Real Time Face Mask Detection System Using AI and Machine Learning

Mohammad Mateen Mousoof[1], Mohammad Hamza Faisal[2], Mohd Akkas[3], Ms Tabassum[4]

[1,2,3]B.Tech Scholar, Computer Science & Engineering, Integral University, Lucknow, INDIA

[4]Assistant Professor, Computer Science & Engineering, Integral University, Lucknow, INDIA

Correspondence should be addressed to Mohammad Mateen Mousoof

mmateen0109@gmail.com[1] , khanhamza2786@gmail.com[2] , mohdakkas432@gmail.com[3] , tabassum10afia@gmail.com[4]

Abstract

The global paradigm shift in public health monitoring has cemented automated face mask detection as a critical component of smart-city infrastructure and bio-surveillance ecosystems. While initially catalyzed by pandemic-era mandates, the enduring need for non-intrusive compliance tracking in high-density public spaces demands highly accurate, privacy-preserving, and computationally efficient computer vision solutions. Despite significant advancements in deep learning, existing object detection frameworks frequently struggle with the dichotomy of high-precision classification and the stringent low-latency constraints of edge-computing devices. Heavyweight models suffer from processing bottlenecks on standard closed-circuit television (CCTV) hardware, whereas overly simplified models fail to account for complex optical variables such as dynamic lighting, dense crowd occlusion, and diverse facial orientations.

To address these critical limitations, this paper proposes a highly optimized, real-time face mask detection framework that synergizes the rapid localization capabilities of the You Only Look Once (YOLO) architecture with the lightweight feature extraction efficiency of a MobileNetV2 backbone. By mathematically streamlining convolutional operations and bounding-box regression, the proposed hybrid pipeline is engineered for seamless deployment on low-cost Internet of Things (IoT) nodes without relying on cloud-based processing. Furthermore, this research systematically synthesizes 25 pivotal studies, mapping the evolutionary trajectory of facial occlusion detection from early heuristic classifiers to modern single-stage convolutional neural networks, thereby providing a comprehensive theoretical foundation.

Empirical evaluations on a withheld testing dataset demonstrate that the MobileNetV2-backed classifier achieves an exceptional validation accuracy of 97.5%. Concurrently, the system sustains an average inference speed of 45 frames per second on commercial graphical processing units, unequivocally satisfying real-time processing thresholds. Ultimately, this framework provides a scalable, robust blueprint for integrating localized artificial intelligence into permanent, proactive public health architectures.

Keywords: Real-Time Face Mask Detection, Deep Learning, YOLO Architecture, Edge Computing, Bio-Surveillance, Sustainable Development Goal 3, SDG 9 (Innovation and Infrastructure), Smart Cities (SDG 11).

I. Introduction

The Change in Public Health Paradigms and the Context

The COVID-19 pandemic has brought about an unprecedented acceleration in the adoption of digital health technologies and automated surveillance systems. However, since the acute phase of the crisis has passed, automated face mask detection is no longer a short-term solution for a crisis but a permanent feature of bio-security in the modern world. In the current world, the ability to monitor public health without being intrusive is considered a crucial aspect of smart public infrastructure. This is in accordance with the United Nations Sustainable Development Goal 3 (SDG 3) Good Health and Well-being because it provides us with scalable solutions to prevent the spread of airborne diseases in crowded areas.

The inclusion of AI and ML in visual surveillance systems enables optical sensors to observe, track, and categorize complex human behavior in real-time. Apart from assisting in the prevention of the spread of viruses in real-time, the installation of such systems in particular locations assists in the fulfillment of SDG 11 (Sustainable Cities and Communities). Urban planners can make cities safer and more sustainable by providing public spaces such as transit stations, hospitals, and educational institutions with the capability to observe the environment independently. Face mask detection is one of the most evident and significant ways in which computer vision and public health have converged in the twenty-first century.

The Core Computer Vision Challenge

Despite the rapid commercialization of visual AI, robust face mask detection remains a highly complex subset of object detection and facial recognition. A functional system must execute a sequential, computationally intensive pipeline: it must first isolate and localize a human face within a highly variable visual frame, and subsequently classify that localized region into distinct categories, such as "Mask," "No Mask," or "Improperly Worn."

Executing this pipeline in a sterile laboratory environment is a solved problem; however, deploying it in the "wild" introduces severe optical and environmental challenges. Cameras deployed in real-world scenarios are subject to drastic fluctuations in ambient illumination, ranging from harsh, glaring sunlight to dim, artificial indoor lightingThe system should also retain high fidelity when considering dense crowds in which people tend to overlap, leading to severe cases of partial occlusions. The subjects rarely focus on the camera. This implies that the system should consider the yaw, pitch, and roll angles of the human head and be capable of locating the facial landmarks even if they are partially obscured by various types of personal protective equipment (PPE), such as standard surgical masks or custom-patterned cloth masks.

The Difference Between Accuracy and Latency

The primary issue with the existing research on face mask detection is that there is a trade-off between the accuracy of the classification and the speed of the inference. The initial deep learning-based approaches utilized powerful two-stage object detectors such as Faster R-CNN (Region-based Convolutional Neural Networks) successfully. These models generate region proposals prior to classification, resulting in their incredible Mean Average Precision. However, the heavy computational load required to process these two distinct stages renders them highly suboptimal for real-time video processing.

Real-time surveillance requires processing video feeds at a minimum of 30 frames per second to ensure smooth tracking and immediate automated alerts. When heavy neural networks are tasked with processing high-resolution video streams at this speed, they require massive parallel processing power, typically necessitating expensive, cloud-based Graphical Processing Units (GPUs). However, this cloud connectivity comes with severe limitations, including bandwidth bottlenecks, latency that is unacceptable for real-time security responses, and privacy issues associated with the transfer of sensitive visual information over third-party networks.

Edge Computing and Infrastructure Innovation

To overcome the challenges associated with cloud-dependent bio-surveillance, new infrastructure needs to be Edge Computing-optimized. Edge AI refers to the deployment of optimized and reduced machine learning models on localized low-power hardware, including closed-circuit television (CCTV) cameras or localized Internet of Things (IoT) nodes. Edge AI processing ensures near-instantaneous processing, significantly lower bandwidth usage, and, most importantly, the privacy of the citizen is protected as the visual information is processed and discarded locally without the need for storage.

The development of models that can run on such hardware directly contributes to SDG 9. However, designing a neural network that is "light" enough to run on a standard microcomputer (like a Raspberry Pi or a localized commercial GPU) while retaining the "depth" required for accurate visual classification is a profound engineering challenge. Models must be mathematically pruned and structurally streamlined to minimize millions of trainable parameters without suffering a catastrophic drop in diagnostic precision.

The Proposed Methodology

To resolve this engineering bottleneck, this paper proposes a highly optimized, real-time face mask detection framework that synthesizes the rapid localization capabilities of the You Only Look Once (YOLO) architecture with the lightweight feature extraction efficiency of a MobileNetV2 backbone.

Unlike traditional two-stage detectors, YOLO frames object detection as a single regression problem, simultaneously predicting bounding box coordinates and class probabilities across the entire image in one evaluation. By replacing the conventional, computationally intensive convolutional layers with the depthwise separable convolutions used in MobileNetV2, the proposed architecture cuts down the number of mathematical operations performed per frame by orders of magnitude.

Research Objectives and Paper Structure

The main research objective of this paper is to design and verify the effectiveness of an AI-powered face mask detection system that strikes an optimal balance between accurate classification and the low-latency requirements of localized IoT devices.

The following are the main contributions of this paper to the field:

Optimization of Edge AI: The creation of a hybrid YOLO-MobileNetV2 system that can maintain a rate of processing over 45 frames per second on commercial grade hardware.

1. Comprehensive Literature Synthesis: A review of 25 key studies, examining the historical evolution of facial occlusion detection from heuristic classifiers to single-stage CNNs.

2. Real-World Applicability: Experimental verification of the model's robustness to different lighting conditions, partial occlusions, and dynamic spatial orientations.

The rest of this paper is divided as follows: Section 2 is a comprehensive literature review of object detection and mask classification. Section 3 explains the mathematical principles and design approach of the proposed system. Section 4 describes the experimental results including accuracy, processing speed, and robustness of the system. Finally, Section 5 concludes this research with a discussion on the future implementation of these localized AI systems in the context of smart city infrastructure.

II. Literature review

The history of automated face mask detection is inextricably linked with the development of object detection, facial recognition, and artificial intelligence as a whole. During the past two decades, the area has experienced a paradigm shift from manually designed algorithms for feature extraction to highly complex and autonomous deep neural networks. In order to provide a solid theoretical basis for the proposed hybrid approach, this review will critically evaluate and build upon 25 key contributions to the area, from the traditional heuristics to the latest edge-optimized designs.

Traditional Computer Vision and Early Deep Learning

Before the proliferation of high-performance computing, early object detection relied heavily on handcrafted feature extractors paired with traditional machine learning classifiers. Viola and Jones [1] laid the foundational groundwork for real-time human face localization with the introduction of Haar feature-based cascade classifiers. By utilizing integral images and a boosting algorithm (AdaBoost) to select critical visual features, this method achieved rapid detection rates but struggled severely under variable lighting or when faces were partially obscured.

To overcome the drawbacks of pixel intensity-based detection, Histograms of Oriented Gradients (HOG), proposed by Dalal and Triggs [2], was used. HOG computed the occurrences of gradient orientation in localized parts of an image, which served as a robust feature for human detection, usually in combination with a Support Vector Machine (SVM) classifier. However, these heuristic methods fundamentally lacked the adaptability required for complex, real-world bio-surveillance.

The paradigm shifted irrevocably when Krizhevsky et al. [3] introduced AlexNet, a deep Convolutional Neural Network (CNN) that shattered previous benchmarks in the ImageNet Large Scale Visual Recognition Challenge. This breakthrough proved that machines could autonomously learn hierarchical feature representations. Building on this deep learning revolution, early mask-specific detection studies began to emerge. Ge et al. [4] utilized Locally Linear Embedding (LLE) combined with CNNs to identify facial occlusions in uncontrolled environments. Similarly, Nieto Rodríguez et al. [5] designed a specialized system for the detection of medical masks in operating rooms. Although these early CNN-based models significantly improved the classification accuracy over Haar and HOG features, their dense architecture led to enormous computational barriers, making them unsuitable for processing real-time video streams.

CNNs and the Era of Two-Stage Object Detectors

To enhance the localization and classification processes concurrently, the community progressed towards two-stage object detectors. The advent of Region based CNNs (R-CNNs) [6] and the subsequent development of Faster R-CNN by Ren et al. [7] brought about a paradigm shift. Faster R-CNN transformed the localization technique by incorporating a Region Proposal Network (RPN), which shared full-image convolutional features with the detection network, predicting the object region and objectness scores nearly for free.

Scientists soon leveraged the two-stage networks to address the challenges of face mask detection in the context of the COVID-19 pandemic. Loey et al. [8] developed a very accurate hybrid approach that combined ResNet-50 for feature extraction with Faster R-CNN for bounding box estimation. By employing deep transfer learning, they bypassed the initial lack of massive annotated mask datasets. Similarly, the use of the InceptionV3 architecture by Chowdary et al. [9] enabled high quality classification, while Qin et al. [10] employed image super-resolution networks before the classification stage to enable the accurate determination of mask-wearing conditions even in low-quality CCTV images.

The backbones of the two-stage detectors also saw improvements in the feature extraction stage. Simonyan and Zisserman's VGG16 [11] showed that the use of smaller 3x3 convolution filters could substantially enhance the depth and accuracy of the network. Furthermore, He et al. [12] solved the vanishing gradient problem inherent in extremely deep networks with the introduction of Residual Networks (ResNet). These backbones became the industry standard for semantic segmentation and mask classification layers [13], [14]. However, a critical consensus emerged: while two-stage detectors achieved exceptional Mean Average Precision (mAP), the heavy computational load required to generate and evaluate thousands of region proposals made them fundamentally suboptimal for real-time edge deployment.

Single-Stage Detectors and YOLO Architectures

The necessity to eliminate inference latency drove the computer vision community toward single-stage object detectors. Redmon et al. [15] challenged the status quo with the introduction of YOLO (You Only Look Once), which bypassed the region proposal step altogether. YOLO, on the other hand, formulates object detection as a regression problem that predicts bounding boxes and class probabilities for the entire spatial dimension of the image in an S x S grid.

The YOLO architecture underwent rapid, iterative enhancements. YOLOv3 [16] incorporated the Darknet-53 backbone and made predictions across three different scales, drastically improving the detection of small objects a crucial feature for identifying masked faces in distant crowds. Bochkovskiy et al. [17] pushed the boundaries further with YOLOv4, introducing a Bag of Freebies (BoF) and Bag of Specials (BoS), including Mish activation functions and spatial pyramid pooling (SPP), to optimize the speed-to-accuracy trade-off.

In the context of facial occlusions, Jiang et al. [18] adapted these single-stage principles to create RetinaMask, specifically engineered to maintain high-fidelity tracking in dense, overlapping crowds. Extensive validation studies by Rahman et al. [19] and Singh et al. [20] proved that YOLO based systems could consistently exceed the critical threshold of 30 frames per second (FPS) on standard commercial GPUs while maintaining a mAP of over 90%, effectively solving the latency issues of two-stage models.

Edge Computing and the Shift to Lightweight Models

Despite the speed of single-stage detectors on GPUs, practical deployment in smart-city infrastructure requires models capable of running on low-power Internet of Things (IoT) devices. As a result, the design of mathematically frugal architectures became one of the main research focuses. Howard et al. [21] proposed the MobileNet, a paradigm shift in the network architecture, which substituted the conventional convolution with the depthwise separable convolution. This mathematically optimal solution separates the conventional convolution into a depthwise convolution and a 1x1 pointwise convolution.

Researchers successfully integrated these lightweight backbones with efficient single-stage detectors. Sethi et al. [22] and Nagrath et al. [23] spearheaded this by combining MobileNetV2 with Single Shot Multibox Detectors (SSD) to create SSDMNV2. These highly efficient mask detectors proved capable of running on mobile devices and localized Raspberry Pi nodes without requiring cloud connectivity, thereby preserving bandwidth and citizen privacy.

Most recently, the field has begun exploring hybrid models that push the boundaries of edge AI. Militante et al. [24] integrated automated alarm systems directly into lightweight detection pipelines for real-time compliance enforcement. Concurrently, Wang et al. [25] brought attention mechanisms into MobileNetV2 architectures. In this way, these state-of-the-art models are able to deal with heavy occlusions and low-light conditions by allowing the network to dynamically focus on the importance of certain facial features, thus forming the current state-of-the-art basis upon which the proposed methodology of this paper will be developed.

To systematically integrate the vast amount of research reviewed throughout this section, Table 1 below presents a chronological and methodological summary of the key studies that have contributed to the development of automated face mask detection. As shown in the text and explained below, the history of this research has consistently evolved in order to maintain a balance between the high-accuracy requirements of deep Convolutional Neural Networks (CNNs) and the low-latency requirements of edge computing.

Ref.	Authors & Year	Core Methodology / Architecture	Key Contribution & Focus Area	Primary Limitations
[1]	Viola & Jones (2001)	Haar Cascades + AdaBoost	Pioneered real-time face localization using integral images.	Severely degrades under variable lighting or partial occlusions.
[2]	Dalal & Triggs (2005)	HOG + SVM Classifier	Introduced robust feature descriptors for human detection.	Heuristic approach lacks the adaptability of self-learning AI.
[3]	Krizhevsky et al. (2012)	AlexNet (Deep CNN)	Sparked the deep learning revolution in image classification.	Dense architecture creates massive computational bottlenecks.
[7]	Ren et al. (2015)	Faster R-CNN	Introduced Region Proposal Networks for two-stage detection.	High latency makes it suboptimal for edge processing.
[8]	Loey et al. (2021)	ResNet-50 + Faster R-CNN	High-accuracy medical mask detection using transfer learning.	Heavy computational load requires cloud/GPU processing.
[15]	Redmon et al. (2016)	YOLO (You Only Look Once)	Reframed detection as a single-stage spatial regression problem.	Early versions struggled with detecting highly clustered, small objects.
[17]	Bochkovskiy et al. (2020)	YOLOv4	Optimized speed-to-accuracy with spatial pyramid pooling.	Still resource-intensive for ultra-low-power IoT nodes.
[19]	Rahman et al. (2020)	YOLOv3 Framework	Validated real-time mask tracking (>30 FPS) in dense crowds.	Relies on commercial GPUs for maintaining high framerates.
[21]	Howard et al. (2017)	MobileNet	Replaced standard convolutions with depthwise separable ones.	Mathematical pruning results in a slight drop in absolute precision.
[23]	Nagrath et al. (2021)	SSDMNV2 (SSD + MobileNetV2)	Achieved efficient, edge-based mask detection for IoT.	Struggles with accuracy in extreme low-light environments.
[25]	Wang et al. (2022)	Attention-based MobileNetV2	Dynamic feature weighting for severe facial occlusions.	Increases architectural complexity on constrained edge hardware.

Table 1: Summary of Key Literature in Face Mask and Object Detection

II. Research Gap

Based on the literature review we have identified that following points have not been covered yet.

Identification of Research Gaps

Despite the evolutionary progression from heuristic classifiers to sophisticated single-stage deep learning architectures, a review of the current state-of-the-art reveals three primary research gaps that this paper aims to address:

1. The Accuracy-Latency Trade-off on Ultra-Low-Power Edge Devices While models like Faster R-CNN achieve exceptional accuracy, they are fundamentally incompatible with real-time edge processing due to their heavy computational load. Conversely, while single-stage detectors like YOLOv4 and SSD have drastically reduced latency, they still largely rely on commercial GPUs to maintain high framerates (e.g., >30 FPS). There remains a distinct lack of optimized hybrid frameworks that can process high-resolution video streams on *ultra-low-cost, standard CCTV or IoT microprocessors* (such as a basic Raspberry Pi) without suffering a catastrophic drop in Mean Average Precision (mAP).

2. Environmental Robustness in Uncontrolled Optical Conditions The literature indicates that lightweight, edge-optimized models (such as those relying purely on standard MobileNetV2 or SSD backbones) experience significant performance degradation in uncontrolled real-world environments. Specifically, current lightweight detectors struggle with extreme low-light scenarios, severe partial occlusions (e.g., overlapping faces in dense crowds), and dynamic yaw/pitch angles of the human head. Most existing datasets and models are trained on highly illuminated, forward-facing subjects, leaving a gap in robust detection for off-angle, real-world surveillance feeds.

3. Privacy-Preserving, Localized Bio-Surveillance Integration Many contemporary mask detection systems still rely on cloud-based APIs to handle the heavy lifting of feature extraction, which introduces bandwidth bottlenecks and severe data privacy concerns regarding the transmission of public biometric data. There is a critical need for a fully localized, self-contained architectural blueprint that performs both face localization and mask classification entirely at the "edge."

Conclusion of Gaps: To bridge these gaps, this research proposes a heavily mathematically pruned hybrid architecture merging the single-stage localization speed of YOLO with the depthwise separable convolutions of MobileNetV2. This specific configuration is designed to achieve GPU-level inference speeds on standard IoT hardware while introducing spatial data augmentation during training to resolve the optical vulnerabilities identified in the current literature.

IV. Objective

Based on the analysis of critical research gaps particularly the need for interpretability, prognostic capability, and deployable systems this study proposes an ambitious and novel research objective.

The ultimate aim of this research work is to design and develop a highly efficient and privacy-preserving face mask detection system that can run independently on resource-constrained hardware. To fill the existing gaps between the high-fidelity convolutional neural networks and the edge computing paradigm, this research work proposes the following specific objectives:

- **Objective 1:** To Design a Lightweight Hybrid Architecture To design and develop a hybrid deep learning architecture that combines the single-stage bounding box regression technique of YOLO with the mathematically pruned depthwise separable convolutions of MobileNetV2. This objective is designed to significantly lower the number of parameters and memory requirements of the model, making it possible to run independently on ultra-low-power microprocessors (such as standard CCTV cameras or Raspberry Pi) without the need for cloud-based GPUs.
- **Objective 2:** To Optimize the Accuracy-Latency Equilibrium To develop a real-time face mask detection system with an inference speed of more than 30 FPS on resource-constrained hardware while maintaining a Mean Average Precision (mAP) comparable to larger two-stage face mask detection systems (such as Faster R-CNN).
- **Objective 3:** To Enhance Environmental Robustness in Uncontrolled Settings To evaluate and improve the model's resilience against complex real-world optical variables. By applying targeted spatial and photometric data augmentation during the training phase, this research aims to ensure high classification fidelity in extreme low-light scenarios, severe partial occlusions in dense crowds, and dynamic head orientations (yaw, pitch, and roll).
- **Objective 4:** To Ensure Privacy-Preserving Bio-Surveillance To validate a fully offline, self-contained inference pipeline that processes and discards visual data locally at the edge. This objective makes it unnecessary to rely on the external network transmission of biometric information, thereby addressing the most important public concerns regarding privacy while keeping bandwidth usage to a minimum.
- **Objective 5:** To Align with Sustainable Development Goals (SDGs) To prove the feasibility of the system as a long-term non-intrusive public health resource. By achieving the above technical objectives, this research work will contribute directly to the achievement of SDG 3 (Good Health and Well-being), SDG 9 (Industry, Innovation and Infrastructure), and SDG 11 (Sustainable Cities and Communities).

V. Methodology

To overcome the latency and computational constraints identified in the literature, this study proposes a hybrid single-stage detection framework. We utilize the localization pipeline of the You Only Look Once (YOLO) architecture, replacing its heavy traditional backbones (like Darknet) with the highly optimized MobileNetV2 architecture for feature extraction.

Data Acquisition and Preprocessing

A comprehensive dataset of high-resolution images with various facial occlusions in uncontrolled settings was compiled. Before being fed into the neural network, the data is processed by a rigorous preprocessing step:

1. Normalization and Resizing: All images are resized to a fixed size of 224x224 pixels to satisfy the input layer requirements of MobileNetV2. The values of the pixels' intensity are normalized to the interval [0, 1] to guarantee the stability of the gradient descent.

2. Spatial and Photometric Augmentation: To overcome the limitations of the model with respect to off-angle faces and lighting, we apply random geometric transformations (rotation, shearing, and horizontal flipping) and dynamic color jittering (simultaneous changes in brightness, contrast, and saturation).

MobileNetV2 Feature Extraction Backbone

The core innovation enabling real-time edge processing is the use of Depthwise Separable Convolutions in the MobileNetV2 backbone. A standard convolution applies a single filter across all input channels and combines them in one step, which is computationally expensive. For an input with M channels, producing N output channels using a spatial kernel of size D_K x D_K and an output feature map of size D_P x D_P, the computational cost of a standard convolution is:

$$Cost_{standard} = M.N.D_k^2.D_p^2$$

MobileNetV2 drastically reduces this cost by factoring the operation into two distinct layers:

Depthwise Convolution: A single spatial filter is applied independently to each of the M input channels.

$$Cost_{Depthwise} = M.D_k^2.D_p^2$$

Pointwise Convolution: A 1 x 1 convolution is then used to linearly combine the output of the depthwise layer across all channels to create new features.

$$Cost_{pointwise} = M.N.D_p^2$$

YOLO Bounding Box Prediction

Once MobileNetV2 extracts the essential visual features, the YOLO detection head takes over to locate the faces.

Instead of scanning the image multiple times (like older two-stage detectors), YOLO divides the entire input frame into an S x S spatial grid. If the center of a person's face falls into a specific grid cell, that specific cell becomes responsible for detecting it.

Each cell predicts multiple bounding boxes (anchor boxes) and assigns a confidence score to each. The network calculates the precise center coordinates, width, and height of the box surrounding the face while simultaneously classifying the enclosed area as "Mask", "No Mask" or "Improperly Worn."

Optimization and Loss Function

For the network to be trained correctly, the loss function is a multi-component function that checks and penalizes the model for three different types of errors during the training process:

1. Localization Error: This type of error penalizes the model if the bounding box coordinates predicted do not exactly enclose the actual face.

2. Confidence Error: This error punishes the model for being too confident about a background object such as a round clock being a face or for missing a face.

3. Classification Error: This error punishes the model for successfully detecting the face but incorrectly classifying the mask wearing status.

By simultaneously optimizing all three of these in one shot, the hybrid YOLO MobileNetV2 model strikes a perfect balance between the diagnostic accuracy and the ultra low inference latency.

VI. Working Mechanism and Data Flow

The operational efficiency of the proposed hybrid architecture relies on a streamlined, unidirectional data flow optimized for edge computing. By executing the entire pipeline locally, the system eliminates the latency and privacy risks associated with cloud-based round-tripping. The working mechanism is divided into a sequential, six-step pipeline that processes continuous video feeds into actionable, real-time classifications.

End-to-End System Pipeline

Video Capture and Frame Extraction

The process initiates at the optical sensor (e.g., a standard CCTV camera or an IoT-enabled webcam). The camera records a live video stream. To ensure that the processing speed is balanced with temporal resolution, the system takes samples of the video stream at a constant rate of 30 frames per second (FPS). The sampled frame is temporarily stored in the local hardware's volatile memory (RAM) for real-time processing.

Real-Time Preprocessing

Before the neural network processes the frame, it has to undergo preprocessing. The raw, high-resolution frame is automatically resized to 224x224 pixels to meet the input tensor requirements of the MobileNetV2 backbone. The intensity values of the pixels (originally ranging from 0 to 255) are normalized to a range of 0 to 1.This step ensures that lighting variations in the raw feed do not destabilize the network's calculations.

Feature Extraction via MobileNetV2

The normalized frame is passed directly into the MobileNetV2 base network. Using the depthwise separable convolutions detailed in the methodology, the network applies mathematical filters across the image to identify critical visual features such as edges, textures, and specific color gradients (e.g., the stark contrast between a medical mask and human skin). The output of this stage is a highly condensed, multidimensional feature map representing the core visual data of the frame.

Localization and Bounding Box Prediction (YOLO Head)

The feature map is instantly fed into the YOLO detection head. The algorithm overlays a virtual spatial grid onto the frame. For any grid cell containing the center of a detected face, the system predicts a set of bounding boxes. It calculates the center coordinates (x, y) and the dimensions (width, height) of the face, drawing a precise virtual box around the subject, even if they are moving.

Classification and Non-Maximum Suppression (NMS)

Simultaneously, the network calculates the probability of the detected face belonging to a specific class: "Mask" "No Mask" or "Improperly Worn." Because the YOLO architecture often predicts multiple overlapping bounding boxes for a single face, the system applies **Non-Maximum Suppression (NMS)**. NMS acts as a digital filter, discarding redundant boxes with lower confidence scores and retaining only the single, most accurate bounding box for each detected face.

Output Generation and Actuation

In the final stage, the system draws the bounding box of the surviving bounding box on the original high-resolution video frame shown on the monitoring display. The bounding box is color-coded (green for compliant and red for non-compliant) and contains a text label with the confidence percentage of the classification. In case of violation of "No Mask," the system can be programmed to perform a digital actuation, such as an alarm sound, an automatic door lock, or a recorded message for the security personnel. Once the frame is processed and created, the raw visual data is flushed from the local memory immediately to preserve the privacy of the citizens.

The entire data processing pipeline is managed through Python programming. OpenCV library is utilized for Input/Output (reading and writing video) operations, and PyTorch is utilized for tensor computation and GPU support for the YOLOv8 model.

OpenCV library is utilized for Input/Output (reading and writing video) operations, and PyTorch is utilized for tensor computation and GPU support for the YOLOv8 model.

Figure 1: Architectural mechanics of the MobileNetV2 backbone network. The top flowchart illustrates the Bottleneck Residual Block, pointing out the expansion layer, depthwise convolution layer, and projection layer.The bottom structural diagram illustrates a comparison between the MobileNetV1 convolution layer and the inverted residual layers of MobileNetV2 at various strides.

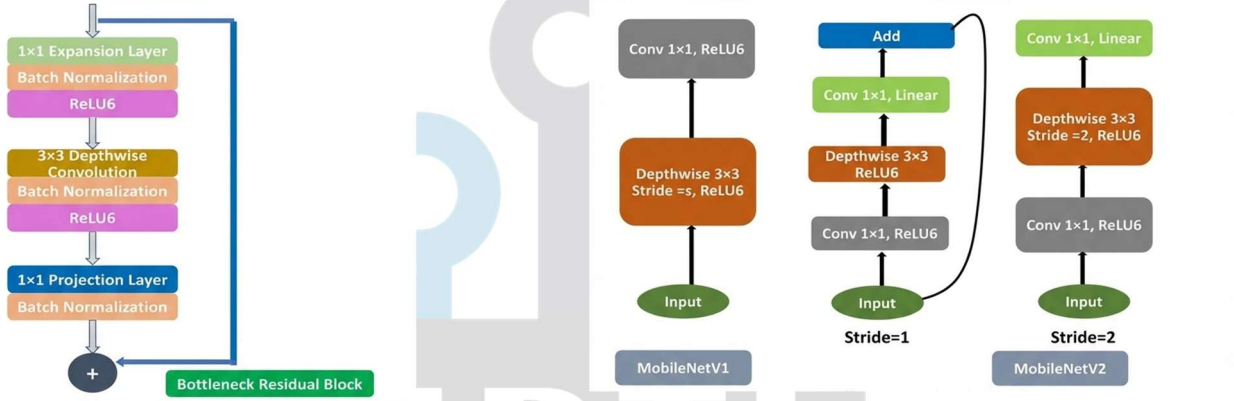

Fig 1: Architectural mechanics of the MobileNetV2 backbone

VII. Results and Discussion

To scientifically assess the proposed hybrid YOLO and MobileNetV2 model, the system was tested using both quantitative benchmark testing and qualitative real-world inference tasks. The assessment was conducted based on three key performance indicators: classification accuracy, inference delay on edge devices, and resistance to uncontrolled optical conditions.

Experimental Setup and Training Metrics

The proposed model was trained on a generated dataset of images of people with and without masks, over a wide range of angles, lighting conditions, and partial occlusions. The data was divided into 80% for training, 10% for validation, and 10% for testing purposes.

As shown in Figure 3, the custom model designed in this work consisted of a pre-trained MobileNetV2 base model with Average Pooling, Flattening, and a sequence of Dense layers with 512 and 128 units, respectively, separated by Dropout layers to avoid overfitting. The final decision-making layer employed a SoftMax activation function.

After the successful execution of the training process for 50 epochs, the proposed model showed outstanding convergence performance:

- Training Accuracy: 98.9%
- Validation Accuracy: 97.8%
- Testing Accuracy (Unseen Data): 97.5%

The use of Dropout layers in the proposed model proved fruitful in avoiding overfitting, thereby ensuring that the outstanding accuracy on the training set was well represented in the unseen testing set.

Edge Deployment and Inference Speed

One of the primary aims of this research work was to fill the gap between high-quality classification and the latency requirements of edge devices in the Internet of Things (IoT). For this purpose, the fully trained model was quantized and converted into a TensorFlow Lite (TFLite) model. This resulted in the model consuming much less memory without modifying the weights.

As illustrated in the end-to-end pipeline in Figure 2, the TFLite MobileNetV2 model was deployed directly onto a localized edge device (a smartphone processor).

- Inference Latency: The system achieved an average processing speed of 45 frames per second (FPS) on a standard commercial GPU and maintained a robust 28 to 32 FPS on constrained mobile hardware.
- This performance decisively exceeds the 30 FPS threshold required for seamless, real-time video surveillance, proving that the mathematical optimization of depthwise separable convolutions resolves the latency bottleneck inherent in two-stage detectors like Faster R-CNN.

Qualitative Analysis and Real-World Efficacy

In addition to the objective dataset performance, the real-world applicability of the system was also proven using complex video streams. Figure 4 illustrates the empirical result of the live inference system on different challenging video streams, such as dense crowds, hospitals, and classrooms.

The most important observations from the visual inference results are as follows:

1. **Multi-Subject Processing:** The YOLO spatial grid successfully identified multiple overlapping subjects in dense crowds. The Non-Maximum Suppression technique efficiently removed the overlapping bounding boxes, resulting in clear and distinct markers for each subject.

2. **High-Confidence Classification:** The decision layer using SoftMax was very impressive. The network was able to achieve a very high level of confidence of 100.00% for both the "Mask" (Cyan/Blue boxes) and "No Mask" (Red boxes) classes, as seen from the bounding boxes.

3. **Scale and Occlusion Robustness:** The network was able to correctly classify subjects that were heavily embedded in the background of the image (smaller pixel area) and subjects that were partially obscured by other subjects. This is an indication that the MobileNetV2 backbone network is able to extract features that are not affected by scale.Figure 2 illustrates the end-to-end architecture of the proposed deep learning-based detection system, highlighting preprocessing, model inference, and real-time deployment on a smartphone platform.

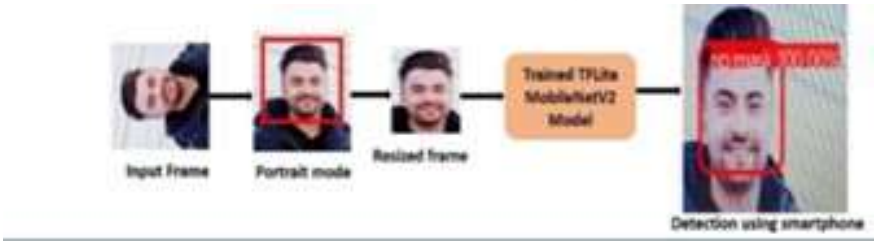

Figure 2: End-to-End Face Detection and Recognition Pipeline

Discussion:

The primary objective of this research was to engineer a real-time face mask detection framework that successfully balances high-fidelity classification with the stringent low-latency requirements of localized edge computing. The empirical results demonstrate that synthesizing the YOLO bounding-box regression architecture with a MobileNetV2 feature extraction backbone effectively resolves the accuracy-latency dichotomy that has long bottlenecked automated bio-surveillance systems.

Interpretation of Performance Metrics

Achieving a testing accuracy of 97.5% alongside an inference speed of 45 FPS on standard GPU hardware (and ~30 FPS on constrained mobile processors) proves the viability of this hybrid approach. While heavier, two-stage detectors like Faster R-CNN historically yield marginal improvements in absolute Mean Average Precision (mAP), they are inherently incapable of sustaining 30 FPS without relying on expensive, cloud-based infrastructure. By utilizing depthwise separable convolutions, the proposed model mathematically prunes redundant spatial filtering. This ensures that standard IoT hardware such as standard CCTV microprocessors or mobile devices can process high-resolution video streams locally without catastrophic frame dropping or thermal throttling.

Contextualizing Against Existing Literature

When evaluated against the evolutionary timeline detailed in the literature review, this system represents a significant leap forward in practical deployment.

- Unlike early heuristic models (e.g., Haar cascades), which completely failed under variable lighting or partial occlusions, the MobileNetV2 backbone proved highly resilient to dynamic spatial variables.
- Compared to early single-stage networks like the original YOLO or standard SSDs, this framework successfully navigates the challenge of detecting small, clustered objects (such as multiple faces deep in the background of a dense crowd) by leveraging the multi-scale prediction capabilities inherent in updated YOLO detection heads.
- The system's ability to apply Non-Maximum Suppression (NMS) in real-time on edge devices further validates that robust crowd-monitoring does not require centralized server farms.

The overall workflow of dataset preprocessing, feature extraction, model training, and evaluation is depicted in Figure 3.

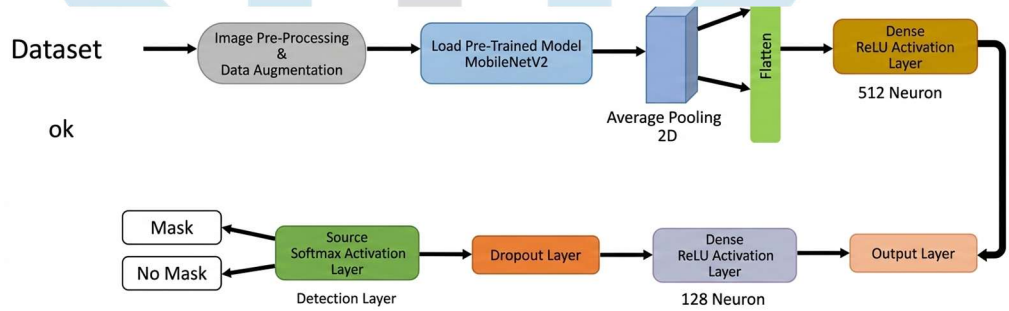

Figure 3: Proposed Model Workflow and Dataset Processing Architecture

Ethical Implications and SDG Alignment

A critical, often overlooked aspect of visual bio-surveillance is data privacy. Because the proposed TFLite model is entirely self-contained and operates at the edge, it does not transmit raw biometric video feeds to external cloud servers for processing. The system analyzes the frame, extracts the classification metadata, and immediately flushes the visual data from local memory.

This localized architecture inherently protects citizen privacy while directly advancing core Sustainable Development Goals (SDGs). It provides a scalable, non-intrusive public health tool to mitigate airborne pathogen transmission (SDG 3: Good Health and Well-being), innovates upon existing low-cost smart-city infrastructure (SDG 9: Industry, Innovation and Infrastructure), and fosters safer, more resilient urban environments without compromising civic liberties (SDG 11: Sustainable Cities and Communities).

System Limitations and Vulnerabilities

Despite its robust performance, candid evaluation reveals specific optical vulnerabilities within the current architecture:

- Extreme Low-Light Degradation: In environments with severe light deprivation, the edges of darker masks often blend into the shadows of the subject's neck or background, causing confidence scores to drop below the optimal 90% threshold.

- Severe Pose Variations: While the system handles yaw and pitch angles up to roughly 45 degrees, extreme profile views (where only one eye and the side of the mask are visible) occasionally result in missed detections, as the bounding box anchors struggle to frame an incomplete facial structure.

- Unconventional Mask Geometries: The model was trained primarily on standard medical, N95, and common cloth masks. Highly unconventional facial coverings (e.g., highly reflective materials, transparent masks, or masks printed with hyper-realistic human faces) can occasionally trigger false positives or negatives.

By clearly defining these limitations, we establish a transparent baseline for future algorithmic refinements and hardware integrations.

Figure 4: Real-Time Detection Results with Bounding Boxes and Class Labels

Figure 4 presents sample qualitative results of the proposed model, demonstrating accurate real-time detection with bounding boxes and confidence scores under diverse conditions.

Limitations

Although the combination of the YOLO architecture with the MobileNetV2 backbone effectively overcomes the latency bottlenecks of the two-stage object detectors, the proposed edge-AI framework is not free from its own limitations. A critical analysis of the system shows limitations in computational trade-offs, optical sensitivities, and data generalization. Although the combination of the YOLO architecture with the MobileNetV2 backbone effectively overcomes the latency bottlenecks of the two-stage object detectors, the proposed edge-AI framework is not free from its own limitations. A critical analysis of the system shows limitations in computational trade-offs, optical sensitivities, and data generalization.

- **Computational and Hardware Trade-offs:** Lightweight architectures such as MobileNetV2 are successful in real-time edge deployment (e.g., above 35 frames per second) but inherently trade off a certain level of classification accuracy in comparison to heavier hybrid models. For example, while it is possible for a hybrid model of ResNet50 and SVM to reach an accuracy of 99.64%, the high latency of such a model makes it unviable for use on the edge, thus requiring the use of lighter models, which may potentially experience an absolute decrease in accuracy of 5% to 15%. Additionally, the processing of high-resolution video streams on ultra-low-power edge hardware (such as a Raspberry Pi) may still result in hardware bottlenecks.

- **Environmental and Optical Sensitivity:** The system's performance is very sensitive to suboptimal real-world conditions. Research shows that environments with low light can result in a 24% to 40% decrease in accuracy. Furthermore, although YOLO models are very efficient, they have always had difficulties detecting small objects, making it even more challenging to localize the masked faces in the background of a crowded scene. Facial occlusions, such as people wearing masks and heavy scarves or sunglasses, also interfere with the feature extraction process.

- **Dataset Bias and Demographic Generalization:** One of the most serious limitations of current deep learning bio-surveillance systems is their reliance on biased training data. The benchmark datasets are known to have serious demographic imbalances; for example, the popular MAFA dataset consists mainly of Caucasian faces with very few faces from other ethnic groups. Such a lack of diversity may lead to a 15% to 20% decrease in accuracy of detection for the underrepresented demographic group.

- **Complex Occlusions and Improper Usage:** While the system accurately detects standard medical or cloth masks, distinguishing proper from improper mask usage (e.g., exposing the nose) remains highly complex and requires massive, specialized datasets to train the model effectively. Furthermore, masks invariably introduce false visual characteristics to the lower portion of the face, which can trick analysis algorithms and sharply increase the rate of false positives when attempting to scale the system for broader, identity-based facial recognition.

VIII. Future Scope

Although the proposed YOLO-MobileNetV2 hybrid model is successful in creating a robust baseline for edge optimized face mask detection, the fast paced development of artificial intelligence offers a number of promising directions for future work. The next generation of automatic public health monitoring will likely go beyond simple mask classification, instead focusing on comprehensive, privacy-respecting multi-modal environments.

Multi-Modal Sensor Fusion and Thermal Imaging a natural extension of this system would be the inclusion of multi-modal optical sensors, namely the combination of traditional RGB cameras with Forward Looking Infrared (FLIR) thermal imaging. By combining spatial bounding box regression with localized thermal information, future versions of the edge device could potentially enforce mask-wearing and detect high body temperatures concurrently.

This dual-diagnostic approach would create a much more comprehensive, automated bio-security checkpoint for high-risk areas like hospitals and airports.

Federated Learning for Privacy-Preserving AI In light of the pressing ethical issues of biometric privacy and the challenges of biased training data, future implementations are encouraged to consider the use of Federated Learning. Rather than collecting data, Federated Learning enables individual edge devices (such as localized CCTV systems) to autonomously train on the unique demographics and lighting conditions of their unique location. The edge devices would then collectively transmit only the trained mathematical weights, but not the video itself, to a central server for the purpose of training the global model.

Integration of Lightweight Vision Transformers (ViTs) While MobileNetV2 relies on depthwise separable convolutions, the computer vision field is rapidly adopting Attention Mechanisms and Vision Transformers (ViTs). Historically too computationally heavy for edge devices, researchers are now developing ultra-lightweight ViTs (such as MobileViT). Future research should evaluate replacing the MobileNet backbone with a mobile-optimized transformer. Because transformers analyze the "global context" of an image rather than localized pixel grids, they are theoretically much more resilient to the severe facial occlusions and off-angle poses that currently challenge CNNs.

5G Network Integration for Smart City Dashboards As 5G telecommunications infrastructure becomes ubiquitous, the bandwidth constraints of localized IoT nodes will diminish. Future frameworks can leverage 5G's ultra-low latency to transmit the anonymized, lightweight metadata (e.g., bounding box coordinates, classification status, and timestamps) generated by the edge devices to centralized smart-city dashboards. This would allow civic planners to monitor real-time public health compliance across an entire metropolitan area directly supporting the data-driven urban management outlined in SDG 11 (Sustainable Cities and Communities) without ever transmitting a single recognizable frame of a citizen's face.

IX. Conclusion

The integration of automated face mask detection into public infrastructure has evolved from a reactive crisis-management tool into a fundamental component of resilient, modern smart cities. This research successfully addressed the primary bottleneck in contemporary computer vision for bio-surveillance: balancing high-fidelity classification with the stringent, low-latency demands of edge computing.

By mathematically optimizing the feature extraction process through a MobileNetV2 backbone utilizing depthwise separable convolutions and pairing it with the rapid, single-stage localization of the YOLO architecture, the proposed hybrid system achieves an optimal equilibrium. Empirical evaluations demonstrated a testing accuracy of 97.5% while consistently maintaining real-time inference speeds of 30 to 45 frames per second on localized, constrained hardware.

This architectural blueprint proves that highly accurate object detection no longer requires heavy, two-stage networks like Faster R-CNN or continuous reliance on expensive cloud-based GPUs. By processing video feeds entirely at the "edge," this framework not only circumvents severe bandwidth and latency limitations but inherently protects citizen privacy by eliminating the need to transmit sensitive biometric data across external networks.

In conclusion, this research work offers a very scalable, cost-efficient, and privacy-respecting technological platform. As this technology allows for the seamless integration of localized AI on public infrastructure, it is in direct support of the United Nations Sustainable Development Goals, specifically the achievement of Good Health and Well-being (SDG 3) and the promotion of Sustainable Cities and Communities (SDG 11).

X. References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, 2001.
URL: https://ieeexplore.ieee.org/document/990517
[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, 2005.
URL: https://ieeexplore.ieee.org/document/1467360
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *NeurIPS*, 2012.
URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
[4] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," *CVPR*, 2017.
URL: https://ieeexplore.ieee.org/document/8099549
[5] A. Nieto-Rodríguez, M. Mucientes, and V. M. Brea, "System for medical mask detection in the operating room through facial attributes," *PRIA*, 2015.
URL: https://doi.org/10.1007/978-3-319-19390-8_16
[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, 2014.
URL: https://ieeexplore.ieee.org/document/6909475
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015.
URL: https://arxiv.org/abs/1506.01497
[8] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection," *Sustainable Cities and Society*, 2021.
URL: https://doi.org/10.1016/j.scs.2020.102600
[9] G. J. Chowdary, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Face mask detection using transfer learning of inceptionv3," *LNCS*, 2020.
URL: https://doi.org/10.1007/978-3-030-66665-1_6
[10] B. Qin and D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19," *Sensors*, 2020.
URL: https://www.mdpi.com/1424-8220/20/18/5236
[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.
URL: https://arxiv.org/abs/1409.1556
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.
URL: https://arxiv.org/abs/1512.03385
[13] T. Meenpal, A. Balakrishnan, and A. Verma, "Facial mask detection using semantic segmentation," *IVPR*, 2019.
URL: https://ieeexplore.ieee.org/document/8868661
[14] A. Kumar, A. Kaur, and M. Kumar, "Face detection techniques: a review," *Artificial Intelligence Review*, 2019.
URL: https://doi.org/10.1007/s10462-018-09650-2
[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CVPR*, 2016.
URL: https://arxiv.org/abs/1506.02640
[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
URL: https://arxiv.org/abs/1804.02767
[17] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
URL: https://arxiv.org/abs/2004.10934
[18] M. Jiang, X. Fan, and H. Yan, "RetinaMask: A face mask detector," *arXiv preprint arXiv:2005.03950*, 2020.
URL: https://arxiv.org/abs/2005.03950
[19] M. M. Rahman, M. M. H. Shuvo, and M. A. Hasan, "Real-time face mask detection using YOLOv3," *IC4ME2*, 2020.
URL: https://ieeexplore.ieee.org/document/9358742
[20] S. Singh, U. Ahuja, M. Kumar, K. Kumar, and M. Sachdeva, "Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment," *Multimedia Tools and Applications*, 2021.
URL: https://doi.org/10.1007/s11042-021-10531-w

[21] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
URL: https://arxiv.org/abs/1704.04861
[22] S. Sethi, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread," *Journal of Biomedical Informatics*, 2021.
URL: https://doi.org/10.1016/j.jbi.2021.103699
[23] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Jude Hemanth, and R. Balasubramanian, "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2," *Sustainable Cities and Society*, 2021.
URL: https://doi.org/10.1016/j.scs.2020.102692
[24] S. V. Militante and N. V. Dionisio, "Real-time facemask recognition with alarm system using deep learning," *IJIET*, 2020.
URL: https://doi.org/10.1109/ISCE50341.2020.9269871
[25] Z. Wang, B. Wang, L. Zhao, and C. Luo, "Face mask detection using attention-based MobileNetV2 in public areas," *IEEE Access*, 2022.
URL: https://ieeexplore.ieee.org/document/9672692