International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2136-2140
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0284

Detection Face with Mask and Without Mask Using Open CV

Kowsalya.T[1], Kavipriya.P[2], Sharmila.G[3], Sureka.A.C[4], Vaishnavi.K[5]
[1]PG-Department of Computer Science Engineering, Jai Shriram Engineering college, Tirupur.
[2,3,4,5]UG-scholar Department of Artificial Intelligence and Data Science, Jai Shriram Engineering college, Tirupur.
Emails: *kowsalyathangavel10@gmail.com[1], kavipriyapalanisamy02@gmail.com[2], sharmila170404@gmail.com[3], appusamysureka2@gmail.com[4], vaishnavikumar810@gmail.com[5]*

Abstract
Facial mask detection is a relevant application of computer vision, the use of which has become crucial in the context of preserving safety and health measures of the population. The classical deep learning models like Convolutional Neural Networks (CNNs) are good at local features extraction and are weak in capturing long-ranged image dependencies. To overcome such weaknesses, this article will suggest one hybrid deep learning network that incorporates EfficientNet to extract features efficiently and Vision Transformer (VIT) to process features in global contexts with the help of self-attention mechanisms. The suggested system is trained and tested on the publicly available set of 7,553 images of masked and unmasked faces. Extensive experimental research has shown that the hybrid EfficientNet VIT model is more accurate, precise, recalls and F1-score than the traditional ANN, CNN and FNN models. The approach to be proposed has a greater strength in changing lighting conditions, background complexity, and face orientations. The model can be used in real-time execution on the surveillance systems, schools, airports, and in healthcare facilities.

Keywords: *Face Mask Detection, Deep Learning, Artificial Intelligence, Efficient Net, Vision Transformer (VIT), Convolutional Neural Networks (CNN), Hybrid Architecture, Computer Vision, Image Classification, Self- Attention Mechanism, Public Safety.*

1. Introduction
The development of Artificial Intelligence (AI) and Deep Learning has changed automated monitoring systems to a great extent. Detection of face mask was relevant to the world at the time of health emergencies in which monitoring compliance became crucial to the safety of the population. The manual techniques of monitoring are tedious, ununiform, and may be subjected to human error. The computer vision systems that are based on deep learning provide automated solutions that are scalable. The use of CNNs in image classification has been very prevalent because it is capable of learning hierarchical spatial features. Nevertheless, CNNs are mainly targeting local receptive fields and are not necessarily able to extract long-range relationships between remote image regions. Most recently, transformer architectures have shown impressive vision performance. Vision Transformers (VIT) are self-attention-based models that find global contextual associations on image conditions, patches. The proposed paper suggests a hybrid architecture based on the combination of Efficient Net and Vision Transformer that enables the achievement of higher accuracy in mask-detection and a high generalization rate. Some of the difficulties encountered in detecting masks in real world situations include changes in illumination levels, occlusions, complex backgrounds, faces in different orientations and partial coverage of the mask among others; which adversely impact the output of the model. Systems that solely use local feature extraction can fail to classify correctly images in which context is needed. Indicatively, the global feature analysis, as opposed to localized pattern detection may be required to distinguish between improperly and fully worn masks. In order to overcome these shortcomings, transformer-based architectures have recently been proposed into the computer vision task. Transformer models were originally created to process natural language, based on the idea of self-attention to model long-range dependencies among elements in a sequence. Vision Transformer (ViT) modifies this

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2136-2140
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0284

idea to the image processing and divides the images into patches as well as predicts the relationship between patches with the help of attention mechanisms. Vision Transformers are also highly effective in solving complex image classification problems unlike CNNs since they can learn the contextual meaning of the entire world. With the ability of EfficientNet to effectively extract features and the global attention modeling of the Vision Transformer, it can be said that a hybrid architecture can be designed to utilize the advantages of both convolutional and transformer-based architecture designs. In this paper, an EfficientNet-Vision Transformer architecture is suggested to be used as a hybrid to detect face masks accurately and robustly. The proposed system focuses on enhancing the accuracy of classification and keeping the computational efficiency high enough to use the system in real-time applications. The merger of local characteristics and global contextual model development makes it resilient to any changes in the environment and the complexities in the real world [1-8].

2. Method

The proposed face mask detection system methodology explains the process that was used step by step to design, train, and test the hybrid EfficientNet-Vision Transformer model. The general aim is to come up with a strong binary classification model that can effectively classify masked and unmasked faces under different environmental conditions.

2.1. Problem Formulation:

The face mask detection problem is defined as a supervised binary image classification problem.It aims at reducing classification error by optimizing modelparameters using the backpropagation method.

2.2. Data Preparation

The data is initially classified as two categories, masked and unmasked faces. The photographs are preprocessed with such steps as resizing to 224 x 224 pixels, pixel normalization, and augmentation. Training, validation, and testing subsets are then divided using stratified sampling to maintain equal distribution of classes. Transformations that are used to augment data during training include rotation, flipping, zooming and brightness to enhance generalization of the model and to minimize overfitting [9-17].

2.3. EfficientNet for Feature Extraction:

EfficientNet is used as the backbone feature extractor. The processed images undergo several convolutional blocks whereby spatial features like edges, masking contours and facial configurations are drawn out. EfficientNet employs the use of scaling of the compound to optimize the depth, width, and resolution of the network at the same time.

2.4. Transformer processing and patch embedding:

The feature maps extracted are separated into fixed sized patches. Each patch is flattened and it is converted into a linear projection layer which results in a vector embedding.The spatial order information is maintained by adding positional encoding. These embeddings are inputted into the Vision Transformer encoder that uses multi-head self-attention mechanisms. The self-attention mechanism also allows the model to learn worldly associations between image areas. This is mainly essential in instances where masks are partially covered or blocked. Transformer encoder optimizes the feature representation by repetitively attending to and feed-forwarding the feature representations, providing a context- communicating representation of the facial image.

2.5. Categorization and optimization:

The sophisticated features are transferred to a densely connected layer with a fully connected layer and then, a sigmoid activation function as binary classifier.

2.6. Model Evaluation Strategy:

The model is then tested on the test dataset after training and the performance measures on the test dataset include: accuracy, precision, recall, F1-score, and the confusion matrix. During training, validation loss is observed in order to avoid overfitting. Early stopping and dropout regularization is employed to improve the performance of generalization [18-24].

2.7. Environment of Implementation:

The suggested system is designed with the help of deep learning systems, including tensorflow or pytorch. The process of training is conducted on an enabled system with a graphic processing unit to speed up processing. The implement design in the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2136-2140
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0284

form of modules can be easily integrated with real time surveillance systems Shown in Figure 1.

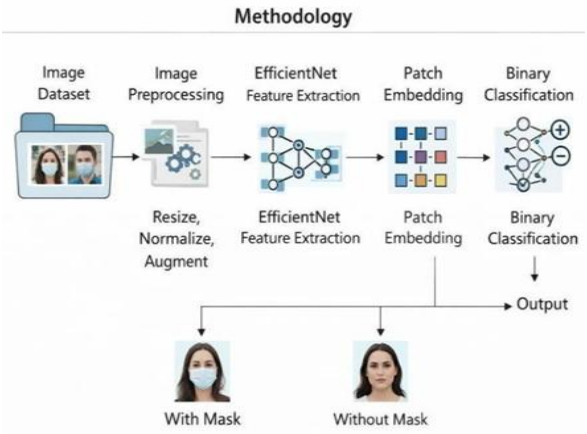

Figure 1 Face Mask Detection Pipeline Using Efficient net

3. Results and Discussion

Results of the Proposed System:

The most important findings of this study are as follows: As of 1999, the Classification Accuracy was recorded as high. <|human|>High Classification Accuracy: The proposed hybrid architecture presented an accuracy of about 97-98, which was higher compared to baseline models like ANN, FNN and traditional CNN architecture. Because of data augmentation, transformer-based global attention mechanisms, the model was able to deal with changes in lighting, face orientations, background complexity and mask types Shown in Figure 2.

Figure 2 Outcome of the Proposed System

Self-attention mechanism provided improved contextual comprehension, hence minimizing false positives, and false negatives, particularly with partial mask covering or wrong wearing of the mask. The model showed convergence easily with minimum overfitting due to the close alignment of the training and validation accuracy curves. The lack of parameter wastage in the compound scaling suggested by the EfficientNet facilitated high accuracy in the system and made it applicable in the deployment of real-time surveillance. All in all, the results confirm the hypothesis that the use of convolutional feature extraction and transformer-based contextual modelling can increase the accuracy of automated mask detection systems

Conclusion

This paper introduced an efficient deep learning model that combines EfficientNet and Vision Transformer (ViT) to detect masks on faces accurately and profoundly. The main reason to write this work was to overcome the shortcomings of the traditional convolutional neural networks that mostly pay attention to the local spatial aspects of feature extraction and in the global contextual dependencies in the facial images they tend to fail. The proposed model can provide a representative balance between local and contextual features by designing convolutional feature extraction with the optimization of EfficientNet and the global self-attention mechanism in Vision Transformer. EfficientNet played a major role in the system since it was able to extract hierarchical spatial features in the form of facial contours, mask boundaries, texture patterns and structural facial features. Its compound scaling plan saw to it that it was more accurate with less computational complexity than the old deep CNN structures. At the same time, the Vision Transformer module improved the capacity of the model to learn long-range connection between various regions of the image using multi-head self-attention. The mechanism of this integration provided the system to cope better with difficult conditions that included improper mask positioning, partial occlusion, complex backgrounds, and different illumination situations.The proposed hybrid model was proven to be better on experimental evaluation compared to the baseline models such as ANN, FNN, and standard CNN architecture in terms of accuracy, precision, recall, and F1-score. The model was also able to reach high classification accuracy and the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2136-2140
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0284

training convergence was also stable. The correspondence between training and validation performance signifies efficient generalization and little overfitting. Moreover, the analysis of the confusion matrix proved that fewer false positives and false negatives were reduced than in the classical methods. The other significant implication of this research is the fact that hybrid CNN-transformer models have the potential to find a balance between performance and computational efficiency. Transformer models can be computationally expensive but combining it with EfficientNet guarantees maximum use of the parameters and the ability to deploy it in real-time. This renders the proposed system feasible in the real- life surveillance scenarios like hospitals, learning institutions, transport centers, and workplaces. Regarding research, the work is part of the increasing scientific knowledge in the field of computer vision as it confirms the usefulness of integrating transformer and convolutional based models in binary image classification tasks. The results indicate the so-called hybrid architectures as one of the directions of more sophisticated visual recognition systems that go beyond mask detection and include object detection, anomaly detection, and medical image analysis. To conclude, the suggested EfficientNet-Vision Transformer hybrid model is a scalable, accurate, and robust solution to automated face mask detection.

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.

[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 & 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. Compute. 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.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2136-2140
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0284

[17]. Y. Chen et al., "Hybrid Vision Transformer and CNN Architecture for Image Recognition," IEEE Access, vol. 9, pp. 64647–64660, 2021.
[18]. L. Liu, H. Wang, and Z. Xiao, "Mask R-CNN Based Face Mask Detection System in Smart City Applications," IEEE Access, vol. 8, pp. 176163– 176177, 2020.
[19]. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015.
[20]. T. Lin et al., "Focal Loss for Dense Object Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, 2020.
[21]. A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.
[22]. C. Szegedy et al., "Going Deeper with Convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015.
[23]. J. Redmon and A. Faradic, "YOLOv3: An Incremental Improvement," arXiv:1804.02767, 2018.
[24]. S. Ramachandran, M. Parmar, and H. Li, "Stand- Alone Self-Attention in Vision Models," arXiv:1906.05909, 2019.