on the face of the user.**

Keywords—Neural Network, Facemask Detection, ResNet50, Video Stream.

I. Introduction

Automatic Teller Machines plays an imperative role in modern social and economic society. Surveillance cameras have been installed into the ATMs to identify the person who is carrying out a transaction and to ensure that a fraudulent transaction does not occur. The use of a surveillance camera helps to capture the image of the person who is carrying out the transaction. Therefore, in case of any fraudulent activities occurring in an ATM, the person can be recognised quickly. And hence for a long time, it was a standing rule to not to cover the face of the person who is carrying out a transaction.

There has been extensive use of surveillance devices in banks, shops, ATMs etc in the last few decades[1]. A surveillance device, installed in the ATM, helps in storing the facial information of the person who is conducting a transaction. This information provides great help in investigating people doing illicit activity within the ATM premise. If any fraudulent activity such as ATM skimming is carried out, then this surveillance video can serve as a vital piece of evidence to track back to the fraudster.

However, with the rise of the global pandemic in December 2019, the majority of the countries has made it mandatory that a face mask should be worn in public places. Wearing a face mask in public places can help in controlling the spread of the disease. When it comes to ATM security, people accessing ATMs with their face masks on is an increasing concern for ATM security. As per the Interim European Association for Secure Transactions (EAST), held on 10th June 2020, the impact of COVID-19 on the future of ATMs was discussed as an important concern. A detailed analysis conducted on the current scenario by EAST shows that the COVID-19 pandemic has proved that the criminals are quick enough to rise and change their criminal strategies to exploit a vulnerable situation. The strategic report published by EAST for the year 2020 has also shown an increase in the number of crimes in ATM among the European member states [2].

To wear a mask in a public place is a matter of safety and it will help to control the pandemic too. But, considering the number of fraudulent activities happening, taking the pandemic situation as an advantage, it is better if a person carries out his/her ATM transaction after removing their face mask. In this project, we are introducing a system to notify any person who is conducting a transaction to remove their mask first before proceeding further. The system will identify if the person is wearing a mask and if he/she is wearing a mask, the system generates a pop-up message asking him/her to remove the mask. The banking services will be available to the person only after the person removes their mask and complete facial recognition.

In this project, we are building a Convolutional Neural Network (CNN) using TensorFlow with Keras and OpenCV to detect if a person is wearing a facemask or not. CNN's are a feed-forward class of neural networks with demonstrated efficient results in areas such as image recognition and classification. CNN's encompass filters or neurons with weights. The base-model used is Residual Network known as Resnet50 which is a deep 50-layer model which is a classic network used for multiple computer vision tasks. The Haar Cascade Classifier which was initially proposed by Paul Viola and Michael Jones was used a base model classifier to train the model [3].

The dataset selected for the project is a collection of real-time images which was made publicly available to carryout face detection studies. The dataset consists of around 3000 images which are labelled as 'with mask' and 'without mask'. The model was trained to detect the people who have a face mask on and who does not have one. This can help in identifying the people who are carrying out an ATM transaction with a face mask and by integrating the system into the ATM, it can be redesigned such that to carry out a transaction a person must remove his/her face mask. The above-mentioned model was trained and produced an accuracy of 97%.

With an accuracy which could detect the cases of each class considered in the research, the model is designed to reduce the number of illicit activities which is happening during the pandemic situations where the current social situations make it difficult to identify a fraudster.

II. Literature Review

A. Object Detection.

Paul Viola and Michael Jones created the Viola-Jones System back in 2001 and, alongside many of their CNN's, this is still a leading face-detection framework. It works well by detecting objects in images easily and accurately and works with the human face especially well. This architecture incorporates Haar-like design definitions, integral images, and the AdaBoost algorithm to construct a rapid and precise object detection system [3]. The major drawback of this model is that it is computationally expensive.

An alternative to the Viola-Jones framework is the Histogram of Oriented Gradients feature extractor. The HOG function extractor calculates the directed gradient directions and magnitudes over the cell. The deformable partial model (DPM) later detects and connects objects to assess classes to which objects belong [4].

Instead of using handcrafted features, a deep learning-based model can be used which provides valuable feature extraction capabilities. There are two types of object detectors namely one stage and two-stage object detectors. Region proposals are generated in the primary stage for two-stage detectors and then in the second stage, these proposals are fine-tuned. In the initial stage of a Region-based Convolutional Neural Network(R-CNN) model proposed by R. Girshick et al. [5] a region proposal network is used to generate a region of interest. In the second stage, object classification and bounding box regression are done by sending these region proposals down the pipeline.

The two-stage object detectors are high in accuracy but relatively slower. Whereas, in a single stage detector a single neural network is used for object detection. The mainly used single-stage detectors include You Only Look Once (YOLO) and Single Shot Multibox Detector (SSD) that works by using predefined anchor boxes. These anchor boxes contain the specification of the image including the width and height. The YOLO detector achieves high speed by partitioning the image into cells and matching it with an anchor box. Since this approach is not good for small objects [6], multi-scale detection has been proposed in SSD. This method is done for the detection of faces of different sizes on various feature maps [7]. Later, a RetinaNet model was proposed by Lin et. al [8] which combines the functionalities of an SSD and FPN to mitigate class imbalance problems and improve detection accuracy.

B. Face Detection.

Face Recognition is a process by which a system recognizes a person's face using computer vision technology. It is one of the most important advancements in technology mostly used in the fields of access control, security systems, and video surveillance [9]. Studies show that the conventional shallow learning methods have been facing pose variation, facial disguise, and lighting issues since they make use of only basic features of an image and are more reliable on artificial intelligence [10].

More complicated face features can be extracted using Deep Learning methods [11]. Deep learning is making significant strides in addressing problems that have for many years limited the best efforts of the artificial intelligence community. It has proven excellent in exposing complex structures in large-scale data and thus extends to other areas of research, industry, and government [12].

C. Deep Learning Models.

The mostly used deep learning methods include Convolutional Neural Networks (CNN), Stacked Autoencoders, and Deep Belief Network (DBN) [13]. CNN is used in this project for facial recognition. CNN is a type of artificial neural network which uses a convolution procedure for converting input data to extract the characteristics to increase features. LeCun introduced the idea of CNN and was first used in recognition of handwriting [14]. Owing to its superior space extraction and less processing costs, CNN plays an important role in computer vision-related pattern detection tasks [15]. [16] proposes an inception network model that lets the network to learn the best kernel combinations. much deeper networks can be trained using a Residual Network (ResNet) model [17]. Another proposed model includes using Mobile Network (MobileNet)[18] which can be used for object detectors deployed on portable devices or mobiles.

In research done at NIT Rourkela, India, VGG face deep CNN is used for gender detection. This is done by training the VGGNet-face detector to recognize more than 10000 images of celebrities. The process was done by learning a new SoftMax layer above the FC layer to differentiate the gender to male or female. Among the two convolutional layers, the output from each layer is activated by an activation function named RELU. Finally, the output of the previous layer is needed to do max-pooling to reduce the size of the images [19]. A similar approach is used in our research paper as well.

Another approach to face detection includes a proposal in which the face images are highlighted onto the principal components of the actual set of training data. The resulting faces are known as eigenfaces and they are classified by comparing them with known individuals. This approach was proposed by Turk and Pentland [20]. Turk and Pentland present data on 16 faces with different head orientations, scaling, and lighting topics. For illumination, orientation, and size variations their method achieves 96 percent, 85 percent, and 64 percent correctly categorized respectively. They do have a minor difference in the facial appearance and facial features and location of facial images. The scale is renewed to its surface size based on the head size calculation. The middle of the faces is stressed to minimize the adverse impact of shifting hairstyles and backgrounds.

Another approach called Attention Mechanism is used to simulate human attention that can concentrate on important details. Attention Mechanism is used in the recurrent neural network (RNN) with an additional encoder-decoder approach. In the Convolutional Block Attention Module (CBAM), A basic but successful attention mechanism is proposed with spatial and channel attention [21].

In this project, the Convolutional Neural Network method is used along with the functionalities of ResNet. The dataset consists of Masked and Non-Masked images. The base model is trained using the training dataset. Optimization is done using Adam optimizer. Advancements of deep learning technology are utilized to solve a real-world obstacle faced during the pandemic situation. The project makes use of CNN functionality to prevent fraudulent activity happening within the ATM premise by detecting and advising the customer to remove the mask.

Research Question:

"What is the maximum accuracy and precision of detecting the presence of mask on the face can be attained using Deep CNN?"

III. METHODOLOGY

The methodology used in this research paper is Knowledge Discovery in Databases (KDD). KDD is a methodology used to identify novel patterns from broad and complex datasets. The procedure starts with the steps to determine the goals of the KDD process and ends with implementing the determined knowledge.

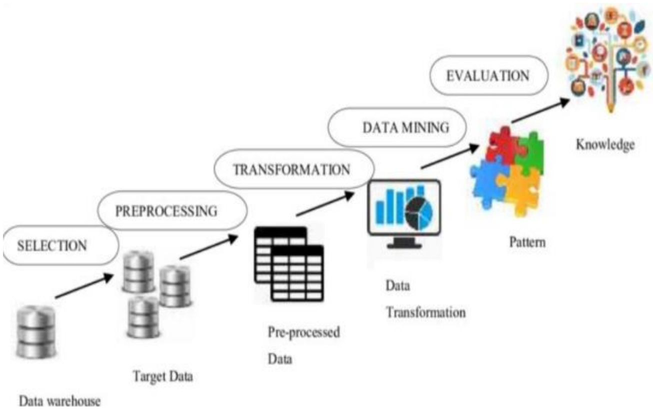


Figure 1: KDD Methodology

A. Data Selection

The dataset used here is a real-time dataset which is available to the public to develop and research on problems related to face detection. The dataset consists of 2700 images that belong to two classes namely 'mask' and 'no mask'. The former class consists of 1281 images and the latter has 1476 images. The dataset was created using computer vision by applying facial landmarks. Faces were captured in real-time without proper orientation.

B. Data Preprocessing

The dataset has been saved to google drive and then imported to google collab using the 'import' command. Then the dataset has been labelled as '0' for 'without_mask' and '1' for 'with_mask'. Then the images are converted to a NumPy array and then resized and normalized to avoid overfitting. The labels of the images are then converted to categorical values.

C. Transformation

The pre-processed data is split into training and testing data with a test size of 30 and a training size of 70. The split data is saved into four NumPy files using 'np.save' function. One file for the training dataset and one for test dataset for masked and non-masked dataset each.

D. Data Modelling

a) Proposed Approach: The base-model used in this project is ResNet. ResNet stands for Residual Networks, The major advantage with ResNet is that it can be used for training 150+ deep learning layers. In our project, Resnet50 is used which has 5 stages with convolutional and identity blocks. Here we have used the 'weights=None' for initializing the model with random predefined weights as the dataset that we intend to provide as input. An alternative to this is to use the pre-trained weight as 'ImageNet'. The 'include_top' is initialized as 'False' to disclude the pooling and fully connected layers in the base model. Instead of that, we have added Global Average Pooling and the 'Dense' Activation layer to the model. The deeply connected neural network dense layer has 512 layers, 'Relu' activation layer is used in the dense layer. The final output layer is named as 'predictions' and 'softmax' activation function is used in this layer. The 'predictions' have 2 outputs namely 'with_mask' and 'without_mask'. The model is defined in the next step, The input of the base model is provided as input to the model and the predictions are provided as an output of the model. Adam Optimization Algorithm is an extension of a stochastic downward gradient algorithm, recently seen wider use in computer vision and natural linguistic processing for deep learning applications. For optimizing the model, the ADAM optimization algorithm is used which is a stochastic downward gradient algorithm, recently seen wider use in computer vision and natural linguistic processing for deep learning applications. Then the model is fitted with a batch size of 64 and a total of 20 Epochs.

b) Model Architecture:

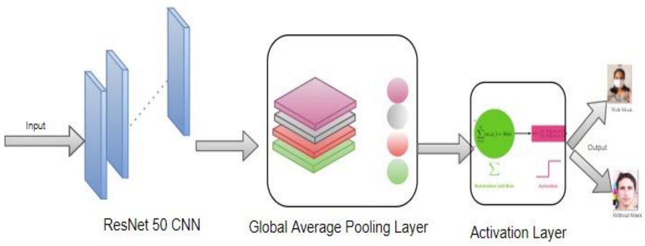


Figure 2: Implemented Model Architecture

Pooling Layer: The size of each activation map is reduced but the most important information will still be available. The input images are divided into a set of rectangles that do not overlap. Sampling each region is done by specific operations such as average or maximum.

Base Model: Resnet50 which is one of the major advancements in deep learning is used in this project. Before Resnet came into existence, training the model seemed to be one of the hardest tasks due to fading gradients. Skip connection is one of the major strengths of the ResNet model. ResNet50 has 50 neural network layers. The network takes the input image in multiples of 32 as width and height.

Global Average Pooling: The size of each activation map is reduced but the most important information will still be available. The input images are divided into a set of rectangles that do not overlap. Sampling each region is done by specific operations such as average or maximum. To minimize the overfitting of the model, Global Average Pooling Layers are used in our model. The spatial dimensions of a 3-dimensional tensor are reduced by using GAP layers.

Activation Layer: The layers are reduced into 512 and a pyramid-like shape**.** The major role of this layer is the reduction and the activation function used here is the RELU. The activation function is responsible in a neural network for converting the total weighted input from the node into the node or output activation for this input.

Output Layer: The output layer here is named as ‘Prediction’, in this layer, the output of the dense layer is fed as input and the output is obtained as a binary value of either masked or non-masked. The output of this layer is fed directly as output to the model.

IV. Implementation

Implementation of face mask detection using the convolution neural network will be explained in this section of the report. This implementation of this model takes place in the process of five steps as shown in the Figure below.

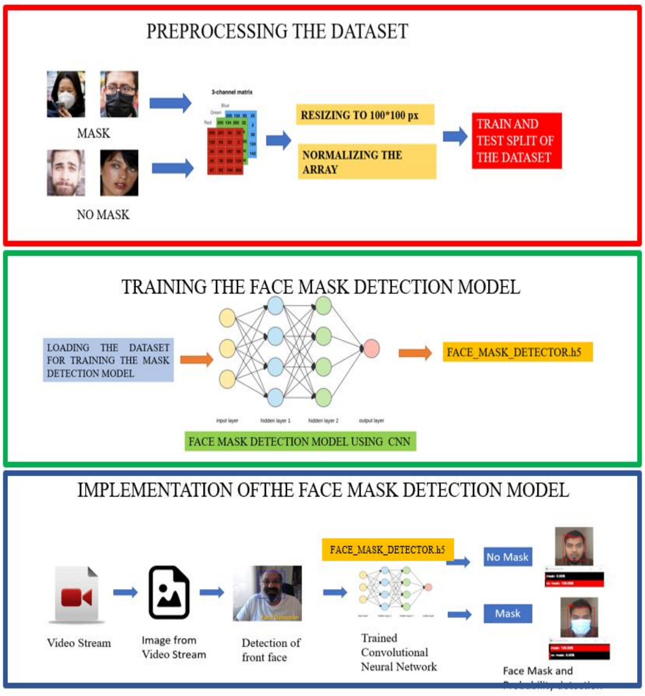


Figure 3: Implementation of Face Mask Detection

1. *Capturing the frames from Video Stream.*

This implementation is initialized from capturing the frames from the live stream from the webcam which is captured at a rate of 30s per frame. The image from the video stream is resized to the width of 600px. To capture the multiple frames, the webcam has been processed to multiple loops and the video is a capture using the primary camera source which is chosen by cv2.VideoCapture(0).

2. *Detection of Region of Interest from the Images.*

Next step of capturing the images from the video stream, the images are converted into grayscale and then resized to the width of 600px. A pretrained model which is publicly available in the OpenCV repository to detect the front face of the person is used to detect the presence of face in the image frame is used in this model, which helps the model to initiate the frame capture once the face is detected in the video stream. The pre-trained HaarCascade Classifier model available in the Open repository ‘haarcascade_frontalface_default.xml’. The cascade face classifier is customized to detect the face of any scale in the image. The canvas window showing the probability is made to replicate the frame of the video stream. The feature extraction is done from the area of Region of Interest.

3. *Processing the Image from Video Stream.*

The trained face mask detection model of Convolutional Neural Network has been trained by the resizing the images to 100*100 pixel and which is resized by dividing the array with 255.0, so the image with the ROI is resized and normalized the same way so it could be suitable for the model to predict the face mask from the face in the frame of the image. The Probability of the face mask detection from the ROI in the image is presented in an individual window. Once the task of the face mask detection is performed the windows of video stream and probability were made to close when the classification is done.

4. *Trained Convolutional Neural Network Model.*

The features extracted from the image is provided as input to the convolutional neural network layers. The Resnet50 is the neural network, which is utilized here as base model, it provides a deep network of 50 layers which helps to train the model in a better way. The base model is not trained with any predefined weights. The shape of input is defined same as the image size from the processed image which is 100*100 pixels.

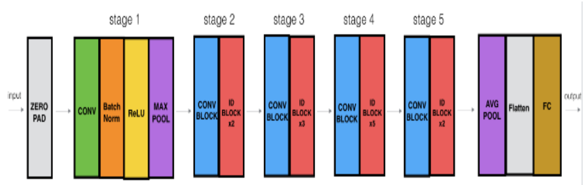


Figure 4: Layers of ResNet50

The output obtained from the base model is made to undergo global average pooling and for the input of the model the ReLU activation function is used and for the output, the SoftMax is used. The model is then optimized with the Adam optimizer with the learning rate of 0.0001, the loss here considered is the categorical cross-entropy, and the evaluation metrics of every training epoch is calculated with the accuracy. The model is then trained with the train and validation dataset with the batch size of 64 and epochs considered here is 20.

5. Model Extraction.

The trained model is saved in the format of Hierarchical Data Format version 5, which is mentioned as '.h5' to the file name. This model can now be utilized for the purpose of face mask detection in any field of application. Since this format provides flexibility and support to data of any size this saved model can be utilized for detecting the face mask on the users' face.

V. EVALUATION

This section discusses about the evaluation procedure carried out on the developed model to analyze its efficiency. Scikit python library is utilized for this purpose.

A. Accuracy and Loss

Accuracy of the model represents the percentage of the classifying accurately the images in the test dataset set. The model is trained with a sample of 2205 samples from the training dataset and tested with 828 images which are split with the proportion of 80:20. The graph in the Figure below represent the accuracy, Validation accuracy, loss and validation loss plot against every epoch while training the model. The Validation accuracy achieved at the end of the Training the model is 95%. Accuracy of the model has started to increase after the second epoch and continues exponentially and reach the max accuracy of around 96% by the end of the epochs and the validation accuracy is slightly above the accuracy in every epoch. And the loss is exponentially reduced in consecutive epochs and reaches is minimum point lesser than 15% by the end of training. This is shows that the model is well trained, and the detection of face mask can be performed well.

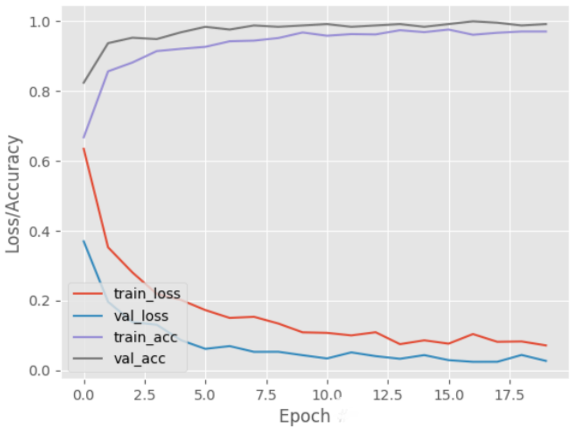
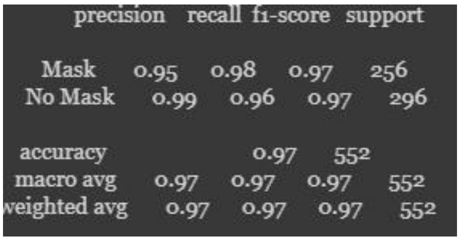


Figure 5: Accuracy and Loss vs Epoch

B. Classification Report

The Classification Report is obtained for the model with the help of 'scikit.metrics' package in python programming, which is provided in the Figure below.



	precision	recall	f1-score	support
Mask	0.95	0.98	0.97	256
No Mask	0.99	0.96	0.97	296
accuracy			0.97	552
macro avg	0.97	0.97	0.97	552
weighted avg	0.97	0.97	0.97	552

Fig.4 Classification Report

The confusion matrix of the classification of face mask detection is plotted and the precision, recall and F1 scores were identified as the evaluation metrics of the model. Precision represents the correct prediction of the true positive cases of both the classes. This model has produced the precision of 0.97 and 0.99 in predicting the mask and no mask classes respectively from the test dataset. Recall represent the prediction of actual positive cases in both the classes and it is found to be 0.98 and 0.96 for mask and no mask classes, respectively.

F1 Score represent the relation between the precision and recall. F1 score extends from 0 to 1, where 0 represent the worst model and 1 represents the best model. We got the F1 score to be 0.97, which shows the goodness of the model.

C. Cohen's Kappa Coefficient

Kappa coefficient provides the inter-relationship between the two categorical classes considered for the classification, which is mask and no mask in our case. It provides the insight about the robustness of the model, since it provides the possibilities of classification that occurred by chance. This value is obtained by the formulating the values with the formula given below.

$$k = \frac{p_0 - p_e}{1 - p_e}$$

where P0 is the value of observed outcome which has the value same as accuracy and Pe is the probability of hypothetically getting the outcome desired. The Kappa value for the model is obtained as 0.811, which represent the model's reliability in detecting the face mask.

V. RESULTS

The model has been trained with the preprocessed images from dataset and the model has been saved as a separate file which is utilized to classify the image obtained from the video stream. The model has been trained with the image which is preprocessed and obtained in RGB format. The model has been compiled and the fitted with the preprocessed data and the model is trained with the epoch count of 20. Every epoch is monitored for the accuracy and the loss and the validated with the test set and the validated accuracy and loss is obtained, which has been shown with the plotted graph in Fig.5. The model has been evaluated and the accuracy is measured to be 97% and with the precision of around 96% in detecting the binary classes of the images.

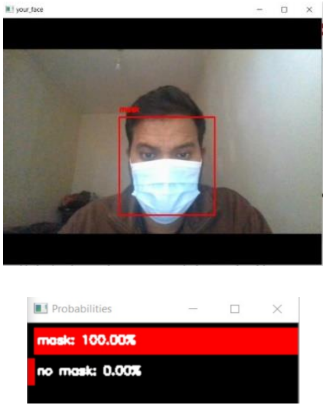


Fig.6 Detection of face with mask and the percent of face covered by mask

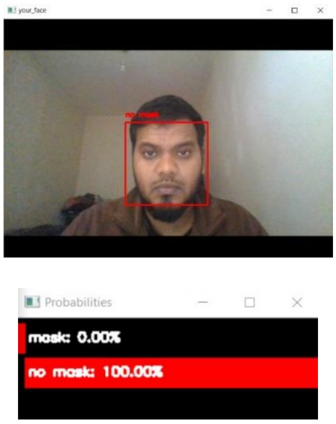


Fig.7 Detection of face without mask and the percent of face without mask-on.

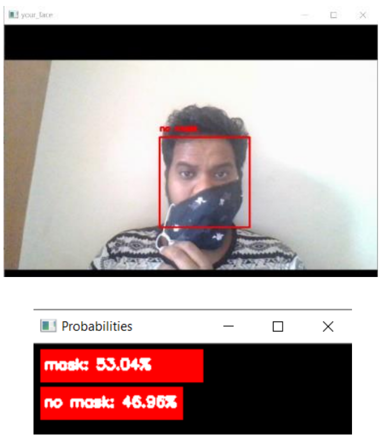


Fig.8 Detection of partial face cover and the probability of the covered face.

VI. CONCLUSION AND FUTURE SCOPE

This model is developed to detect the face cover of the users of the ATM machine which is accessible for the public to withdraw the amount. With reportedly many anomaly activities recorded at the ATM stations by the users with the mask on the face. This face mask detection model can be deployed to ATM software as a pre-authorization step before initiating the transaction, especially during the pandemic situation which has made usage of face mask mandated everywhere this detection will make users to not use the mask until the transaction is complete so that the users face can be recorded from the surveillance camera placed at the top of the machine at every station.

VII. REFERENCES

[1] Q. Rajpoot and C. Jensen, "Security and Privacy in Video Surveillance: Requirements and Challenges", ICT Systems Security and Privacy Protection, pp. 169-184, 2014. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] "Interim EAST Meeting - National and Global Members | EAST", EAST, 2020. [Online]. Available: https://www.association-secure-transactions.eu/interim-east-meeting-national-global-members/. [Accessed: 08- Aug- 2020]K. Elissa, “Title of paper if known,” unpublished.

[3] P. Viola and M. Jones, "Robust Real-Time Face Detection", International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004. Available: 10.1023/b:visi.0000013087.49260.fb.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan, "Visual object detection with deformable part models", Communications of the ACM, vol. 56, no. 9, p. 97, 2013.

[5] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 2016.

[6] D. H C, "An Overview of You Only Look Once: Unified, Real-Time Object Detection", International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 6, pp. 607-609, 2020.

[7] A. Kolekar and V. Dalal, "Barcode Detection and Classification using SSD (Single Shot Multibox Detector) Deep Learning Algorithm", SSRN Electronic Journal, 2020.

[8] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal Loss for Dense Object Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 2020.

[9] "Face recognition based on convolutional neural network - IEEE Conference Publication", Ieeexplore.ieee.org, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8248937. [Accessed: 08- Aug- 2020]

[10] Z. Zong, "Efficient Human Face Recognition Method under Subtle SIFT Features Using Optimized K-means", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 10, no. 7, pp. 195-204, 2017.

[11] K.CVA,"https://www.medwinpublishers.com/JOBD/JOBD16000139.pdf", Journal of Orthopedics & Bone Disorders, vol. 1, no. 7, 2017.

[12] T. Le, C. Duong, L. Han, K. Luu, K. Quach and M. Savvides, "Deep contextual recurrent residual networks for scene labeling", Pattern Recognition, vol. 80, pp. 32-41, 2018.

[13] Y. Bengio, "Learning Deep Architectures for AI", Foundations and Trends® in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[14] G. Hinton, S. Osindero and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[15] L. Fei-Fei, J. Deng and K. Li, "ImageNet: Constructing a large-scale image database", Journal of Vision, vol. 9, no. 8, pp. 1037-1037, 2010.

[16] "IEEE Computer Society Conference on Computer Vision and Pattern Recognition Fontainebleau Hilton, Miami Beach, Florida June 22-26, 1986", Computer, vol. 18, no. 12, pp. 22-22, 1985.

[17] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang and Q. Sun, "Deep learning for image-based cancer detection and diagnosis − A survey", Pattern Recognition, vol. 83, pp. 134-149, 2018.

[18] J. Xiong, L. Ye, D. Jiang, T. Ye, F. Wang and L. Zhu, "Efficient Traffic Sign Recognition Using Cross-Connected Convolution Neural Networks Under Compressive Sensing Domain", Mobile Networks and Applications, 2019

[19] A. Dhomne, R. Kumar and V. Bhan, "Gender Recognition Through Face Using Deep Learning", Procedia Computer Science, vol. 132, pp. 2-10, 2018.

[20] M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[21] J. Park, S. Woo, J. Lee and I. Kweon, "A Simple and Light-Weight Attention Module for Convolutional Neural Networks", International Journal of Computer Vision, vol. 128, no. 4, pp. 783-798, 2020.