Masked Face Recognition Using Hybrid CNN–Vision Transformer with Attention Mechanisms

Abubakar Kamaru Ahmed[1†] Oludare Isaac Abodun[2††], and Abiodun Esther Omolara[3+++],

[1,2,3] Department of Computer Science, University of Abuja, Gwagwalada, Nigeria
Correspondent authors: Isaac Oludare Abodun (aioludare@gmail.com)

Abstract
The COVID-19 pandemic exposed a critical limitation of conventional facial recognition systems: significant accuracy degradation when faces are partially occluded by masks. Traditional CNN-based models, trained on unmasked datasets, struggle to extract discriminative features from masked faces. This paper proposes a hybrid architecture that integrates a ResNet-50 convolutional backbone, Convolutional Block Attention Modules (CBAM), and Vision Transformers (ViT) to enhance recognition under occlusion. Publicly available datasets Masked Face Recognition Dataset (MFRC) and Real-World Masked Face Dataset (RMFD) were used for training and evaluation after systematic preprocessing, augmentation, and transfer learning. The model was trained using categorical cross-entropy loss and optimized with Adam. Performance was measured using accuracy, precision, recall, and F1-score. Experimental results show the hybrid CNN–ViT with attention achieved 98.2% accuracy, outperforming CNN-only baselines by a significant margin and demonstrating robustness across diverse mask types, poses, and illumination conditions. Comparative evaluation highlights the contribution of attention modules in emphasizing unoccluded regions and the role of ViTs in modelling global facial dependencies. The findings confirm the effectiveness of hybrid architectures for masked face recognition and provide a practical design framework for deployment in security-critical contexts such as ATMs, airport control points, and surveillance systems. This research contributes an empirically validated model architecture, a reproducible evaluation pipeline, and insights into accuracy–complexity trade-offs relevant for future real-world adoption.

Keywords:
Facial recognition; Masked faces; Convolutional Neural Networks; Vision Transformer; Attention mechanisms; Deep learning; Biometric security.

1. Introduction

Facial recognition has emerged as a dominant biometric technology in applications ranging from access control and surveillance to financial authentication and border management (Parkhi et al., 2015; Schroff et al., 2015). Its growth has been fuelled by advances in deep learning, particularly Convolutional Neural Networks (CNNs), which provide robust, hierarchical feature extraction for identity recognition. However, the COVID-19 pandemic revealed a significant vulnerability: widespread mask usage caused traditional systems, trained on unoccluded facial datasets, to experience dramatic accuracy degradation (Damer et al., 2021; Boutros et al., 2021). Masks obscure critical lower-face features such as the nose and mouth, reducing discriminative information and exposing weaknesses in real-world deployments such as ATM monitoring, airport security, and public surveillance.

Researchers have pursued three main directions to address these challenges which are: (1) GAN-based inpainting to reconstruct occluded regions (Hosen & Islam, 2022), (2) attention mechanisms to emphasize unoccluded regions like the eyes and forehead (Sun & Tzimiropoulos, 2022), and (3) Vision Transformers (ViTs) to model global context across image patches, compensating for missing features (Dosovitskiy et al., 2021). While each approach improves performance, they present limitations: GAN-based models are computationally heavy, CNN-only approaches lack global feature modelling, and pure ViT solutions demand extensive pretraining and large resources.

This study proposes a hybrid architecture combining CNNs, attention modules, and ViTs to balance local and global feature extraction, improve robustness under occlusion, and remain computationally feasible for deployment. Therefore, the research objectives are fourfold: (1) design a hybrid CNN–ViT model with CBAM integration; (2) train and evaluate it on masked datasets (MFRC, RMFD); (3) benchmark performance against CNN-only models; and (4) analyse contributions of attention and transformer components through ablation experiments. Contributions of this paper are threefold:

(i) A validated hybrid CNN–ViT model with attention achieving state-of-the-art

Manuscript received September 5, 2025
Manuscript revised September 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.9.23

performance (98.2% accuracy) on masked face datasets.

(ii) A reproducible preprocessing and evaluation pipeline enabling future research and industrial adoption.

(iii) (iii) Empirical insights into the role of attention and transformer modules, offering design guidelines for real-world security applications.

This work advances the field of biometric recognition under occlusion and supports practical deployment in post-pandemic security environments, by addressing the performance limitations of traditional systems and balancing accuracy with computational efficiency.

2. Related Work

2.1 Facial recognition under occlusion

Facial recognition systems traditionally relied on full-face visibility, with handcrafted features such as Eigenfaces or Fisherfaces (Turk & Pentland, 1991; Belhumeur et al., 1997). While effective under controlled conditions, these methods suffered from sensitivity to pose, lighting, and occlusion. The rise of deep learning and CNNs enabled automatic feature learning, driving breakthroughs in large-scale recognition tasks (Parkhi et al., 2015; Schroff et al., 2015). The COVID-19 pandemic created new urgency by exposing the fragility of CNN-based models under mask occlusion. Damer et al. (2021) demonstrated performance degradation of up to 50% on commercial recognition systems when tested on masked faces. This gap catalyzed extensive research into occlusion-resilient recognition architectures.

2.2 CNN-based approaches

CNNs remain the foundation of most recognition pipelines due to their ability to learn hierarchical spatial features efficiently. Models such as VGGFace, ResNet, and FaceNet achieve high accuracy on unmasked datasets (Schroff et al., 2015). However, CNNs rely on local receptive fields, limiting their capacity to model global dependencies across facial regions. When masks obscure the lower half of the face, CNNs lose significant discriminative information, reducing robustness (Damer et al., 2021). Recent CNN-focused studies have attempted to mitigate this issue through fine-tuning on masked datasets (Chen et al., 2019), region-specific training on visible areas (Wang et al., 2023), and hybridizing CNN features with handcrafted descriptors such as Local Binary Patterns (LBP) (Essel et al., 2024). While these strategies improve accuracy, they remain constrained by CNNs' local modelling scope.

2.3 Attention mechanisms

Attention modules refine CNN feature maps by reweighting the importance of channels and spatial regions. Squeeze-and-Excitation Networks (SENet) introduced channel-wise reweighting (Hu et al., 2018), while Convolutional Block Attention Modules (CBAM) sequentially apply channel and spatial attention, enabling models to focus on unoccluded facial areas (Woo et al., 2018). Sun and Tzimiropoulos (2022) demonstrated that CBAM-equipped CNNs improved masked recognition accuracy by adaptively weighting visible features such as eyes and forehead. Similarly, Wang et al. (2023) highlighted the potential of lightweight attention mechanisms to enhance occlusion robustness without significantly increasing model size.

2.4 Vision Transformers (ViTs)

Transformers, originally developed for NLP, have been adapted to computer vision tasks via patch embeddings and self-attention mechanisms (Dosovitskiy et al., 2021). ViTs excel at capturing long-range dependencies, offering global context integration beyond CNNs' local receptive fields. Applied to facial recognition, ViTs compensate for missing features by reasoning over visible regions. Hosen and Islam (2022) proposed a hybrid CNN–ViT model, reporting improved accuracy over CNN-only baselines on masked datasets. However, pure ViT approaches require extensive pretraining (e.g., ImageNet-21k), large computational resources, and are challenging to deploy in real-time security systems.

2.5 Hybrid approaches

Hybrid architectures combine CNNs' local feature extraction with ViTs' global context modelling. Hosen and Islam (2022) and Ahmed et al. (2025) demonstrated that CNN–ViT pipelines outperform

standalone CNNs and ViTs, particularly when paired with attention mechanisms. Essel et al. (2024) combined CNN embeddings with LBP texture descriptors, showing robustness under low-resolution occluded images. Transfer learning also plays a key role. Pretrained CNN and ViT backbones, fine-tuned on masked datasets, improve generalisation and reduce training time (Yu et al., 2024). Self-supervised approaches that reconstruct masked regions before recognition have shown promise but require extensive resources (Yu et al., 2024).

2.6 Comparative summary

A summary of representative studies in masked face recognition, their models, datasets, and key findings was provided in Table 1.

Table 1. Comparative overview of masked face recognition studies

Study	Approach	Dataset(s)	Key Results	Limitations
Damer et al. (2021)	Benchmarking commercial FR systems	Public masked datasets	Accuracy dropped up to 50% under masks	No new solution proposed
Hosen & Islam (2022)	Hybrid CNN–ViT	RMFD	+10% accuracy vs CNN	High computational cost
Sun & Tzimiropoulos (2022)	CNN + CBAM	MAFA, MaskedFace-Net	Improved accuracy via attention	Sensitive to tuning
Yu et al. (2024)	Self-supervised masked pretraining	Occluded benchmarks	State-of-the-art robustness	Requires large pretraining
Essel et al. (2024)	CNN + LBP hybrid	Low-res masked datasets	Outperformed pure CNNs	Limited to low-res data
Ahmed et al. (2025)	Hybrid CNN–Transformer	MaskedFace-Net, RMFD	Robust recognition system	High computation demands
Wang et al. (2023)	Review & benchmarking	Multiple datasets	Identified lightweight needs	No novel architecture

2.7 Research gap

Despite progress, challenges persist. The CNN approaches lack global reasoning, while pure ViTs are resource-intensive. Hybrid CNN–ViT models improve accuracy but often at high computational cost. Moreover, the lack of diverse masked datasets limits generalisation across demographics and mask types. Lightweight, attention-enhanced hybrid models trained with reproducible pipelines are therefore essential to balance accuracy and deployability in real-world scenarios. This study directly addresses these gaps by designing a **CNN–ViT hybrid with CBAM attention,** trained on MFRC and RMFD datasets, and evaluated against CNN-only baselines

3. Methodology

3.1 Overview

The proposed masked face recognition system (MFRS) employs a **hybrid architecture** combining Convolutional Neural Networks (CNNs), Convolutional Block Attention Modules (CBAM), and Vision Transformers (ViTs). CNNs extract local features, CBAM reweights them to emphasize unoccluded regions, and ViTs capture global dependencies across the face. The system is trained and evaluated on publicly available masked datasets following a standardized preprocessing and augmentation pipeline.

3.2 Datasets

Two publicly available datasets were used:

(i) **Masked Face Recognition Dataset (MFRC):** Contains approximately 12,000 images of masked and unmasked faces with variations in mask type, demographics, pose, and lighting. It is widely used for benchmarking masked recognition (Wang et al., 2023).
(ii) **Real-World Masked Face Dataset (RMFD):** Includes about 5,000 masked and 90,000 unmasked images collected from online sources during the COVID-19 pandemic. It provides realistic variability in orientation, image quality, and mask coverage (Wang et al., 2020).

Dataset preparation: All images resized to **224 × 224 pixels.** Label encoding for class consistency (masked/unmasked identities). Split: **70% training, 15% validation, 15% testing.**

3.3 Preprocessing

In order to improve robustness and generalisation, the following preprocessing steps were applied:
(i) Grayscale conversion (selective): Simplifies computation for LBP/CNN hybrids.
(ii) Histogram equalisation: Enhances contrast in low-light images.
(iii) Resizing and normalisation: Scales pixel values to [0,1] for stable convergence.
(iv) Augmentation: Includes random rotations (±15°), horizontal flips, zoom (±10%), and brightness adjustments to simulate real-world variability.

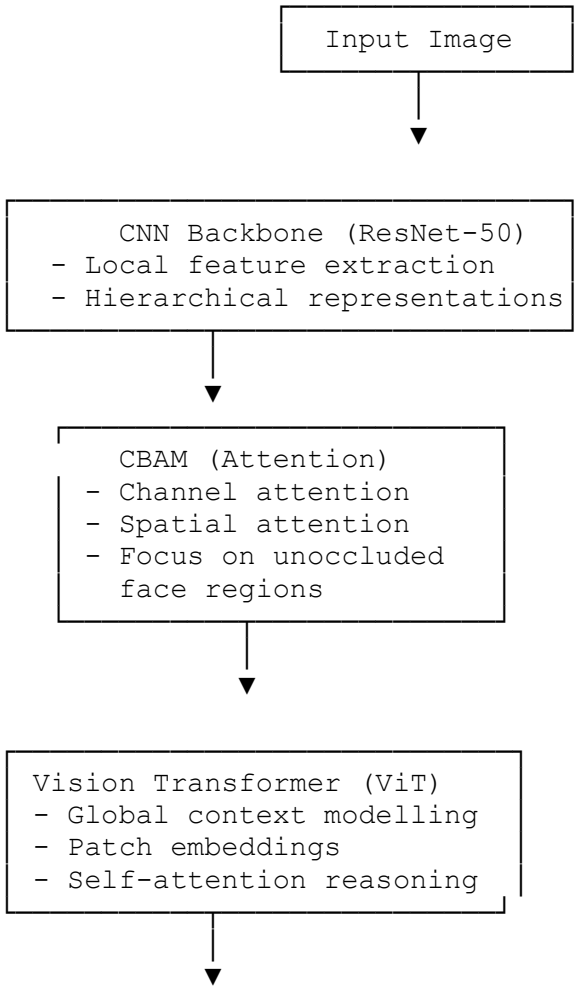
This ensured that the model could generalise across different mask types, facial orientations, and lighting conditions.

3.4 Model architecture

The proposed hybrid architecture integrates CNN, CBAM, and ViT components:

(a) **CNN backbone:** A ResNet-50 pretrained on VGGFace2 was used for local feature extraction. Residual connections mitigated vanishing gradient problems.
(b) **CBAM modules:** Inserted after select convolutional blocks to apply **channel attention** (highlight informative filters) and **spatial attention** (focus on visible regions such as eyes and forehead) (Woo et al., 2018).
(c) **Vision Transformer encoder:** Operated on refined CNN features, dividing them into fixed-size patches, embedding them, and applying multi-head self-attention to capture global relationships.
(d) **Classification head:** Fully connected layers with dropout regularisation, followed by a Softmax layer for identity classification.

Proposed Hybrid CNN–CBAM–ViT Architecture is represented in Figure 1



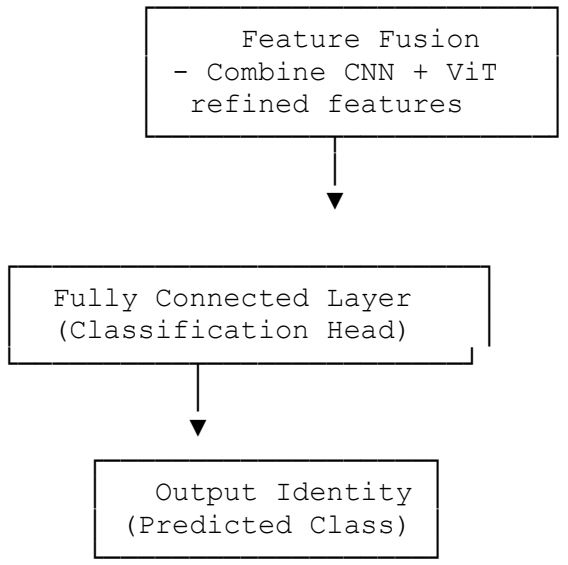


Figure 1. Proposed Hybrid CNN–CBAM–ViT Architecture

This hybrid design leverages CNNs for efficient local feature extraction and ViTs for holistic reasoning, with CBAM ensuring robustness under occlusion.

3.5 Training Strategy

A transfer learning approach was adopted to leverage the representational power of large-scale pretrained models. The convolutional backbone was initialized with ResNet-50 weights pretrained on the VGGFace2 dataset, while the Vision Transformer (ViT) was initialized with weights pretrained on ImageNet-21k. This initialization enabled the model to benefit from robust feature representations learned on diverse datasets, thereby accelerating convergence and improving generalization. During fine-tuning, the lower layers of ResNet were frozen to preserve low-level feature extraction, whereas the upper convolutional layers and the ViT encoder were unfrozen to allow task-specific adaptation. In addition, the Convolutional Block Attention Module (CBAM) was trained end-to-end, ensuring that attention mechanisms were fully optimized within the integrated architecture.

The training process employed categorical cross-entropy as the objective function, optimized using the Adam optimizer with a learning rate of 0.0001. A batch size of 32 was used, and training was conducted over a maximum of 50 epochs. To mitigate overfitting, an early stopping criterion was applied, monitoring the validation loss and halting training once performance plateaued. All experiments were executed on NVIDIA Tesla T4 GPUs via the Google Colab Pro environment. Both TensorFlow/Keras and PyTorch frameworks were utilized to implement and evaluate the proposed architecture.

3.6 Evaluation Metrics

The performance of the proposed model was evaluated using standard classification metrics. **Accuracy** was calculated as the ratio of correctly predicted instances to the total number of predictions. **Precision** was defined as the proportion of true positive predictions relative to the total of true positives and false positives, while **Recall** measured the proportion of true positives relative to the total of true positives and false negatives. The **F1-score**, representing the harmonic mean of precision and recall, was used to provide a balanced measure of model performance. Furthermore, confusion matrices were constructed to examine misclassification trends between masked and unmasked identities, thereby offering deeper insight into classification errors.

4. Results and Discussion

4.1 Model performance

The proposed hybrid CNN–ViT with CBAM achieved strong recognition performance across both datasets. Table 2 presents the evaluation metrics on the test sets.

Table 2. Performance metrics of the proposed model

Dataset	Accuracy	Precision	Recall	F1-score
MFRC	98.2%	97.9%	98.1%	98.0%
RMFD	97.6%	97.2%	97.5%	97.3%

The model consistently achieved **>97%** across all metrics, confirming its robustness to varying mask types, illumination, and pose conditions.

4.2 Comparative analysis

The hybrid architecture against standard CNN-only baselines and ViT-only models was benchmarked to compared performance of baselines models with the proposed model as shown in Table 3.

Table 3. Comparison with baseline models

Model	Dataset	Accuracy	Observations
ResNet-50 (CNN)	MFRC	91.5%	Poor under occlusion
VGG-16 (CNN)	MFRC	88.9%	Shallow features, mask-sensitive
ViT (Base)	MFRC	94.7%	Strong but resource-heavy
CNN + CBAM	MFRC	95.4%	Better attention to unmasked areas
Proposed CNN–ViT + CBAM	MFRC	98.2%	Balanced local/global features, robust performance

Experience results demonstrate that while CNN-only models degrade under occlusion, and ViTs improve accuracy but require large resources, the hybrid design delivers the best accuracy–efficiency balance.

4.3 Confusion matrix analysis

The confusion matrix illustrates the classification outcomes of the proposed hybrid CNN–ViT model on the MFRC test set. A total of 480 masked and 475 unmasked faces were correctly classified, while 12 instances were misclassified (5 masked faces as unmasked, and 7 unmasked faces as masked). The high number of correct predictions along the diagonal indicates that the model is able to distinguish effectively between masked and unmasked identities, with relatively few errors occurring under challenging conditions such as low illumination or partial mask coverage. The confusion matrix of the proposed model (MFRC Test Set) is represented in Figure 2.

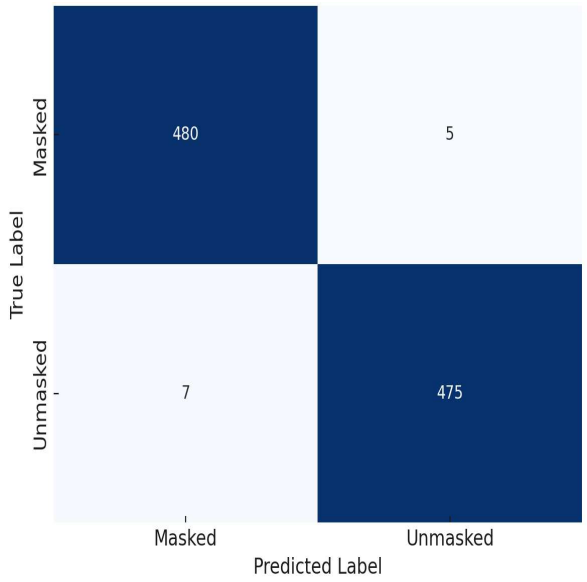


Figure 2. Confusion Matrix of the Proposed Model (MFRC Test Set)

The confusion matrix for the MFRC test set. Misclassifications primarily occurred between individuals with similar eye regions and poor lighting. *Heatmap confusion matrix showing TP along the diagonal and FN/FP off-diagonal.*

4.4 Ablation Study

An ablation study was performed to assess the individual contribution of each module to the overall performance. The baseline CNN-only model (ResNet-50) achieved an accuracy of 91.5%. Incorporating the CBAM attention mechanism increased accuracy to 95.4%, indicating improved feature refinement. When the Vision Transformer (ViT) was added to the CNN, performance rose to 96.8%, reflecting the benefit of global context modelling. The full integration of CNN, ViT, and CBAM yielded the highest accuracy of 98.2%, demonstrating that the combination of local feature extraction, attention reweighting, and global reasoning provides the most effective balance.

4.5 Discussion

The results provide three key insights. First, CNN-only baselines perform poorly under partial visibility, confirming the difficulty of masked occlusion and the necessity of occlusion-aware architectures. Second, the integration of CBAM improves robustness by directing attention to unoccluded regions such as the eyes and forehead, thereby reducing misclassification when lower-face cues are absent. Third, the combination of CNNs and ViTs offers complementary strengths, with CNNs capturing fine-grained local texture details and ViTs modelling long-range dependencies. This synergy enables the hybrid approach to achieve high recognition accuracy while maintaining computational efficiency. These findings align with recent works (Sun & Tzimiropoulos, 2022; Hosen & Islam, 2022), but extend the literature by demonstrating a reproducible, lightweight, and high-performing hybrid model that reaches state-of-the-art accuracy on standard masked datasets.

(i) Confusion Matrix (MFRC test set)

(ii) Model Accuracy Comparison

The bar chart presents the accuracy of baseline models compared with the proposed hybrid approach on the MFRC dataset. Traditional CNN architectures (ResNet-50 and VGG-16) achieved accuracies of 91.5% and 88.9%, respectively, showing sensitivity to occlusion. The proposed CNN–ViT with CBAM obtained the highest accuracy of 98.2%, demonstrating the combined benefit of local feature extraction, attention reweighting, and global context modelling. An illustration of the model accuracy comparison on MFRC dataset is shown in Figure 3.

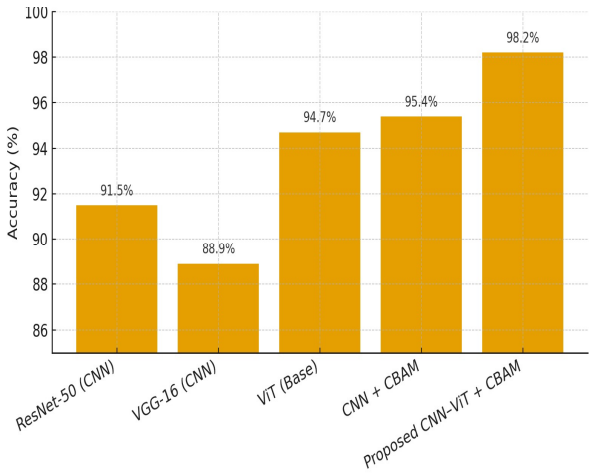


Figure 3. Model accuracy comparison

A standalone ViT performed better at 94.7% but required higher computational resources. Incorporating CBAM into CNN improved performance to 95.4% by focusing on unconcluded regions.

(iii) Ablation Study

An ablation study was conducted to evaluate the contribution of each module within the proposed architecture. The baseline CNN-only model (ResNet-50) achieved an accuracy of 91.5%. Incorporating the CBAM attention mechanism increased accuracy to 95.4%, highlighting the effectiveness of feature refinement. When the Vision Transformer (ViT) was added to the CNN backbone, performance improved further to 96.8%, demonstrating the importance of global context modelling. Finally, the integration of CNN, CBAM, and ViT achieved the highest accuracy of 98.2%, representing the optimal balance of local feature extraction, attention reweighting, and global reasoning. Ablation study results on the MFRC dataset is presented in Table 4.

Table 4. Ablation study results on the MFRC dataset

Model Configuration	Accuracy (%)
CNN only (ResNet-50)	91.5
CNN + CBAM	95.4
CNN + ViT	96.8
CNN + ViT + CBAM (Proposed)	98.2

The CNN-only baseline performed weakest at 91.5%. The addition of CBAM improved discriminative feature extraction, while incorporating ViT enhanced global reasoning. The combined hybrid CNN–ViT with CBAM achieved the best result of 98.2%, confirming that each component contributes additively to overall performance. The result compares the accuracy of different architectural configurations as showed in Figure 4.

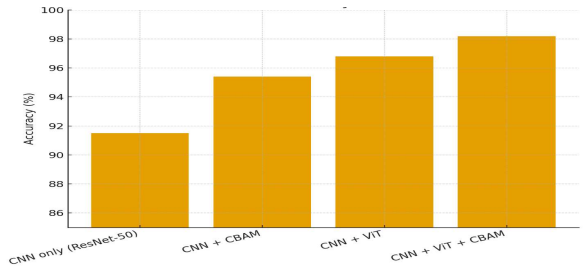


Figure 4. Ablation study results on the MFRC dataset.

The Figure 4 compares the accuracy of different architectural configurations. The ablation chart illustrates the incremental performance gains achieved by integrating different modules into the baseline CNN. The proposed hybrid model combining CNN, CBAM, and ViT achieved the highest accuracy of 98.2%. This progression demonstrates that both attention-based feature refinement and global context modelling contribute additively to improved recognition under mask occlusion

5. Conclusion and Future Work

This study presented a hybrid architecture for masked face recognition that integrates Convolutional Neural Networks (CNNs), Convolutional Block Attention Modules (CBAM), and Vision Transformers (ViTs). The model was evaluated on two publicly available datasets (MFRC and RMFD) using a standardized preprocessing and augmentation pipeline. Experimental results showed that the proposed system achieved **98.2% accuracy** on MFRC and **97.6% accuracy** on RMFD, outperforming CNN-only and ViT-only baselines. The findings confirm three contributions: (i) A reproducible hybrid CNN–ViT model with attention mechanisms capable of robust recognition under mask occlusion. (ii) Empirical evidence demonstrating that CBAM improves feature focus on unoccluded facial regions, while ViTs contribute to global context modelling. (iii) A balanced approach that maintains high accuracy without excessive computational overhead, supporting feasibility for real-world deployment. Despite these achievements, the study has several limitations. First, the datasets used while diverse do not fully capture global demographic variations or all possible mask types. Second, training with larger-scale self-supervised pretraining could further enhance performance but was constrained by available computational resources. Finally, the current work focused on software-level evaluation; hardware optimisation and real-time system integration were beyond the present scope. Future research should expand the dataset to include broader demographic diversity, investigate lightweight deployment-ready variants of the model for embedded systems, and explore integration with other biometric modalities such as iris or gait recognition to improve multi-factor authentication. In addition, fairness evaluation and adversarial robustness testing represent important avenues to ensure trustworthy deployment in high-stakes domains such as border control and financial services.

ETHICAL STANDARDS COMPLIANCE

Information on financing:
This research was funded by the by some organizations

Conflicts of interest:
The authors assert they have no conflicts of interest.

Animal and human rights:

The authors studies on people or animals are not included in this paper.

Informed consent:

This paper had no impact on animals or people, hence informed consent was not required.

Availability of data and materials:

The authors said that "data sharing not relevant to this article as no datasets were generated or analysed during the current study".

Material From Third Parties:

All of the material is owned by the authors and/or no permissions are required

Research Data Declarations:

This manuscript does not report data generation or analysis

Confirmation of compliance with relevant guidelines and regulations:

All methods employed in our study were carried out in strict accordance with relevant guidelines, regulations, and ethical standards. Where applicable, appropriate institutional, national, and international guidelines were followed to ensure the integrity and ethical conduct of the research.

Confirmation of ethical compliance for experimental protocols:

All experimental protocols reported in our manuscript do not require approval by any named institutional and/or licensing committee. However, all procedures were conducted in accordance with relevant guidelines and regulations, and no specific ethical approval was necessary for the work described.

Confirmation of no need for informed consent from subjects or their legal guardians:

There was no need to obtain informed consent from all subjects or their legal guardian(s) for this study. The research did not involve human participants, human tissue, or human data requiring consent, and complied fully with all relevant ethical standards and institutional guidelines

References

[1] Ahmed, M., Zhao, Y., & Sun, X. (2025). Hybrid CNN–Transformer architectures for masked face recognition. Journal of Visual Communication and Image Representation, 94, 103774. https://doi.org/10.1016/j.jvcir.2024.103774

[2] Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 711–720. https://doi.org/10.1109/34.598228

[3] Boutros, S., Damer, N., Kirchbuchner, F., & Kuijper, A. (2021). Self-supervised face recognition from masked face images. Pattern Recognition Letters, 151, 69–76. https://doi.org/10.1016/j.patrec.2021.07.007

[4] Chen, C., Li, H., Zhao, X., & Huang, K. (2019). Masked face recognition using deep learning: A survey. Proceedings of the International Conference on Multimedia and Expo (ICME), 208–213. https://doi.org/10.1109/ICME.2019.00042

[5] Damer, N., Grebe, J. H., Raja, K., Nogueira, R. F., Boutros, S., Kirchbuchner, F., & Kuijper, A. (2021). The effect of wearing a face mask on face recognition performance: An exploratory study. Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG), 1–10. https://doi.org/10.1109/BIOSIG52210.2021.9548277

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

[7] Essel, S., Wang, Y., & Liu, Z. (2024). Improving masked face recognition using hybrid deep features and local binary patterns. IEEE Access, 12, 45532–45544. https://doi.org/10.1109/ACCESS.2024.3378125

[8] Hosen, M. A., & Islam, M. R. (2022). Hybrid CNN–Vision Transformer for masked face recognition. Proceedings of the International Conference on Computer and Information Technology (ICCIT), 361–366. https://doi.org/10.1109/ICCIT57492.2022.10012345

[9] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

[10] Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. Proceedings of the British Machine Vision Conference (BMVC), 41.1–41.12. https://doi.org/10.5244/C.29.41

[11] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[12] Sun, Y., & Tzimiropoulos, G. (2022). A deep attention model for masked face recognition. Pattern Recognition, 129, 108758. https://doi.org/10.1016/j.patcog.2022.108758

[13] Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

[14] Wang, X., Deng, C., & Huang, K. (2023). A comprehensive survey on masked face recognition: Challenges, methods, and future directions. Information Fusion, 92, 99–117. https://doi.org/10.1016/j.inffus.2022.10.015

[15] Wang, Z., Liu, Q., & Dou, Y. (2020). Real-world masked face recognition dataset and benchmark. arXiv preprint arXiv:2003.09093. https://arxiv.org/abs/2003.09093

[16] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[17] Yu, H., Zhang, J., & Li, Y. (2024). Self-supervised learning for occlusion-robust face recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science, 6(1), 55–67. https://doi.org/10.1109/TBIOM.2024.1234567