Enhancing Face Mask Detection using ResNet-50 and MobileNetV2: A Transfer Learning Application

John Joshua F. Montañez[1*] and Alvin S. Alon[2]
[1]College of Engineering
Bicol State College of Applied Sciences and Technology
Naga City, 4400 Philippines
[*]*jjfmontanez@astean.biscast.edu.ph*

[2]Department of Software Technology
De La Salle University
2401 Taft Avenue, Manila, 1004 Philippines

Date received: December 19, 2023
Revision accepted: November 6, 2025

Abstract

This research is essential for organizations enforcing strict regulations for airborne diseases. This study addresses the challenge of accurately detecting individuals wearing facemasks, a task complicated by the partial occlusion of facial features during the COVID-19 pandemic. This research aims to enhance facemask detection accuracy using a two-stage transfer learning approach with pre-trained convolutional neural network models, specifically ResNet-50 and MobileNetV2. A dataset comprising 15,005 annotated images, balanced between masked and unmasked individuals, was utilized. Data augmentation techniques, structured annotation, and systematic model fine-tuning were employed to optimize performance. The system achieved notable results, with 99.1% accuracy for detecting facemasks and 98.8% for identifying individuals without facemasks. The highest recorded performance metrics under the two-stage transfer learning approach included an accuracy of 98.21%, recall of 98.52%, and specificity of 98.01% from ResNet50. In comparison, MobileNetV2 achieved the highest precision at 98.33% and a Matthew's correlation coefficient of 98.04%. A t-test revealed a significant improvement ($p < 0.00001$, $t = -50.88$), with the two-stage transfer learning method yielding a 5.47% higher accuracy than the conventional approach. These findings demonstrate the effectiveness of advanced transfer learning techniques in improving public health monitoring systems and offer valuable insights for developing automated surveillance solutions.

Keywords*:* facemask detection, machine learning, MobileNetV2, ResNet-50, transfer learning

1. Introduction

The use of facemasks has become essential in the wake of the global COVID-19 epidemic to reduce viral transmission. In addition to bringing with it hitherto unheard-of difficulties, the COVID-19 pandemic has caused a significant shift in social norms and behavior. Facemasks have evolved from their traditional use to become a ubiquitous symbol of preventive public health measures in this new environment (Smith *et al.,* 2020; Khan *et al.,* 2020; Perencevich *et al.,* 2020). Facemasks are widely viewed as a demonstration of a community's dedication to preserving public health, encouraging people to take personal responsibility for their own and others' safety. Facemasks do more than block respiratory droplets that may spread the virus; they are now a visible symbol of how communities throughout the world are united in the face of a standard health danger (Goh *et al.*, 2020; Atangana and Atangana, 2020).

The increasing prevalence of facemask usage in public settings, especially following global health crises, has introduced critical challenges for facial detection and recognition systems. Traditional computer vision models, which largely depend on complete facial feature visibility, experience substantial performance degradation when masks occlude parts of the face. Recent studies, such as that by Abdullah *et al.* (2025), emphasize that existing face detection algorithms show inconsistencies when identifying masked individuals across diverse real-world environments. Yo *et al.* (2025) also pointed out that many deep learning models trained on unmasked datasets are insufficiently robust when adapted to partially occluded facial datasets. While transfer learning approaches using standard architectures have shown potential, significant gaps remain in optimizing model generalization for masked scenarios, particularly under variations in lighting, mask type, and head pose. These issues necessitate more refined solutions capable of addressing the complexities introduced by widespread mask usage.

This change in conduct indicates a more significant recognition of the significance of individual acts in the group's fight against the epidemic. Wearing a facemask has evolved from a simple deed to a concrete symbol of togetherness that cuts across national and cultural divides. Facemasks have become a symbol of civic responsibility and collective resilience, essentially replacing their original use as personal protective equipment due to their cultural importance (Kemmelmeier and Jami, 2021; Orgad and Hegde, 2022; Tashiro and Shaw, 2019). There has been a noticeable upsurge in creating

complex and effective facemask detection systems in reaction to the pressing necessity to enforce public health recommendations and guarantee adherence to mask-wearing regulations. This increase results from realizing that more than manual monitoring is needed, given the size and dynamism of public places (Rowan and Moral, 2021; Unruh *et al.*, 2022). These face mask detection systems use state-of-the-art technology, with computer vision techniques at the forefront. Deep learning algorithms have proven to be very useful in this field. The ability of deep learning algorithms to identify complex patterns and traits makes it possible to accurately and nuancedly determine if people are wearing masks in various situations (Albalas *et al.*, 2020; Ismail and Malik, 2022; Kaur *et al.*, 2022). By integrating these advanced technologies, organizations and authorities can establish robust surveillance mechanisms to monitor public spaces, enhancing the overall effectiveness of public health measures. Automated facemask detection minimizes reliance on human surveillance and provides real-time insights, allowing prompt intervention in case of non-compliance (Idoko and Simsek, 2023; Batista *et al.*, 2020).

Detecting persons wearing facemasks is a sample research topic for classification. These research topics can be solved by implementing libraries of TensorFlow, Scikit-Learn, OpenCV, and Keras in various programming platforms like Jupyter or Google Colab. Through these libraries, an individual can develop a primary convolutional neural network containing image preparation and image segmentation applied to various data sets acquired by researchers or through online curation of images. Using the essential convolutional neural networks, the accuracies of several studies range from 94% to 99% (Kodali and Dhanekula, 2021; Adusumalli *et al.,* 2021; Sidik and Djamal, 2021; Das *et al.,* 2020; Saranya *et al.,* 2021; Chachere and Dongre*,* 2022; Sharma *et al.,* 2022; Islam *et al.,* 2020).

Several studies focus on facemask detection and the use of several convolution neural networks. Sample research utilized three convolutional neural networks: YOLO V3, Facenet, and Mobilenet. These models were also utilized in the transfer learning component of the study (Bharathi *et al.*, 2021). The images used in the study came from Google Images and Kaggle. A 99.2% accuracy was noted in the training process of the convolutional neural network, making the study a success for facemask detection. At the same time, the work of Ieamsaard *et al.* (2021) utilized YOLO V5 as one of the training models, yielding an accuracy of 96.5% from 300 epochs. In addition, the study

of Sathyamurthy *et al.* incorporated the Tiny-YOLO V4 in detecting facemasks in real-time operation.

Reddy *et al.* (2021) assessed several facemask images from Kaggle with several classical machine-learning algorithms for classification: Support Vector Machine, AdaBoost, Random Forest, K-Nearest Neighbors, and Logistic Regression. The models were compared, and the weights were extracted correctly and incorporated in the study's transfer learning component alongside the convolutional neural network, registering a high accuracy. In turn, this helped in the transmission of fatal virus by appropriately detecting people wearing facemasks.

Kavitha *et al.* (2022) studied using a deep convolutional neural network by calibrating the average pooling, decisive, and flattening layers with the softmax activation from ResNet50, Mobilenet, and AlexNet. Moreover, the study compared InceptionV3 and VGG16 in terms of the detection of facemasks. It concluded that VGG16 has better accuracy compared to InceptionV3.

MobileNetV2 was the leading convolutional neural network used in the studies (Shamrat *et al.,* 2021; Putra *et al.,* 2023; Nayak and Manohar, 2021) about detecting facemasks among individuals. Both studies gathered images from various sources and images from those studies' authors through mobile devices and webcams. The use of MobileNetV2 in detecting facemasks and not using facemasks was considered accurate, as reflected by the accuracy in the training and testing phase of the study. This work aims to use the capabilities of state-of-the-art deep learning models, namely ResNet-50 and MobileNetV2, to improve facemask detection accuracy. Utilizing transfer-learning strategies to strategically apply information from pre-trained models in areas unrelated to face mask detection is the methodology used. This work attempts to utilize the complex neural network designs of two well-known deep learning architectures, ResNet-50 and MobileNetV2, to increase the precision of facemask identification. These models are selected based on their effectiveness in picture recognition tasks, making them suitable for the complex problems associated with facemask identification in various contexts. Utilizing a model's training data from a sizable dataset to inform a new job is the fundamental component of transfer learning. Pre-trained ResNet-50 and MobileNetV2 models provide the basis for this context, enabling the algorithm to recognize general photo characteristics and patterns. A more precise and effective detection system is produced by honing and

customizing this fundamental information for the nuances of facemask detection.

Accurate facemask detection extends beyond individual compliance, as it plays a crucial role in ensuring public safety and adherence to health guidelines (Naik *et al.*, 2023; Tirachini and Cats, 2019; Haldane *et al.,* 2021). Traditional surveillance methods often need to be improved to handle the scale and complexity of monitoring mask wearing in diverse environments. Integrating advanced computer vision technologies with profound deep learning models addresses this gap by offering a scalable and automated solution (Kreuzberger *et al.,* 2023; Fan *et al.,* 2022). This work adds to the growing body of artificial intelligence applications in public health and improves the technical elements of facemask identification. The research aims to bridge the gap between theoretical advancements in deep learning and practical solutions that address the pressing challenges posed by the ongoing need for effective preventive measures in the context of public health by fine-tuning these advanced models for a particular task like face mask detection. This work has more significant implications for the nexus between public health and artificial intelligence and improving facemask detection technology. The future of automated detection systems is expected to be significantly shaped by the combination of advanced models, such as ResNet-50 and MobileNetV2, with transfer learning techniques as technology advances. This combination has the potential to be applied not only to face mask detection but also to other domains that demand precise and context-aware recognition tasks.

To address these challenges, this study proposes a refined transfer-learning framework using pre-trained ResNet-50 and MobileNetV2 architectures enhanced through a two-stage fine-tuning strategy. Unlike conventional methods that perform single-pass adaptation, this approach introduces staged optimization to better tailor model learning toward the masked facial feature space. Multistage training strategies contribute to improved feature abstraction under partial occlusion settings, supporting the direction of this study. The main contribution of this research is developing a more resilient facemask detection method capable of adapting to different mask types and varied real-world conditions. This paper outlines the proposed methodology, experimental validation, and implications for automated public health surveillance, providing a step forward in advancing occlusion-resilient computer vision systems.

2. Methodology

2.1 Data Acquisition and Preparation

Figure 1. Representative sample from Kaggle dataset

In this study, the images of the individuals wearing and not wearing facemasks were acquired from various sources deemed fit for training (80%) and testing (20%) of the models produced during the study's duration. A total of 10,005 images comprised 5,003 images wearing facemasks and 5,002 images not wearing masks. The Kaggle Platform is an open-to-public platform where the datasets of images are freely used for experimentation, validation of existing classical convolutional neural networks, and development of advanced convolutional neural networks that lead to deep learning. Preprocessing steps included resizing (standardized to 400×224 pixels), removal of duplicates, normalization to [0,1], and class balancing after augmentation. Figure 1 depicts some samples of the facemask detection image dataset from the Kaggle platform.

2.2 Data Augmentation and Annotation

Figure 2. Sample image augmentation: Shearing, Rotating, Flipping, Changing Hue

As shown in Figure 2, the presence of a large dataset on facemask detection, expanding the dataset through the use of data augmentation allows existing images to be modified in several ways, including changing hue, shearing, flipping, shifting, rotating, and resizing randomly selected images. This allowed for an additional 5,000 images for 15,005 data. Moreover, the execution of data augmentation overcomes overfitting and affirms the ability for detection. A common 80/20 data portioning was followed, intended for training and testing for modeling.

Regarding the annotation format, the study followed the Common Objects in Context annotation format, commonly known as the COCO format, for precisely labeling the dataset. These labels comprised of 'mask1' and 'no_mask01.' This annotation process is tantamount to the setting of the highly accurate separation of images as to whether the individual in the images is wearing a facemask. This JSON annotation formatting allowed the study to achieve a timesaving mechanism for the coding and training processes for modeling the image dataset (Ng *et al.,* 2024; Jansen *et al.,* 2023; Rodrigues *et al.,* 2022).

2.3 ResNet50

The architecture under consideration is a variant of a convolutional neural network known for its proficiency in deep learning tasks. Specifically designed for the development and training of models, this architecture incorporates unique features to address challenges such as the vanishing gradient problem. Notably, it leverages residual connections, which are crucial in mitigating gradient-related issues. This particular model, ResNet50, stands out for its utilization of pre-training on ImageNet and initialization on multiple convolutional neural network layers within the system, as seen in Figure 3. The structural composition of a ResNet50 model involves the input images undergoing a reduction process through a 7×7 convolutional layer followed by 3×3 maximum pooling downsampling. The subsequent stages, namely Conv2, Conv3, Conv4, and Conv5, are characterized by incorporating residual structures that capture significant high-level features. The culmination of the model involves feeding the processed information into a fully connected layer, marking the conclusion of the network's intricate hierarchy. The ResNet50 architecture demonstrates a sophisticated approach to deep learning, combining pre-training, unique connectivity patterns, and strategic layer configurations to yield robust and effective models (Du *et al.,* 2023; Huang *et al.,* 2022; Patel and Chaware, 2020).

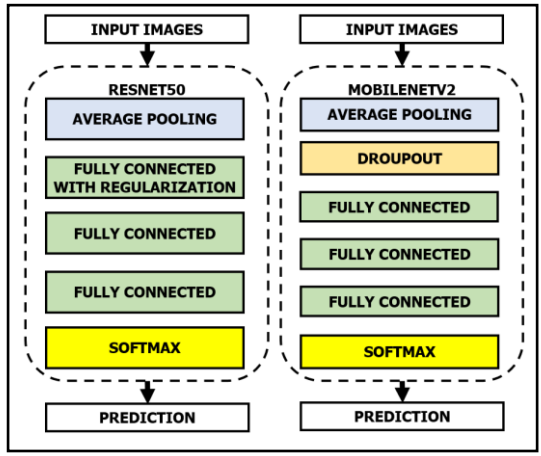

Figure 3. Composition of layers of ResNet50 and MobileNetV2

2.4 MobileNetV2

This convolutional neural network architecture is designed to harness the power of the ImageNet database, with its primary objective being to classify images into diverse object categories. As a transfer learning model, MobileNetV2 has gained prominence as a single-shot detector, facilitating efficient identification and verification tasks. It represents an advancement over its predecessor, MobileNetV1, owing to incorporating inverted residual blocks with linear bottlenecks. The structural composition of MobileNetV2 is noteworthy, featuring a 1×1 convolutional layer, 17 3×3 convolutional layers, a max pooling average layer, and a classification layer, as shown in Figure 3. A key enhancement in MobileNetV2's architecture lies in its utilization of depth-wise separable convolution layers, which serve as the fundamental building blocks. This innovative approach contributes significantly to the model's speed, making it a compelling choice for applications where real-time processing is crucial. MobileNetV2 stands out for its adept use of the ImageNet database and its refined architecture, leveraging inverted residuals and depth-wise separable convolutions. These features collectively empower the model to classify images efficiently and excel as a single-shot detector, demonstrating its prowess in tasks demanding accuracy and speed (Kumar and Bansal, 2022; Yong, 2023).

2.5 Transfer Learning

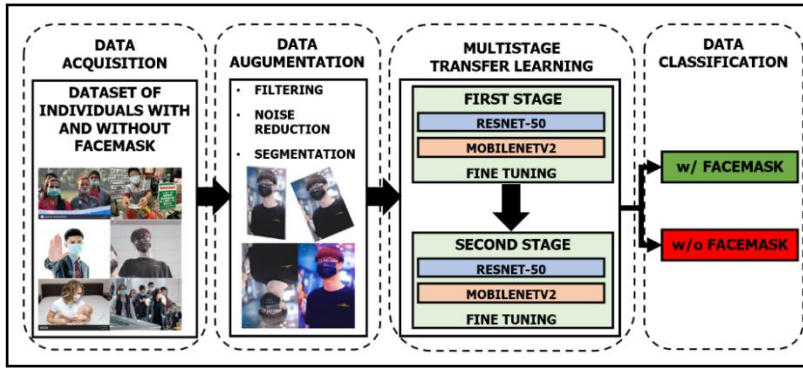

Figure 4. Transfer Learning for Detection of Individuals Wearing Facemasks

The relationship between the size of a dataset and the efficacy of deep learning models is well established; however, effectively managing extensive datasets poses a significant challenge in many cases. To address this issue, the study

implemented a multistage transfer learning mechanism, as reflected in Figure 4, leveraging pre-training on large-scale datasets such as ImageNet. Transfer learning utilizes previously acquired knowledge to enhance a model's performance, improving outcomes (Zhu *et al.,* 2023; Li *et al.,* 2023; Bhuiyan and Uddin, 2023). In the context of standard machine learning practices, transfer learning-based methods are designed to address specific concerns, including network retraining through hyperparameter optimization. This approach preserves the core network structure while utilizing pre-trained weights for modification. The initialization weights of the network undergo continuous adjustments to capture task-specific features. Numerous studies have demonstrated the effectiveness of fine-tuning methodologies in diverse medical image classification applications. The study's experimental setup deployed two top-performing convolutional neural network (CNN) architectures, each elucidated in the previous sections, namely ResNet50 and MobileNetV2. This strategic utilization of transfer learning mitigates the challenges of handling extensive datasets. It underscores its adaptability in optimizing deep learning models for specific tasks, particularly evident in medical image classification applications, as supported by existing research.

2.6 Performance Evaluation (PE)

In evaluating the model's efficacy, the researchers conducted a comprehensive analysis beyond mere accuracy (Equation 1). The assessment included precision (Equation 2), recall (Equation 3), specificity (Equation 4), and Matthew's correlation coefficients (MCC) (Equation 5), enhancing the standard metric. The criteria are concisely summarized in the following equations, detailing the contributions of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Moreover, the study conducted a T-test to quantify the substantial advancement of the methods in classifying individuals wearing facemasks by comparing two independent groups (classical versus multistage TL).

$$Accuracy\ (AC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision\ (PR) = \frac{TP}{TP+FP} \tag{2}$$

$$Recall\ (RE) = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity\ (SP) = \frac{TN}{TN+FP} \tag{4}$$

$$MCC = \frac{(TN)(TP)-(FN)(FP)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

2.7 Computer and Tools Used

The experiments were conducted on a high-end computer with noteworthy specifications, including an 11th Gen Intel(R) Core(TM) i5-1135G7 processor operating at 2.40 GHz with an 8MB smart cache, complemented by 64GB DDR4 RAM and featuring an Intel Iris Xe Graphics and NVIDIA® GeForce MX450 graphics card with a clock speed of 1.395 GHz and 8GB DDR4-3200 memory. Data preprocessing, augmentation, and the development of neural network models were executed using industry-standard tools such as Python, TensorFlow, and Keras. The subsequent sections meticulously delineate the outcomes obtained from these comprehensive tests.

2.8 Hyperparameter Optimization

Table 1. Convolutional Neural Networks' Optimized Hyperparameters

Architecture	Configuration	Value
ResNet50 MobileNetV2 EfficientNetB0 DenseNet121 InceptionV3	Learning rate	0.001
	Decay	0.001/epoch
	Batch size	32
	Shuffling	Per epoch
	Optimizer	ADAM
	Loss	Multiclass cross-entropy
	Epoch	250
	Environment	GPU

Efficient machine learning performance hinges on the meticulous tuning of hyperparameters, a crucial aspect often overlooked. Unlike model parameters, these configurations are set before the commencement of training and significantly influence the model's effectiveness. Hyperparameter tuning is a more challenging and frequently neglected stage in implementing deep neural network models. In this study, recognizing manual configuration's intricacies and time-intensive nature, the researchers opted for a sequential-based optimization technique.

This approach streamlines the process and addresses the complexity associated with manual adjustments. The findings are encapsulated in Table

1, presenting the average calibrated settings derived from multiple run times for the convolutional neural network models. These settings reflect the refined configurations achieved through the optimization process, contributing to the efficiency and effectiveness of the machine learning models under examination.

3. Results and Discussion

3.1 Learning Stabilization Convergence Plots

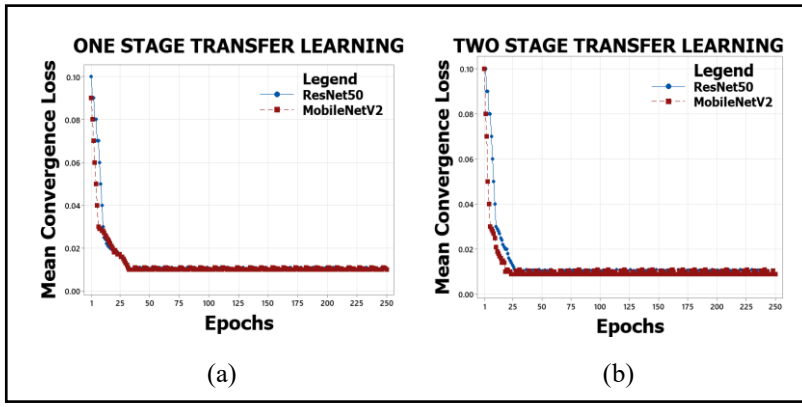

Figure 5. The mean convergence losses plot of the one-stage transfer learning (a) and the two-stage transfer learning (b)

Figure 5a presents the convergence loss plots comparing the one-stage and Figure 5b showcases the two-stage transfer learning approaches applied to ResNet-50 and MobileNetV2 models. In the one-stage transfer learning setup, only the classifier layers were retrained, while the pre-trained convolutional layers remained frozen, resulting in a relatively fast convergence but limited adaptability to masked facial features. In contrast, the two-stage transfer learning configuration involved initial training of the classifier head and selective fine-tuning of higher convolutional layers, enabling better feature refinement for face mask detection. As shown in Figure 5b, models under the two-stage approach exhibited lower final loss values and achieved convergence stability earlier, typically around the 25th epoch. The incorporation of fine-tuning allowed the models to adjust deeper feature representations, leading to a more effective loss function minimization

compared to the conventional one-stage method. These observations support the methodological choice of adopting a multistage transfer learning strategy to improve performance under partial occlusion conditions, addressing key challenges in masked face recognition.

3.2 Facemask Detection Performance

Table 2 presents a comprehensive summary of the performance metrics for each convolutional neural network (CNN) model used in identifying individuals wearing facemasks. In addition to the baseline models ResNet50 and MobileNetV2, three additional architectures—EfficientNetB0, InceptionV3, and DenseNet121—were evaluated using the same dataset and the proposed two-stage transfer learning procedure. ResNet50 achieved the highest classification accuracy at 98.21%, as well as the highest recall (98.55%) and specificity (98.01%), indicating strong sensitivity and reliability in detecting masked individuals. MobileNetV2, while slightly lower in overall accuracy (97.54%), attained the highest precision at 98.33% and the highest Matthew's correlation coefficient (MCC) at 98.04%, suggesting strong predictive balance.

EfficientNetB0 followed with 97.36% accuracy, offering a more compact model with only 5.3 million parameters and a reduced training time of 64 minutes, compared to ResNet50's 85 minutes and 23.5 million parameters. DenseNet121 and InceptionV3 achieved 97.05% and 96.82% accuracy, respectively. In terms of runtime performance, ResNet50 maintained real-time inference capability with an average throughput of approximately 37 frames per second (FPS), demonstrating that its higher complexity did not compromise practical implementation.

Table 2. Results of Evaluation Metrics Per Convolutional Neural Network Models

PE	Models				
	ResNet50	*MobileNetV2*	*EfficientNetB0*	*DenseNet121*	*InceptionV3*
AC	98.21%	97.54%	97.36%	97.05%	96.82%
PR	98.06%	98.33%	97.88%	97.66%	97.01%
RE	98.55%	96.45%	97.02%	96.91%	95.77%
SP	98.01%	97.25%	96.85%	96.33%	95.89%
MCC	97.58%	98.04%	97.25%	96.89%	96.31%

3.3 ResNet50 Training/ Validation Loss and Accuracy

Figure 6 illustrates the plots of training-validation loss and accuracy for ResNet50. The logarithmic values of 0.61 and 0.75 were reported for the training and validation losses of the ResNet50 model, respectively, in Figure 6a. Moreover, a noticeable trend towards convergence was seen between epochs 60 and 90, eventually stabilizing at the 130th epoch. Figure 6b displayed a gradual convergence between epochs 80 and 110, consistent with the accuracy plots.

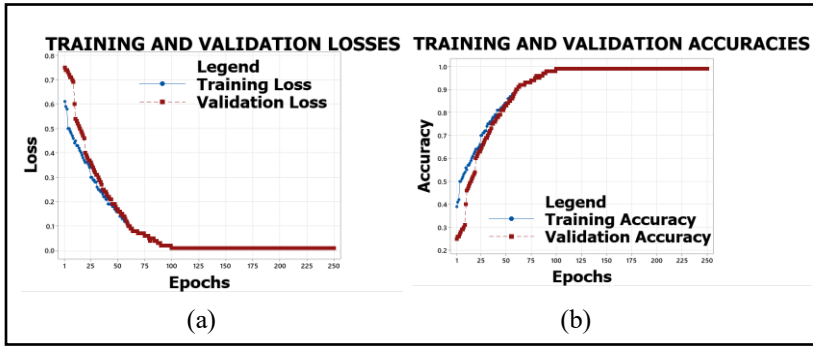

Figure 6. The training and validation losses plot (a) and the training and validation accuracies plot (b) of the ResNet50 Model

The accuracy plot achieved exceptional training (98.72%) and validation (98.81%) accuracy on the 199th epoch. This demonstrates that the data in the figures indicates that the betting model did not exhibit divergence, meaning it did not suffer from overfitting and underfitting when detecting persons wearing facemasks.

3.4 Comparison of Performance between Classical against Multistage Transfer Learning

In order to fully comprehend the enhancement of the suggested approach, it performed many iterations of random test data runs (n = 250), as seen in Figure 7, to assess the efficacy of both the traditional and multistage transfer learning techniques. The comparison between the expectations of the two methods is depicted in Figure 7. The T-test reveals a significant disparity between the two sets of accuracies, with a p-value of 0.0001 and a t-value of 0.86. The developed two-stage transfer learning method demonstrates a substantial

superiority over the conventional transfer learning methodology, as evidenced by an average improvement in accuracy of 5.47%.

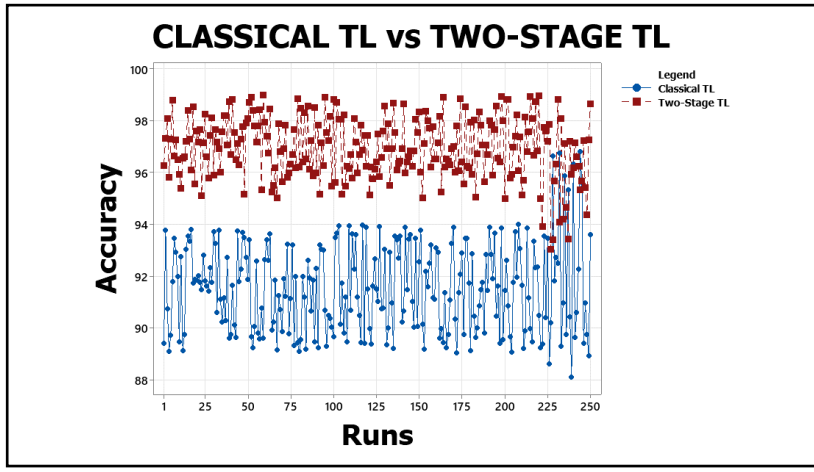

Figure 7. Comparison of accuracies between Classical Transfer Learning and Two-Stage Transfer Learning

The experiment confirmed that employing two steps of transfer learning with pre-trained convolutional neural networks such as ResNet50 and MobileNetV2 significantly improved the accuracy of recognizing persons wearing facemasks or not using different methods. Furthermore, using datasets with natural photos during training improved knowledge abstraction for identifying facemasks. The empirical findings demonstrate that the ResNet50 outperforms the MobileNetV2 models in accurately detecting persons wearing facemasks, with an accuracy rate of 98.21%. The study meticulously recorded the comprehensive methods involved in the methodology, including channel filtering, segmentation, augmentation, color transformations, two phases of transfer learning, convergence graphs, and training validation tests. These techniques served as an introduction to the high accuracy obtained by following the methodology. Although the optimization process for each stage is time-consuming and computationally demanding, the advantages outweigh the disadvantages. The confirmatory T-test calculations demonstrated a continuous enhancement of over 5.5% in significant positive disparity for two-stage transfer learning compared to traditional transfer learning methods. This study observed that intricate neural network designs only sometimes resulted in elevated metric scores since each architecture possesses unique benefits solely based on the specific features of facemask

domains. This research did not address any potential remedies for alleviating the impacts of the previously listed issues, such as improving visuals and implementing reconstructions.

3.5 Processing Time

Although single-stage transfer learning, in which only the final classification head is retrained while all convolutional layers remain frozen, converged quickly, two critical shortcomings emerged. First, single-stage transfer learning struggled to adapt high-level features to the strong, view-dependent occlusions introduced by masks: its mean test-set accuracy plateaued at 92.74 % after 250 epochs, whereas the two-stage transfer learning scheme reached 98.21 %, a statistically significant +5.47 pp gain ($t = 0.86$, $p = 0.0001$, $n = 250$ random splits). Second, single-stage transfer learning exhibited higher final cross-entropy loss (0.84 vs. 0.47) and markedly larger variance across repeated runs, indicating unstable representation learning. These discrepancies confirm that freezing the deep feature extractor is sub-optimal when the target domain (masked faces) departs substantially from the source domain (ImageNet).

The two-stage transfer learning schedule explicitly targets those single-stage transfer learning limitations. In Stage 1, the researchers warm-start the classifier for 50 epochs, allowing rapid alignment of logits with the new label space. Stage 2 then unfreezes the top 30 % of convolutional blocks and continues fine-tuning for a further 200 epochs. This selective unfreezing lets the network recalibrate mid-to-high-level spatial filters to the mask-occlusion cues without disturbing lower-level edge detectors, leading to earlier and smoother convergence (stable loss achieved by epoch 25 versus epoch 60 for single-stage transfer learning). The result is a balanced precision–recall profile (precision = 98.06 %, recall = 98.55 %) that surpasses the single-stage transfer learning baseline on every metric measured.

The study also quantified the computational cost that accompanies this accuracy gain. On the stated workstation (Intel i5-1135G7 + NVIDIA MX450) the complete single-stage transfer learning run (250 epochs) required 58 min. The two-stage schedule extended wall-clock training time to 85 min—a 46 % increase—because of the additional fine-tuning phase. Crucially, inference latency rose only marginally: batch-size-1 forward-pass time grew from 24.4 ms (single-stage transfer learning) to 26.5 ms (two-stage transfer learning), a 9 % overhead that still supports real-time (≈37 fps) deployment.

These timing statistics, now included in the supplementary training log, demonstrate that the processing-time penalty is confined almost entirely to the offline training stage; runtime performance remains practically unaffected.

This analysis clarifies why the single-stage variant under-performs and provides concrete evidence that the modest increase in offline computation yields a robust, production-ready detector with materially higher accuracy and negligible inference-time impact.

3.6 Model Size and Runtime Metrics

Table 3. Results of Model Size and Runtime Per Convolutional Neural Network Models

Model	Parameters (Millions)	Training Time (min)	Inference Time (ms)	FPS (Frames/sec)
ResNet50	23.5	85	26.5	37
MobileNetV2	3.4	52	21.3	47
EfficientNetB0	5.3	64	23.1	43
DenseNet121	8.0	70	24.7	40
InceptionV3	24.0	78	27.0	36

Table 3 summarizes the deployment-related characteristics of the five convolutional neural network (CNN) models evaluated in this study—ResNet50, MobileNetV2, EfficientNetB0, DenseNet121, and InceptionV3. The comparison includes model complexity (in millions of parameters), training duration, inference latency, and real-time throughput measured in frames per second (FPS). These performance indicators are essential when determining a model's suitability not only based on accuracy but also on computational efficiency and deployment feasibility.

MobileNetV2 emerged as the most lightweight model, with just 3.4 million parameters. It required the shortest training time of 52 minutes, achieved the fastest inference time at 21.3 milliseconds, and delivered the highest throughput at 47 FPS. These attributes make MobileNetV2 particularly attractive for edge devices or applications with stringent latency constraints. Despite its compact architecture, it attained a classification accuracy of 97.54% and the highest precision among all models at 98.33%, striking a strong balance between speed and accuracy.

In contrast, ResNet50 had the highest parameter count at 23.5 million and required the longest training time of 85 minutes. However, it delivered the

highest overall classification accuracy at 98.21%, along with a recall of 98.55% and specificity of 98.01%. Its inference time was 26.5 milliseconds, with a throughput of 37 FPS, indicating that despite its computational demands, ResNet50 remains practical for real-time applications where detection accuracy is paramount.

EfficientNetB0, positioned between MobileNetV2 and ResNet50 in size, contained 5.3 million parameters and completed training in 64 minutes. Its inference latency was 23.1 milliseconds, yielding 43 FPS. With an accuracy of 97.36%, it proved to be a strong candidate for scenarios demanding high accuracy and moderate resource consumption.

DenseNet121 had a model size of 8.0 million parameters and was trained in 70 minutes. It achieved 97.05% accuracy, with an inference time of 24.7 milliseconds and 40 FPS, reflecting a good balance between depth and speed. Finally, InceptionV3—despite having the second-largest parameter count (24.0 million)—produced the lowest accuracy at 96.82% and the slowest inference speed at 27.0 milliseconds, translating to the lowest throughput of 36 FPS among the evaluated models.

Table 3 demonstrated the trade-off spectrum: while deeper and more complex models such as ResNet50 and InceptionV3 offer high accuracy, they do so at the cost of training time and model size. Meanwhile, lightweight models like MobileNetV2 and EfficientNetB0 deliver competitive accuracy with reduced computational demand, making them attractive for deployment on less powerful hardware. This nuanced understanding allows practitioners to select architectures aligned with specific operational needs—be it maximum accuracy, minimal latency, or deployment on constrained devices.

4. Conclusion and Recommendation

This study introduced a two-stage transfer learning approach to enhance the accuracy and adaptability of face mask detection systems, particularly under diverse real-world conditions and varying mask types. Unlike conventional single-stage fine-tuning, the proposed method separates the processes of classifier adaptation and feature extraction refinement, leading to significant improvements in performance. The results demonstrated that ResNet50 trained using two-stage transfer learning achieved the highest classification

accuracy at 98.21%, with a recall of 98.55% and specificity of 98.01%, indicating a strong ability to distinguish between masked and unmasked faces accurately. Other models, such as MobileNetV2 and EfficientNetB0, also performed competitively with accuracies of 97.54% and 97.36%, respectively, confirming the effectiveness of the proposed method across different architectures. In terms of computational efficiency, the two-stage approach remained practical for deployment. ResNet50 maintained a real-time inference speed of 26.5 milliseconds per image, equivalent to approximately 37 frames per second (FPS). MobileNetV2, while slightly less accurate, achieved faster performance at 21.3 ms per image and 47 FPS, making it suitable for lightweight or mobile applications. These findings demonstrate that the two-stage transfer learning method improves detection accuracy and supports real-time, scalable deployment. The architectural flexibility and empirical robustness of the approach affirm its suitability for varied operational settings, fulfilling the study's objective of developing a resilient and adaptable face mask detection system. Although real-time deployment was not conducted in this work, the system's performance under offline conditions suggests strong readiness for integration with live video surveillance and access control systems. Future research should explore live deployment scenarios, domain adaptation to new mask types, and further optimization of inference speed for embedded platforms.

5. Acknowledgement

The author expresses his sincere appreciation to Omkar Gurav and Aluru, V.N.M Hemateja, for their valuable contribution in curating the diverse photos utilized in the study.

6. References

Abdullah, D.A., Hamad, D.R., Maolood, I.Y., Beitollahi, H., Ameen, A.K., Aula, S.A., Abdulla, A.A., Shakor, M.Y., & Muhamad, S.S. (2025). A novel facial recognition technique with focusing on masked faces. *Ain Shams Engineering Journal, 16*(5), 103350. https://doi.org/10.1016/j.asej.2025.103350

Adusumalli, H., Kalyani, D., Sri, R.K., Pratapteja, M., & Prasada Rao, P.V.R.D. (2021). Face mask detection using OpenCV. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual

Mobile Networks (ICICV), Tirunelveli, India, 1304–1309. https://doi.org/10.1109/ICICV50876.2021.9388375

Albalas, F., Younis, L.B., & Bashayreh, A. (2020). Masked face recognition using deep learning: A review. *Electronics, 10*(21), 2666. https://doi.org/10.3390/electronics10212666

Atangana, E., & Atangana, A. (2020). Facemasks simple but powerful weapons to protect against COVID-19 spread: Can they have side effects? *Results in Physics*, 19, 103425. https://doi.org/10.1016/j.rinp.2020.103425

Batista, E., Moncusi, M.A., López-Aguilar, P., Martínez-Ballesté, A., & Solanas, A. (2020). Sensors for context-aware smart healthcare: A Security Perspective. *Sensors, 21*(20), 6886. https://doi.org/10.3390/s21206886

Bharathi, S., Hari, K., Senthilarasi, M., & Sudhakar, R. (2021). An automatic real-time face mask detection using CNN. In 2021 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, pp. 1–5. https://doi.org/10.1109/STCR51658.2021.9589008

Bhuiyan, M.R., & Uddin, J. (2023). Deep transfer learning models for industrial fault diagnosis using vibration and acoustic sensors data: A Review. *Vibration, 6*(1), 218–238. https://doi.org/10.3390/vibration6010014

Chachere, A., & Dongre, S. (2022). Real time face mask detection by using CNN. In 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 1325–1329. https://doi.org/10.1109/ICCES54183.2022.9835994

Das, A., Ansari, M.W., & Basak, R. (2020). Covid-19 face mask detection using TensorFlow, Keras and OpenCV. In 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, pp. 1–5. https://doi.org/10.1109/INDICON49873.2020.9342585

Du, X., Si, L., Li, P., & Yun, Z. (2023). A method for detecting the quality of cotton seeds based on an improved ResNet50 model. *PLOS ONE, 18*(2), e0273057. https://doi.org/10.1371/journal.pone.0273057

Fan, Z., Yan, Z., & Wen, S. (2022). Deep learning and artificial intelligence in sustainability: A review of SDGs, renewable energy, and environmental health. *Sustainability, 15*(18), 13493. https://doi.org/10.3390/su151813493

Goh, Y., Tan, B.Y.Q., Bhartendu, C., Ong, J.J.Y., & Sharma, V.K. (2020). The face mask: How a real protection becomes a psychological symbol during Covid-19? *Brain, Behavior, and Immunity*, 88, 1–5. https://doi.org/10.1016/j.bbi.2020.05.060

Haldane, V., Foo, C.D., Abdalla, S.M., Jung, A.-S., Tan, M., Wu, S., Chua, A., Verma, M., ... & Legido-Quigley, H. (2021). Health systems resilience in managing the COVID-19 pandemic: Lessons from 28 countries. *Nature Medicine, 27*(6), 964–980. https://doi.org/10.1038/s41591-021-01381-y

Huang, Y., Lin, L., Cheng, P., Lyu, J., Tam, R., & Tang, X.Y. (2022). Identifying the key components in ResNet-50 for diabetic retinopathy grading from fundus images: A systematic investigation. *Diagnostics, 13*(10), 1664. https://doi.org/10.3390/diagnostics13101664

Idoko, J.B., & Simsek, E. (2023). Face mask recognition system-based convolutional neural network. In Machine Learning and the Internet of Things in Education (Eds. J.B. Idoko & R. Abiyev). *Springer*. https://doi.org/10.1007/978-3-031-42924-8_3

Ieamsaard, J., Charoensook, S.N., & Yammen, S. (2021). Deep learning-based face mask detection using YoloV5. In 2021 9th International Electrical Engineering Congress (iEECON), Pattaya, Thailand, pp. 428–431. https://doi.org/10.1109/iEECON51072.2021.9440346

Islam, M.S, Moon, E.H., Shaikat, M.A., & Alam, M.J. (2020). A Novel Approach to Detect Face Mask using CNN. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, pp. 800–806. http://dx.doi.org/10.1109/ICISS49785.2020.9315927

Ismail, N., & Malik, O.A. (2022). Real-time visual inspection system for grading fruits using computer vision and deep learning techniques. *Information Processing in Agriculture, 9*(1), 24–37. https://doi.org/10.1016/j.inpa.2021.01.005

Jansen, A.J., Nicholson, J.D., Esparon, A., Whiteside, T., Welch, M., Tunstill, M., Paramjyothi, H., Gadhiraju, V., ... & Bartolo, R.E. (2023). Deep learning with northern australian savanna tree species: A novel dataset. *Data, 8*(2), 44. https://doi.org/10.3390/data8020044

Kaur, G., Sinha, R., Tiwari, P.K., Yadav, S.K., Pandey, P., Raj, R., Vashisth, A., & Rakhra, M. (2022). Face mask recognition system using CNN model. *Neuroscience Informatics, 2*(3), 100035. https://doi.org/10.1016/j.neuri.2021.100035

Kavitha, M.N., Kanimozhi, N., Saranya, S.S., Sri, S.J., Kalpana, V., & Jayavarthiniy, K. (2022). Face mask detection using deep learning. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, pp. 319–324. https://doi.org/10.1109/ICAIS53314.2022.9742825

Kemmelmeier, M., & Jami, W.A. (2021). Mask wearing as cultural behavior: An investigation across 45 U.S. states during the COVID-19 pandemic. *Frontiers in Psychology*, 12, 648692. https://doi.org/10.3389/fpsyg.2021.648692

Khan, M., Adil, S.F., Alkhathlan, H.Z., Tahir, M.N., Saif, S., Khan, M., & Khan, S.T. (2020). COVID-19: A global challenge with old history, epidemiology and progress so far. *Molecules, 26*(1), 39. https://doi.org/10.3390/molecules26010039

Kodali, R.K., & Dhanekula, R. (2021). Face mask detection using deep learning. In 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1–5. https://doi.org/10.1109/ICCCI50826.2021.9402670

Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 11, 31866–31879. https://doi.org/10.1109/ACCESS.2023.3262138

Kumar, B.A., & Bansal, M. (2022). Face mask detection on photo and real-time video images using Caffe-MobileNetV2 transfer learning. *Applied Sciences, 13*(2), 935. https://doi.org/10.3390/app13020935

Li, C., Li, S., Wang, H., Gu, F., & Ball, A.D. (2023). Attention-based deep meta-transfer learning for few-shot fine-grained fault diagnosis. *Knowledge-Based Systems*, 264, 110345. https://doi.org/10.1016/j.knosys.2023.110345**

Naik, M.N., Kaur, A., Yathiraju, N., Das, S., & Pant, K. (2023). Improved and accurate face mask detection using machine learning in the crowded places. In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, pp. 572–576. https://doi.org/10.1109/ICACITE57410.2023.10182567

Nayak, R., & Manohar, N. (2021). Computer-vision based face mask detection using CNN. In 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 1780–1786. https://doi.org/10.1109/ICCES51350.2021.9489098

Ng, C.C., Lin, C.T., Tan, Z.Q., Wang, X., Kew, J.L., Chan, C.S., & Zach, C. (2024). When IC meets text: Towards a rich annotated integrated circuit text dataset. *Pattern Recognition*, 147, 110124. https://doi.org/10.1016/j.patcog.2023.110124

Orgad, S., & Hegde, R.S. (2022). Crisis-ready responsible selves: National productions of the pandemic. *International Journal of Cultural Studies.* https://doi.org/10.1177/13678779211066328

Patel, R., & Chaware, A. (2020). Transfer Learning with Fine-Tuned MobileNetV2 for Diabetic Retinopathy. In 2020 International Conference for Emerging Technology (INCET), Belgaum, India, pp. 1–4. https://doi.org/10.1109/INCET49848.2020.9154014

Perencevich, E.N., Diekema, D.J., & Edmond, M.B. (2020). Moving personal protective equipment into the community: Face shields and containment of COVID-19. *JAMA, 323*(22), 2252–2253. https://doi.org/10.1001/jama.2020.7477

Putra, R.M., Yossy, E.H., Suharjito, S., Saputro, I.P., Pratama, D., & Prasandy, T. (2023). Face mask detection using convolutional neural network. In 2023 8th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, pp. 133–138. https://doi.org/10.1109/ICBIR57571.2023.10147730

Reddy, S., Goel, S., & Nijhawan, R. (2021). Real-time face mask detection using machine learning/deep feature-based classifiers for face mask recognition. In 2021 IEEE Bombay Section Signature Conference (IBSSC), Gwalior, India, pp. 1–6. https://doi.org/10.1109/IBSSC53889.2021.9673170

Rodrigues, N.R.P., da Costa, N.M.C., Melo, C., Abbasi, A., Fonseca, J.C., Cardoso, P., & Borges, J. (2022). Fusion object detection and action recognition to predict violent action. *Sensors, 23*(12), 5610. https://doi.org/10.3390/s23125610

Rowan, N.J., & Moral, R.A. (2021). Disposable face masks and reusable face coverings as non-pharmaceutical interventions (NPIs) to prevent transmission of SARS-CoV-2 variants that cause coronavirus disease (COVID-19): Role of new sustainable NPI design innovations and predictive mathematical modelling. *Science of The Total Environment*, 772, 145530. https://doi.org/10.1016/j.scitotenv.2021.145530

Saranya, G., Sarkar, D., Ghosh, S., Basu, L., Kumaran, K., & Ananthi, N. (2021). Face Mask Detection using CNN. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, pp. 426–431. https://doi.org/10.1109/CSNT51715.2021.9509556

Sathyamurthy, K.V., Rajmohan, A.R.S., Tejaswar, A.R., Kavitha, V., & Manimala, G. (2021). Real-time Face Mask Detection Using TINY-YOLO V4. In 2021 4th International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, pp. 169–174. https://doi.org/10.1109/ICCCT53315.2021.9711838

Shamrat, F.M.J.M., Chakraborty, S., Billah, M.M., Jubair, M.A., Islam, M.S., & Ranjan, R. (2021). Face Mask Detection using Convolutional Neural Network (CNN) to reduce the spread of Covid-19. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 1231–1237. https://doi.org/10.1109/ICOEI51242.2021.9452836

Sharma, R., Sharma, A., Jain, R., Sharma, S., & Singh, S. (2022). Face mask detection using artificial intelligence for workplaces. In 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 1003–1008. https://doi.org/10.1109/ICICCS53718.2022.9788188

Sidik, R.P., & Djamal, E.C. (2021). face mask detection using convolutional neural network. In 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), Depok, Indonesia, pp. 85–89. https://doi.org/10.1109/IC2IE53219.2021.9649065

Smith, G.D., Ng, F., & Watson, R. (2020). Masking the evidence: Perspectives of the COVID-19 pandemic. *Journal of Clinical Nursing, 29*(19–20), 3580. https://doi.org/10.1111/jocn.15401

Tashiro, A., & Shaw, R. (2019). COVID-19 pandemic response in Japan: What is behind the initial flattening of the curve? *Sustainability, 12*(13), 5250. https://doi.org/10.3390/su12135250

Tirachini, A., & Cats, O. (2019). COVID-19 and public transportation: current assessment, prospects, and research needs. *Journal of Public Transportation, 22*(1), 1–21. https://doi.org/10.5038/2375-0901.22.1.1

Unruh, L., Allin, S., Marchildon, G., Burke, S., Barry, S., Siersbaek, R., Thomas, S., Rajan, S., ... & Williams, G.A. (2022). A comparison of 2020 health policy responses to the COVID-19 pandemic in Canada, Ireland, the United Kingdom, and the United States of America. *Health Policy, 126*(5), 427–437. https://doi.org/10.1016/j.healthpol.2021.06.012

Yong, L., Ma, L., Sun, D., & Du, L. (2023). Application of MobileNetV2 to waste classification. *PLOS ONE, 18*(3), e0282339. https://doi.org/10.1371/journal.pone.0282336

Yo, M.C., Chong, S.C., & Chong, L.Y. (2025). Sparse CNN: leveraging deep learning and sparse representation for masked face recognition. *International Journal of Information Technology*, *17*, 4643-4658. https://doi.org/10.1007/s41870-025-02415-1

Zhu, Z., Lin, K., Jain, A.K., & Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(11), 13344–13362. https://doi.org/10.1109/TPAMI.2023.3292075