Face Mask Compliance Evaluation Model for COVID-19 Risk Assessment using Machine Learning

Shruthi Ravichandran

Under the direction of

Mr. Richard Hamlin
Quantum Risk Analytics, Inc.

Research Science Institute
July 27, 2020

Abstract

As of July 2020, there have been over 16.2 million COVID-19 cases worldwide. Knowledge of facemask usage is imperative in creating a holistic risk assessment for COVID-19, an airborne virus. In order to accurately evaluate risks associated with face mask usage, we must identify the type of mask being worn and whether it is worn correctly. Since types of masks, faces, and locations vary, it has been difficult to establish models to accomplish this task. Thus, in this work, we create a modular pipeline for this problem, where each module uses a deep neural network to analyze images of people; the modules are able to detect a face, determine whether a mask is present, classify the type of mask, and evaluate whether the mask is worn correctly, when given an image or video as input. A You-Only-Look-Once (YOLO) object detection model successfully is able to detect the visible faces in both images and videos. These faces are fed as input into a binary classifier which predicts whether a mask is being worn with over 99% accuracy. The image is then fed into a mask type classifier which predicts what type of mask is being worn with approximately 84% accuracy. This work demonstrates the potential of combining such a model with a more robust risk assessment model to accurately provide populations with critical and accurate information.

Summary

When evaluating the risk in visiting an establishment while the COVID-19 pandemic continues to rapidly spread, there are many factors that must be taken into account. One critical factor in reflecting the containment of COVID-19 is mask usage - a metric that cannot be easily measured because there are many different types of masks, with different characteristics. One way to measure this is employing artificial intelligence techniques, specifically classification methods. Here, we create and end-to-end model pipeline, in which each step solves a task: determining whether a mask is present, classifying the type of mask, and evaluating the fit of the mask. These results can then be correlated with the relative effectiveness of each type of mask and fed into a more general risk assessment model to enhance public precaution in mitigating viral transmission. This work, ultimately, demonstrates a combination of cutting edge deep learning techniques with the public health application of tracking and enhancing safeguards for COVID-19.

1 Introduction

In the age of a global pandemic, it becomes imperative to create technology-based solutions to the wide-scale healthcare crisis. In November of 2019, a novel coronavirus was detected in the city of Wuhan in the Hubei province of China, officially deemed a pandemic under the name "COVID-19" [1]. As of the evening of July 26, 2020, Johns Hopkins University reported almost 16.2 million confirmed cases worldwide, with just under 650,000 of those as fatal cases. [2]

Airborne transmission has been shown to be the dominant route for COVID-19 to spread [3]. In response to this, the Center for Disease Control (CDC) and World Health Organization (WHO) propose two main forms of precautions in order to slow the spread: social distancing and mask-wearing [4]. Covering a potentially infected mouth and nose with a cloth mask has been shown to reduce to reduce particle transmission by up to 90% [5]. Research has also shown that up to 25% of individuals confirmed to have COVID-19 are asymptomatic, yet are just as likely to infect others as symptomatic patients, as carriers. [6] Though masks have been the primary means of slowing transmission, yet metrics regarding their public usage remain minimal.

As countries and local legislation worldwide gradually lift strict quarantining guidelines, it becomes imperative for institutions and establishments to accurately assess their risk of exposure and spread [7]. To create a comprehensive risk assessment, many factors must be taken into consideration: location, local fatality rate, relative effectiveness of social distancing, percentage of people traveling to neighboring areas, percentage of mask usage in public, and average mask effectiveness. Such factors do not have an existing database and must be evaluated for a specific location.

One specific factor, percentage of mask usage, proves especially difficult, given the lack of standardized data. Additionally, the problem becomes more complex when evaluating dif-

ferent types of masks, each with varying amounts of protection. There are four main types of masks that are used by the general public: surgical masks, cloth masks, scarfs/bandanas, and N-95 respirators. Though all four offer some protection against the spread of the disease, it's important to be able to distinguish between each, as each are widely different with respect to effectiveness. Finally, the problem is especially nuanced, given that individuals may take their mask off for various activities, so one "counted" metric is not holistic enough.

One solution to this problem of mask detection and classification is applying artificial intelligence to detecting mask usage from live-stream feeds, such as security cameras or from geo-tagged images from social media or traffic cams.

Machine Learning Background

Convolutional Neural Networks

Recent advances in computer vision techniques have produced a revolutionary technology that aims to solve image classification's most nuanced challenges: convolutional neural networks. Aptly named, CNNs take inspiration from the connectivity between neutrons of the Human Brain and the organization of the Visual Cortex. In the brain, a visual area is collapsed into restricted regions called "receptive fields." Neurons respond to stimuli only in their respective receptive fields. Similarly, a CNN is able to break down an inputted image into smaller "convolved features," while still preserving spatial and temporal dependencies. The Convolution Operation allows the model to extract high-level characteristics from the input image. Subsequent layers are able to extract lower-level characteristics. This separation allows the model to gain a holistic understanding of an image. Figure 1 depicts a general model architecture of a CNN.

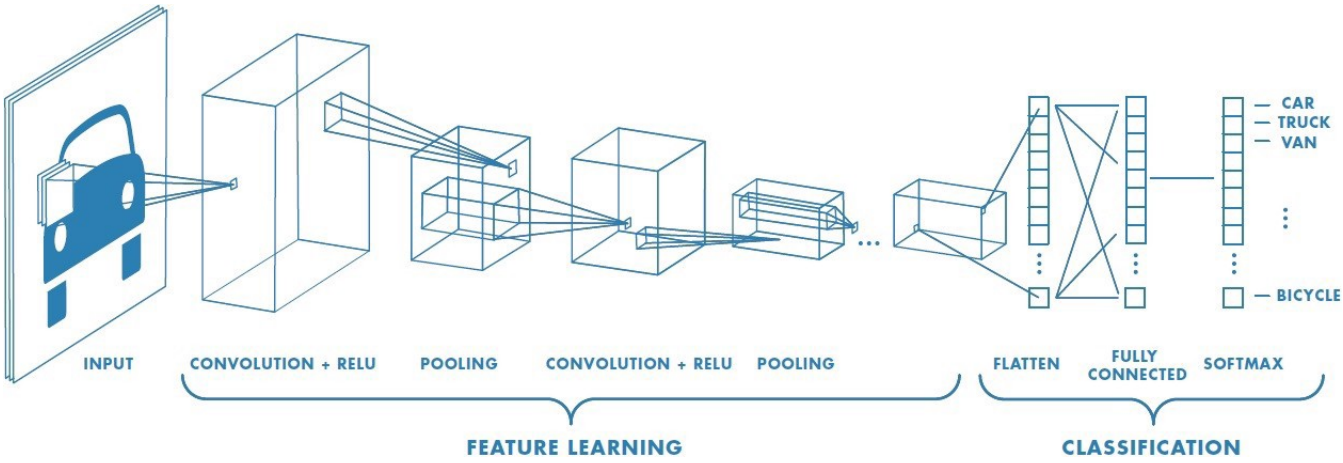

Figure 1: Visual Depiction of Architecture of CNN. The model can be broken into two parts: feature learning and classification. Feature Learning includes Convolution Layers and Pooling Layers, while classification includes Flatten, Fully Connected, and Softmax layers. [8]

The activation functions in a neural network allow them to do non-linear mappings from inputs to outputs. The most widely used activation function for a neural network is the Rectified Linear Unit, or ReLu function. The ReLu function determines which neuron to activate for an output. Following a non-linearity, like the ReLu function, a Pooling layer is added to summarize the features detected in the input. This allows small changes in each example to not drastically affect the overall model. The produced "feature map" from the Pooling layer is converted into a single column by the Flatten layer. This is fed into the Fully Connected layer, allowing the model to utilize the results of the previous layers and classify them into a label to output. Finally, a Dense layer adds the fully connected layer to the neural network.

In a neural network, a loss function is a measure of error that informs the model how far it was from a correct guess and in which direction it should correct itself. Loss can be optimized in order to accurately train the model. This allows for an accurate, self-learning model.

Object Detection

Real-time object detection is rapidly evolving sector of artificial intelligence that has seen many new advances in the recent years. Object detection, as a field, can be broken into two main approaches: two-shot (or region based) and single-shot. Two-shot detection involves two stages: first, the region proposal, then the classification of the region and fine-tuning the location prediction. Alternatively, single shot prediction is able to skip the region proposal and instead identifies the final location and predicts the content simultaneously. Although single-shot is therefore generally faster than two-shot, the accuracy is often lower as a trade-off.

YOLO: You Only Look Once

One of the most widely cited and used single-shot object detection model is You Only Look Once, or YOLO [9]. This ground-breaking object detection software allows for real-time processing. The YOLO architecture includes 24 convolutional layers (layers with specific, learnable filters), followed by two fully connected layers [9]. Figure 2 visually represents the architecture of YOLO.

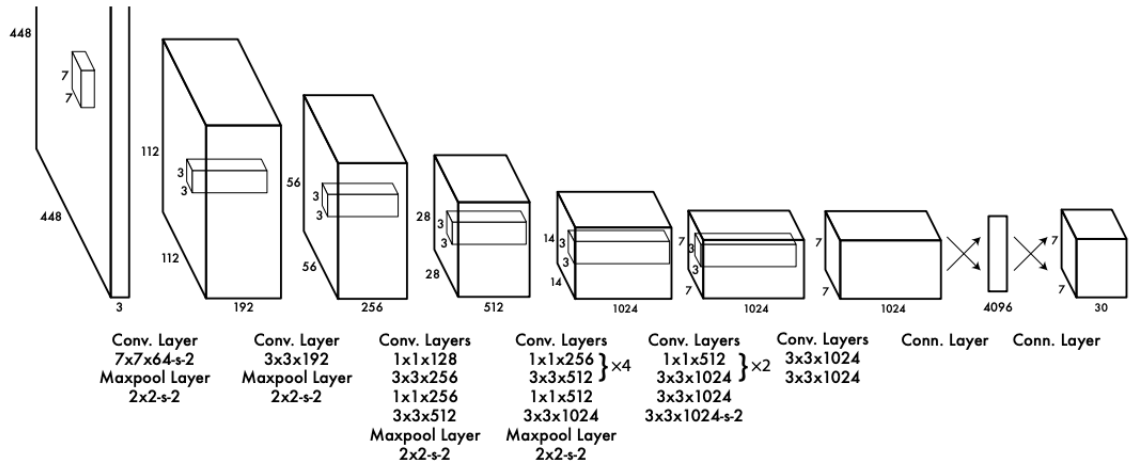

Figure 2: Visual Depiction of Architecture of YOLO: the YOLO layers can be explained as a CNN network built to predict a (7, 7, 30) tensor. The CNN reduces the dimensions to 7 x 7 with 1024 output channels. YOLO then uses two fully connected layers to make a bounding box prediction, using linear regression. [10]

This enables several advantages, including: as a single-shot model, YOLO is fast and therefore suitable for real-time processing; since both the object location and class are predicted from the same model, YOLO can be trained end-to-end to improve accuracy; additionally, because YOLO is able to see the whole image, rather than limit itself to a proposed region (like two-shot models do), YOLO is more generalized and also predicts fewer false positives in the background [11]. Here, we employ a YOLO model for face detection as a prerequisite to mask usage evaluation.

Thus, this study will aim to create a modular pipeline, utilizing CNN and YOLO models to estimate average mask usage in a given location. This estimate will be used to develop a distribution of risk that will be incorporated in a final risk assessment model, which will be distributed to governments and institutions for their use in reopening measures.

2 Methods

In order to effectively approach the task of quantifying mask compliance, we split the task into 4 main components: detecting a face, detecting whether a mask is present, classifying the type of mask, and evaluating mask fit. Each sub-task was independently developed and the final model will combine each into a modular pipeline, shown in Figure 3. In stage one, we use a YOLO object detection model to identify faces. In stages 2-4, we use a convolutional neural network model to classify existence, type, and fit. Then, we count those labels and relate them to the effectiveness of each type of mask, before outputting a probability that can then be used in a risk assessment model.

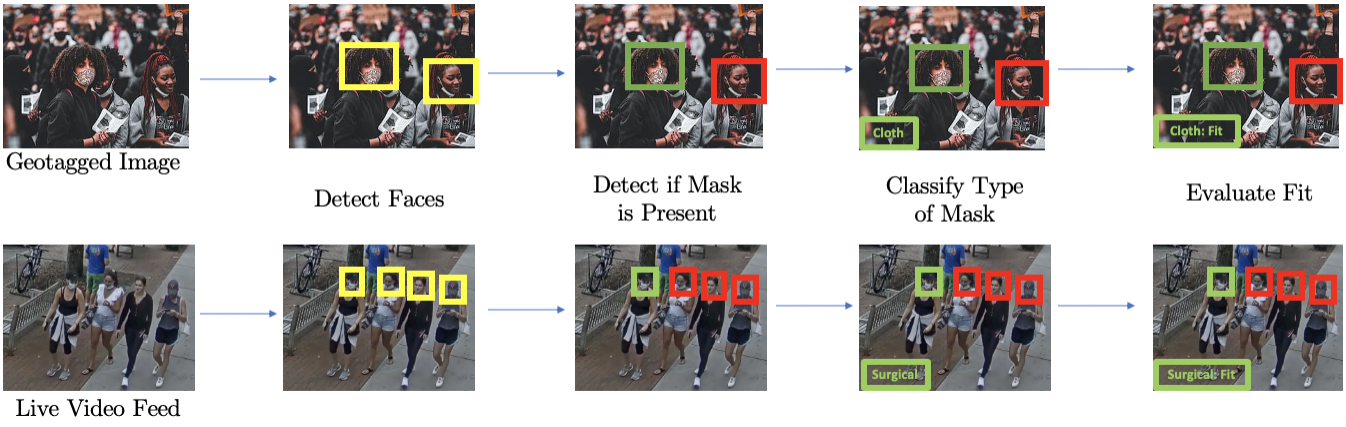

Figure 3: Proposed Workflow Schematic

2.1 Face Detection Model

Prior to any evaluation of mask usage, faces must be detected in order to be inputted to any further algorithm. In particular, any faces in the image should have bounding boxes, mapped around them that can later be cropped to and fed as input into the model. This is especially important when the input is real time video footage, such as a security camera feed, where it is less likely that a face is the only object in frame.

A pre-trained YOLOv3 model was used to detect faces from both video and audio files [12]. The model is able to place bounding boxes delineating the four corners of a face, as well as count the total number of faces in the frame at a time [12]. We modified the model to output tuples of the bounding boxes in json files, rather than the output of a .jpg or .avi file with bounding boxes which the original work produced.

2.2 Mask Detection and Identification

Once a face has been detected, the next step is to determine whether a mask is present. The following model aims to use classification techniques to do this. In order to create a more accurate model, it is imperative to know the type of mask being worn, as this can then be related to the effectiveness of each mask type. The following work aims to solve these two tasks.

2.2.1 Evaluation Metrics

In this work, we utilize two types of loss functions: binary cross entropy and categorical cross entropy. We use the binary cross entropy loss function for the mask detection classification task. The binary cross entropy loss function calculates the average shown below as the loss metric for each example.

$$\text{TotalLoss} = -\frac{1}{\text{n}} \sum_{i=1}^{n} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

where n is number of samples, y_i is the ground truth label, and $\hat{y}_i$ is predicted label.

We use the categorical cross entropy loss function for the mask type classification task. The categorical cross entropy loss function calculates the average shown below as the loss metric for each sample.

$$\text{SampleLoss} = -\sum_{i=1}^{k} y_i \cdot \log \hat{y}_i$$

where k is the number of classes, y_i is the ground truth label, and $\hat{y}_i$ is predicted label..

Given a loss function, in order to improve accuracy, the weights are optimized using an RMSprop optimizer, an adaptive learning rate method. RMSprop is a gradient-based, stochastic technique that uses a moving average of squared gradients for normalization, therefore balancing the step size, or momentum. This enables it to decrease the step size for large gradients and increase the step size for small gradient. This technique of using an adaptive learning rate, rather than treating the learning rate as a hyperparameter allows for a more precise and accurate model.

2.2.2 Data Preprocessing

Mask Presence Classification

Data for this model was acquired in two sets: masked faces and unmasked faces. Approximately 700 unmasked face images were obtained from a GitHub repository of images scraped from the web [13]. The masked face images were acquired from the "Humans in the Loop Medical Mask" dataset [14]. Further data processing was done before use of this data in the mask type classification model.

Mask Type Classification

We obtained data for this model from the "Humans in the Loop Medical Mask" dataset. The raw data contains 6000 images of people from the public domain, intentionally featuring people of all ethnicities, ages, and regions. The dataset covers 20 classes of facial accessories (type of mask, glasses, etc), as well as classification of each face with, without, and with an incorrectly worn mask. Each image was annotated with a bounding box by the refugee workforce of Humans in the Loop in Bulgaria. In order to process the data, we extracted the image to the bounding box of the face and assigned it the label associated with the type of face covering. Our processed data included approximately 23,000 faces, each with the correct label. The four labels used were cloth mask, surgical mask, N95 mask, and scarf/bandana.

These categories were grouped and determined based on effectiveness of each respective facial covering.

Once each image was processed and saved, images were randomly assigned to train, validation, and test batches in a 70:15:15 split.

2.2.3 Training Details

Each model was trained with approximately 4000 images for 1000 epochs, where each epoch consisted of 100 training iterations each, with the early stopping callback included in order to minimize the validation loss function (patience = 15 epochs). The batch size used for training was 20, and padding was employed in order to minimize loss of edge details of the input images. At the conclusion of each epoch, 20 validation images were randomly selected and the validation accuracy and loss for that set were recorded. The weights from the model with the highest validation accuracy were saved.

The model implementation was done in Python using Keras and Tensorflow [15] [16]. The model was trained on the GPU hardware-accelerator provided by Google Colab [17].

2.2.4 Model Architecture

The baseline model's architecture is shown in Figure 4. This architecture was chosen for its simplicity which limits the risk for overfitting of the data.

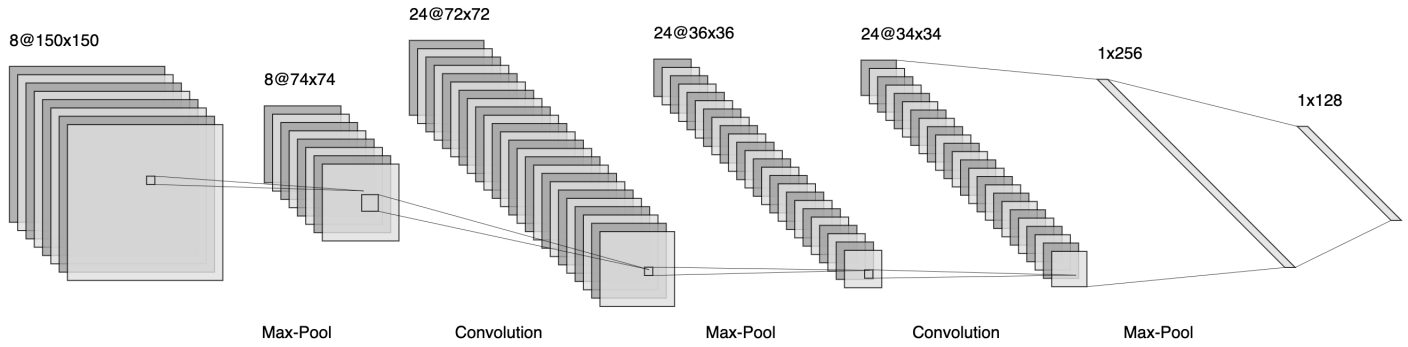

Figure 4: Proposed Baseline Model Architecture

2.2.5 Transfer Learning

Because of the novelty of the task, data sets of relevant images are limited in scope, and therefore, it becomes difficult to extensively train a model to continue learning.

Though expanding the dataset would certainly aid this issue, transfer learning is another method that allows learned knowledge from a larger dataset and image database to be transferred to a specific task - in the case, mask presence and type classification. Transfer learning utilizes models trained on large datasets, e.g. ImageNet, and uses the weights from those models as the initialization for the model trained on a smaller dataset. The features learned on the large dataset are "transferred" to the features in the smaller dataset [18]. In our work, we use the architecture and ImageNet weights of the residual neural network (ResNet50) in Keras [18]. Figure 5 shows a visual depiction of the architecture of the ResNet50 Model.

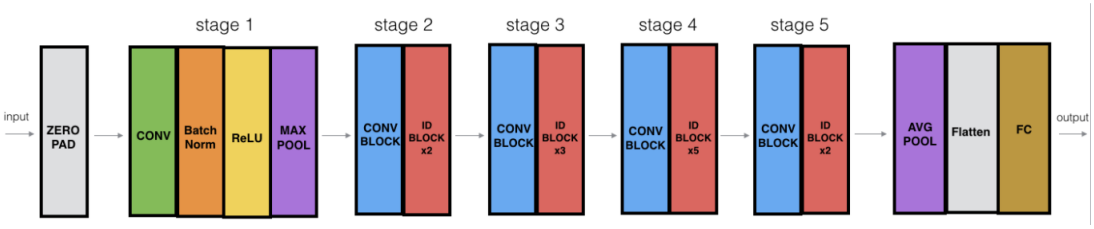

Figure 5: Architecture Diagram of ResNet50 Model. The model has skip-connections which allow it to train that give the model more complexity. [18]

Additionally, we also use the architecture of the pretrained convolutional neural network, VGG19. [19]. Figure 6 shows a visual depiction of the VGG19 architecture and layers.

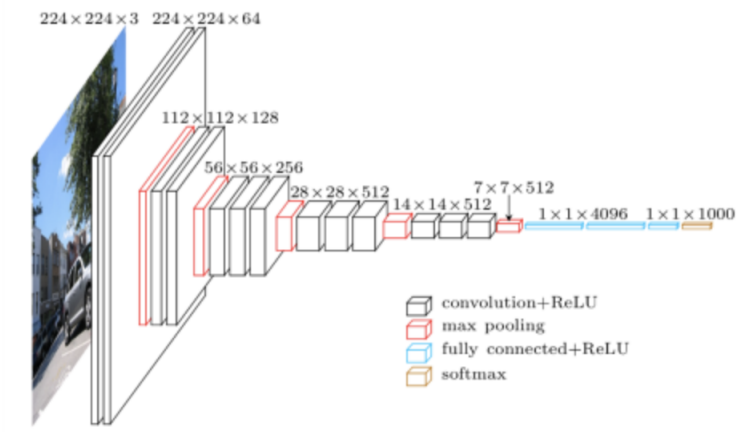

Figure 6: Architecture Diagram of VGG Model. The model is characterized by its simplicity: it uses only stacked 3x3 convolutional layers which increase in depth; the Max Pooling layer reduces the volume size and two fully-connected layers, each with 4,096 nodes are fed into a softmax activation function. [20]

3 Results

3.1 Object Detection Results

The YOLOv3 object detection model was successfully able to place bounding boxes around faces in both stereo images and videos. Figure 7(A) shows a still image fed into the

YOLO model, with four faces - none of which face the camera. Image (B) shows the output of the model, with bounding boxes places around all four faces, as well as a counter in the top left corner.

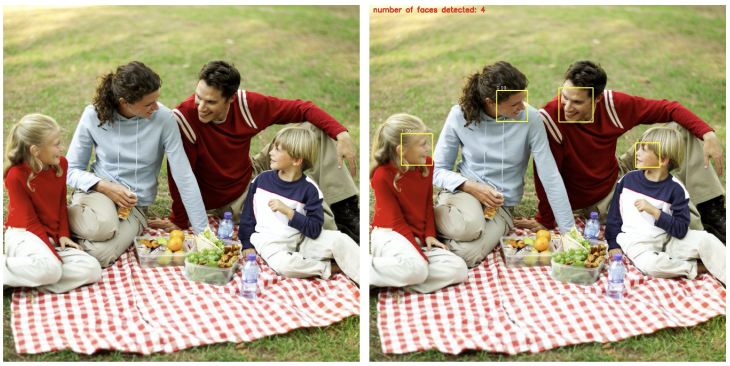

Figure 7: YOLO model is able to detect faces, even when they are turned away from the camera. (A) An image inputted into the model. (B) The bounding boxes placed on the inputted image by the model. [12]

When a video file of a livestream of a street was fed into the model, the model once again captured all visible faces in the frame. Figure 8 shows a frame of the output of the model, including yellow bounding boxes placed on the faces in the frame.

Figure 8: YOLO model is able to identify faces realtime from a video. YOLO tracks faces with bounding boxes as they move, in order not to repeat people in an image. [12]

3.2 Mask Detection and Identification Results

3.2.1 Mask Presence Classification is Optimized with Weighted ResNet50

For each model trained, the weights with the highest validation accuracy was saved as the best model. The testing accuracy for each of these models are shown in the table below. The ResNet50 Model with ImageNet weights performed the best and had the highest testing accuracy. This is likely because of its detailed architecture, as well as the input from the ImageNet weights.

Model	Testing Accuracy
Guessing	0.50
Baseline (CNN)	0.9699
Baseline (CNN) with Dropout	0.9616
VGG19 with weights	0.9199
ResNet50 with ImageNet weights	**0.9987**

Table 1: Classification Accuracy of Mask Existence Models. If the model guessed between the two classes, it would have a 0.5 probability of correctly guessing the label. The Baseline model is the basic CNN architecture shown in figure 4. The Baseline model with Dropout drops some neurons each epoch in order to not overfit the data. The VGG19 model is a pretrained model whose architecture was applied to this dataset. The ResNet50 model with ImageNet weights is also a pretrained model whose architecture and weights were applied to the dataset.

3.2.2 ResNet50 Pretrained on ImageNet Has Highest Test Accuracy

Similar to the previous classification model, the model weights which had the highest validation accuracy were saved and their testing accuracy is plotted in the table below. Similar to the binary classification results shown in Table 1, the ResNet50 Model with ImageNet weights performed the best and had the highest testing accuracy. Again, this is likely due to its detailed architecture, as well as the input from the ImageNet weights.

Model	Testing Accuracy
Guessing	0.25
Baseline (CNN)	0.7850
Baseline (CNN) with Dropout	0.7867
VGG19 with weights	0.6813
ResNet50 without weights	0.7829
ResNet50 with ImageNet weights	**0.8043**

Table 2: Results of Mask Type Classification Models. If the model guessed between the four classes, it would have a 0.25 probability of correctly guessing the label. The Baseline model is the basic CNN architecture shown in figure 4. The Baseline model with Dropout drops some neurons each epoch in order to not overfit the data. The VGG19 model is a pretrained model whose architecture was applied to this dataset. The ResNet50 model without weights was trained with random initialization. The ResNet50 model with ImageNet weights is also a pretrained model whose architecture and weights were applied to the dataset.

4 Discussion

In this work, we sought to develop a mask compliance evaluation model for COVID-19 risk assessment. Our method used object detection and classification techniques, which performed well and show promise for use in conjunction with each other. In this section, we discuss our results and consider limitations of our methods.

The results demonstrate the applicability of these models to evaluate mask usage when assessing COVID-19 risk. Unlike other models which are far more limited in scope and generalizability, the models demonstrated in this study can be widely generalized to both images and videos, to faces of different features (age, race, ethnicity), and to varying levels of image clarity. The combination of the object detection model, YOLO, and the image classification model, the CNN, allows an end-to-end system that can be applied to a wide range of image sources. The binary classification's high accuracy shows promise in its immediately

applicability to security cameras and geotagged images in order to evaluate mask usage of a population. The mask classification model shows promise with a larger dataset. The dataset acquired from the Humans in the Loop project placed emphasis on ensuring there was diversity in the dataset. This could be improved upon more intentionally if a larger dataset were present.

The ideal situation for input into the model would be high resolution security cameras or street cameras that are able to capture candid videos of the general population. Though this is optimal, geo-tagged images are far more common, especially given the prevalence of social media platforms, such Facebook, Instagram, and Flickr. For images, though, context of the photo must be taken into account. A photo that was staged is likely more unreliable to draw any conclusions from, as the individual in the photo could have taken their mask off for the photo, or worn it just for the picture. These challenges show the need for an image context classifier that can provide useful insight into whether data from an image can be trusted and then generalised.

As object detection and image classification techniques evolve, the accuracy and the ability to predict from sparser data will improve. This work has demonstrated the potential for applying machine learning techniques to population estimation problems.

5 Future Work

In this work, we do create a modeular pipeline of models which evaluate mask compliance in order to assess risk for COVID-19. While the model has already yielded useful data, there are numerous areas for improvement. The next steps fall into three categories: increasing accuracy, incorporating a wide variety of data sources, and streamlining the model. We address those in this section.

Improving accuracy is of the highest importance and therefore the immediate next step.

To do this, there are multiple methods. Recent work in the fields of classification and object detection have provided significant increases in the ability of these models to generalize. One important factor essential in generalizability lies in the quality of the dataset. Next steps would involve finding a more comprehensive dataset to train the models on, especially with varying background images.

The second category of work involves a wide variety of data sources. A major source of images for assessing compliance in this setting would be from social media; thus, it is also important to be able to understand the context of an image and gauge whether it has been staged or not. As discussed above, if staged, it is likely to skew the results and therefore, that information should be recorded separately.

The final category of work involves steam-lining the models into an end-to-end pipeline. While the models are built and have been trained, the final step still remains: counting the probability of mask usage and relating this to the effectiveness of each type of mask. This will require extracting the number of bounding boxes (faces) and then, once the other models have run, extracting information about the type and fit of each mask. Once this is done, any image or video can be fed into the model with the final output being a probability.

6 Conclusion

The aim of this work was to create a holistic risk assessment model that could be distributed to the public, as well as institutions and the government. This work has created a modular pipeline in order to evaluate mask compliance in a location. To do this, we employed machine learning techniques, such as classification and object detection, in order to create a model that can be utilized in many situations. Though these technologies existed before, the combination and application to this problem in a modular approach is novel. The output of the comprehensive model can be inputted into a risk assessment model for a

particular location, along with input for effectiveness of each type of mask, as well as factors not covered in this paper, such as effectiveness of social distancing. Using such techniques, we identify risk-reducing behaviors and use them to mitigate the spread of COVID-19.

7 Acknowledgments

This work would be impossible without the mentorship of Mr. Richard Hamlin and the team at Quantum Risk Analytics, Inc. I would also like to thank Ms. Ana Lyons for her tutelage, support, and guidance throughout the research process, as well as her invaluable feedback on presentations and this paper. My deepest gratitude to the RSI alumni who have so graciously helped me through the drafting process of this paper, especially Abdulrahman Alhamdan, Rhea Malhotra, Pratham Soni, Basheer AlDajani, Aditya Saligrama, and Alan Zhu. A special thanks to my co-mentee, Emily Kang, for her constant support and friendship and also Aditya Bora for his invaluable feedback on a prior version of this work. Furthermore, I would like to thank the Research Science Institute, the Center for Excellence in Education, the Massachusetts Institute of Technology, and their sponsors for their generous support of young minds in science and all of their work towards such an incredible opportunity and research experience. I would also like to thank my family, especially my sister, and my teachers for supporting me in every way possible. Without them, none of this would be possible. Finally, I would like to thank my fellow RSI '20 students for their constant moral support, encouragement, and inspiration. I have the sincerest appreciation for such a life-changing experience.

References

[1] C. P. E. R. E. Novel et al. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china. *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi*, 41(2):145, 2020.

[2] J. H. C. R. Center. Covid-19 map.

[3] R. Zhang, Y. Li, A. L. Zhang, Y. Wang, and M. J. Molina. Identifying airborne transmission as the dominant route for the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(26):14857–14863, 2020.

[4] How to protect yourself & others.

[5] C. M. Clase, E. L. Fu, M. Joseph, R. C. Beale, M. B. Dolovich, M. Jardine, J. F. Mann, R. Pecoits-Filho, W. C. Winkelmayer, J. J. Carrero, and et al. Cloth masks may prevent transmission of covid-19: An evidence-based, risk-based approach. *Annals of Internal Medicine*, 2020.

[6] S. Whitehead and C. Feibel. Cdc director on models for the months to come: 'this virus is going to be with us', Mar 2020.

[7] Opening up america again.

[8] C. C. Chatterjee. Basics of the classic cnn, Jul 2019.

[9] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[10] O. O. D. Science. Overview of the yolo object detection algorithm, Sep 2018.

[11] J. Hui. Real-time object detection with yolo, yolov2 and now yolov3, Aug 2019.

[12] T. Nyugen. YOLOface: Deep learning based Face detection using the YOLOv3 algorithm, Aug. 2019.

[13] P. Bhandary. Mask Classifier, Apr. 2020.

[14] H. in the Loop. Medical mask dataset.

[15] K. Team.

[16] G. Brain.

[17] G. Colab.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[20] D. Frossard. Vgg in tensorflow, Jun 2016.