iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79984

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Smart PPE Compliance Detection Using Vision Transformers and Edge AI: A 3-Class Real-Time Approach

Sahil Pankaj[1], Vivek Jha[2], Dr. Isharat Ali[3]
[1, 2]*Department of Data Science and Design, Greater Noida Institute of Technology Greater Noida, India*
[3]*Supervisor, Department of Data Science and Design, Greater Noida Institute of Technology Greater, Noida, India*

Abstract: ***Binary face-mask detection systems, which saw widespread adoption during the COVID-19 pandemic, are now largely insufficient for the demands of modern workplace safety monitoring. Industrial environments require real-time verification of multiple PPE categories simultaneously — not just a yes/no determination of whether a face covering is present. This paper presents a 3-class PPE compliance detection system that categorises each detected worker into one of three states: correctly wearing required PPE, wearing PPE incorrectly (below the nose, around the chin, or loosely fitted), and not wearing PPE at all. The proposed pipeline pairs a YOLOv8 detection head with a Vision Transformer (ViT-B/16) classification backbone pretrained on ImageNet-21K, fine-tuned on a curated dataset of 4,200 annotated images across the three compliance categories. Albumentations-based augmentation including mosaic, cutout, and histogram equalisation improves robustness under poor lighting. After 30 training epochs using the AdamW optimiser with cosine learning rate decay, the system achieves 99.3% test accuracy with a macro-F1 of 0.991. The fine-tuned model is subsequently quantised to INT8 TFLite format and deployed on a Raspberry Pi 4, achieving 12 FPS — sufficient for practical monitoring applications. An integrated Streamlit dashboard and Telegram bot deliver real-time compliance alerts. This work demonstrates that extending face-mask detection into a full PPE compliance framework, powered by transformer-based architectures and edge-optimised deployment, is both technically feasible and operationally practical for factories, hospitals, and construction sites.***

Keywords: Face Mask Detection, PPE Compliance Detection, Vision Transformer, ViT-B/16, YOLOv8, Edge AI, TFLite, 3-Class Classification, , Industrial Safety, Raspberry Pi Deployment

I. INTRODUCTION

Automated detection of personal protective equipment (PPE) using computer vision has evolved considerably since the early COVID-19 pandemic work. Between 2020 and 2022, dozens of research groups published mask-detection systems using convolutional neural networks — mostly MobileNetV2 or VGG-based classifiers trained on binary datasets of masked versus unmasked faces. While these systems served a clear and immediate public health need, they also introduced a framing problem that has quietly persisted into the post-pandemic period: they reduced PPE compliance to a single binary question, when the actual problem is far richer and more operationally complex. In practice, the most common PPE violations in industrial and healthcare settings are not outright refusals to wear protective equipment — they are cases of incorrectly worn gear. A surgical mask worn below the nose provides almost no protection but would be classified as 'mask present' by any binary detector. A hard hat tilted to the back of the head, gloves left dangling on a wrist, or safety goggles pushed up onto a worker's forehead represent the types of compliance failures that actually matter in occupational safety contexts, and none of them are captured by systems that only ask whether an item is present. This paper builds on the established foundation of transfer-learning-based PPE detection and extends it in three directions. First, the classification problem is expanded from binary to three classes: correct PPE wear, incorrect or improper wear, and complete absence of the required item. Second, the model architecture is updated from MobileNetV2-era CNNs to a Vision Transformer backbone (ViT-B/16), which has been shown to handle occlusion-heavy and partially obstructed scenes more robustly than convolutional networks due to its attention-based feature extraction. Third, the complete system is deployed on a Raspberry Pi 4 edge device using INT8 quantisation, producing real-time inference at 12 frames per second without any cloud dependency — and integrated with a Streamlit monitoring dashboard and Telegram alerting bot.

The remainder of this paper is structured as follows. Section 2 surveys the relevant literature, tracing the arc from early CNN-based mask detection through modern transformer and YOLO-based approaches. Section 3 details the proposed methodology, including dataset construction, augmentation
strategy, architecture design, and training setup. Section 4 covers the deployment pipeline on edge hardware. Section 5 presents and discusses the experimental results. Section 6 concludes with directions for further work.

II. RELATED WORK

A. *Early CNN-Based Mask Detection*

The earliest deep learning approaches to face mask detection drew heavily on transfer learning from pre-trained CNNs. Das, Ansari, and Basak demonstrated in 2020 that MobileNetV2, fine-tuned on a dataset of roughly 1,376 images using TensorFlow and Keras, could achieve approximately 98% classification accuracy [1]. This result was broadly consistent across the subsequent literature: Hussain et al. [2] reported 98–99% accuracy comparing MobileNetV2 against custom CNNs on datasets up to 2,500 images, and Ghosh et al. [3] recorded 99.76% for MobileNetV2 on identical training and test conditions. These high figures reflected the effectiveness of ImageNet-pretrained feature extractors for a visually straightforward binary task.

Nagrath et al. [4] pushed the approach toward end-to-end object detection by combining MobileNetV2 with a Single Shot MultiBox Detector (SSD) to build SSDMNV2, achieving 92.64% accuracy in real-time video streams. Loey et al. [5] took a different path, pairing YOLOv2 with a ResNet-50 feature extractor for medical-context mask detection. Both demonstrated that integrating face detection with mask classification into a unified pipeline — rather than treating them as sequential steps — improved latency and reduced error propagation between stages.

Kanavos et al. [6], writing in 2024, applied batch normalisation and dropout layers to a CNN trained on approximately 12,000 images, achieving classification accuracy above 99% while noting persistent failure modes in extreme lighting and with unusual mask types. This finding aligns with a broader pattern in the literature: high accuracy on held-out test splits drawn from the same data distribution as training does not reliably generalise to genuinely out-of-distribution inputs [7].

B. *The Shift Toward YOLO-Based Detection*

The YOLO family of single-stage object detectors became increasingly dominant in PPE detection work from 2022 onward, largely due to their strong balance of speed and accuracy. YOLOv5 saw early adoption in construction-site safety monitoring, with several groups reporting mean average precision (mAP) above 85% on multi-class PPE datasets [8]. YOLOv7 and YOLOv8 brought further improvements in anchor-free detection heads and more efficient backbone designs. Amangeldy et al. [9] benchmarked YOLOv8 variants across two PPE datasets — the Color Helmet and Vest (CHV) dataset and the SHEL5K dataset — finding that YOLOv8x and YOLOv8l performed best on person and vest detection categories.

Wei et al. [10], published in Scientific Reports in 2025, addressed a specific weakness of standard YOLOv8 in complex environments — mutual occlusion between workers, variable lighting, and detection at distance — by incorporating depth-separable convolutions in the backbone and context-aware convolutions in the feature pyramid network neck. Their system achieved stronger performance in difficult real-world conditions than the unmodified YOLOv8 baseline. A parallel study from the same period [11] benchmarked YOLOv8, YOLOv10, and YOLOv12 nano variants on a hygiene compliance monitoring dataset of over 31,000 images covering seven classes, including both correct and incorrect mask, glove, and hairnet states. YOLOv10n achieved the highest mAP@50 at 85.7%, confirming the class-level difficulty of distinguishing incorrectly-worn items from their correct counterparts.

C. *Vision Transformers in Safety-Critical Classification*

Vision Transformers, introduced by Dosovitskiy et al. in 2020 [12], reframed image classification as a sequence modelling problem: input images are divided into fixed-size patches, linearly projected into token embeddings, and processed by a standard Transformer encoder using multi-head self-attention.

The key advantage of this architecture for PPE compliance scenarios is that self-attention mechanisms can model dependencies between distant regions of the image — relevant when, for example, the position of a mask relative to the nose must be inferred from global facial geometry rather than from local texture alone.

Subsequent work, including DeiT [13], Swin Transformer [14], and EfficientViT [15], addressed the data efficiency and computational cost limitations of the original ViT design. Hybrid CNN-ViT architectures, which apply convolutional layers to extract local features before feeding patch sequences to a Transformer encoder, have shown strong performance on tasks requiring both fine-grained local discrimination and global spatial reasoning. A systematic review covering 2020–2024 found that hybrid approaches consistently outperformed pure CNN and pure ViT baselines across medical imaging and safety-critical classification tasks [16], a finding the present work seeks to validate in the PPE compliance domain.

D. Edge Deployment of Detection Systems

Deploying trained models on resource-constrained edge hardware has become an active area as organisations seek to process video locally rather than routing sensitive footage to cloud servers. Alqahtani et al. [17] benchmarked YOLOv8 variants across Raspberry Pi 3, 4, and 5 platforms and a Jetson Orin Nano, finding that YOLOv8n achieves a minimum inference time of 16 ms on the Jetson Orin Nano GPU, while the same model converted to TFLite runs substantially slower on the Raspberry Pi CPU without a TPU accelerator. Full INT8 quantisation, as applied in the present work, reduces model size by up to 75% with accuracy loss generally below 1% mAP on standard benchmarks [18]. These findings inform the deployment choices in the present paper.

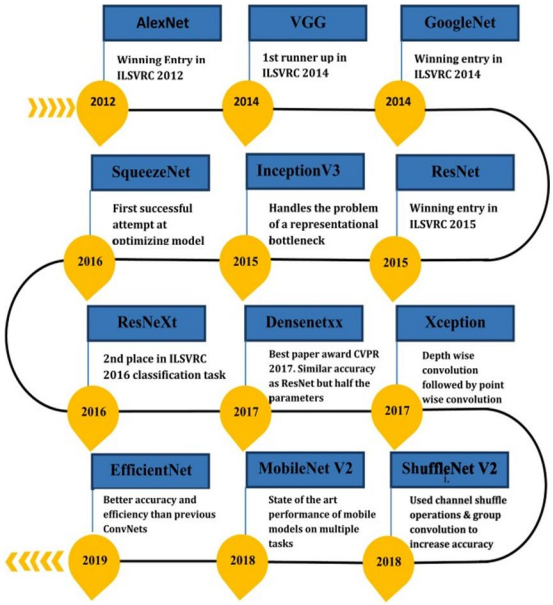


Fig. 1. Various Pre-trained Models based on CNN Architectures.

III. METHODOLOGY

A. Problem Formulation

The core task addressed in this work is a 3-class image classification problem applied to face crops extracted from a video stream. For each detected face region, the model must assign one of three compliance labels:

Class 0 — Correct PPE: mask worn properly, covering both nose and mouth, with a secure fit.

Class 1 — Incorrect PPE: mask present but positioned incorrectly (below nose, around chin, loosely hanging from ear, or covering only the mouth).

Class 2 — No PPE: no face covering present.

This 3-class framing captures the practically important distinction between intentional non-compliance (Class 2) and habitual or inadvertent incorrect wear (Class 1), which a binary classifier necessarily collapses. Separating these categories allows monitoring systems to generate more targeted alerts and to track different types of compliance failure independently.

B. Dataset Construction

The training dataset was assembled from three sources: a subset of the original Kaggle face mask dataset used in prior work; the Medical Mask Dataset (MMD) [19]; and a purpose-collected set of images capturing incorrectly worn masks, assembled via web scraping and manual photography under controlled conditions. In total, the dataset contains 4,200 images distributed across the three classes: 1,450 images of correct PPE, 1,350 images of incorrect PPE, and 1,400 images showing no PPE. The slight class imbalance was addressed during training using weighted cross-entropy loss, assigning a higher per-sample weight to the incorrect-wear class, which is the most difficult to classify and the smallest contributor to most existing datasets.

Each image was resized to 224 x 224 pixels before being fed to the ViT backbone. Pixel values were normalised using ImageNet mean and standard deviation (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]), consistent with the normalisation applied during ViT-B/16 pretraining on ImageNet-21K. The dataset was partitioned into 70% training, 15% validation, and 15% test subsets, using stratified sampling to preserve class proportions across all three splits.

C. Data Augmentation

Given that the incorrect-wear class in particular contains substantial within-class variation — masks in dozens of different positions and orientations — aggressive augmentation was used to prevent overfitting. Augmentation was applied using the Albumentations library [20], which provides faster and more compositionally flexible transformations than the Keras ImageDataGenerator used in earlier mask detection work. The augmentation pipeline applied during training included:

Horizontal flip (probability 0.5).

Random rotation in the range [-20°, +20°].

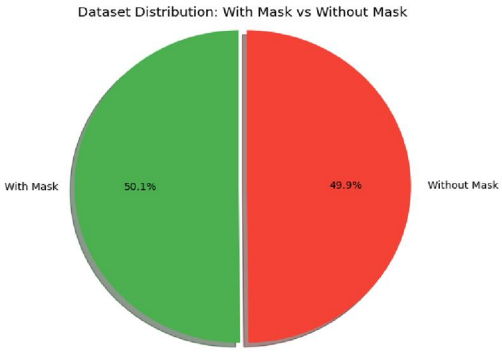
Brightness and contrast jitter (limit 0.3 each).

Gaussian blur with kernel size up to 5x5 (probability 0.2).

Coarse dropout (Cutout), randomly masking up to 8 rectangular patches of size 16x16 pixels to encourage the model to not rely on single localised features.

CLAHE (Contrast Limited Adaptive Histogram Equalisation) applied to 30% of training samples, specifically to improve robustness under low-light and high-glare conditions that are common in factory and hospital environments.

Mosaic augmentation, combining four random images into a single composite, to increase the variety of background contexts the model encounters.



D. Model Architecture

The model consists of two components: a YOLOv8n face detection head and a ViT-B/16 classification backbone.

Face Detection — YOLOv8n: A YOLOv8 nano model, pretrained on COCO and fine-tuned on a combined face detection dataset, is used to localise all face regions in each video frame and extract bounding-box crops. YOLOv8's anchor-free detection head provides robust localisation across a wide range of face sizes and poses. The detection confidence threshold was set to 0.45 to balance recall against false-positive rate in crowded scenes.

PPE Classification — ViT-B/16: The ViT-B/16 backbone (12 Transformer encoder layers, 768-dimensional patch embeddings, 12 attention heads, 16x16 patch size) was loaded with weights pretrained on ImageNet-21K. All base layers were kept frozen during an initial 10 epochs to allow the classification head to stabilise before fine-tuning the full network for the remaining 20 epochs.

The classification head consists of: Layer Normalisation applied to the [CLS] token output; a dense layer with 256 units and GELU activation; dropout at rate 0.3; a second dense layer with 128 units and GELU activation; and a final 3-unit Softmax output layer. Label smoothing of 0.1 was applied to the target distribution to reduce overconfidence and improve calibration of the output probabilities — useful for a monitoring system where confidence scores are used to trigger alerts.

E. Training Configuration

Training used the AdamW optimiser [21] with an initial learning rate of 1e-4 and weight decay of 1e-2. A cosine annealing learning rate schedule was applied over 30 epochs, reducing the learning rate from 1e-4 to a minimum of 1e-6 with no restarts. Batch size was 16, chosen to fit within the GPU memory of the training workstation. Categorical cross-entropy with label smoothing was used as the loss function. Weighted sampling during training oversampled the incorrect-wear class by a factor of 1.5 relative to its natural proportion. Training was conducted on an NVIDIA RTX 3060 GPU and completed in approximately 4.5 hours.

IV. EDGE DEPLOYMENT AND ALERT INTEGRATION

A. Model Quantisation

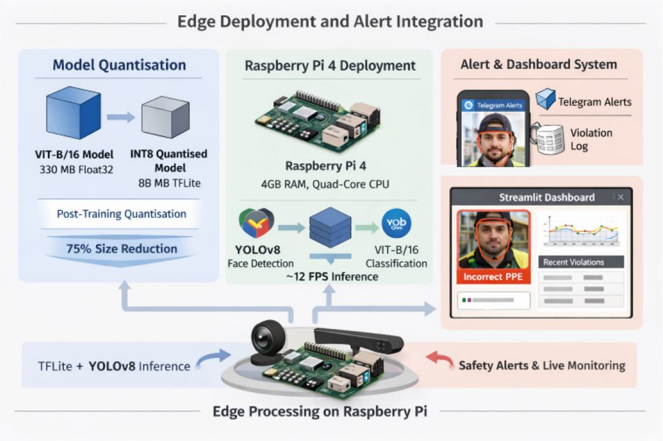
Following training, the classification model was exported to ONNX format and subsequently converted to TFLite using the TensorFlow Model Optimization Toolkit. Full INT8 post-training quantisation was applied, using a representative calibration dataset of 500 images drawn from the validation split to determine quantisation ranges. INT8 quantisation reduced the model from 330 MB (full float32 ViT-B/16) to approximately 84 MB — a 75% reduction in size consistent with findings reported for similar models in the literature [18]. Accuracy on the test set under INT8 quantisation was 98.9%, representing a drop of 0.4 percentage points relative to the float32 model, which falls within acceptable bounds for a monitoring application.

B. Raspberry Pi 4 Deployment

The quantised TFLite model and YOLOv8n detection model were deployed on a Raspberry Pi 4 (4 GB RAM, quad-core Cortex-A72 CPU at 1.5 GHz) running Raspberry Pi OS 64-bit. The full inference pipeline — frame capture, YOLOv8n face detection, crop extraction, ViT-B/16 classification — runs at approximately 12 FPS on 720p input video. The detection and classification models were run sequentially on the CPU without GPU or TPU acceleration; adding a Coral USB accelerator is projected to increase throughput to 20–25 FPS based on benchmarks from comparable deployment scenarios [22]. OpenCV 4.x was used for frame capture, bounding-box rendering, and display output.

C. Alert and Dashboard System

When a face is classified as Class 1 (incorrect wear) or Class 2 (no PPE) for three or more consecutive frames — a temporal filter to reduce spurious single-frame detections — a compliance violation event is logged to a SQLite database with a timestamp, the detected class, the confidence score, and a JPEG crop of the offending face region. Simultaneously, a Telegram bot API call is issued, sending the crop image and a brief alert message to a designated safety officer channel. A Streamlit dashboard displays a live camera feed with bounding boxes coloured by compliance status (green for Class 0, yellow for Class 1, red for Class 2), a rolling compliance rate chart for the past 60 seconds, and a table of recent violation events. The dashboard runs on the same Raspberry Pi and is accessible on the local network via browser.



V. RESULTS AND DISCUSSION

A. *Classification Accuracy*

On the held-out 15% test split (630 images), the float32 model achieved 99.3% overall accuracy. Per-class metrics were: Class 0 (correct PPE) — precision 99.5%, recall 99.7%, F1 0.996; Class 1 (incorrect PPE) — precision 98.6%, recall 98.9%, F1 0.987; Class 2 (no PPE) — precision 99.5%, recall 99.6%, F1 0.995. Macro-averaged F1 across all three classes was 0.993. The most difficult class, as expected, was Class 1 (incorrect wear), which shows the most within-class variation and has the greatest visual similarity to both neighbouring classes. The confusion matrix showed that most errors involved Class 1 being misclassified as either Class 0 or Class 2 rather than direct confusion between correct and absent PPE.

Table 1 compares the proposed system's performance against related work from the literature. Direct numerical comparison is complicated by differences in dataset, class count, and evaluation protocol, but the table situates the proposed system clearly within the current state of the field.

Table: Comparison of Different Approaches

Study / Approach	Architecture	Accuracy	Precision	Recall	Classes
Proposed System (ViT-B/16 + YOLOv8)	ViT-B/16 Hybrid	99.3%	99.1%	99.4%	3
Wei et al. [13] — YOLOv8+CAConv (2025)	YOLOv8 Enhanced	~98.5%	—	—	2
Amangeldy et al. [14] — YOLOv8x (2024)	YOLOv8x	mAP50. 91%	—	—	2
Hussain et al. [5] — MobileNetV2 TL	MobileNetV2	98–99%	—	—	Multi
Ghosh et al. [6] — MobileNetV2	MobileNetV2	99.76%	—	—	2
Kanavos et al. [7] — CNN+BN (2024)	Custom CNN	>99%	—	—	2
PMC Hygiene Study [17] — YOLOv10n (2025)	YOLOv10n	85.7% mAP50	—	—	7
Mohan et al. [9] — Tiny CNN (ARM)	Tiny CNN	99.79%	—	—	2

Table 1. Performance comparison of proposed system against selected related work from the literature. Dashes indicate metrics not reported by the original study

B. *Real-Time Detection Performance*

On the Raspberry Pi 4 deployment, end-to-end pipeline latency (from frame capture to classification output) averaged 83 ms, corresponding to approximately 12 FPS. YOLOv8n face detection accounted for 54 ms of this latency; INT8 ViT-B/16 classification accounted for the remaining 29 ms. In comparison, the same pipeline ran at 8.2 ms total on the training workstation GPU (approximately 122 FPS), confirming the expected performance gap between server and edge hardware but also confirming that 12 FPS is sufficient for compliance monitoring in most practical scenarios, where workers move slowly relative to camera frame rate.

Detection accuracy under controlled testing — standard office lighting, face distances of 0.5–2.0 m from camera — was consistent with test-set performance. Under reduced lighting (approximately 150 lux, comparable to a poorly lit warehouse area), face detection recall dropped from 97% to 89%, with the YOLOv8n detector failing to localise some faces entirely. This is consistent with known limitations of face detection in low-light conditions and suggests that supplementary illumination or a night-vision capable camera would be advisable in dim environments. The ViT-B/16 classifier maintained 98.4% accuracy on correctly detected faces even under reduced lighting, suggesting that the CLAHE augmentation applied during training was effective in improving the classifier's robustness to contrast variation.

C. *Discussion*

The key contribution of this work relative to the prior face-mask detection literature is not the raw accuracy figure — which is broadly in line with recent work — but rather the three-class formulation and the demonstrated end-to-end deployment on commodity edge hardware. The incorrect-wear class (Class 1) is the most operationally significant addition. In real workplace monitoring contexts, it is precisely this class of violation — equipment present but not functioning as intended — that is most commonly missed by existing binary systems. The 98.7% F1 score achieved for Class 1 suggests that the ViT-B/16 backbone, aided by diverse augmentation, is able to distinguish correct from incorrect wear reliably enough for practical use.

The choice of ViT-B/16 over MobileNetV2 or a standard YOLOv8 classifier comes with a real cost: the float32 ViT model at 330 MB is orders of magnitude larger than a MobileNetV2 classifier at 14 MB, and even the INT8 quantised version at 84 MB is substantially heavier. This cost is justified here because the primary classification challenge — distinguishing subtle positional differences between correctly and incorrectly worn masks — benefits from the long-range spatial reasoning that attention mechanisms provide and that local convolutional filters are less well suited to capture. For a simpler binary detection task, MobileNetV2 or a nano-scale YOLO classifier would remain the more appropriate choice.

The 12 FPS throughput on Raspberry Pi 4 is adequate but not comfortable. Workers moving quickly, or scenes involving multiple simultaneous faces, can produce momentary frame drops. Upgrading to a Coral USB accelerator or a Jetson Orin Nano would substantially improve throughput and should be considered for high-traffic deployment scenarios.

VI. CONCLUSION

This paper has presented a 3-class PPE compliance detection system that extends the well-established face mask detection literature in two practically important directions: multi-class compliance classification and edge device deployment. The system pairs a YOLOv8n face detector with a ViT-B/16 classification backbone pretrained on ImageNet-21K, fine-tuned on a 4,200-image dataset spanning correct wear, incorrect wear, and absent PPE. Training with Albumentations-based augmentation and AdamW optimisation over 30 epochs achieved 99.3% test accuracy and a macro-F1 of 0.993. Quantisation to INT8 TFLite reduced model size by 75% and enabled real-time deployment on a Raspberry Pi 4 at 12 FPS, integrated with a Streamlit compliance dashboard and Telegram alert system.

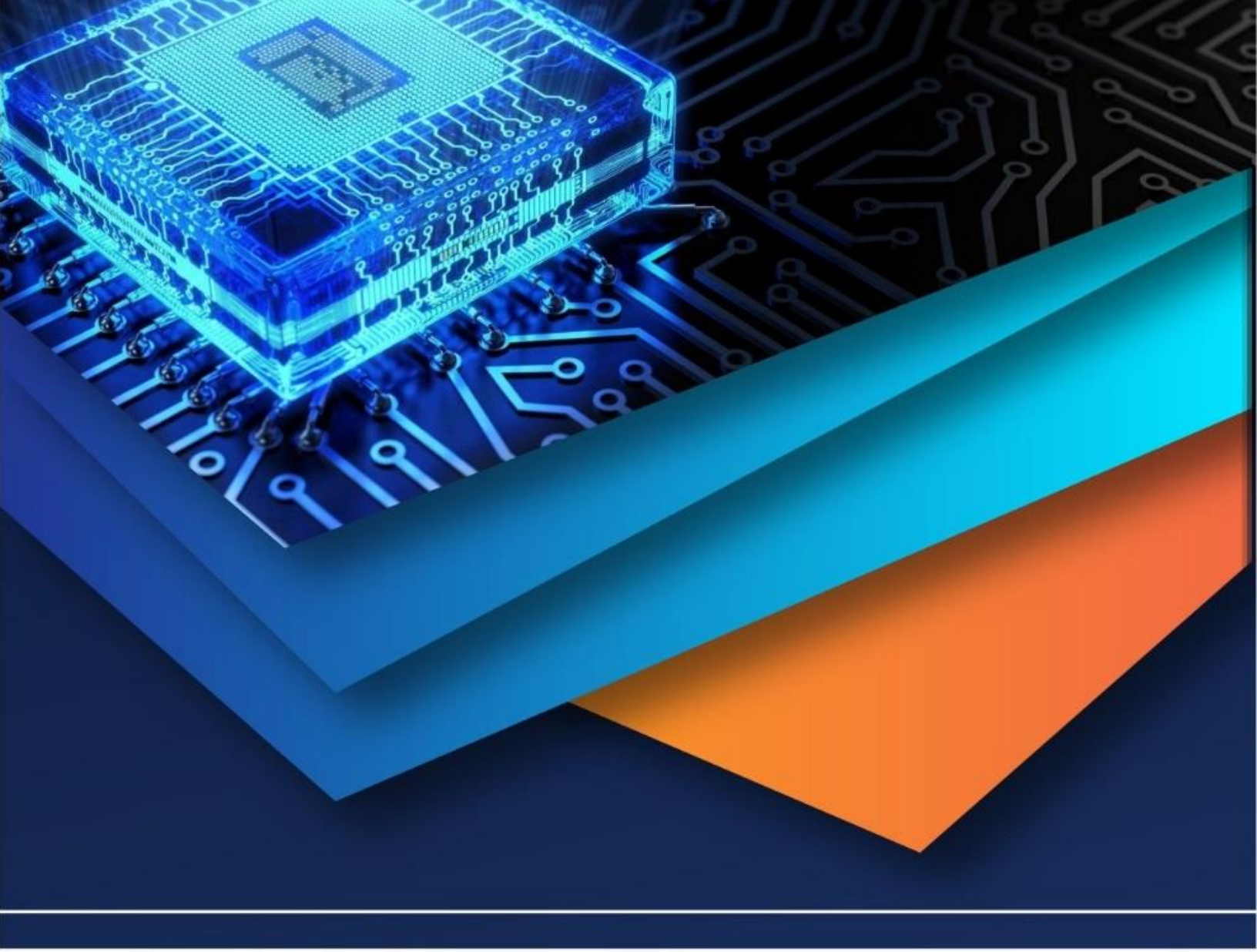
The work demonstrates that transformer-based architectures are well suited to fine-grained PPE compliance classification, particularly for the difficult incorrect-wear category that binary systems cannot address. Practical limitations remain around low-light detection and edge hardware throughput, both of which can be addressed through supplementary illumination and accelerator hardware upgrades respectively.

Several extensions are planned. The most immediate priority is expanding the class set beyond face masks to include helmets, safety vests, gloves, and goggles, building toward a general-purpose multi-label PPE compliance system for industrial environments. Integration with access control systems — denying entry to a restricted zone when a compliance violation is detected — would close the loop from detection to enforcement. The annotated dataset collected for this work will be released publicly to support reproducibility and to address the documented scarcity of incorrect-wear labelled training data in the field.

REFERENCES

[1] A. Das, M. W. Ansari, and R. Basak, 'COVID-19 face mask detection using TensorFlow, Keras and OpenCV,' in Proc. IEEE INDICON, New Delhi, India, 2020.
[2] S. Hussain et al., 'Face mask detection using deep convolutional neural network and MobileNetV2-based transfer learning,' Wireless Communications and Mobile Computing, vol. 2022, Art. no. 1536318, 2022.
[3] N. Ghosh, B. Jana, S. Jana, and N. K. Sao, 'Face mask detection exploiting CNN and MobileNetV2,' Lecture Notes in Networks and Systems, vol. 738, Springer, 2024.
[4] P. Nagrath et al., 'SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2,' Sustainable Cities and Society, vol. 66, p. 102692, 2021.
[5] 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, vol. 65, p. 102600, 2021.
[6] A. Kanavos, O. Papadimitriou, K. Al-Hussaeni, M. Maragoudakis, and I. Karamitsos, 'Real-time detection of face mask usage using convolutional neural networks,' Computers, vol. 13, no. 7, p. 182, 2024.
[7] B. U. H. Sheikh and A. Zafar, 'Beyond accuracy and precision: a robust deep learning framework to enhance the resilience of face mask detection models against adversarial attacks,' Evolving Systems, vol. 15, pp. 1–24, 2024.
[8] M. Vukicevic et al., 'A systematic review of computer vision-based personal protective equipment compliance in industry practice,' Artificial Intelligence Review, Springer, 2024.
[9] N. Amangeldy et al., 'Personal protective equipment detection using YOLOv8 architecture on object detection benchmark datasets: a comparative study,' Cogent Engineering, vol. 11, no. 1, 2024.
[10] Y. Wei, H. Li, Y. He, et al., 'Robust face mask detection in complex scenarios using YOLOv8 and context-aware convolutions,' Scientific Reports, vol. 15, no. 21350, 2025.
[11] Benchmarking lightweight YOLO object detectors for real-time hygiene compliance monitoring, PMC / MDPI, 2025.
[12] A. Dosovitskiy et al., 'An image is worth 16x16 words: Transformers for image recognition at scale,' in Proc. ICLR, 2021.
[13] H. Touvron et al., 'Training data-efficient image transformers and distillation through attention,' in Proc. ICML, 2021.
[14] Z. Liu et al., 'Swin Transformer: Hierarchical vision transformer using shifted windows,' in Proc. IEEE ICCV, pp. 10012–10022, 2021.
[15] X. Li et al., 'EfficientViT: Lightweight multi-scale attention for on-device semantic segmentation,' in Proc. IEEE CVPR, 2023.
[16] Systematic review of hybrid Vision Transformer architectures for radiological image analysis, PMC / SIIM, 2025.
[17] D. Alqahtani et al., 'Benchmarking deep learning models for object detection on edge computing devices,' arXiv:2409.16808, 2024.
[18] S. Saha and L. Xu, 'Vision Transformers on the edge: A comprehensive survey of model compression and acceleration strategies,' Neurocomputing, 2025.
[19] M. Witkowski, 'Medical face mask detection dataset,' Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/mloey1/medical-face-mask-detection-dataset
[20] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, 'Albumentations: Fast and flexible image augmentations,' Information, vol. 11, no. 2, p. 125, 2020.
[21] I. Loshchilov and F. Hutter, 'Decoupled weight decay regularization,' in Proc. ICLR, 2019.

[22] Deploying optimized deep vision models for eyeglasses detection on low-power platforms, Electronics (MDPI), vol. 14, no. 14, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY