Springboard Capstone Report

September 20, 2021

# Monitoring face mask compliance using computer vision

**Submitted by: Frances Nikki Amurao**

linkedin.com/in/nikki-amurao amurao.frances@gmail.com

# I. The problem

As the battle against Covid-19 continues worldwide, wearing face masks has become an essential component in the umbrella of preventive measures to combat the virus transmission. Since the Covid-19 is transmissible via small respiratory droplets from presymptomatic and asymptomatic individuals, the use of face masks can provide high levels of protection and is most effective when compliance is high.[1] In many areas, mandatory compliance to face mask regulations has extended beyond airports and train stations and into offices and malls where security protocols are put in place to ensure that the face masks are worn upon entry.

Despite these efforts, the effectiveness of face mask protection can be severely diminished by improper usage. The mask must conform to the face without gaps such that the airflow is through the mask, rather than around or through the gaps at the sides, top or bottom. Hence, "a mask that is frequently pulled down to breathe or talk, or is worn under the nose, is not effective." And several studies confirm, this is commonplace.

- "…25% wore masks improperly, on their necks, or covering only their mouths, but not noses.." - Izet Mašić, Bosnia and Herzegovina
- "Also, one can observe many cases of half-compliance or sham compliance. For instance, people do wear masks, but slide them down onto their chins or take them off completely while talking to someone on the street or speaking on the phone. And this is all a performance, keeping their masks somewhere within reach in case of the sudden emergence of police officers, who are indeed issuing fines for not wearing a mask." - Aleksandra Głos, Poland[2]
- 51.5% reported removing the face mask if they needed to talk to someone[3].
- While the prevalence of wearing face masks reached approx. 80% among the Japanese general public, rates of compliance with appropriate measures for correct face mask usage recommended by the WHO ranged from 38.3% to 83.5% and only 23.1% of participants were following all recommended measures.[4]

[1]https://www.preprints.org/manuscript/202004.0203/v1

[2]Martinelli L, Kopilaš V, Vidmar M, et al. Face Masks During the COVID-19 Pandemic: A Simple Protection Tool With Many Meanings. *Front Public Health*. 2021;8:606635. Published 2021 Jan 13. doi:10.3389/fpubh.2020.606635

[3]Sikakulya FK, Ssebuufu R, Mambo SB, Pius T, Kabanyoro A, Kamahoro E, et al. (2021) Use of face masks to limit the spread of the COVID-19 among western Ugandans: Knowledge, attitude and practices. PLoS ONE 16(3): e0248706. https://doi.org/10.1371/journal.pone.0248706

[4]Machida M, Nakamura I, Saito R, et al. Incorrect Use of Face Masks during the Current COVID-19 Pandemic among the General Public in Japan. *Int J Environ Res Public Health*. 2020;17(18):6484. Published 2020 Sep 6. doi:10.3390/ijerph17186484

In fact, the health risks from incorrectly wearing a face mask have been presented as an important argument against the use of face masks as a public health measure[5], especially where it provides a sense of false confidence to the public.

This challenge presents an interesting application of computer vision which the project aims to explore. Can we use deep learning to differentiate between a person not wearing a face mask and wearing one? And if so, is he/she wearing it properly?

Developing an image classifier to discriminate between these cases will be valuable to authorities and even private establishments who would like to enforce the use of face masks more strictly. Serving primarily as a proof-of-concept, this project seeks to demonstrate the potential of using computer vision to actively monitor proper compliance of wearing face masks.

# II. The approach

## Objective and scope of the project

While other projects on face mask detection have focused on identifying the location of the face mask or developing a model to predict whether the subject is wearing a mask or not, the focus here is to include a third category for the 'incorrectly worn masks', where the nose, mouth or chin are not properly covered.

The aim is to develop an image classifier that can distinguish not only between a person wearing a face mask or not, but also indicate whether the face mask is correctly being worn. Achieving a good level of accuracy for all three cases will be crucial for implementation. Hence, the main objective is to reach at least 90% accuracy across all three classes ('without mask', 'with mask', 'incorrectly worn'). Special attention will be given to the type of mistakes the model makes. Specifically, there is some challenge expected in discriminating between the proper and the improper use of masks but less in detecting whether the subject is wearing a face mask or not.

In improving the implementation of preventive measures against Covid, there are undeniably other factors to consider such as the type of face mask worn. However, this will be out of scope and can be considered in a further study following the success of this project. Additionally, the scope is restricted

[5]Makovicky N. The political lives of masks: citizenship, civility and covering up during the COVID-19 pandemic. Allegra. (2020). https://allegralaboratory.net/the-political-lives-of-masks-citizenship-civility-and-covering-up-during-the-covid-19-pandemic/

to the detection of correctly worn face masks on individuals rather than on groups of people. The main deliverables of the project are code, model and accompanying report for documentation.

## The dataset

While there are many available datasets containing images of people wearing masks, there are not enough for the improper use of face masks (i.e. not covering nose, mouth, chin). Hence, a combination of a few sources was used to train the model.

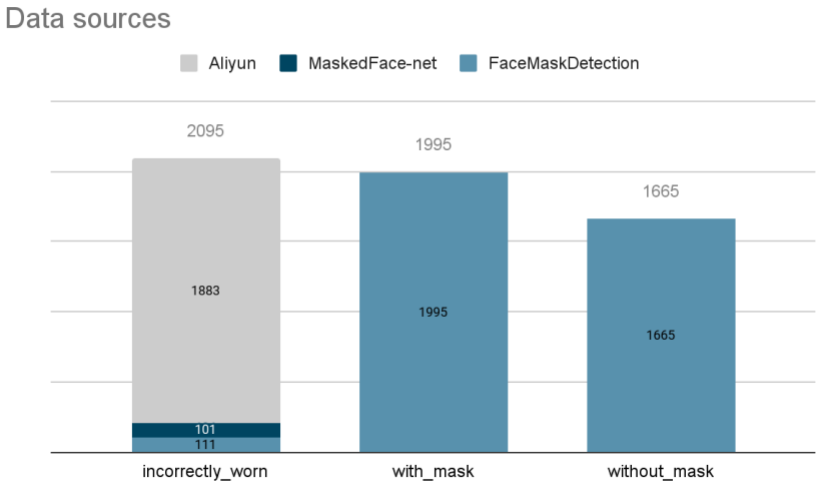



Figure 1. Data sources for each label

### *Chankdrika Deb's Face Mask Detection*[6]

The photos in this dataset are real images of subjects wearing masks and not wearing masks, sourced from Bing Search API, Kaggle datasets and RMFD dataset. The project uses the full dataset, creating the majority of the samples for the two labels (with_masks, without_masks).

Upon manual inspection of the images, several photos were reclassified from 'with_masks' to 'incorrectly_worn' and a few others were excluded as they contained several subjects in one photo.

### *Alibaba Tianchi Dataset*[7]

From this collection of augmented images, only the original (i.e. non-augmented) images of incorrectly worn masks were used in the project.

[6] https://github.com/chandrikadeb7/Face-Mask-Detection

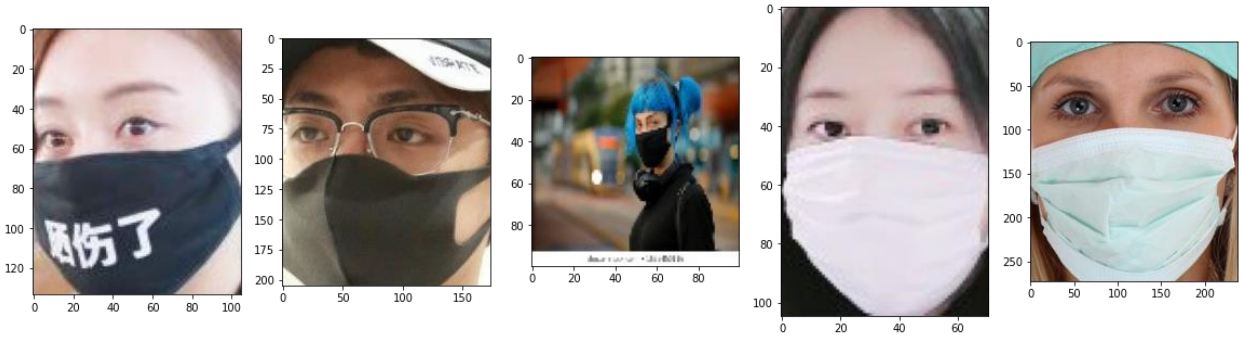
[7] https://tianchi.aliyun.com/dataset/dataDetail?dataId=93724

### *MaskedFace-Net*[8]

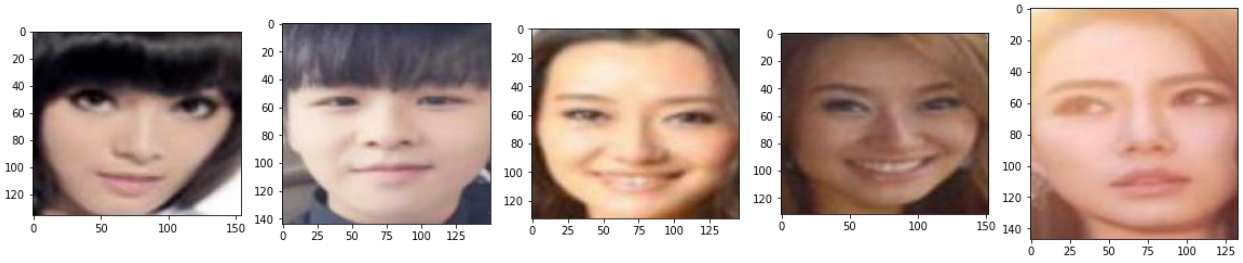
Motivated by the lack of datasets for masked faces, MaskedFace-Net is an attempt to create realistic masked faces. The images of faces have been taken from Flickr-Faces-HQ (FFHQ) and overlaid with 1) properly positioned surgical masks and 2) uncovered chin, nose and/mouth for the incorrectly worn face masks. From this large database of >60k images, only the first two folders of images in the IMFD were used (folders 0000 and 0100) for the project.

Unlike the first two sources, the photos here are synthetic and lack the variety in the type of face masks worn (exclusively surgical face masks). However, without this, the c. 200 images for the 'incorrectly_worn' label will create a highly imbalanced class when compared to >1,500 images for both the 'with_mask' and 'without_masks' labels.

A few examples for each label are shown below.

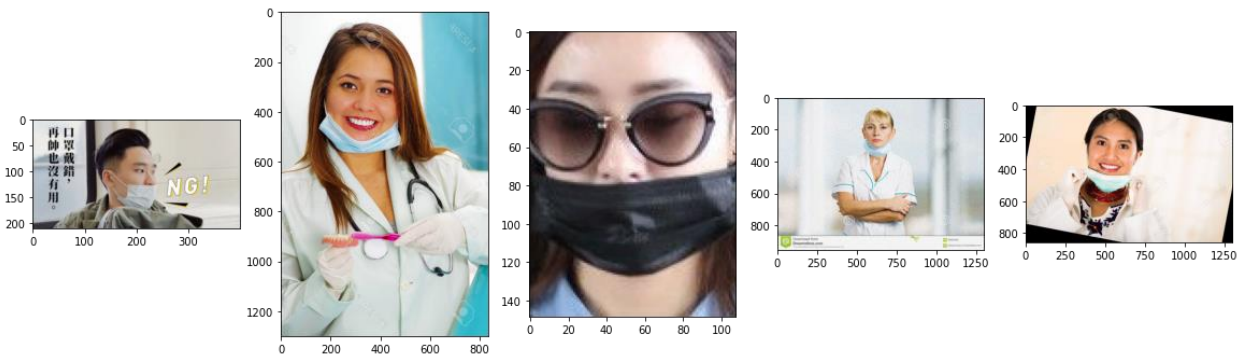


CORRECTLY WORN MASKS (label: 'with_mask')



NO MASK WORN (label: 'without_mask')

[8] https://github.com/cabani/MaskedFace-Net




IMPROPER USE OF MASKS (label: ‘incorrectly_worn’’)

Figure 2. Sample images for each class label

The combined final dataset created a total of 6,041 images (2,015 examples of correctly worn masks, 2,096 examples of incorrectly worn masks, and 1,930 examples of no mask worn) which were then split 60/20/20 between train, development and test sets.

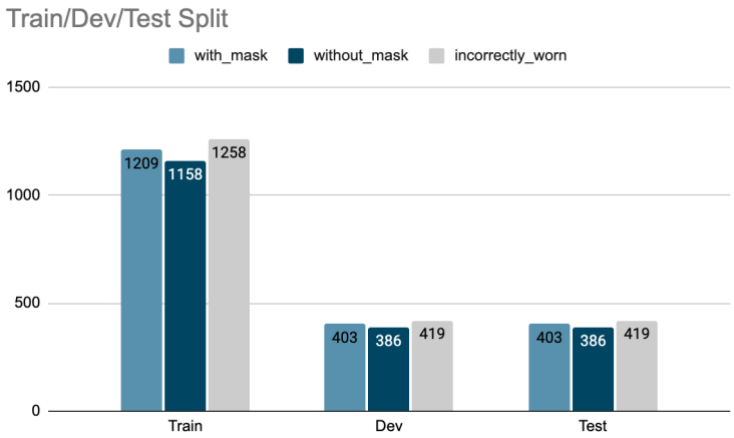



Figure 3. Train/Dev/Test split for each class

## Modelling

The images were uniformly resized to 224 x 224 prior to model training and processed in batches of 128. The models were trained with an Adam optimizer for 50 epochs although early stopping was implemented based on decreasing model loss.

Before training on more advanced neural architectures, I used a 3-block VGG-style architecture to establish a baseline performance level, upon which the succeeding models will be benchmarked against. Specifically, the layers are as follows:

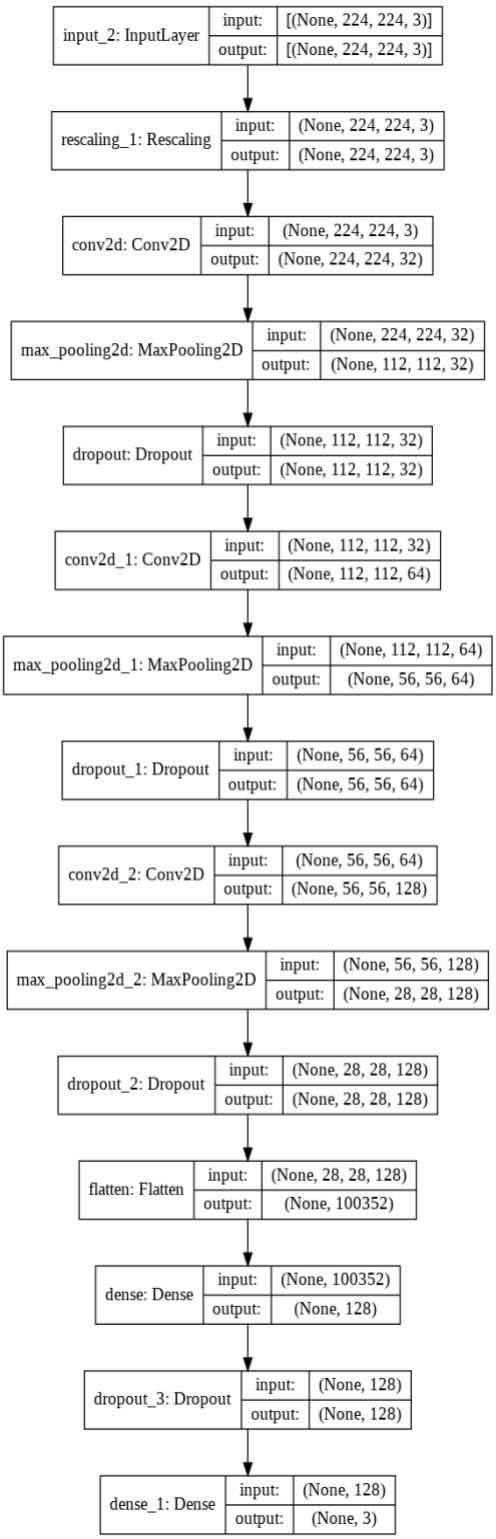



Figure 4. Baseline model architecture (3-layer VGG style)

After training on the baseline, we explored how much performance improvement can be gained by leveraging on some of the top performing pre-trained models for image classification, namely VGG, Inception, ResNet and EfficientNet.

**VGG16** model architecture (also called OxfordNet) was developed by Oxford and gained popularity after winning the ILSVR (ImageNet) competition in 2014 where it achieved a 92.7% top-5 test accuracy on the ImageNet dataset of over 14 million images belonging to 1000 classes. It utilizes multiple 3x3 kernel-sized filters stacked on top of each other.

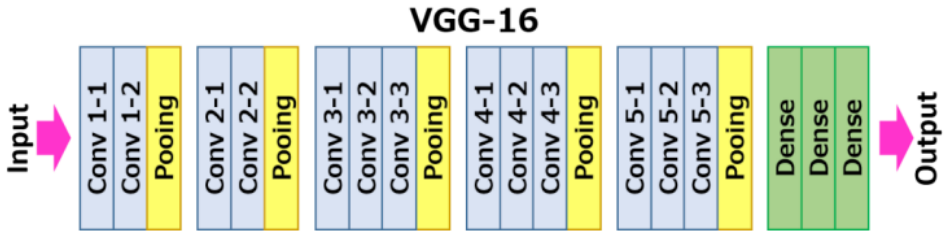



Figure 5. VGG16 architecture[9]

**ResNet50**, a 50-layer deep residual network developed by Microsoft, overcomes the 'degradation problem' and allows the network to grow deeper while achieving greater accuracy by leveraging on the concept of 'skip connections'. Compared to VGG, the ResNet architecture also uses far fewer filters and lower complexity during the training.

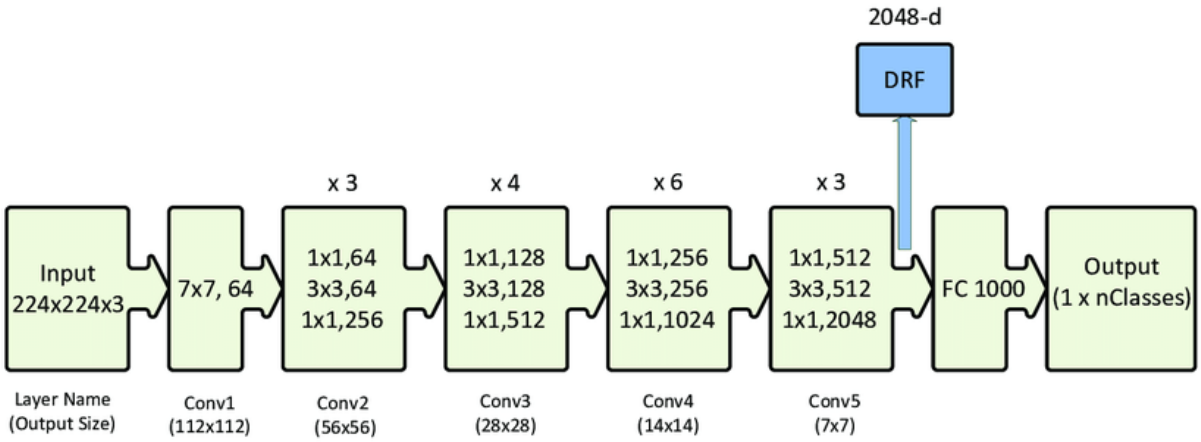



Figure 6. ResNet50 architecture[10]

[9] https://medium.com/nerd-for-tech/vgg-16-easiest-explanation-12453b599526

[10] https://www.researchgate.net/figure/ResNet-50-architecture-26-shown-with-the-residual-units-the-size-of-the-filters-and_fig1_338603223

Another popular model architecture used in the field of image classification and object detection is **InceptionV3**, a 48-layer network developed by Google which has proven to be computationally efficient in terms of the number of parameters generated by the network and the economic cost incurred (memory and other resources). The architecture is characterized by factorizing convolutions, auxiliary classifiers, and efficient grid size reduction.

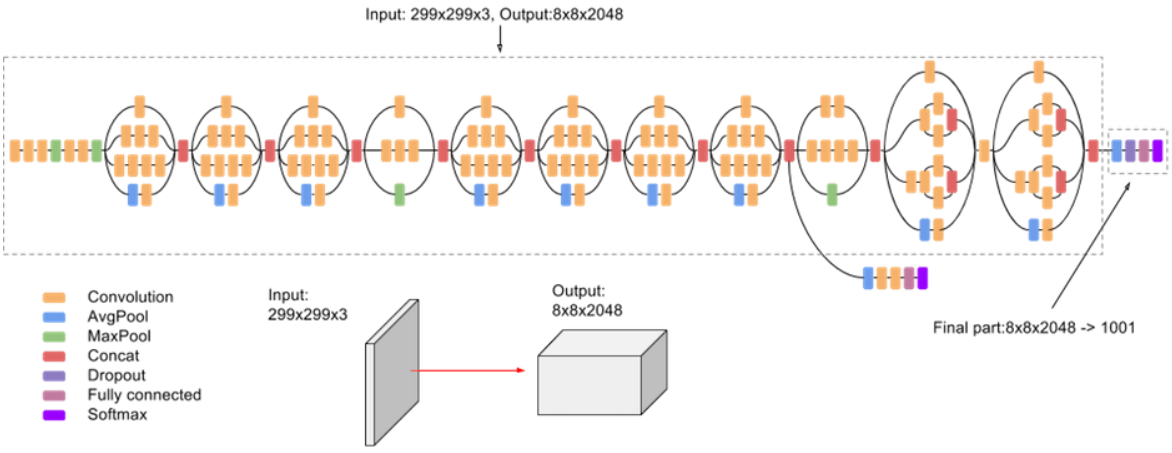



Figure 7. Inception model architecture[11]

**EfficientNet** is a family of convolutional neural network models which uniformly scales network depth, width and resolution (not arbitrarily, as in conventional practice). The rationale behind this is that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image.[12]

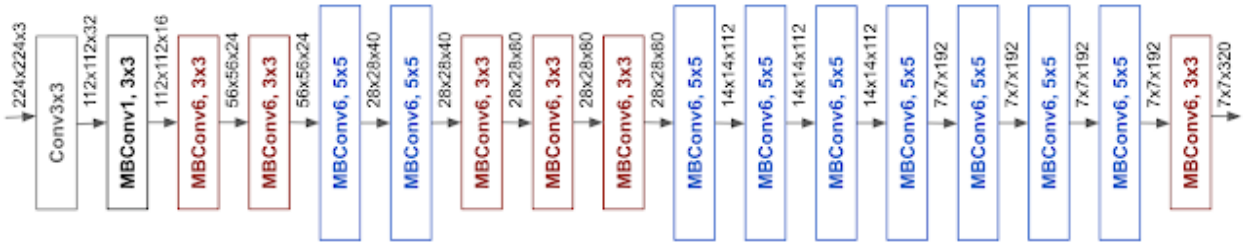



Figure 8. EfficientNet model architecture[13]

## Pre-processing

[11] https://paperswithcode.com/method/inception-v3

[12] https://paperswithcode.com/method/efficientnet

[13] https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

The recommended pre-processing for model architecture was followed prior to model training.

- VGG16, ResNet50: convert the input images from RGB to BGR, then zero-center each color channel with respect to the ImageNet dataset, without scaling
- InceptionV3: scale input pixels between -1 and 1
- EfficientNet: a rescaling layer is already included in the model architecture
- Base / Benchmark model: a rescaling layer is included in the model architecture

## Selecting the model architecture and fine-tuning the model

To select the best model architecture for this problem, I evaluated the model performance under the following scenarios:

1) Using the pre-trained model and weights as a feature extractor;
2) Fine-tuning the pre-trained model by unfreezing and retraining the last two layers; and
3) Learning the weights of the model architecture by unfreezing the base model

While the classes are not exactly equal, the imbalance is not significant so the models were compared based on validation accuracy. Other performance metrics such as precision, recall, area under the curve, precision-recall curve, and (categorical cross-entropy) loss were also monitored throughout the training phase to check for potential overfitting.

After selecting the model architecture, the model can be fine-tuned further by optimizing parameters. In this case, the focus is on the number of layers to unfreeze, as well as the learning rate.

# III. Results

## Baseline model

With early stopping, training ended on the 12th epoch as the model reached 91.57% accuracy on the validation set. Model improvement started to plateau quickly after about 2 epochs as shown by the loss and accuracy charts below.

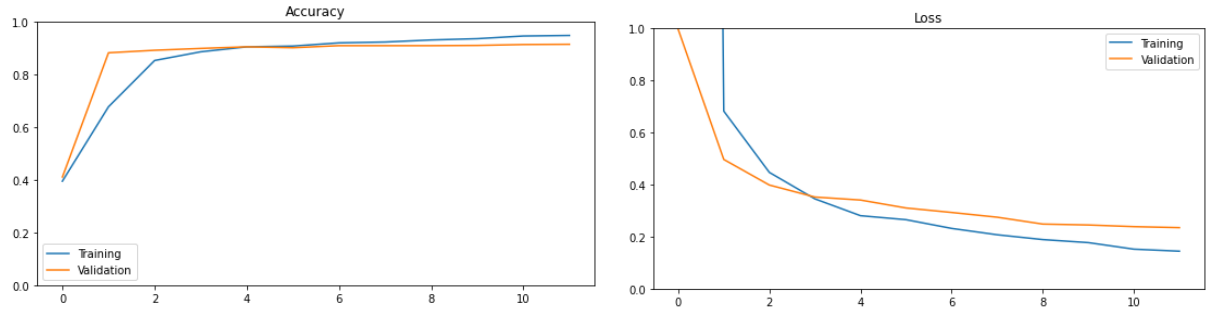
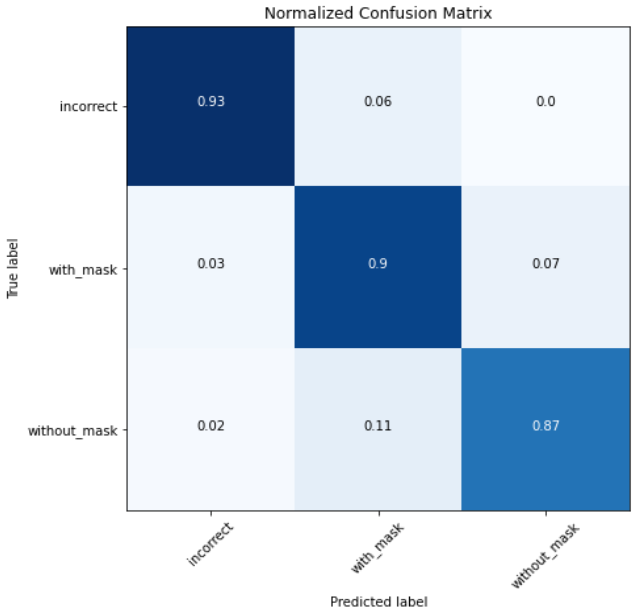



Figure 9. Baseline model training

On the test set, the model was able to generalize well, reaching a 90% accuracy; however, there is clearly space for improvement, especially in distinguishing subjects without masks as it misclassified 13% of the samples as wearing a mask.




| | Precision | Recall | F1 |
|---|---|---|---|
| **incorrect** | 0.96 | 0.93 | 0.94 |
| **with_mask** | 0.85 | 0.90 | 0.87 |
| **without_mask** | 0.91 | 0.87 | 0.89 |

Figure 10. Baseline model performance on test set

## Comparison of model architectures

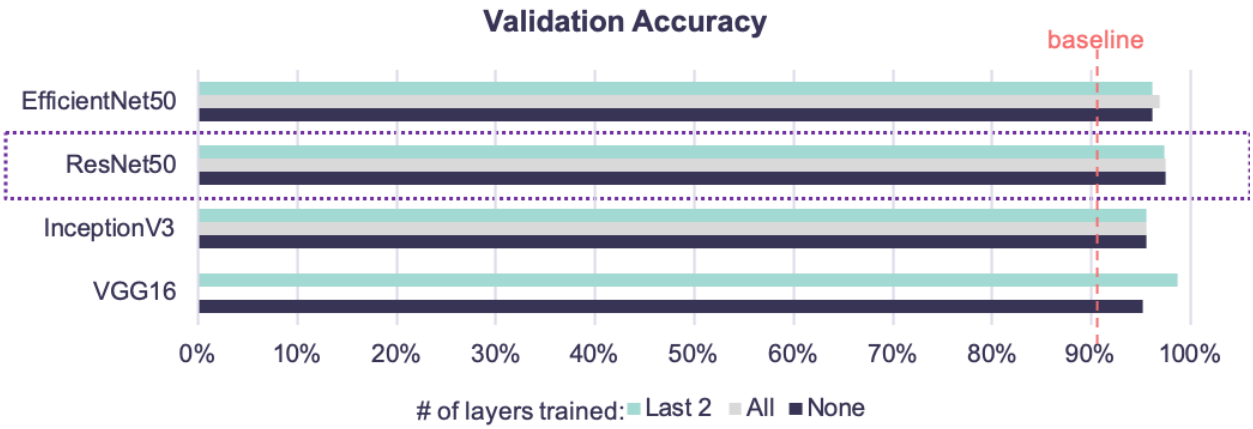
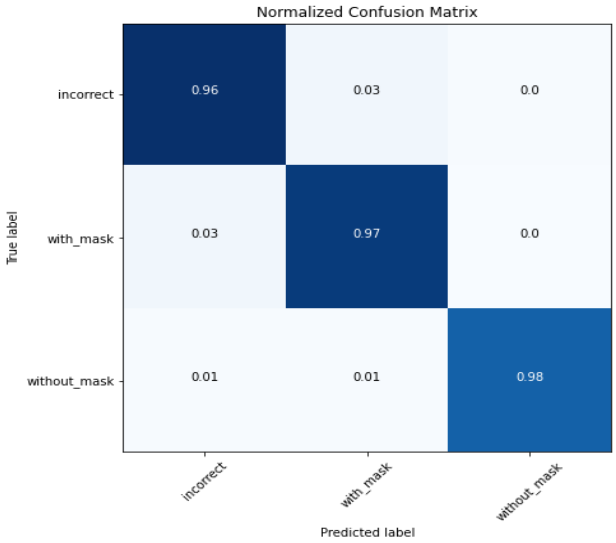



Figure 9. Validation accuracy of various model architectures

While all model architectures surpassed the baseline accuracy, ResNet50 consistently outperformed the other networks on all three cases (although by a slight margin only) so we have chosen to create the final model based on a ResNet50 architecture.

## Final model

By experimenting with the number of layers to be fine-tuned and the learning rate, we have achieved a better performance by fine-tuning the last three layers using a learning rate of 0.001. Predictive accuracy on the test is very high, landing at 97%.




| | Precision | Recall | F1 |
|---|---|---|---|
| **incorrect** | 0.96 | 0.96 | 0.96 |
| **with_mask** | 0.96 | 0.97 | 0.96 |
| **without_mask** | 0.99 | 0.98 | 0.99 |

Figure 10. ResNet model performance on test set

Looking at the predictions for each class, it is important to highlight the following:

1) Model achieved top performance on identifying subjects without masks (f1 score = 0.99)
    - Out of the subjects who are not wearing masks, the model misses to identify only 2% of the group. Keeping this number to a minimum is crucial from a practical standpoint as these are the individuals who need to be prompted to comply with the protocols.
2) The lowest predictive performance on a class still surpassed our benchmark, achieving a 96% recall
    - The error mainly comes from misclassifying the 3% of subjects with properly worn masks as not having worn them properly. If implemented, making this type of mistake is more 'forgivable' as it leans towards a more cautious reaction.

# IV. Conclusion

Transfer learning has been key to unlocking superior model performance. Using the ResNet50 architecture, we were able to achieve a 97% accuracy rate on the test set, demonstrating the potential of using computer vision to enforce stricter compliance to face mask regulations.

It is worth noting, however, that this performance level is based on the specific dataset used for the project which has its limitations. Majority of incorrectly worn face masks samples were synthetic, edited surgical face masks overlaid on the subjects' faces, raising possible challenges in the generalizability of the model, specifically in detecting non-surgical incorrectly worn masks. Moreover, the training dataset could benefit from increased diversity in terms of age and race. Hence, before implementation, model performance should be tested against these possible shortcomings.

With deployment in mind, future projects can focus on extending the model to make classifications for multiple subjects at one as this will be valuable in monitoring crowded places like train stations and airports. Moreover, future models can also aim at making predictions in real-time (video input instead of images) and improving the efficiency of the model.