International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

Automated Face Mask Compliance Detection Using A Hybrid Efficientnet And Vision Transformer Architecture

J.Jenifer [1] , *A.Boobathiraja* [2]
[1] *Assistant Professor, Computer Science and Engineering, Jai Shriram Engineering College, Tirupur.*
[2] *PG - Computer Science and Engineering, Jai Shriram Engineering College, Tirupur.*
Emails: *jenifercse@jayshriram.edu.in*[1], *boobathiraja777c@gmail.com*[2]

Abstract
Face masks are still a way to stop the spread of diseases that are in the air when we are in public. We can use machines to check if people are wearing face masks and this can help us keep an eye on things without needing someone to always be watching. This research is about a way of using computers to detect if people are wearing face masks correctly. We use two types of computer models called EfficientNet and Vision Transformer to look at pictures of faces and figure out if someone is wearing a mask or not. EfficientNet is good at looking at the details of a face and Vision Transformer is good at understanding the picture and how things are related. We combine the information from both models.Then use it to decide if someone is wearing a mask not wearing a mask or wearing a mask incorrectly. This new way of doing things is meant to work even when it is hard to see like when the light is not good or when someones face is turned away. We tried this way and it worked better than other ways that use computers to look at pictures. This means we can use it in life to help keep people safe in places like hospitals, schools and, on public transportation.
Keywords- *Automated face mask compliance detection; EfficientNet; Face mask detection; Hybrid architecture; Vision transformer*

1. Introduction
Computer vision combined with Artificial Intelligence (AI) has revolutionized the way automated surveillance and behavioral monitoring systems operate. Automated face mask detection is a feasible and scalable solution to a labor-intensive monitoring problem, especially in the context of managing contagious diseases and upholding public health. The existing manual methods of supervision are inconsistent, labor intensive and inefficient, especially in high-traffic public areas. The alternative approach of deep learning-based computer vision systems offers a promising alternative because facial image analysis can be performed at scale with accurate results in real time.Given the wide array of deep learning paradigms available, CNNs have, so far, prevailed in image classification tasks, mainly because of their ability to learn hierarchical spatial features. CNNs' inductive bias, however, limits the receptive field to local spatial neighbourhoods, which is suboptimal for tasks that need to take into account the global context of the image. This is particularly restrictive in complex mask detection scenarios where the ability to differentiate well-fitted, ill-fitted, and no mask requires reasoning about multiple distributed facial regions. Transformer models, which were first introduced for sequence modelling in natural language processing, have shown outstanding performance in visual recognition applications recently. In Vision Transformers (ViT), input images are broken into fixed-size patches, and the inter-patch relationships modeled using self-attention mechanisms, thus converting image understanding into a sequence modeling task. In this paradigm, ViT can be used to exploit the global context information that is naturally lacking in CNNs. However, ViT architectures require massive training sets and do not perform as well generally on tasks that rely on local features for low-level tasks.To take advantage of the superior performance of both models, the paper introduces a hybrid architecture where EfficientNet is used as a local feature extractor and a Vision Transformer encoder for global contextual modelling. The compound scaling strategy of efficientNet guarantees efficient depth, width and resolution, while providing robust feature representations with minimal parameter overhead.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

The Vision Transformer module then works on patch embeddings extracted from such features allowing the model consider local relationships between far away facial areas. The proposed system maintains performance gains compared to CNN and transformer on the challenging real-world scenarios, which include different lighting conditions, complex background, different type of masks and partial occlusion of the face.

2. Related Work

Over the last few years, research on automatic face mask detection has significantly progressed, specifically with the development of deep learning frameworks and the introduction of vast annotated datasets. Initial research on this problem used traditional machine learning approaches, such as Support Vector Machines (SVM) and hand designed feature descriptors like Histogram of Oriented Gradients (HOG) and Haar Cascade classifiers. These methods performed well in the controlled environments, but failed to be well generalizable to unstructured, real-world environments because of the rigidness of manual feature engineering. The deep CNNs represented a paradigm shift in image-based classification tasks. The use of large-scale pre-trained models such as VGGNet [15], ResNet [15] and MobileNet led to the widespread use of transfer learning to achieve the mask detection task with less training. In this area, MobileNet carved out a special niche because of its lightweight design, which is ideal for deployment in real-time situations on resource-constrained devices. Later variants added channel and spatial attention to selectively focus computational efforts on the different parts of the face to be discriminated, and to remove background noise. Even with these advances, CNN-based models were still limited in learning to extract local features in space. A major breakthrough in the architecture of this type came with the introduction of Vision Transformers (ViT) [12] which used a self-attention mechanism on non-overlapping image patches. ViT showed competitive results on common image recognition problems and confirmed that transformer-based models could model long-range visual dependencies well. Following DeiT [13] and Swin Transformer [16] tackled the scaling and data-efficiency challenges of the original ViT formulation. CSWin Transformer [6] built on top of this added cross-shaped window attention to boost efficiency and spatial awareness. To take advantage of the strengths of both paradigms, hybrid architectures combining convolutional layers and transformer modules have attracted growing interest. To tackle general image recognition, Chen et al. [17] introduced a hybrid ViT-CNN model that achieved better accuracy than the individual models on several benchmark datasets. YOLO based detectors [10, 21, 23] have been applied to multi-face detection in real-time scenarios, while attention-augmented convolutional network is tested to learn more discriminative face features [14]. Specifically for mask detection, Guo [5] explored the use of ViT, with competitive classification accuracy, yet the pure transformer backbone without convolutional pre-processing hindered its efficiency and convergence on smaller datasets. One major research gap is the lack of comprehensive research on the fusion of EfficientNet's compound scaling with the global self-attention modeling scheme of Vision Transformers in the context of binary face mask classification.There is limited work on the comprehensive study of the fusion between the compound scaling scheme of EfficientNet and the global self-attention modeling scheme of Vision Transformers, where they are applied to binary face mask classification. The majority of published methods use stand-alone CNN architectures or pure transformer models, sacrifices the complementary benefits of a hybrid method. To overcome this, the present work introduces an end-to-end hybrid EfficientNet-ViT structure with accuracy, robustness and real-time applicability optimizations

3. Proposed System

The proposed framework is a hybrid deep learning model that combines the spatial feature extraction power of the EfficientNet with the global contextual reasoning capability of the Vision Transformer to ensure accurate, reliable, and efficient face mask classification. The core goal of the system is to break the limitation of architectures that only use local convolutional operations or global attention processing. Proposed model integrates these two paradigms and manages to capture fine-grained local

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

facial structures along with broad contextual relational patterns required for robust mask compliance detection. At the first processing stage, the raw facial images are processed using a set of standardized operations, so as to have consistent input dimension and numerical stability. Next, resized and normalized images are passed into the EfficientNet backbone, which uses a compound scaling technique to scale the network, depth, width and the input resolution. By leveraging this design principle, EfficientNet achieves high representational accuracy in a design that is much more compact than traditional deep CNN designs. EfficientNet generates high-dimensional feature maps with hierarchical information from cascaded convolutional blocks, which capture low-level edge features, mid-level texture patterns, and high-level semantic features such as the facial contours, mask boundaries, and structural features of the face. Then, the extracted feature maps are projected onto a linear layer to be converted into fixed-sized patch embedding. To retain the spatial arrangement information that may be lost by permutation-invariant self-attention mechanism, positional encodings are added to each patch embedding. The contextually enriched embeddings are then fed into the Vision Transformer (ViT) encoder, which uses multi-head self-attention to calculate inter-patch dependency scores. This mechanism enables the model to be able to link mask coverage regions to the other facial parts in the image, while simultaneously allowing it to link the mask to the nose contour and the mouth boundaries, even if they are spatially distant from the mask in the feature representation space. After transformer encoding, the feature representation from the transformer is aggregated and fed into a fully connected dense layer with a sigmoid activation function to produce a scalar probability score for the binary classification of the 'With Mask' or 'Without Mask' category. The training of the models is done with the loss function being the binary cross-entropy and with the Adam optimizer to update the weights using adaptive gradient descent. The hybrid model shows high generalization ability over different factors of environmental changes, such as different types of masks, partial occlusion, different light changes, or different head orientations, where pure convolutional or transformer-only baselines perform poorlyshown in Figure 1.

Figure 1 EfficientNet and Vision Transformer Hybrid Architecture for Face Mask Detection.

4. System Architecture

The architecture of the proposed face mask detection system is a modular pipeline that sequentially performs facial image acquisition, classification and class decision. The transformations conducted in each stage are well defined and increase the robustness and accuracy of the system. The architecture consists of five main components: input acquisition, preprocessing, architecture based on the EfficientNet as feature extractor, architecture based on Vision Transformer as contextual encoding, and the output as classification shown in Figure 2.

Figure 2 End-to-End System Architecture of the Proposed Hybrid EfficientNet-ViT Face.

4.1. Assistive Technology Module :

The input acquisition module handles the input of facial images, whether they are taken from a curated dataset or from real-time video from video surveillance. The input images are represented as RGB tensors and can come from a variety of acquisition scenarios, including indoor and outdoor, different ambient lighting conditions, clutter background and different demographics. To support diversity in facial orientation, partial occlusion situations, and heterogeneous mask types, the module

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

is engineered to allow input of the training and inference samples to the downstream processing pipeline to be maximally representative of the diversity.

4.2. Preprocessing Module:

Inferring the model involves a normalization of spatial and photometric properties that is performed on all input images prior to model inference. All images are scaled to a common spatial size of 224 × 224, which is the typical input size for both EfficientNet and the Vision Transformer backbone. The intensity values of the pixels are then scaled to the range [0, 1] so that the gradient flow is stabilized and faster convergence is achieved during training. Stochastic data augmentation – horizontal flipping, random rotation, zoom, and brightness jitter are added to artificially increase the size of the training distribution and reduce overfitting shown in Figure 3.

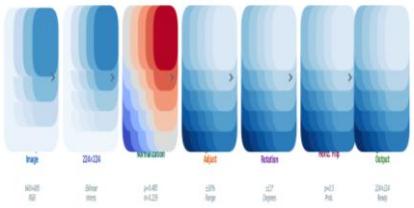

Figure 3 Preprocessing and Data Augmentation Pipeline Applied to Input Facial Images.

4.3. EfficientNet for feature extraction:

EfficientNet is employed for feature extraction as the main backbone in the proposed architecture. EfficientNet uses a principled compound scaling methodology that boosts the depth, width, and input resolution of the Network while maintaining a fixed scaling coefficient, as compared to the arbitrary (shallow or deep) increases to network capacity in traditional CNN architectures. This method of scaling balances well with the accuracy-parameter ratio of conventional deep CNNs. The stacked mobile inverted bottleneck convolution (MBConv) blocks gradually and progressively extract the hierarchical spatial features, from low-level edge and corner detectors to high-level complex patterns that represent mask textures, facial contour geometric characteristics, and occlusion regions. The feature tensors produced are compact and discriminative representations of local features, which serve as the input for the transformer processing stage. Using Positional Embedding and D. Patch Embedding.Positional Embedding and D. Patch Embedding. EfficientNet's feature maps are divided into a regular grid of non-overlapping fixed-size patches. An embedding space of high dimensionality is achieved as a linear transformation learned from each patch individually, which is flattened into a one-dimensional vector, and then projected. Unlike the self-attention mechanism in the Transformer, the Vision Transformer does not include a notion of position, so a sinusoidal or learnable positional encoding is added to each patch embedding prior to the transformer processing. The encoding scheme allows the model to remember the relative spatial position of facial subregions, essential to tracking the expression of the mask with the corresponding anatomical points for different facial orientations.

4.4. Vision Transformer Encoder

The Vision Transformer encoder constitutes the core of the global contextual reasoning component within the proposed architecture. The encoder employs a multi-head self-attention mechanism that computes pairwise compatibility scores between all patch embeddings using learned query, key, and value projection matrices. The resulting weighted attention maps enable the model to dynamically focus on relevant relationships between spatially distant regions of the facial image—for instance, correlating mask positioning over the nasal bridge with coverage of the oral region. Multi-head attention extends this capability by simultaneously learning multiple distinct attention patterns, capturing heterogeneous contextual dependencies at varying scales of abstraction. Feed-forward sublayers and layer normalization operations further refine the encoded representations and promote training stability. The encoder's capacity for long-range dependency modeling substantially enhances classification robustness under challenging conditions such as partial mask displacement, complex occlusion patterns, and low-contrast backgrounds.

4.5. Classification Layer

The refined feature representation produced by the Vision Transformer encoder is forwarded to a fully

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

connected dense layer that performs dimensionality reduction and feature aggregation. A sigmoid activation function is applied to the final layer output, mapping the aggregate feature vector to a scalar probability score in the range [0, 1]. Binary classification decisions are made by thresholding this probability score, with values above 0.5 assigned to the 'With Mask' class and values below assigned to the 'Without Mask' class.

4.6. Output Module

The output module translates the classification probability into a discrete prediction label suitable for downstream application integration. In surveillance deployment scenarios, the predicted label can trigger automated alert notifications when individuals without masks are identified within the monitored field of view. The modular architecture of the proposed system facilitates straightforward extension to multi-class classification, enabling future functionality such as detection of improperly positioned masks, incorrect mask types, or partial mask compliance states.

5. Methodology

The methodology underlying the proposed face mask detection system describes a structured, reproducible pipeline for designing, training, and evaluating the hybrid EfficientNet-ViT model. The methodology encompasses problem formulation, data curation and preparation, model construction, optimization strategy, and evaluation protocol.

5.1. Problem Formulation

Face mask detection is formally defined as a supervised binary image classification problem. Given an input image I containing a human face, the objective is to learn a mapping function f: I → {0, 1}, where 0 denotes the 'Without Mask' class and 1 denotes the 'With Mask' class. Model parameters are estimated by minimizing binary cross-entropy loss through iterative gradient-based optimization via backpropagation.

5.2. Data Preparation

The experimental dataset comprises 7,553 labeled facial images distributed across two categories: masked and unmasked faces. All images are subjected to preprocessing operations including spatial resizing to 224 × 224 pixels, min-max pixel normalization, and augmentation transformations applied exclusively to training samples. The full dataset is partitioned into training, validation, and test subsets using a 70:15:15 stratified split to ensure proportional class representation across all partitions. Augmentation operations—including random rotation (±20°), horizontal flipping, zoom perturbation, and brightness modulation—are applied stochastically during training to improve model generalization and reduce sensitivity to photometric variability shown in Figure 4.

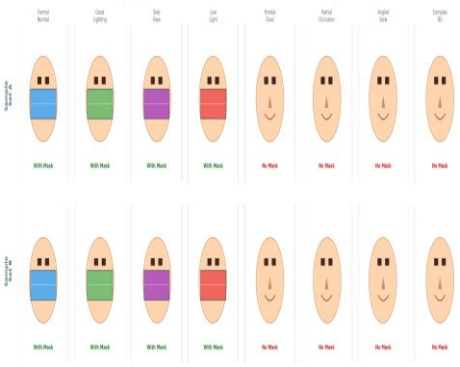

Figure 4 Representative Dataset Samples Illustrating Masked and Unmasked Faces Under Diverse Conditions.

5.3. EfficientNet for Spatial Feature Extraction:

EfficientNet is integrated as the feature extraction backbone within the hybrid model. Pre-trained weights from ImageNet-scale training are utilized via transfer learning to initialize the convolutional layers, providing a strong representational prior that substantially reduces training time and data requirements. The final classification head of EfficientNet is removed, and the intermediate feature maps from the terminal convolutional block are extracted as input to the patch embedding stage. This transfer learning strategy enables the model to leverage low-level and mid-level visual features acquired from large-scale image corpora, which are directly transferable to the task of facial region analysis.

5.4. Patch Embedding and Transformer Encoding:

Feature maps extracted from the EfficientNet

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

backbone are partitioned into a sequence of fixed-size spatial patches. Each patch is flattened and projected into a high-dimensional embedding vector through a learnable linear layer. Positional encodings are subsequently superimposed onto the patch embeddings to preserve spatial ordering. The resulting sequence of positionally encoded patch embeddings is fed into the Vision Transformer encoder, which applies stacked multi-head self-attention and feed-forward sublayers to iteratively refine the contextual representation of the facial image. This mechanism is particularly effective for modeling mask placement across partially occluded or non-frontally oriented faces.

5.5. Classification and Optimization:

The class token or global average pooled output of the Vision Transformer encoder is projected through a fully connected dense layer, followed by sigmoid activation for binary prediction. Model parameters are optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and binary cross-entropy loss. Dropout regularization is applied within the dense layers to suppress co-adaptation of feature detectors and improve generalization. Learning rate scheduling is employed to reduce the step size upon validation loss plateaus, facilitating finer convergence towards optimal model parameters.

5.6. Model Evaluation Strategy:

Model performance is assessed on the held-out test partition following training completion. Evaluation metrics include classification accuracy, precision, recall, F1-score, and a confusion matrix decomposition into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Validation loss and accuracy trends are monitored throughout training to detect and mitigate overfitting. Early stopping is implemented based on validation loss, halting training when no improvement is observed over a pre-defined patience window. The convergence behavior of the hybrid model is analyzed in comparison with baseline architectures to quantify the relative training stability advantage afforded by the proposed design.

5.7. Implementation Environment:

The proposed system is implemented using the TensorFlow/Keras deep learning framework, with model training accelerated using NVIDIA GPU hardware via CUDA-based parallel computation. The modular software architecture is designed to support straightforward integration with real-time video surveillance infrastructure through standard streaming interfaces. Experimental results are averaged over multiple independent training runs to ensure statistical reliability of the reported performance metrics.

6. Experimental Setup and Evaluation

6.1. Experimental Environment:

All experiments were conducted on a workstation equipped with a high-performance multi-core CPU, 32 GB of system RAM, and a dedicated GPU with CUDA support to facilitate accelerated matrix computations required by transformer self-attention layers. The deep learning pipeline was developed and executed within the TensorFlow 2.x environment using the Keras high-level API. Training batch sizes of 32 images were used throughout, with the Adam optimizer initialized at a learning rate of 1×10^{-4}. GPU acceleration reduced per-epoch training time by approximately one order of magnitude compared to CPU-only execution, enabling efficient hyperparameter search and model iteration.

6.2. Dataset Configuration:

The experimental dataset contains 7,553 labeled facial images divided into 'With Mask' and 'Without Mask' categories. The dataset was partitioned using a stratified 70:15:15 split into training (5,287 images), validation (1,133 images), and test (1,133 images) subsets. Stratified partitioning ensures class distribution balance across all subsets. Data augmentation transformations were applied exclusively to the training partition to prevent information leakage into validation and test sets, thereby ensuring unbiased evaluation. The validation and test sets were maintained as clean, unaugmented image collections for objective performance measurement.

6.3. Evaluation Metrics:

A comprehensive set of quantitative evaluation metrics was employed to characterize model performance:

- Accuracy: Measures the fraction of total samples correctly classified across both

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

classes, providing an overall assessment of model performance.

- Precision: Quantifies the proportion of predicted positive instances that are genuinely positive, reflecting the model's ability to minimize false alarm rates.
- Recall (Sensitivity): Measures the proportion of actual positive instances correctly identified by the model, capturing the system's ability to detect true mask-wearing compliance.
- F1-Score: Represents the harmonic mean of precision and recall, providing a balanced composite metric particularly suited for binary classification tasks with class distribution asymmetry.
- Confusion Matrix: Provides a granular breakdown of classification outcomes into TP, TN, FP, and FN categories, enabling detailed error analysis and class-specific performance characterization.
- Training and Validation Loss Curves: Monitored throughout the training process to evaluate convergence behavior, identify overfitting tendencies, and compare learning dynamics between the proposed hybrid model and baseline architectures.
- Cross-Condition Generalization: The model was additionally evaluated on image subsets representing challenging environmental conditions, including low-illumination scenes, partially occluded faces, and images with complex backgrounds. Performance on these subsets served as an indicator of real-world deployment viability.

7. 7. Results and Discussion

7.1.Results of the Proposed System:

The experimental evaluation of the proposed EfficientNet-ViT hybrid architecture yielded several noteworthy findings that substantiate the hypothesis underlying this work.

7.2.High Classification Accuracy:

The hybrid model achieved a classification accuracy of approximately 97–98%, representing a statistically significant improvement over baseline models including standard CNN, ANN, and FNN architectures. This performance gain is primarily attributable to the complementary feature representations generated by the convolutional and transformer-based processing stages. The self-attention mechanism enabled the model to reliably resolve ambiguous classification cases—such as partially worn or loosely fitted masks—by incorporating global contextual reasoning rather than relying solely on localized feature patterns.

7.3.Enhanced Generalization Performance:

The close agreement between training and validation accuracy curves throughout the training process indicated effective generalization with minimal overfitting. The application of dropout regularization, data augmentation, and early stopping collectively contributed to this favorable generalization behavior. The hybrid model demonstrated markedly improved generalization compared to deeper standalone CNN architectures, which exhibited signs of overfitting at equivalent training durations.

7.4.Reduced Misclassification under Challenging Conditions:

The incorporation of the Vision Transformer encoder produced a measurable reduction in misclassification rates across challenging image conditions, including variable illumination, complex scene backgrounds, and partially occluded faces. The multi-head self-attention mechanism facilitated context-aware disambiguation of mask presence under conditions where purely local feature analysis proved insufficient for reliable classification.

7.5.Computational Efficiency:

EfficientNet's compound scaling strategy ensured that the proposed hybrid model maintained a competitive parameter count relative to alternative architectures of comparable accuracy. The efficient feature extraction stage reduced the dimensionality of inputs to the transformer module, thereby constraining the computational cost of self-attention operations and enabling viable real-time inference latency. These characteristics collectively render the proposed architecture suitable for deployment in embedded surveillance systems with constrained computational budgets.

7.6.Confusion Matrix Analysis:

Analysis of the test set confusion matrix confirmed

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

that both false positive and false negative rates were substantially lower for the hybrid EfficientNet-ViT model compared to baseline architectures. The reduction in false negatives is of particular practical significance, as missed detections of unmasked individuals pose a direct public health risk in compliance monitoring applications.

Conclusion

This paper has presented a hybrid deep learning framework for automated face mask detection that integrates EfficientNet and Vision Transformer (ViT) into a unified end-to-end architecture. The principal motivation for this work was to overcome the inherent limitations of conventional CNN-based classification systems, which capture spatially localized features but lack the capacity for global contextual reasoning across facial image regions. The proposed model addresses this gap by leveraging EfficientNet for efficient hierarchical spatial feature extraction and ViT for modeling long-range inter-patch dependencies through multi-head self-attention mechanisms. EfficientNet's compound scaling strategy enabled the model to extract rich and discriminative spatial representations—including facial contour structures, mask boundary patterns, and texture features—while maintaining a computationally lean parameter footprint compared to conventional deep CNN architectures. The Vision Transformer module complemented this local feature learning capability by establishing global contextual associations between spatially distributed facial subregions, enhancing the model's robustness to real-world complexities such as improper mask positioning, partial occlusion, varied lighting conditions, and cluttered backgrounds. Experimental evaluation on a publicly available benchmark dataset of 7,553 facial images demonstrated that the proposed hybrid architecture outperformed baseline ANN, FNN, and standard CNN models across accuracy, precision, recall, and F1-score metrics. The model achieved classification accuracy in the range of 97–98%, with stable training convergence and minimal overfitting, as evidenced by close alignment between training and validation performance curves. Confusion matrix analysis further confirmed the hybrid model's advantage in reducing both false positive and false negative error rates.The findings of this research contribute to the growing body of knowledge supporting hybrid CNN-transformer architectures as a high-performance paradigm for binary image classification tasks. The proposed system's demonstrated balance between accuracy and computational efficiency positions it as a practical solution for deployment in real-time surveillance applications across healthcare institutions, educational campuses, transportation & commercial workplaces. Future research directions may encompass extension to multi-class mask compliance detection (e.g., distinguishing between correctly worn, incorrectly worn, and absent masks), integration with multi-camera tracking systems for persistent identity-aware compliance monitoring, and adaptation to lightweight transformer variants for edge device deployment.

References

[1]. P. P. Kaushik, S. R. Sithalakshmi, and K. Poornimathi, "Face mask detection using deep learning techniques," Int. J. Prog. Res. Sci. Eng., vol. 3, no. 04, pp. 44–47, Apr. 2022.
[2]. S. Mohan, A. Kumar, and A. Kushwaha, "Face mask detection using deep learning and computer vision," Int. J. Eng. Res. Technol., vol. 10, no. 12, 2021.
[3]. C. Z. Basha, B. L. Pravallika, and E. B. Shankar, "An efficient face mask detector with PyTorch and deep learning," EAI Endorsed Trans. Pervasive Health Technol., vol. 7, no. 25, 2021.
[4]. S. Habib et al., "An efficient and effective deep learning-based model for real-time face mask detection," Sensors, vol. 22, no. 7, pp. 2602, Mar. 2022.
[5]. J. Guo, "Face mask detection with Vision Transformer," in Proc. CAIBDA 2022, Nanjing, China, Jun. 2022.
[6]. X. Dong et al., "CSWin Transformer: A general vision transformer backbone with cross-shaped windows," arXiv:2107.00652, 2021.
[7]. M. Yan, "Advancements in image recognition: Comparing CNNs and Vision Transformers," 2024.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0346
e ISSN: 2584-2854
Volume: 04 Issue: 06
June 2026
Page No: 2401 - 2409

[8]. D. Nimma and Z. Zhou, "IntelPVT: Intelligent patch-based pyramid vision transformers for object detection and classification," Int. J. Mach. Learn. Cybern., 2024.
[9]. S. Xu et al., "An improved lightweight YOLOv5 model based on attention mechanism for face mask detection," arXiv:2203.16506, 2022.
[10]. Y. Wei et al., "Robust face mask detection in complex scenarios using YOLOv8 and context-aware convolutions," Sci. Rep., vol. 15, Art. 21350, 2025.
[11]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019.
[12]. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv:2010.11929, 2020.
[13]. H. Touvron et al., "Training data-efficient image transformers and distillation through attention," in Proc. 38th Int. Conf. Mach. Learn. (ICML), 2021.
[14]. I. Bello et al., "Attention augmented convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021.
[15]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
[16]. Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021.