OPEN ACCESS

International Journal of Innovations in Science & Technology

RESEARCH & INNOVATION DIVISION

IJIST

Enhancing Face Mask Detection in Public Places with Improved Yolov4 Model for Covid-19 Transmission Reduction

Muhammad Ayaz[1], Adnan Hameed[1], Said Khalid Shah[1], Muhammad Javed Khan[1], Sadiq Nawaz Khan[1], Fouzia Idrees[2], Zamar Ali Shah[1],

[1] Department of Computer Science University of Science and Technology Bannu, KP, Pakistan.

[2] Department of Computer Science, Shaheed Benazir Bhato Women University of Peshawar, Peshawar, Khyber Pakhtunkhwa, Pakistan.

***Correspondence**: ayazkhanttt@gmail.com

Citation | Ayaz. M, Hameed. A, Shah. S. K, Khan. M. J, Khan. S. N, Idrees. F, Sha. Z.A "Enhancing Face Mask Detection in Public Places with Improved Yolov4 Model for Covid-19 Transmission Reduction", IJIST, Vol. 6 Issue. 1 pp 275-286, Mar 2024

Received | Feb 20, 2024 : **Revised |** March 21, 2024 : **Accepted |** March 26, 2024 : **Published |** March 28, 2024.

Over the past decade, computer vision has emerged as a pivotal field, focusing on automating systems through the interpretation of images and video frames. In response to the global impact of the COVID-19 pandemic, there has been a notable shift towards utilizing computer vision for face mask detection. Face masks, endorsed by international health authorities, play a crucial role in preventing viral transmission, prompting the development of automated monitoring systems in various public settings. However, existing artificial intelligence (AI) technologies' effectiveness diminishes in congested environments. To address this challenge, the study employs a meticulously fine-tuned YOLOv4 model for identifying instances of mask non-compliance in accordance with COVID-19 Standard Operating Procedures (SOPs). A distinctive feature of the training dataset is its inclusion of images featuring Muslim women with both half and full-face veils, considered compliant with face mask guidelines. The dataset, comprising 5800 images, including veil images from various sources, facilitated the training process, achieving a comparatively good 97.07% validation accuracy using transfer learning. The adaptations, coupled with a custom dataset featuring crowded images and advanced pre-processing techniques, enhance the model's generalization across diverse scenarios. This research significantly contributes to advancing computer vision applications, particularly in enforcing COVID-19 safety measures within public spaces. The tailored approach, involving model adjustments, underscores the adaptability of computer vision in addressing specific challenges, highlighting its potential for broader societal applications beyond the current global health crisis.

Keywords: COVID-19, SOPs, Yolov4, Face mask detection, Offline-augmentation, Transfer learning, Veil images

Introduction:

Scientists initially identified a human coronavirus in 1965, marked by symptoms akin to the common cold. A decade later, a group of related human and animal viruses was identified and named coronaviruses due to their crown-like appearance. Seven variants of coronaviruses can impact humans, with COVID-19 emerging in Wuhan, China, in December 2019, initially linked to a large animal and seafood market [1][2]. The virus spread rapidly during the Chinese Spring Festival celebrations, becoming a nationwide concern [1]. COVID-19, initially named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS CoV-2), was officially designated as COVID-19 by the World Health Organization (WHO) on February 11, 2020 [3]. The situation escalated to a pandemic on March 11, 2020, affecting millions of people globally [3]. The virus, an RNA structure measuring 30-32 kb, is visible only under an electron microscope due to its glycoprotein envelope, categorized into Alpha, Beta, Delta, and Gamma genera [4].

WHO emphasized precautionary measures, including hand hygiene, social distancing, and, notably, face mask usage, considered a crucial preventive step [5]. As COVID-19 symptoms range from severe fever to respiratory distress, wearing face masks has emerged as a simple yet effective preventive measure, especially for vulnerable groups with weak immune systems or underlying health conditions. Despite the implementation of various strategies like curfews and lockdowns globally, the economic toll prompted the search for alternatives. WHO issued SOPs, stressing the importance of face masks even post-vaccination due to the risk of breakthrough infections [6]. However, challenges persist, with a significant portion of the population remaining unvaccinated and the emergence of new variants like Omicron. After the announcement of WHO, a new variant of COVID-19 called Omicron was detected in South Africa, and also on December 13, 2021, in Karachi, the National Command Operation Centre (NCOC), an organization of Pakistan working against the Coronavirus, spread the message to follow SOPs and use face masks in public places.

The monitoring of face mask compliance remains a critical issue for government organizations in controlling the disease, given the availability and affordability of face masks in markets. In the ongoing battle against COVID-19, the effective use of face masks stands out as a key component in mitigating the spread of the virus in public spaces. Face mask detection is a sub-part of object detection and classification in the area called digital image processing and computer vision. It was first adopted by the government of France in a Paris metro station using CCTV cameras [7]. Face mask detection is the latest research area under consideration after the outbreak of COVID-19 at the end of 2019. Most of the researchers used deep learning for face mask detection, which works well but fails when there is a crowd in an image or video frame. Also, new shapes of masks that are not part of the training dataset also affect the model's accuracy. In the proposed method, a deep neural network-based model, YOLOv4, was modified and fine-tuned for face mask detection and trained on a custom dataset containing 5800 images. Images were taken from a real environment as well as from the internet. Full-face and half-face veil images used by Muslim women are also added and annotated as masks. This model performed well on images as well as on videos, with high accuracy compared to existing systems. Key contributions of the study are:

- Collection of 3000 images of both classes from various sources, including veil images as face masks.
- Offline data augmentation techniques were used, and the training dataset was enlarged to 5800 images to enhance and generalize the model performance.
- Offline and online pre-processing techniques were used before training the model.
- The model was modified and fine-tuned to fit the training dataset well.
- The model was also validated on three public datasets.

Literature Review:

Face-mask surveillance is part of object detection and recognition using digital image processing techniques. There are two main categories of digital image processing: traditional image processing and deep learning-based image processing. In traditional image processing, mathematical formulas are used to detect and analyze the images, while in deep learning-based systems, human brain-like models are trained to detect and analyze images. Most of the previous research was based on deep learning-based models. In the proposed method, the YOLOv4 model is used to detect face masks in congested areas with some advanced image segmentation techniques to improve the results of the existing systems.

C Li et al. [1] used Yolov3 to detect face masks in public places. They used data augmentation techniques called image mix-up and multi-scale training to improve the accuracy of the model. They used the WIDER FACE and MAFA datasets and got 86.3% mAP with a 1.2% improvement compared to the existing literature. The study in [4] used deep learning-based models to classify human faces as wearing masks or not. They used the CNN model with Max pooling and got 96%, and average pooling got 96.5%. They also used MobilenetV2, which got 99% accuracy. They used data augmentation and image processing techniques to improve the models trained for face mask classification. Md. R. Bhuiyan et al. [5] used preprocessing techniques to enhance the training images. Six hundred images were taken from the internet having both categories mask and without mask persons. The labeling tool was used for data annotation. The YOLOv3 model was trained to detect face masks in public places. S. Balaji et al. [6] used the VGG-16 model to detect people without face masks in government offices. They used a custom dataset of images with people labeled "Mask and No-mask."

G. J Chowdary et al. [8] used the inceptionv3 model to classify people in public places with or without face masks. They used transfer learning instead of training from scratch and got 100% testing accuracy. They used data augmentation to reduce error loss. The dataset used for training and testing was the Simulated Face Mask dataset (SMFD) having 1570 images of both categories (with and without face masks). BU Mata et al. [9] used data augmentation to increase the volume of the dataset to improve the performance of the model. The face area was extracted as a region of interest (ROI), and a CNN model was built and trained to recognize the ROI as having a face mask or not. M. M. Rahman et al. [10] used pre-processing techniques to improve the quality of the training pictures. The CCTV image was converted to greyscale and resized to 64x64. Each image was normalized, and a CNN model was created and trained using one input layer, numerous hidden layers, and one output layer. G Yang et al. [11] used an intelligent gate system that allows only those who wear face masks. The YOLOv5 model was trained using a custom dataset. Input images were pre-processed, segmentation was performed to increase detection accuracy, and the YOLOv5 model was used for face mask recognition.

B. Sathyabama et al. [12] used their own dataset of 3800 images collected from different sources and trained the YOLOv3 model to ensure the usage of face masks in public places. They also used Euclidian distance and the Yolov3 model to ensure social distancing between people in crowded areas. Surveillance videos were used to capture videos. These videos were converted to frames, and detection was performed. X Ren and X Liu [13] used the YOLOv3 model to detect face masks with minor changes. They changed the anchor boxes as per their demand, and the IoU was also changed by DIoU. Jiang et al. [14] used a modified version of YOLOv3. They included the squeeze and excitation blocks in Darknet 53 to improve feature extraction. They trained the model using a dataset of 9205 images with three categories to detect face masks to ensure COVID-19 SOPs. S. Saponara et al. [15] used a modified version of YOLOv2 to detect face masks, and the same model was also used for social distancing to control the spread of the Coronavirus in public places. Authors A. El-aziz et al. [16] also utilized deep learning through CNN with Keras and TensorFlow frameworks to detect face masks. They used 2000 images from various sources i.e. Kaggle. They developed a mask detector through which 95% results were obtained as can be seen in the figure. They used some other methodology to preprocess the images for noise removal and for normalization they used reshaping techniques for images before training in order to achieve better performance.

E. Mbunge et al. [17] also provided a comparative study on AI-based classifiers detecting face masks in a densely populated area and the Inceptionv3 outperformed all models with a training accuracy of 99%. Identifying the trajectory and position of the tennis ball during tossing were focused on [17] using the Faster-RCNN, SSD, and Yolov4 models and the Yolov4 was compared and validated the superiority of SSD and Faster-RCNN and obtained a very impressive 93% accuracy rate [18]. These studies deliver a key idea that the model can be effective in diverse areas from face mask detection to object tracking. L. Fang et al. [19] built YOLOv4 with customized adjustments, replacing the CSPDarknet53 backbone with the more efficient Mobilenet-v2 network to reduce redundancies. This upgrade is intended to improve the model's ability to detect ginger shoots prior to cultivation, with a particular emphasis on increasing recognition rates for smaller targets. Tenzin et al. [20] applied image enhancement methods to increase the quality of the image, and Robert's edge detection algorithm was used to detect the vertical lines with a constant digit of binarization (0.0375) to refine the image for the number plate, and the rectangular figure was filtered as expected for the number plate area. The bounding box technique was used for character segmentation, and template matching was used to recognize the plate alphanumeric characters. M Ayaz et al. [21] detected shapes and colors and combined both of these properties with late fusion to detect objects.

According to the background study presented above, it shows that in the short term, the researchers worked on face mask classification with great effort and developed various systems to force people to wear face masks in public places to decrease the chance of COVID-19 virus transmission with high accuracy. The proposed method aims to fill the gap left by the existing systems. The existing systems fail to detect face masks when the image or frame is crowded and at a distance from the surveillance cameras. They also ignore the veils used by Muslim women, and existing systems classify full and half-face veils as without masks. In the proposed method, efforts were used to resolve these two issues.

Proposed Method:

In this study, the YOLOv4 model underwent further changes, fine-tuning, and intensive training specifically aimed at identifying individuals wearing face masks inside facial photos. To address over-fitting problems, online augmentation techniques were substituted by offline augmentation procedures. Significant changes were made, including setting the learning rate to 0.0001 and increasing the input size to 512 × 512. Offline augmentation was assisted by the Keras image data generator toolkit, and the resulting photos were meticulously stored in the training folder. A thorough screening process follows to exclude undesirable photos, followed by preprocessing methods to enhance the quality of the remaining pictures. For model training, a custom dataset of 3000 images was taken, and through offline augmentation, the training dataset expanded to 5800 images. Image labeling was accomplished using Labeling. Transfer learning (TL) was incorporated during the model training process. The workflow of this proposed study is visually represented in Figure 1.

Dataset Acquisition:

In computer-based visionary systems, a dataset is a collection of relevant digital images that developers use to train, test, and validate their model's performance. The success of this learning-by-example method is dependent on a specific scale. Algorithms trained on larger datasets outperform those trained on smaller ones by a wide margin. More data means more variants and the algorithm can learn from the visual world's many differences. A change in the amount of sample data causes a qualitative change in the algorithm's performance in modern machine learning. In data-driven techniques such as deep neural networks and machine learning, data is extremely crucial. The more data you have, the better your results will be. We also require more data, as well as annotated data, in order to operate with YOLO. However, no annotated data was available for this project. We collected 3000 photos of two classes of masks and no-masks from the internet, including Kaggle, a GitHub repository (https://github.com/techzizou), and also from real scenarios using a smartphone. Using offline augmentation, the training dataset was enlarged to 5800 training images.

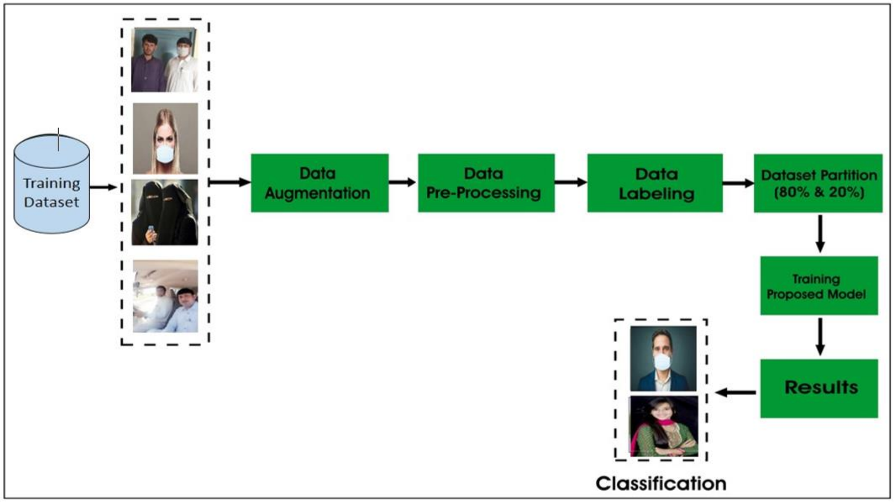

Figure 1: General flow of the proposed system

Data Augmentation:

Training a deep neural network model on an extensive dataset typically yields superior results compared to models trained on smaller datasets. The model's performance is generally contingent on the dataset size; augmenting the data positively impacts the deep learning model's efficacy. Techniques employed for dataset expansion and introducing variations to training images fall under the umbrella of data augmentation. The pivotal aspect for effective and accurate model training lies in the variability of images within the dataset. Sourced from diverse outlets such as the internet and real-life scenarios, these images were captured with attention to lighting conditions. To ensure model accuracy, images were taken at different distances and angles. Following image capture, offline augmentation techniques (i.e., zooming, rotation, shifting, brightening, etc.) were applied to enlarge the training dataset. Augmented images, deemed unsuitable for training, were filtered out.

The remaining augmented and real images were amalgamated, resulting in a comprehensive training dataset comprising 5,800 images.

Image Preprocessing Techniques and Labelling:

Prior to putting images into a CNN, preprocessing techniques must be used to optimize the model's performance. Image resizing and normalization are critical preparatory procedures for training a CNN to perform computer vision tasks. The image resizing process alters input images to a common size so that they can be processed more effectively by the CNN model. This step is crucial since CNN models are created and pre-trained for specific input sizes, and deviations from these sizes can negatively impact model performance. In this study, all input images were equally resized to 512×512. Image normalization, another important technique, involves changing pixel values to a defined scale or range. This normalization ensures that pixel values exhibit similar ranges, a prerequisite for effective learning across various machine learning algorithms, including CNNs. In this particular study, all images were normalized to fall within the range of 0 to 1. Through the implementation of these image preprocessing techniques, we achieved an enhancement in the CNN model's accuracy, facilitating superior performance on new and previously unseen data. The LabelImg tool was instrumental in annotating images, meticulously labeling them according to the specified problem categories, namely, "Masked and No Mask".

YOLOv4:

Object detection and recognition, an automated process in computing, involves two main categories: two-stage detectors and single-stage detectors. Two-stage detectors, like R-CNN, FPN, and Mask-RCNN, initially scan for ROIs and then classify and recognize the target object. In contrast, single-shot detectors, exemplified by Retina-Net, SSD, and YOLO, perform both classification and localization in a single stage. YOLO, coined by Joseph Redmon in 2015, saw subsequent improvements with YOLOv2, YOLOv3, and the latest, YOLOv4. YOLOv4 stands out as the most advanced for real-world applications, demonstrating enhanced accuracy, especially for small targets. It incorporates cross-stage partial (CSP) networks, modifying the neck from Feature Pyramid Network (FPN) to spatial pyramid pooling (SPP) and path aggregation network (PANet), resulting in a 10% map increase and 12% FPS boost compared to previous YOLO versions.

YOLOv4 stands as a significant advancement in object detection models, renowned for its speed and accuracy. Its architecture comprises three pivotal components: the Backbone, Neck (which integrates Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PANet)), and Head. Each of these components is essential for achieving the model's robustness and efficiency in detecting objects within images. The first of these components, the Backbone, employs a CSPDarknet53 architecture, and it is the foundation of YOLOv4, as its primary role is the extraction of hierarchical feature representations from input images. In order to improve training for this component, and to address issues such as vanishing gradients, the Backbone leverages a number of techniques, such as Cross Stage Partial Networks (CSPNet). When training this component, this zero-cost technique creates a parallel route alongside the main convolutional path, and this parallel route helps to enhance gradient flow, which in turn, mitigates vanishing gradient problems during back propagation.

The Neck acts as a bridge between the Backbone and the Head, facilitating feature pooling and fusion. SPP, within the Neck, enables YOLOv4 to better capture contextual information across different scales. PANet, on the other hand, aggregates multi-scale features with adaptive convolutional kernels and greatly improves the model's ability to detect objects of varying sizes and scales. The final component, the Head of YOLOv4, is tasked with generating the network's output, which includes predicting bounding boxes, class probabilities, and confidence scores for each detected object. The Head takes as input feature maps and processes them through convolutional layers to predict the bounding boxes for each grid cell. Furthermore, as it does this, it also calculates both class probabilities and confidence scores. To refine the predicted bounding boxes and reduce redundancy, YOLOv4 applies Non-Maximum Suppression (NMS) to filter out duplicate detections based on their confidence scores and overlapping regions. YOLOv4's detection head is so efficient and accurate that it results in strong robust object detection capabilities, even in cluttered scenarios with multiple objects in close proximity. By combining these elements effectively, YOLOv4 achieves state-of-the-art performance in object detection tasks, making it an ideal choice for real-time applications where speed as well as accuracy are paramount. Moreover, its flexible design allows for flexibility and ease of integration, allowing developers and researchers to modify and extend the model for different scenarios and domains.

Transfer Learning (TL):

TL is a powerful method that builds on initial model weights while training, eliminating the need to flag off a training run from scratch. These initial weights are garnered from prior training on huge image datasets — often Image-Net and MS-COCO. YOLOv4, for instance, implements pre-trained weights fine-tuned from the COCO dataset, which includes a good eighty object classes. The convolutional part of the model, at least, is where the bulk of the pre-trained knowledge can be capitalized, element-wise. The upper layers, which are responsible for concluding the object class, on the other hand, are usually entirely retrained. The weights file in TL contains the kind of low-level features that are universal across different classes of objects — lines, edges, shapes, etc. This technique has several advantages, including reducing training time and preventing over-fitting, which is especially useful in circumstances when the training dataset is limited in size. Integrating TL improves model efficiency by taking on previously learned features, boosting performance across a variety of computer vision applications.

Modifications and Fine-Tuning:

To a large extent, the efficiency of a model depends on the attributes of the dataset, such as object shape, size, and class differences. On-line augmentation is utilized in YOLOv4 to minimize data annotation effort and save disk space. However, this weakens model accuracy compared to offline augmentation as it had to carry out this feature to save time over efficiency [21]. The question arose as to whether the output of online augmentation may change images in such a way that does not alter the image visibly to humans but rather confuses the machine. Our study, thus, discarded on-line augmentation techniques and instead employed off-line techniques, in order to not only introduce variations to training images but also increase the dataset size, thereby ultimately improving the network's detection performance. To customize the YOLOv4 darknet framework, it needs to change the src/data.c or yolov4.cfg files. The configuration file specifies the network architecture, hyper parameters, and settings. Changes to the layer count, input picture size, and anchor boxes necessitate recompilation of the darknet framework as well as model retraining.

In this method, model changes were made using the yolov4.cfg file. Following offline augmentation, the newly generated photos were carefully examined, resulting in the removal of incorrect ones. The training dataset increased from 3000 to 5800 pictures. LabelImg was used for picture annotation. Fine-tuning included changing parameters like learning rate and input image size (416 × 416 to 512 × 512). Following these changes, the model demonstrated relatively good recognition results.

Performance Measurement Metrics:

Accuracy, Precision, Recall, and F1-Score are among the metrics used to evaluate performance. Accuracy is the ratio of accurately predicted instances to the complete dataset, providing a comprehensive assessment of model performance. Precision assesses the accuracy of positive predictions by measuring the ratio of accurately predicted positives to all expected positives. Recall assesses the model's ability to identify all relevant cases by comparing the proportion of accurately predicted positive observations to all observations in the actual class. The F1-Score, the harmonic mean of precision and recall, provides a balanced assessment, which is especially useful in cases with an uneven class distribution. These metrics provide a comprehensive assessment of the model's performance, taking into account correctness, relevance, and the balance of true and false positives. Examining these measures offers insights into the model's strengths and weaknesses across various classification facets. The mathematical formulas for each metric are displayed below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

True positive (TP) shows accurate face mask recognition, whereas false positive (FP) indicates incorrect face mask prediction. True Negative (TN) implies that the No mask prediction is right, whereas False Negative (FN) is to predict open faces as masked faces.

Results and Discussion:

In this study, the YOLOv4 model underwent customization and fine-tuning to detect face masks in public spaces, aiming to enforce COVID-19 SOPs. Leveraging the robust computational capabilities of Google Colab, the model training phase capitalized on the necessity of GPU-powered systems for deep learning models. With a dataset comprising 5800 images encompassing both classes, the model underwent training up to 3000 iterations, achieving a loss of 0.51. The training process, delineated in Figure 3, facilitated monitoring of the model's progression. The dataset included diverse images, ranging from mixed-class compositions to veiled individuals, reflecting the multicultural context where face masks vary. Offline augmentation techniques were employed to enhance dataset diversity, followed by rigorous image screening to eliminate inappropriate samples. Transfer learning with pre-trained weights expedited the model's convergence and performance optimization. For rigorous evaluation, a distinct set of 520 unseen images was utilized for testing, as illustrated in Table 1. This approach ensured the model's generalization ability and readiness for real-world deployment in diverse public settings, aligning with the imperative of adhering to COVID-19 safety measures. Figure 2 is showing an image with both classes.

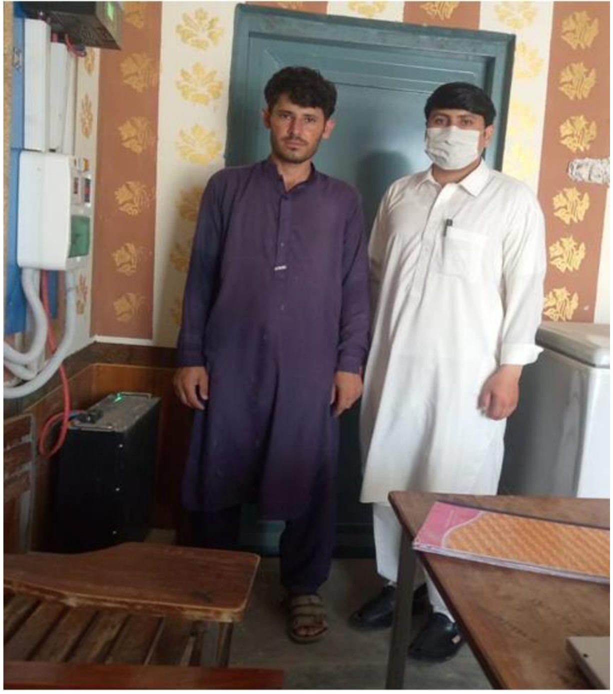

Figure 2: An image with both classes

In an alternate approach, the model was trained using the default setting of YOLOv4, except for changing the necessary setting as per the dataset. The model was trained using online augmentation and tested on the same set of unseen images and stopped after 3300 iterations. The performance results are shown in Table 2. The model was also trained from scratch without leveraging pre-trained weights. This method necessitated an extended training duration of 3600 iterations, resulting in a higher loss value of 0.85. Subsequently, this model underwent testing on the same 520-image test set, yielding lower accuracy compared to its counterpart employing transfer learning. The performance results are outlined in Table 2. This comparative analysis underscores the efficacy of transfer learning in expediting model convergence and enhancing performance, emphasizing its pivotal role in deep learning tasks in cases of limited training dataset size.

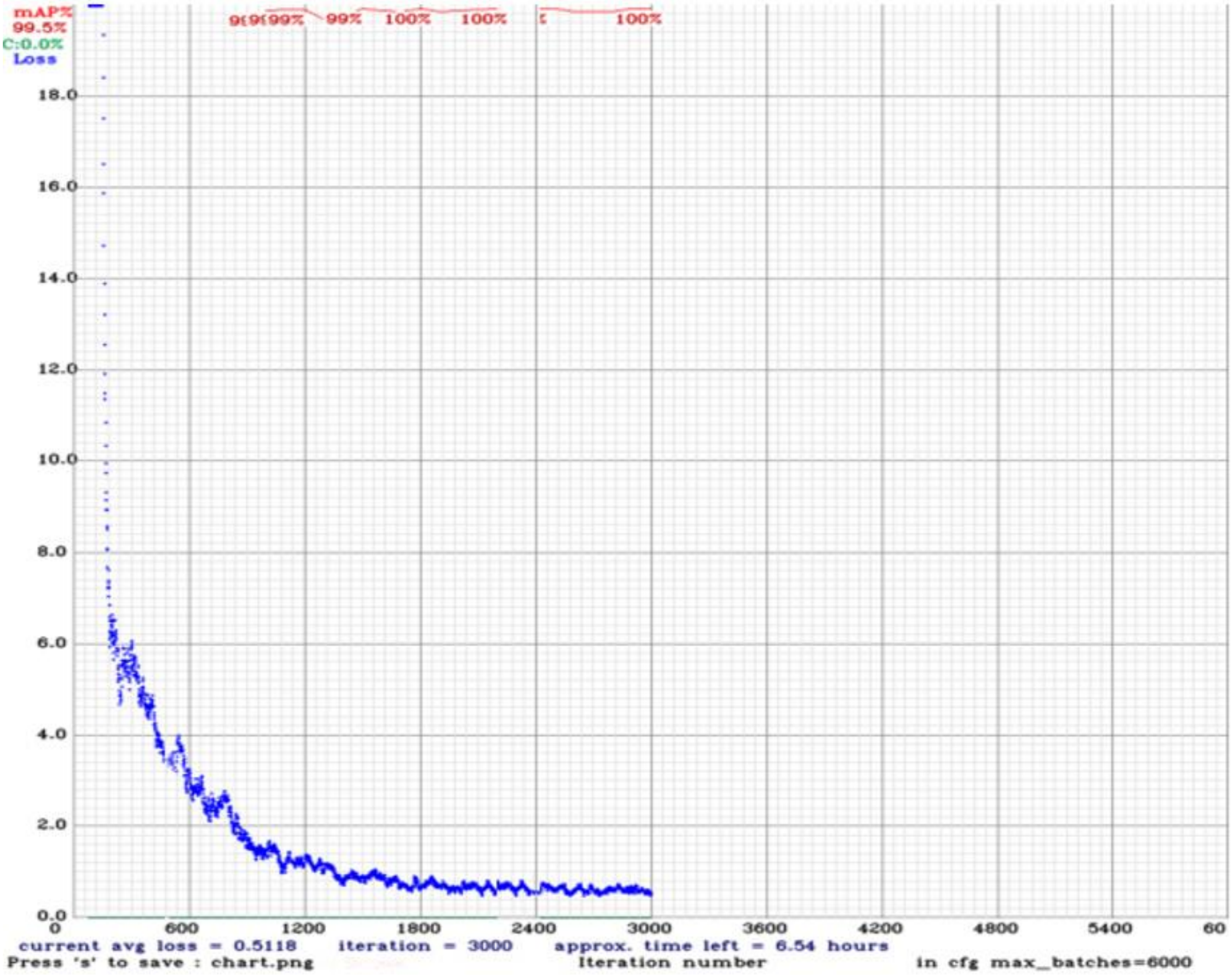

Figure 3: Training loss curve of the proposed model

Table 1: Performance of the proposed model based on TL

S No.	Metrics Name	%Value
1	Accuracy	97.07
2	Precision	96.60
3	Recall	98.10
4	F1-Score	97.34

Table 2: Performance of the YOLOv4 based on various training strategies

Model	Training	Iterations	Augmentation	Accuracy (%)
YOLOv4	Transfer Learning	3300	Online	94.42
	Scratch	3600	Offline	95.19
	Transfer Learning	3300	Offline	97.07

The proposed model's goal was to improve model accuracy when dealing with complicated backgrounds and congested scene pictures. It also had veil images in the collection as face masks. The model was trained using a dataset that included veil photos and complex backdrop images with mixed classes. To train the model on diverse data, data augmentation techniques were applied. Transfer learning was utilized to train the model faster and with better outcomes. The performance results of the suggested model are provided in Table 1. Figure 4 shows some sample images detected by the model accurately.

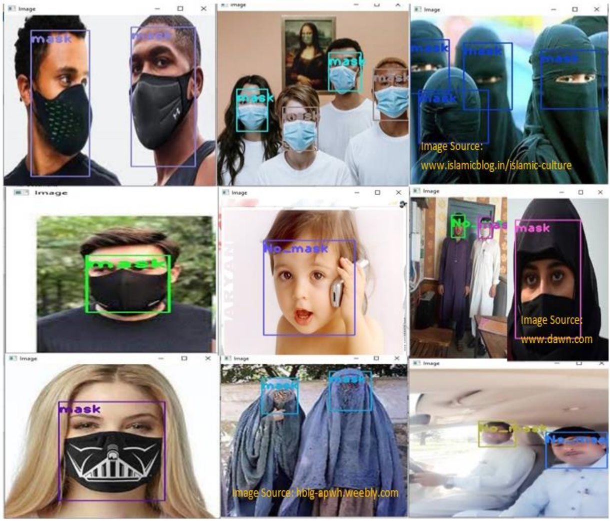

Image Source: www.islamicblog.in/islamic-culture

Image Source: www.dawn.com

Image Source: hblg-apwh.weebly.com

Figure 4: Detection results of the proposed model.

Performance Measurement of the Model on Public Datasets:

The model was carefully trained using the hybrid dataset, which was precisely designed to cover a wide range of scenarios and environments. Subsequently, its robustness was thoroughly tested on three different open-source datasets, as shown in Table 3. These tests were critical in validating the model's ability to generalize and perform accurately over a wide range of datasets and situations. The findings demonstrated the model's effectiveness, with accuracy ranging from 96.7% to 97.37%. This validation not only attests to the model's ability to adapt to varying data distributions but also highlights its reliability in real-world applications across different settings. Such comprehensive assessments are instrumental in bolstering confidence in the model's performance and reinforcing its suitability for deployment in critical tasks such as face mask detection in public places amidst the ongoing challenges posed by the COVID-19 pandemic.

Table 3: Performance of the proposed model on publically available datasets

S No.	Dataset	No. of Images	Sample Selected	%Result
1	MAFA [1]	4357	134	96.70
2	Simulated Masked Face Dataset (SMFD) [7]	1570	170	97.03
3	Kaggle [22]	853	214	97.37

Discussions:

The aim of the proposed study was to develop an automatic model to detect and ensure the wearing of face masks in public places. As there is a lot of work already done in this domain, they face difficulties working in crowded areas. YOLOv4 is the most famous pre-trained model based on one-shot detection with speed as well as accuracy. It can also be fine-tuned and retrained for new tasks. It needs a sound amount of knowledge to be trained and fine-tuned for a new task, and it also needs to change various parameters for accuracy and generalizations. The pre-trained model was trained on 80,000 images, and using it in a new domain requires a large set of training images. In the proposed study, there were not so many training images, so we used transfer learning to reduce the chance of over-fitting. The role of training datasets plays a positive role in training deep learning models. The proposed dataset was collected from various sources, including the half and veil images Data augmentation is another crucial technique to enhance model accuracy and generalization. There are two ways to augment the training dataset [23][24]. In the proposed study, both methods were used and the results were compared. Offline augmentation is a bit more laborious than online augmentation but advantageous because the diversity it brings will be under the control of the developer, and all inadequate images will be removed prior to the model training. A model trained on a good set of images may produce an accurate and efficient model as an output [25]. In the proposed study, both augmentation techniques were used, as well as training from scratch and transfer learning, but the model equipped with transfer learning and offline augmentation showed comparatively good results.

Applications:

The primary aim of the study was to develop a deep learning model that can be deployed in CCTV cameras to detect and ensure the use of face masks in public places to reduce the chance of the COVID-19 virus spreading. It can also be used by health authorities in operation theatres and medical stores to ensure the use of face masks. It can also be used by educational institutions in chemical labs to force students to properly use face masks.

Limitations of the Study: The result section shows the performance of the proposed model on our own hybrid model as well as on three public datasets, but it also has some limitations that need attention in the next research iterations. To overcome the shortage of training datasets, we used data augmentation, but it also causes over-fitting if used on a large scale. Therefore, it is needed to collect images from real scenarios to increase the size of the training and testing datasets, which may ease the generalization of the proposed model.

Conclusion and Future Work:

In the proposed system, we trained a CNN-based model, YOLOv4, with modifications and fine-tuning to detect and recognize people without masks in public places to ensure the SOPs regarding COVID-19. The existing systems perform poorly when the image or frame is crowded or the distance between the camera and the target is greater. In the proposed work, this issue was focused on. Muslim women use half- and full-face veils. They were also included in the dataset to be considered as face masks. Offline data augmentation techniques were used to train the model on a large set of images to reduce over-fitting and increase detection rates. Crowded images were used in training, having both classes in a single image or frame. This approach to face mask detection was tested on new unseen images with a 97.07% classification and detection accuracy. In the future, we will use some other state-of-the-art models to compare their results with the proposed model. We will also add some other functionality to the proposed model, like social distancing and face recognition in masks, to automatically recognize a person who violates the SOPs.

References:

[1] C. Li, J. Cao, and X. Zhang, "Robust deep learning method to detect face masks," *ACM Int. Conf. Proceeding Ser.*, pp. 74–77, Oct. 2020, doi: 10.1145/3421766.3421768.

[2] "WHAT YOU NEED TO KNOW ABOUT CORONAVIRUS (COVID-19) - COVID-19 | Ministry of Health." Accessed: Mar. 13, 2024. [Online]. Available: https://www.health.go.ug/covid/document/what-you-need-to-know-about-coronavirus-covid-19/

[3] F. Di Gennaro *et al.*, "Coronavirus Diseases (COVID-19) Current Status and Future Perspectives: A Narrative Review," *Int. J. Environ. Res. Public Heal. 2020, Vol. 17, Page 2690*, vol. 17, no. 8, p. 2690, Apr. 2020, doi: 10.3390/IJERPH17082690.

[4] F. M. J. Mehedi Shamrat, S. Chakraborty, M. M. Billah, M. Al Jubair, M. S. Islam, and R. Ranjan, "Face Mask Detection using Convolutional Neural Network (CNN) to reduce the spread of Covid-19," *Proc. 5th Int. Conf. Trends Electron. Informatics, ICOEI 2021*, pp. 1231–1237, Jun. 2021, doi: 10.1109/ICOEI51242.2021.9452836.

[5] M. R. Bhuiyan, S. A. Khushbu, and M. S. Islam, "A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, Jul. 2020, doi: 10.1109/ICCCNT49239.2020.9225384.

[6] S. Balaji, B. Balamurugan, T. A. Kumar, R. Rajmohan, and P. P. Kumar, "A Brief Survey on AI Based Face Mask Detection System for Public Places." Mar. 28, 2021. Accessed: Mar. 13, 2024. [Online]. Available: https://papers.ssrn.com/abstract=3814341

[7] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6517–6525, Dec. 2016, doi:

10.1109/CVPR.2017.690.

[8] S. Jignesh Chowdary, G., Punn, N. S., Sonbhadra, S. K., & Agarwal, "Face mask detection using transfer learning of inceptionv3. In Big Data Analytics 8th International Conference, BDA 2020, Sonepat, India, December 15–18, 2020, Proceedings 8 (pp. 81-90).," *Springer Int. Publ.*, pp. 81–90, 2020.

[9] B. M. et Al., "Face mask detection using convolutional neural network," *J. Nat. Remedies*, vol. 21, no. 12, 2021.

[10] 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.*, Sep. 2020, doi: 10.1109/IEMTRONICS51293.2020.9216386.

[11] G. Yang *et al.*, "Face Mask Recognition System with YOLOV5 Based on Image Recognition," *2020 IEEE 6th Int. Conf. Comput. Commun. ICCC 2020*, pp. 1398–1404, Dec. 2020, doi: 10.1109/ICCC51575.2020.9345042.

[12] B. Sathyabama, A. Devpura, M. Maroti, and R. S. Rajput, "Monitoring pandemic precautionary protocols using real-time surveillance and artificial intelligence," *Proc. 3rd Int. Conf. Intell. Sustain. Syst. ICISS 2020*, pp. 1036–1041, Dec. 2020, doi: 10.1109/ICISS49785.2020.9315934.

[13] X. Ren and X. Liu, "Mask wearing detection based on YOLOv3," *J. Phys. Conf. Ser.*, vol. 1678, no. 1, p. 012089, Nov. 2020, doi: 10.1088/1742-6596/1678/1/012089.

[14] X. Jiang, T. Gao, Z. Zhu, and Y. Zhao, "Real-Time Face Mask Detection Method Based on YOLOv3," *Electron. 2021, Vol. 10, Page 837*, vol. 10, no. 7, p. 837, Apr. 2021, doi: 10.3390/ELECTRONICS10070837.

[15] S. Saponara, A. Elhanashi, and A. Gagliardi, "Implementing a real-time, AI-based, people detection and social distancing measuring system for Covid-19," *J. Real-Time Image Process.*, vol. 18, no. 6, pp. 1937–1947, Dec. 2021, doi: 10.1007/S11554-021-01070-6/FIGURES/11.

[16] H. Ahmed, Abd El-Aziz & Azim, Nesrine & Mahmood, Mahmood & Alshammari, "A Deep Learning Model for Face Mask Detection," vol. 2, pp. 101–107, 2021, doi: 10.22937/IJCSNS.2021.21.10.13.

[17] E. Mbunge, S. Simelane, S. G. Fashoto, B. Akinnuwesi, and A. S. Metfula, "Application of deep learning and machine learning models to detect COVID-19 face masks - A review," *Sustain. Oper. Comput.*, vol. 2, no. August, pp. 235–245, 2021, doi: 10.1016/j.susoc.2021.08.001.

[18] R. Laroca *et al.*, "A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, Oct. 2018, doi: 10.1109/IJCNN.2018.8489629.

[19] L. Fang *et al.*, "Ginger Seeding Detection and Shoot Orientation Discrimination Using an Improved YOLOv4-LITE Network," *Agron. 2021, Vol. 11, Page 2328*, vol. 11, no. 11, p. 2328, Nov. 2021, doi: 10.3390/AGRONOMY11112328.

[20] S. Tenzin, P. Dorji, B. Subba, and T. Tobgay, "Smart Check-in Check-out System for Vehicles using Automatic Number Plate Recognition," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, no. July, 2020, doi: 10.1109/ICCCNT49239.2020.9225555.

[21] M. Ayaz *et al.*, "Automatic Early Diagnosis of Dome Galls in Cordia Dichotoma G. Forst. Using Deep Transfer Learning," *IEEE Access*, vol. 11, pp. 59511–59523, 2023, doi: 10.1109/ACCESS.2023.3283568.

[22] "GitHub - prajnasb/observations." Accessed: Mar. 13, 2024. [Online]. Available: https://github.com/prajnasb/observations

[23] "Face Mask Detection." Accessed: Mar. 13, 2024. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/face-mask-detection

[24] "Evaluation of data augmentation of MR images for deep learning." Accessed: Mar. 13, 2024. [Online]. Available: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8952747&fileOId=8952748

[25] S. K. Devalla *et al.*, "A Deep Learning Approach to Denoise Optical Coherence Tomography Images of the Optic Nerve Head," *Sci. Reports 2019 91*, vol. 9, no. 1, pp. 1–13, Oct. 2019, doi: 10.1038/s41598-019-51062-7.