Pakistan Journal of Scientific Research
Multidisciplinary & Peer Reviewed
Volume: 5, Number: 2, Pages: 26-33, Year:2025

Article

Attention-Guided Face Mask Classification Using YOLOv8 with CBAM on MaskedFace-Net Dataset

Sadia Yaseen[1], Nimra Bukhari[2], Hafiz Muhammad Anwar Shahzada[3], Imran Ali Mudassar[2], Shabir Hussain[2,4*]

[1] Department of Computer Science, Virtual University, Pakistan
[2] Department of Computer Science, National College of Business Administration and Economics, Rahim Yar khan, 64200, Pakistan
[3] Department of Computer Science, Khawaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, Punjab, Pakistan
[4] Institute of Biopharmaceutical and Health Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

* Correspondence: Shabir Hussain (e-mail shabir.nicaas@gmail.com; shabir.hussain@sz.tsinghua.edu.cn)

Submitted: 01-10-2025**, Revised:** 25-11-2025, **Accepted:** 10-12-2025

Abstract

The COVID-19 pandemic has highlighted the critical role of face mask classification in safeguarding public health. While existing computer vision approaches primarily focus on binary mask detection, limited attention has been given to fine-grained multi-class face mask classification, which is more representative of real-world scenarios. In this work, we propose a robust and efficient YOLOv8-CBAM based deep learning framework for classifying three mask-wearing conditions: with mask, without mask, and incorrectly worn mask. Experiments on the MaskedFace-Net dataset demonstrate that the proposed YOLOv8-CBAM model achieves a macro F1-score of 99.71%, with ablation and comparative studies confirming the effectiveness of attention mechanisms, enabling accurate real-time face mask classification for surveillance and edge applications.

Keywords: YOLOv8, MaskedFace-Net, CBAM, Face Mask Classification, Deep Learning, COVID-19 Mask, Public Safety AI

1. Introduction

The emergence of the COVID-19 pandemic in late 2019 extremely altered the routine of daily life across the globe [1]. As the virus spread speedily, extraordinary efforts were organized to slow down and control its spread [2]. Global health authorities like WHO and CDC emphasized mandatory mask-wearing and distancing [3,4] to reduce virus spread, while AI and deep learning played a key role in automated pandemic solutions. These innovations helped the improvement of highly accurate and efficient face mask detection [5] and classification systems, a crucial need for automated systems in encouraging adherence with safety rules. By combining CNNs with real-time detectors like YOLO (You Only Look Once), these models identify correctly masked, incorrectly masked, and unmasked individuals in both images and live video streams [6]. CNN-based frameworks are the backbone of face mask classification [7] due to their proven efficiency, while transfer learning [8] with pre-trained models like Inception, MobileNet, VGG, and ResNet. SSD, Fast R-CNN, and YOLO enable accurate, efficient real-time mask detection [9,10] whereas traditional methods like K-NN and SVM [11,12] offer comparatively lower flexibility and

 https://doi.org/10.57041/axs8ff48

performance. The model's contribution lies in enhancing real-time face mask classification through an efficient and attention-guided deep learning framework.

- A novel attention-guided YOLOv8 framework with CBAM is proposed to enhance spatial and channel feature learning for precise face mask classification.
- The model achieves 99.7% accuracy with fast convergence using optimized training.
- Batch-optimized training maintains high accuracy with improved efficiency.

2. Related Work

DeepMaskNet [13] introduced an AlexNet-based framework for face mask detection and recognition on a custom MDMFR dataset, achieving 93.33% accuracy. Building on this, [14] proposed a CNN-based real-time system on the Face Mask 12K dataset, reaching 99% test and 98.83% real-time accuracy for reliable COVID-19 mask monitoring. In [15], a YOLOv3-based approach was implemented with a custom two-class mask dataset in the Darknet framework, achieving 96% real-time detection accuracy. In the post-COVID-19 context, [16] used MobileNetV2 for mask detection, achieving 96.85% accuracy, while [17] combined HSV texture and portrait line features with CNNs to reach 93.64%. The study in [18] applied MobileNetV2 with SVM and K-NN on 1,376 images, achieving 97.1%, and [19] reported 97% accuracy for real-time mask detection using a CNN-based model. The RRFMDS system in [20] combines SSD detection with MobileNetV2 for real-time mask classification on 14,535 images, achieving 99.15% accuracy rate.

3. Methodology

3.1. Dataset

The framework is evaluated on the MaskedFace-Net dataset [21] with 8,982 images evenly split across three mask conditions, featuring diverse illumination, poses, backgrounds, and mask styles for robust real-world generalization. All images are provided in the RGB color space and resized to 128 × 128 pixels for model training. Sample images from each class are shown in Fig. A.

Figure A: Dataset showing face images categorized as masked, incorrectly masked, and without mask.

3.2. Pre Processing

A) Face-centering Cropping

To eliminate background bias and focus learning a lightweight face detector is used to localize facial regions, which are then cropped and aligned before being passed to the classifier. This ensures that critical facial features such as the nose, mouth, and chin are consistently emphasized.

B) Illumination Normalization

To address variations in lighting conditions, a contrast-limited adaptive histogram equalization (CLAHE) is applied to the cropped face images. CLAHE improves local contrast and preserves fine facial details without introducing noise amplification.

C) Image Resizing and Normalization

All preprocessed face images are uniformly resized to 128×128 pixels, balancing spatial detail preservation and memory efficiency. Pixel intensity values are normalized to stabilize gradient updates and improve convergence during training.

D) Compliance-Aware Data Augmentation

A data augmentation approach specific to the task is adopted to enhance generalization and balance classes. Horizontal flips vary viewpoints, while partial occlusion introduces realistic coverage scenarios. This augmentation shifts mask regions to expose the nose or chin, helping the model learn subtle cues of improper mask wearing often underrepresented in real data.

3.3. Proposed Framework

As illustrated in Fig. B, the proposed framework adopts YOLOv8 in classification mode as the backbone network due to its favorable balance between accuracy and computational efficiency. YOLOv8 employs a hierarchical feature extraction backbone and multi-scale feature aggregation, enabling effective representation learning while supporting real-time inference requirements. To enhance fine-grained mask classification, the backbone integrates CBAM, which applies channel and spatial attention to emphasize key facial patterns and regions like the nose and mouth, improving accuracy with minimal computational overhead.

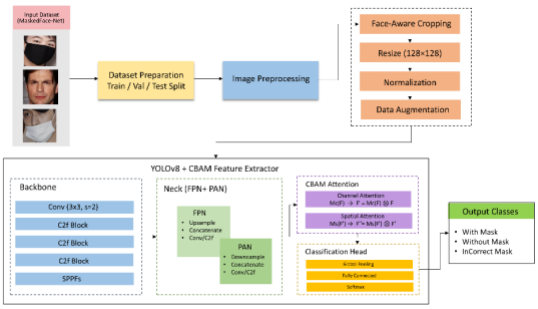

Figure B: Overall framework of the proposed YOLOv8-CBAM–based face mask classification system.

A) Yolov8 Backbone and Neck Module

In the proposed framework, YOLOv8 serves as the core feature extraction network, consisting of a backbone followed by a neck module, as illustrated in Fig. C.

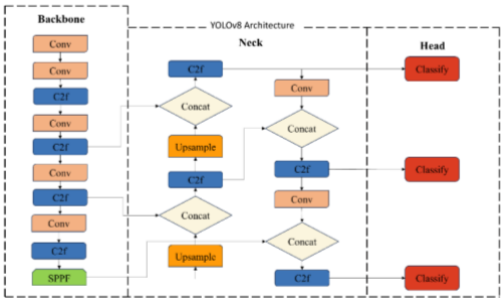

Figure C: Detailed architecture of YOLOv8-based architecture for multi-class face mask classification.

Given a preprocessed input image in Equation.1. The backbone generates deep feature maps in Equation 2 where $B(\cdot)$ denotes the YOLOv8 backbone. The output of the neck in Equation.3 is represented as where $N(.)$ denotes the FPN–PAN neck.

$$I \in \mathbb{R}^{H\times W\times C} \quad (1)$$

$$F_b=B(I), \quad (2)$$

$$Fn = N(Fb), \quad (3)$$

B) Attention-Guided Feature Refinement

While YOLOv8 effectively extracts discriminative facial features, to enhance sensitivity to such fine-grained patterns, CBAM refines the YOLOv8 feature maps by sequentially applying channel attention followed by spatial attention shown in Fig D.

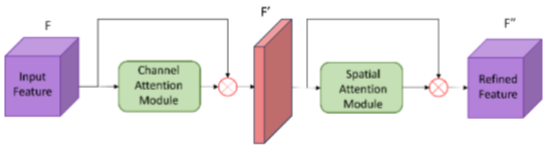

Figure D: CBAM structure illustrating sequential channel and spatial attention for feature refinement.

In Equation. 4 Given an input feature map F, channel attention produces, where Mc represents the channel attention map and , spatial attention highlights discriminative facial regions in Equation 5.

$$Fc = Mc \otimes F, \quad (4)$$

$$Fcs = Ms \otimes Fc, \quad (5)$$

C) Classification Head and Output

The classification head is designed to be lightweight while enabling effective multi-class discrimination. Let $\boldsymbol{Fcs}$ denote the refined feature representation obtained after CBAM processing. In Equation. 6 Global average pooling (GAP) is first applied to reduce spatial dimensions and obtain a compact feature vector $\boldsymbol{z}$, The pooled feature vector is then passed through a fully connected layer followed by a softmax activation in Equation.7 to compute class probabilities.

$$z=GAP(Fcs) \quad (6)$$

$$\hat{Y}=\text{softmax}(Wz+b) \quad (7)$$

Where $\hat{Y}$ denotes the predicted probabilities for “with mask,” “incorrect mask,” and “without mask,”. This classification strategy enables efficient making the proposed framework suitable for face mask classification

3.4. Trainable Parameters

The YOLOv8-CBAM model, with 99 layers and 1.44M parameters, uses convolutional and C2f blocks for efficient feature extraction and classifies face masks into three classes with 3.4 GFLOPs, enabling real-time, high-accuracy inference. As reported in Table I, the attention-enhanced architecture achieves high classification accuracy across all classes.

Table I: Configuration of essential hyperparameters used across all face mask classification models including learning rate, batch size, optimizer, and epoch settings.

From	Layer	Arguments	Params	Module
-1	0	[3, 16, 3, 2]	464	Conv
-1	1	[16, 32, 3, 2]	4672	Conv
-1	2	[32, 32, 1, True]	7360	C2f
-1	3	[32, 64, 3, 2]	18560	Conv
-1	4	[64, 64, 2, True]	49664	C2f

-1	5	[64, 128, 3, 2]	73984	Conv
-1	6	[128, 128, 2, True]	197632	C2f
-1	7	[128, 256, 3, 2]	295424	Conv
-1	8	[256, 256, 1, True]	460288	C2f
-1	9	[256, 3]	334083	Classify

3.5. Systems and Packages Details

The face mask model is implemented with Ultralytics YOLOv8 (v8.3.70) in Python 3.11.11 using PyTorch 2.0.1 and CUDA 11.8 on an NVIDIA Tesla T4 GPU (15 GB), enabling efficient, real-time, and accurate classification.

3.6. Training Configuration and Optimizers Details

The YOLOv8s-cls model trains for 50 epochs on 128×128 images with a batch size of 16, leveraging 8 data loader workers and AMP for efficient, memory-optimized training, while a validation split monitors generalization. The model is fine-tuned using AdamW with 0.0005 weight decay, 0.9 momentum, and a 0.001429 learning rate, promoting rapid convergence, stable learning, and improved generalization.

4. Results and Discussion

4.1. Quantitative Results

A) Class-wise performance evaluation

Table IV compares the proposed YOLOv8-CBAM framework with existing models on the MaskedFace-Net dataset. Conventional CNNs such as MobileNetV2 and EfficientNet-B0 perform well but struggle with fine-grained incorrect mask detection. YOLO-based models improve feature extraction, and YOLOv8-CBAM achieves the highest macro F1-score of 0.9971, demonstrating the benefit of attention-guided feature refinement.

Table II: Class-wise Precision, Recall, and F1 for face mask classification.

Class	Precision (AP)	Recall (AR)	F1-Score
Incorrect Mask	0.9978	0.9978	0.9978
With Mask	0.9978	0.9933	0.9956
Without Mask	0.9955	1.0000	0.9978
Macro Avg.	0.9970	0.9970	0.9971

B) Training Convergence Analysis

Figure E shows the YOLOv8-CBAM training and performance where E(a) presents training and validation loss converge steadily, indicating stable learning and good generalization, with Top-1 accuracy reaching 99.7% and Top-5 at 100% and E(b) higlights the confusion matrix confirms reliable class-wise performance.

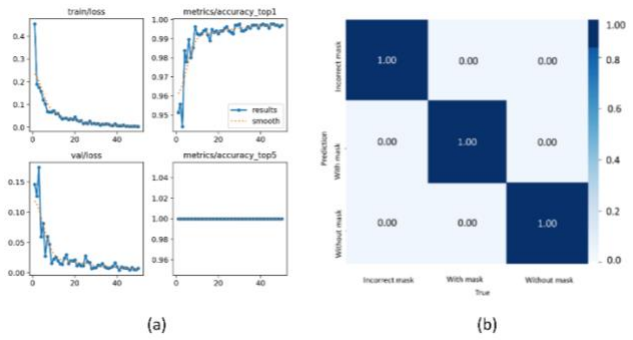

Figure E: YOLOv8 evaluation: (a) training/validation loss, (b) confusion matrix.

4.2. Ablation Study Interpretation

Table III shows the ablation study: the baseline YOLOv8 achieves 0.9912 macro F1, improved by face-aware cropping (0.9936) and task-specific augmentation (0.9951). The full YOLOv8-CBAM model reaches 0.9971, highlighting the incremental benefits of each component.

Table III: Ablation study illustrating the contribution of face-aware preprocessing, task-specific data augmentation, and CBAM-based attention to overall performance.

Configuration	Face-Aware Cropping	Task-Specific Augmentation	CBAM Attention	Macro F1-Score
Baseline YOLOv8	✗	✗	✗	0.9912
YOLOv8 + Face Cropping	✓	✗	✗	0.9936
YOLOv8 + Task-Specific Augmentation	✓	✓	✗	0.9951
YOLOv8-CBAM (Proposed)	✓	✓	✓	**0.9971**

4.3. Comparative Performance Analysis

Table IV compares YOLOv8-CBAM with existing models on MaskedFace-Net. While CNNs like MobileNetV2 and EfficientNet-B0 perform well, YOLO-based models excel at detecting fine-grained incorrect mask patterns, with YOLOv8-CBAM achieving the highest macro F1 of 0.9971 through attention-guided feature refinement.

Table IV: Comparative Performance Analysis on MaskedFace-Net

Method	Architecture	Classes	Macro F1-Score
MobileNetV2 [22]	CNN	3	97%
EfficientNet-B0 [23]	CNN	3	99%
YOLOv5-Cls [24]	YOLO-based	3	99.1%
YOLOv8 (Baseline)	YOLO-based	3	99.4%
YOLOv8-CBAM (Proposed)	Attention-Enhanced YOLO	3	**99.71%**

5. Conclusions

In this study, we presented a deep learning–based face mask classification system built on the YOLOv8-CBAM architecture and evaluated using the MaskedFace-Net dataset. The proposed model classifies facial images into three practical categories: with mask, incorrectly worn mask, and without mask. By integrating efficient feature extraction, multi-scale fusion, and attention-

guided refinement, the system achieves accurate and fast predictions while preserving important facial details. Experimental results demonstrate strong generalization across variations in pose, illumination, and mask placement. The proposed framework offers an effective solution for real-time face mask classification and provides a foundation for future extensions toward video-based monitoring and large-scale deployment in public safety applications.

Author Contributions: Conceptualization, S.Y., N.B., and S.H.; methodology, S.Y. and N.B.; software, N.B.; validation, H.M.A.S. and S.H.; formal analysis, S.Y.; investigation, S.Y.; resources, I.A.M.; data curation, N.B.; original draft preparation, S.Y.; review and editing, N.B., H.M.A.S., and S.H.; supervision, S.H.; project administration, S.H.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hussain, S., Munir, H., Bukhari, N., Yu, Y., Mudassar, I. A., Khan, A., ... & Wahid, J. A. (2025). HybridSVG: Ensemble Framework for Detecting Spatially Variable Genes in Spatial Transcriptomics using Fusion of Global and Local Autocorrelation. *Pakistan Journal of Scientific Research*, *5*(1), 54-62.
2. Zang, L., Liu, J., Zhang, H., Zhu, S., Zhu, M., Wang, Y., ... & Xu, Q. (2025). A deep learning model based on Mamba for automatic segmentation in cervical cancer brachytherapy. *Scientific Reports*, *15*(1), 10152.
3. Husssain, S., Ayoub, M., Wahid, J. A., Khan, A., Alabrah, A., & Amran, G. A. (2024). Cough2COVID-19 detection using an enhanced multi layer ensemble deep learning framework and CoughFeatureRanker. *Scientific Reports*, *14*(1), 25207
4. Traore, M., Hancer, E., Samet, R., Yıldırım, Z., & Nemati, N. (2024). CompSegNet: An enhanced U-shaped architecture for nuclei segmentation in H&E histopathology images. *Biomedical Signal Processing and Control*, *97*, 106699.
5. Hussain, S., Wahid, J. A., Ayoub, M., Tong, H., & Rehman, R. (2023). Automated segmentation of coronary arteries using attention-gated unet for precise diagnosis. *Pakistan Journal of Scientific Research (PJOSR)*, *3*(1), 124-129.
6. Rauf, Z., Khan, A. R., & Khan, A. (2024). Channel Boosted CNN-Transformer-based Multi-Level and Multi-Scale Nuclei Segmentation. *arXiv preprint arXiv:2407.19186*.
7. Hussain, S., Amran, G. A., Alabrah, A., Alkhalil, L., & AL-Bakhran, A. A. (2024). C19-MLE: A Multi-Layer Ensemble Deep Learning Approach for COVID-19 Detection Using Cough Sounds and X-ray Imaging. *IEEE Access*.
8. Gou, F., Tang, X., Liu, J., & Wu, J. (2024). Artificial intelligence multiprocessing scheme for pathology images based on transformer for nuclei segmentation. *Complex & Intelligent Systems*, *10*(4), 5831-5849.
9. Hussain, S., et al., IoT and deep learning based approach for rapid screening and face mask detection for infection spread control of COVID-19. Applied Sciences, 2021. **11**(8): p. 3495.
10. Bukhari, N., Hussain, S., Ayoub, M., Yu, Y., & Khan, A. (2022). Deep learning based framework for emotion recognition using facial expression. *Pakistan Journal of Engineering and Technology*, *5*(3), 51-57
11. Hussain, S., et al., *Ensemble Deep Learning Framework for Situational Aspects-Based Annotation and Classification of International Student's Tweets during COVID-19.* Computers, Materials & Continua, 2023. **75**(3).
12. Bukhari, N., Munir, H., Haider, R., & Hussain, S. (2025). MASKYLO: Hybrid Deep Learning Framework for Detection and Segmentation of HE Stained Histology Image. *Pakistan Journal of Scientific Research*, *4*(2 (Suppl.)), 92-101.

13. Bancila, I. C. (2025). Mini-review: Experimental Approaches for the Biomechanical Testing of Bone. *International Journal of Emerging Engineering and Technology*, *4*(1), 11-20.
14. Awodeyi, A., Ibok, O. A., Ekwemuka, J. U., Idama, O., & Odesa, E. (2025). Enhancing Facial Recognition Performance with Data Augmentation in Occluded Environments. *International Journal of Emerging Engineering and Technology*, *4*(1), 27-32.
15. Huang, H., Gong, T., He, K., Wu, J., Cambria, E., & Feng, M. (2025). Robust Multimodal Sentiment Analysis via Double Information Bottleneck. *Information Fusion*, 103964.
16. Hestrio, Y. F., Mantau, A. J., & Jatmiko, W. (2025). GWSC-SegMamba: Gate Wavelet Spatial Convolution Enhanced State Space Model for Multi-Temporal Agricultural Land Segmentation. *IEEE Access*.
17. Zhu, H., Huang, Y., Yao, K., Shang, J., Hu, K., Li, Z., & He, G. (2025). AttmNet: a hybrid Transformer integrating self-attention, Mamba, and multi-layer convolution for enhanced lesion segmentation. *Quantitative Imaging in Medicine and Surgery*, *15*(5), 4296.
18. Fatima, S., Haider, N. G., & Riaz, R. (2024). YOLOv8 vs RetinaNet vs EfficientDet: A comparative analysis for modern object detection. *International Journal of Emerging Engineering and Technology*, *3*(2), 1-5.
19. Mahbod, A., Polak, C., Feldmann, K., Khan, R., Gelles, K., Dorffner, G., ... & Ellinger, I. (2024). Nuinsseg: a fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. *Scientific Data*, *11*(1), 29
20. Xing, Z., Ye, T., Yang, Y., Cai, D., Gai, B., Wu, X. J., ... & Zhu, L. (2025). Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*.
21. Wang, Z., Zou, Y., & Liu, P. X. (2021). Hybrid dilation and attention residual U-Net for medical image segmentation. *Computers in biology and medicine*, *134*, 104449.
22. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574-584).
23. Zhao, Y., Liu, C., Zhou, X., & Zhang, X. (2024, November). SegUMamba: Integrating Mamba with U_net for Medical Image Segmentation. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (pp. 108-111). IEEE.
24. Zhao, Yang, G., et al. *Face mask recognition system with YOLOV5 based on image recognition*. in *2020 IEEE 6th international conference on computer and communications (ICCC)*. 2020. IEEE.