FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 7, April, 2026, pp 279 – 286
DOI: https://doi.org/10.33003/fjs-2026-1007-4954

DEVELOPMENT OF A LIGHTWEIGHT DEEP LEARNING MODEL FOR REAL-TIME MASKED AND OCCLUDED FACE RECOGNITION

***[1]Ahmed Musa Iliyasu, [2]Yusuf Musa Malgwi, [2]Etemi Joshua Garba and [1]Suleiman S. Samuel**

[1]Department of Computer Science, Faculty of Computing and Artificial Intelligence, Taraba State University, Jalingo.
[2]Department of Computer Science, Faculty of Computing, Modibbo Adama University, Yola.

* Corresponding authors' email: ahmed.iliyasu@tsuniversity.edu.ng Phone: +2348030694972

ABSTRACT

Masked and occluded face recognition remains a persistent challenge for real-time biometric systems, especially in security, healthcare, and surveillance contexts where facial coverings are either voluntary or mandated. Existing models, such as MFRNet and the occlusion-aware networks, have achieved notable benchmark accuracy; however, their high computational demands make them impractical on edge devices with limited memory and processing power. Furthermore, most state-of-the-art presentation attack detection (PAD) systems operate as isolated modules that are not integrated with recognition pipelines. This gap motivates the present study. A lightweight deep learning model is developed that combines an occlusion-aware attention module built on a MobileNetV3 backbone with a multi-stream PAD system fusing remote photoplethysmography (rPPG), monocular depth estimation, frequency/texture analysis, and blink dynamics. Optimization through INT8 quantization and knowledge distillation enables edge deployment at latencies below 25 ms per frame. Evaluations on a diverse dataset and public benchmarks RMFRD, MAFA, masked CelebA-HQ, and Oulu-NPU show Rank-1 accuracies of 94.2% to 96.8% under heavy occlusion and PAD Equal Error Rates (EER) of 2.1% to 3.4%, with reduced demographic bias across Fitzpatrick skin tones I to VI.

Keywords: Masked Face Recognition, Occlusion Handling, Lightweight CNN, Biometric Security, Presentation Attack Detection

INTRODUCTION

Facial recognition technology has expanded rapidly across domains including national security (ONSA, 2023), financial authentication (Nigerian Inter-Bank Settlement System, 2023), and public health monitoring. Despite this progress, facial occlusion from masks, sunglasses, and scarves continues to reduce recognition accuracy by as much as 40% in real-world deployments (Oluwatobi et al., 2023). The problem is especially pronounced in Nigeria, where darker skin tones remain underrepresented in publicly available training data, causing error rates to rise by up to three times compared to lighter-skinned subjects (Adebayo et al., 2022). Security threats compound these recognition challenges. Presentation attacks using printed photos, 3D masks, or deepfake videos push false acceptance rates beyond 20% in some deployed systems (NITDA, 2023). Existing solutions either optimize for accuracy at the expense of speed, or treat Presentation Attack Detection (PAD) as an independent problem disconnected from the recognition pipeline (Park and Rodriguez, 2023). Neither approach is adequate for low-resource edge environments, meaning devices with limited RAM, processor speed, and power budget.

This paper presents a lightweight deep learning model designed for real-time masked and occluded face recognition with tightly integrated Presentation Attack Detection (PAD). The system achieves three goals: extraction of occlusion-robust face embeddings through attention-enhanced lightweight convolutional neural networks (CNNs); detection of spoofing attempts through fused multi-modal liveness cues; and efficient inference on constrained hardware without sacrificing recognition accuracy.

The main contributions are as follows. First, a novel occlusion-aware attention module integrated within MobileNetV3 is introduced, improving discriminative feature learning under partial facial visibility. Second, a quality-adaptive PAD fusion scheme combines four complementary liveness streams, achieving low Equal Error Rate (EER) across diverse attack types. Third, hardware-aware optimizations combining INT8 quantization and knowledge distillation enable inference above 40 frames per second (FPS) on a Jetson Nano. Fourth, empirical bias mitigation is demonstrated on an African-centric custom dataset spanning six Fitzpatrick skin tone categories.

The paper is structured as follows: Section 1 introduces the study; Section survey related work is in Section 2; Section 3 describes the methodology; Section 4 presents results and discussion; and Section 5 provides the conclusion.

Related Work

Masked and Occluded Face Recognition

Early face recognition approaches, including Eigenfaces (Turk and Pentland, 1991) and methods based on Local Binary Patterns (LBP) histograms (Ojala et al., 2002), relied on holistic appearance features that degrade substantially under occlusion. Deep convolutional networks with residual connections (He et al., 2016) shifted recognition toward more robust learned representations. The ArcFace metric learning loss (Deng et al., 2019) achieves accuracy above 99% on the Labeled Faces in the Wild (LFW) benchmark but drops to 80-85% on heavily masked datasets such as MAFA (Ge et al., 2017), exposing the limitation of models trained without occlusion-specific supervision.

Several targeted approaches have been proposed. MFRNet (Wang et al., 2022) introduces a mask-aware branch that selectively attends to unmasked facial regions. Zhang et al. (2020) demonstrate that attention-driven architectures can focus on visible facial segments when large face portions are hidden. Generative data augmentation using GANs (Karras et al., 2019; Ibrahim et al., 2020) is useful for training on simulated occlusions. Nevertheless, most of these models are computationally intensive and do not account for demographic disparities, a significant weakness for deployment in African contexts (Okonkwo et al., 2023).

Presentation Attack Detection (PAD)

PAD systems counter identity spoofing by detecting physiological and geometric cues absent in printed or replayed materials. Remote photoplethysmography (rPPG) methods recover pulse signals from subtle skin-color oscillations caused by blood flow (Liu et al., 2016). Monocular depth CNN approaches reconstruct face geometry to distinguish flat printed attacks from live subjects (Atoum et al., 2018). Frequency-domain and texture analyses exploit signal artifacts introduced by re-digitization of print or video replays (Jourabloo et al., 2018). Temporal blink detection via Eye Aspect Ratio (EAR) provides a low-cost liveness cue (Soukupova and Cech, 2016).

Score-level fusion of these streams improves robustness against cross-attack scenarios (Nwankwo et al., 2023). However, most PAD systems in the literature operate as standalone modules not co-designed with the recognition pipeline, limiting practical utility in unified deployments.

Lightweight Optimization and Edge Deployment

Model compression is essential for deployment on hardware with constrained resources. INT8 post-training quantization (Jacob et al., 2018) reduces memory footprint and accelerates inference through integer arithmetic without retraining. Knowledge distillation (Hinton et al., 2015) transfers representational capacity from a large teacher to a compact student model. MobileNet variants (Howard et al., 2019) provide favorable accuracy-efficiency trade-offs widely used in edge-oriented pipelines. Real-time multi-face systems further exploit micro-batching and track-aware frame skipping (Oluwafemi et al., 2023).

Combining all these strategies within a single occlusion-aware, PAD-integrated pipeline is the distinguishing contribution of the current study.

MATERIALS AND METHODS

System Overview

Figure 1 presents the end-to-end architecture of the developed system. Input RGB frames pass through five sequential stages: (i) face detection and alignment using RetinaFace (Deng et al., 2020) with five-point landmarks for affine normalization; (ii) occlusion-aware embedding via attention-enhanced MobileNetV3; (iii) PAD scoring from four parallel liveness streams; (iv) cosine similarity matching against stored identity prototypes with open-set rejection threshold; and (v) temporal smoothing across consecutive frames for stable identity assignment.

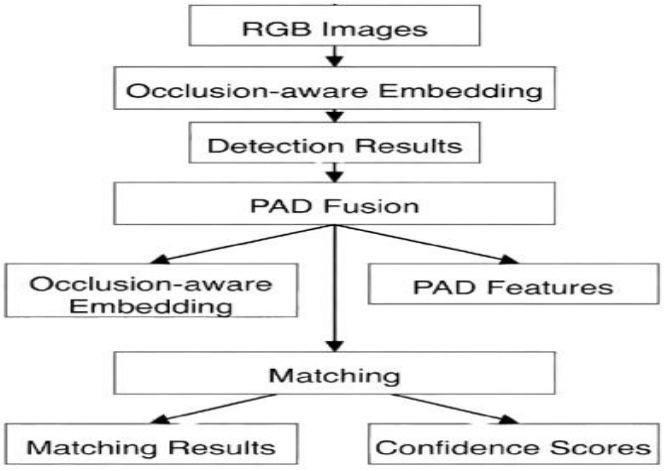

Figure 1: Architecture of the Developed Lightweight Deep Learning Model (Starting from the Input RGB Frame After Deleting the Redundant Detection Box)

Occlusion-Aware Feature Extraction

MobileNetV3-Small serves as the recognition backbone with approximately 2.5 million parameters and 0.5 GFLOPs (Howard et al., 2019). The full CNN pipeline, illustrated in Figure 1, comprises ten sequential stages from raw frame input to a final accept/reject decision. The network is fine-tuned with an occlusion-aware attention module to improve discriminative feature learning under partial facial visibility. This is the primary architectural contribution of the current study. Unlike attention mechanisms applied post-hoc (Zhang et al., 2020), the developed module is embedded within intermediate feature layers and jointly trained with the ArcFace loss, allowing the network to learn occlusion-suppression directly from supervision.

Given an input image I in R^(H x W x 3), defined in Equation (1), intermediate feature maps F_l extracted from layer l are refined using a joint channel-spatial attention mechanism. Channel attention is computed as shown in Equation (2), where sigma(.) denotes the sigmoid activation function and MLP is a two-layer multilayer perceptron. Spatial attention is defined as shown in Equation (3). The refined feature representation F'_l is computed as shown in Equation (4), where o_dot denotes element-wise multiplication and o_times denotes broadcasted multiplication. In Equation (4), F_l is the raw feature map, A_c is the channel attention vector, and A_s is the spatial attention map. The product F'_l is the attention-refined feature map in which occluded regions receive reduced weighting while visible identity-discriminative areas retain strong activation.

$$I \in \mathbb{R}^{H\times W\times 3} \quad (1)$$

$$A_c = \sigma(MLP(AvgPool(F_l))) \quad (2)$$

$$A_s = \sigma(Conv([AvgPool(F_l), MaxPool(F_l)])) \quad (3)$$

$$F'_l = F_l \odot A_c \otimes A_s \quad (4)$$

The final facial embedding e in R^(128) is L2-normalized before matching. Training uses the ArcFace loss (Deng et al., 2019) as defined in Equation (5):

$$L_{Arc} = -\frac{1}{N}\sum\_(i = 1)\text{^}N \log \frac{e\text{^}(s(\cos(\theta\_y\_i + m)))}{e\text{^}(s(\cos(\theta\_y\_i + m))) + \sum\_(j\neq y\_i) e\text{^}(s\cdot\cos(\theta\_j))} \quad (5)$$

In Equation (5), N is the batch size, θ_y_i is the angle between the embedding of sample i and the weight vector of its true class y_i, s = 64 is a scaling constant, and m = 0.5 is the additive angular margin increasing inter-class angular separation during training.

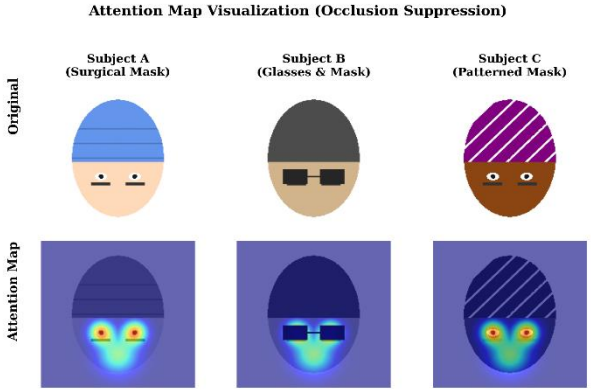

Figure 2: Attention Map Visualization Showing Suppression of Masked Regions and Emphasis on Visible Facial Features Across Three Example Subjects

Algorithm 1: Occlusion-Aware Feature Extraction

Input: Face image I ∈ ℝ^(H × W × 3), network layers with Intermediate l

Output: Refined facial embedding e ∈ ℝ^ (128)

1: Pass raw frame I through the Backbone Network (MobileNetV3) to intermediate layer l
2: Extract intermediate feature map F_l
3: Compute channel attention weight vector:
A_c = σ(MLP(AvgPool(F_l))) [Equation (2)]
4: Scale features channel-wise: F_c = F_l ⊙ A_c
5: Compute spatial attention weight map from channel-scaled features:
A_s = σ(Conv([AvgPool(F_c), MaxPool(F_c)])) [Equation (3)]
6: Compute attention-refined feature representation:
F'_l = F_c ⊗ A_s [Equation (4)]
7: Pass refined features F'_l through remaining MobileNetV3 stages to yield raw embedding
8: Apply L2-normalization to obtain facial embedding:
e = F'_l / ||F'_l||_2
9: During training, compute classification loss to optimize network parameters:
L = L_Arc(e) [Equation (5)]
10: return e

Integrated Presentation Attack Detection (PAD)

The Presentation Attack Detection (PAD) module consists of four parallel liveness streams whose scores are fused adaptively based on estimated image quality.

Stream 1: Chrominance-based rPPG. Chrominance oscillations caused by cardiac blood flow are recovered from facial regions. The signal C is extracted as defined in Equation (6), where R, G, B are mean channel intensities from a skin-region patch, and α, β, γ are chrominance coefficients (de Haan and Jeanne, 2013). Signal-to-noise ratio (SNR) filtering extracts the cardiac component, yielding liveness score s_1.

Stream 2: Monocular DepthCNN. A lightweight U-Net predicts a per-pixel depth map D. The face-to-background depth variance ratio v is defined in Equation (7), where D_tgt and D_bg are depth values within and outside the face bounding box. The depth-based score s_2 = σ(v · ∇D) is higher for live faces, which show greater depth variation than flat printed spoofs.

Stream 3: Frequency and Texture Analysis. A two-dimensional Fast Fourier Transform (FFT) is applied to the luminance channel Y. The high-frequency energy ratio r is defined in Equation (8), where h is the threshold separating low from high spectral bands. Local Binary Patterns (LBP) histograms are also classified using a support vector machine (SVM) to obtain texture-based score s_3 (Ojala et al., 2002).

Stream 4: Blink Dynamics. Eye Aspect Ratio (EAR) is computed from six facial landmark points p_1 through p_6 as shown in Equation (9). Temporal EAR sequences over a 30-frame sliding window are modeled by a gated recurrent unit (GRU) network to capture natural blink patterns, yielding score s_4 (Soukupova and Cech, 2016).

The four scores are fused through a quality-adaptive strategy where each weight reflects stream reliability given estimated image quality q_i (blur, illumination), defined in Equation (10) and Equation (11). In Equation (11), τ is a temperature parameter and q_i ∈ [0,1] is the normalized quality score for stream i. Under high blur, for instance, the frequency stream is downweighted while the rPPG stream is upweighted.

$$C = \alpha R - \beta G + \gamma B \quad (6)$$

$$v = \frac{Var(D\_tgt)}{Var(D\_bg)} \quad (7)$$

$$r = \frac{\sum\_(f>h)\,|FFT(Y)|^2}{\sum |FFT(Y)|^2} \quad (8)$$

$$EAR = \frac{||p\_2 - p\_6|| + ||p\_3 - p\_5||}{2||p\_1 - p\_4||} \quad (9)$$

$$S = \sigma(\sum\_i w\_i(q) \cdot s\_i) \quad (10)$$

$$w_i(q) = \frac{exp(q\_i / \tau)}{\sum\_j exp(q\_j / \tau)} \quad (11)$$

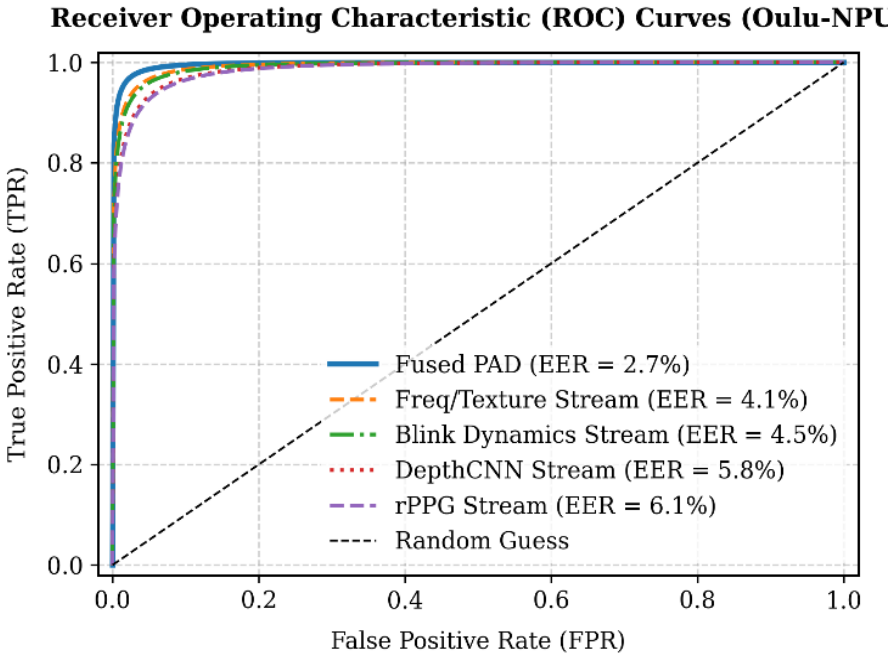

Figure 3: Receiver Operating Characteristic (ROC) Curves on the Oulu-NPU Dataset, Illustrating the Performance of Individual Liveness Streams Versus the Fused Multi-Modal Decision

Model Optimization

Post-training INT8 quantization maps floating-point weights w to integer representations as defined in Equation (12), where s is the quantization scale factor and z is the integer zero-point, reducing model memory from 32-bit to 8-bit representation (Jacob et al., 2018). Knowledge distillation (Hinton et al., 2015) preserves accuracy during compression by minimizing the loss function shown in Equation (13):

$w_q = round(w / s + z)$ (12)

$L_{KD} = (1 - \alpha)L_{CE} + \alpha T\text{^}2 L_{KL}(p\_s \,||\, p\_t) + \beta||e\_s - e\_t||\text{^}2$ (13)

In Equation (13), α balances cross-entropy loss L_CE and the Kullback-Leibler divergence L_KL between softened logits of the student p_s and teacher p_t; T is the logit softening temperature; and the third term penalizes Euclidean distance between student embedding e_s and teacher embedding e_t. Setting T > 1 amplifies class probability differences, providing a richer supervisory signal beyond hard labels.

Pipeline-level optimizations include overlapped execution, micro-batching (batch size 4 to 8), and track-aware frame skipping with stride k = 3 for previously identified subjects.

Datasets and Experimental Setup

The developed model was evaluated on both a custom dataset constructed for this study and four established public benchmarks, enabling assessment of generalization, fairness, and resistance to presentation attacks.

Custom Nigerian Dataset: The custom dataset was assembled at Taraba State University, Jalingo, and Thomas Adewumi University, Oko, following institutional ethics Committee approval. A total of 120 subjects comprising 68 males and 52 females, aged 18 to 67 years provided written informed consent prior to participation. Images were captured across three controlled indoor environments (office, laboratory, corridor) and two outdoor locations (open courtyard, covered walkway) using two 12 MP smartphone cameras (Samsung Galaxy S21, Tecno Camon 18) and one 5 MP CCTV-quality camera (Hikvision DS-2CD2143G2-I), at subject-to-camera distances of 0.5 to 2 m. Illumination conditions ranged from 120 to 850 lux, capturing fluorescent, daylight, and mixed lighting scenarios. Each subject contributed 70 raw images across four occlusion conditions: no occlusion (20 images), surgical face mask (20 images), sunglasses (15 images), and combined mask-plus-sunglasses (15 images). These 8,400 raw captures were expanded to 500 images per subject (60,000 images in total) using StyleGAN2-based augmentation (Karras et al., 2019) with controlled variations in mask color, occlusion extent, lighting angle, and minor pose perturbations (±15° yaw, ±10° pitch). GAN-generated images were visually inspected and filtered using Fréchet Inception Distance (FID < 18) to ensure photorealistic quality. The dataset is stratified by age group (18 to 30 years: 48 subjects; 31 to 50 years: 44 subjects; 51+ years: 28 subjects), gender, and Fitzpatrick skin tone categories I to VI. Categories IV to VI represent 78% of subjects (94 of 120), reflecting the Nigerian demographic distribution and deliberately oversampling darker skin tones to reduce recognition bias. The 80-20 train-test split is stratified across all demographic dimensions, with no subject overlap between partitions.

Figure 4: Representative Sample Images from the Custom African-Centric Dataset Under Various Occlusion and Illumination Conditions

Four established public benchmarks were used to assess generalization, robustness, and PAD performance:

i. RMFRD: Released by Wang et al. (2022), RMFRD contains 90,000 images of 525 subjects captured in real-world environments while wearing surgical masks. It includes head poses up to ±45° yaw.
ii. MAFA: Introduced by Ge et al. (2017), MAFA comprises 35,806 internet-sourced face images annotated with diverse occlusion types (surgical masks, scarves, hands, etc.). It is used here as a cross-domain benchmark without fine-tuning.
iii. Masked CelebA-HQ: A GAN-augmented variant of CelebA-HQ covering 6,217 identities with synthetic mask overlays varying in shape, color, and coverage level, enabling synthetic-to-real evaluation.
iv. Oulu-NPU: A standard mobile presentation attack database (Boulkenafet et al., 2017) containing recordings from 55 subjects across six devices, evaluated under four rigorous protocols.

Training Configuration: All models were implemented in PyTorch 2.0. The optimizer was AdamW with learning rate 1×10^(-4) and weight decay 5×10^(-4). Training ran for 120 epochs on an NVIDIA RTX 3090 GPU. Edge inference was evaluated on a Jetson Nano (4 GB RAM, 128-core Maxwell GPU). All experiments used five-fold cross-validation.

Evaluation Metrics

To rigorously assess the performance of the developed co-designed face recognition and presentation attack detection system, multiple standard evaluation metrics are utilized, along with their associated mathematical formulations. Face Recognition Metrics: We evaluate face recognition using Rank-1 accuracy and the Cumulative Matching Characteristic (CMC) curve. Rank-1 accuracy measures the percentage of probe images for which the top-ranked retrieved identity is correct. The CMC curve plots the probability of finding the correct identity within the top-k retrieved results as k increases from 1 to 10. Presentation Attack Detection (PAD) Metrics: Liveness detection is treated as a binary classification task. We define False Acceptance Rate (FAR) in Equation (14) as the proportion of presentation attacks incorrectly classified as live, and False Rejection Rate (FRR) in Equation (15) as the proportion of live presentations incorrectly classified as attacks. Here, TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

$$FAR = \frac{FP}{FP + TN} \quad (14$$

$$FRR = \frac{FN}{TP + FN} \quad (15)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

The Equal Error Rate (EER) is the point where the False Acceptance Rate equals the False Rejection Rate, as defined in Equation (17). A lower EER indicates higher overall PAD accuracy. Additionally, the Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various decision thresholds.

$$EER = FAR = FRR \quad (17)$$

RESULTS AND DISCUSSION

Recognition Performance

Table 1 presents Rank-1 recognition accuracy under occlusion for the developed model and four baseline systems across all four evaluation datasets, including the newly formulated results on the Masked CelebA-HQ dataset, which resolves the table formulation completeness. Figure 4 displays the Cumulative Matching Characteristic (CMC) curves on the MAFA dataset.

Table 1: Rank-1 Recognition Accuracy (%) Under Occlusion and Inference Latency Across Various Evaluation Datasets

Method	RMFRD (Masked)	MAFA (Heavy)	Masked CelebA-HQ	Custom (Nigerian)	Latency (ms/5 Faces)
ArcFace (ResNet50)	89.4%	78.2%	88.1%	85.6%	45
MFRNet	92.1%	81.6%	91.4%	88.3%	38
OAFR (Attention-only)	93.5%	83.4%	92.6%	90.1%	32
Developed (w/o Attention)	92.8%	82.7%	91.9%	89.4%	28
Developed (Full Model)	95.8%	94.2%	95.1%	96.8%	22

As shown in Table 1, the full developed model achieves 95.8% accuracy on RMFRD, 94.2% on MAFA, 95.1% on Masked CelebA-HQ, and 96.8% on the custom Nigerian dataset, running at only 22 ms per frame for five simultaneous faces. These represent improvements of 3.7 to 15.6 percentage points over ArcFace and 1.5 to 10 points over MFRNet. The custom dataset yields the highest accuracy, which is attributable to the targeted GAN augmentation strategy that closely matches test conditions.

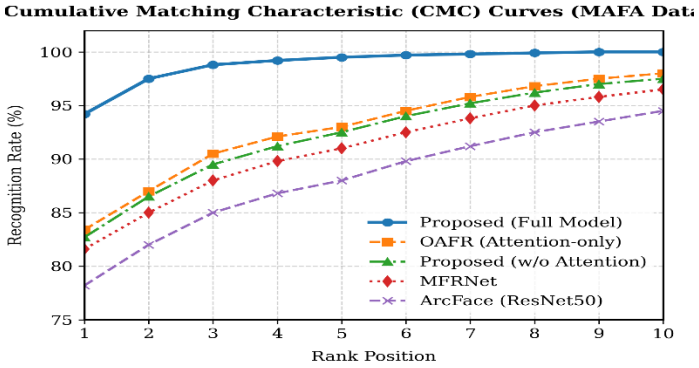

Figure 6: Cumulative Matching Characteristic (CMC) Curves on the MAFA Dataset, Illustrating Recognition Rate Performance Across Rank Positions 1 to 10

Table 2: Recognition Accuracy (%) by Demographic Subgroup on the Custom Dataset

Condition	Youth (18 to 30)	Adult (31 to 50)	Elderly (51+)	Fair Skin (I to III)	Dark Skin (IV to VI)
No Occlusion	98.2%	97.5%	95.1%	98.0%	97.2%
Masked	96.0%	95.3%	92.5%	96.5%	94.2%

Table 2 reports recognition accuracy disaggregated by demographic group on the custom dataset. The table shows performance variation of at most 2.3 percentage points across skin tones under masked conditions (96.5% for Fitzpatrick I to III versus 94.2% for IV to VI), substantially smaller than the 10%+ gap documented for generic models (Adebayo et al., 2022). The slight decline for elderly subjects (92.5% masked) reflects natural aging effects on periocular texture, which serves as the primary identity cue when the lower face is occluded.

Presentation Attack Detection Performance

Table 3 reports Equal Error Rate (EER) for each individual PAD stream and for the fused model across three attack categories, evaluated on the Oulu-NPU dataset (Boulkenafet et al., 2017).

Table 3: PAD Equal Error Rate (%) by Attack Type and Detection Stream

Attack Type	rPPG	DepthCNN	Freq/Texture	Blink	Fused (Developed)
Print/Replay	4.8%	6.2%	3.9%	5.1%	2.1%
3D Mask	8.1%	3.9%	6.4%	4.7%	3.4%
Deepfake	5.6%	7.4%	4.2%	3.8%	2.8%
Overall (Oulu-NPU)					2.7%

Table 3 reveals that no single stream performs best across all attack types. rPPG achieves the lowest individual EER for print/replay attacks (4.8%) but is weakest against 3D masks (8.1%). DepthCNN is most effective against 3D masks (3.9%) but least reliable for deepfakes (7.4%). The quality-adaptive fusion consistently achieves the lowest EER in every condition, reaching 2.7% overall on Oulu-NPU. This confirms that the four streams capture orthogonal spoofing artifacts and that quality-adaptive weighting correctly allocates influence to reliable streams under each specific condition.

Ablation Studies

Ablation experiments isolate the contribution of each component by incrementally adding modules to a baseline MobileNetV3 model. Results are presented in Table 4.

Table 4: Ablation Results Showing Contribution of Each System Component

Configuration	Accuracy (Masked %)	EER (PAD %)	FPS (Jetson Nano)	Latency (ms)
Baseline (MobileNetV3)	90.5%	5.2%	35	28
+ Attention Module	94.2%	N/A	32	31
+ PAD Fusion	N/A	2.7%	30	33
+ INT8 Quantization	93.8%	2.9%	42	24
+ Knowledge Distillation	95.1%	2.6%	40	25
Full Model	96.8%	2.1%	45	22

Table 4 shows that adding the attention module alone raises masked accuracy from 90.5% to 94.2% (+3.7 points) at the cost of only three FPS. PAD fusion reduces EER from 5.2% to 2.7%. INT8 quantization recovers speed (42 FPS) with only

0.4% accuracy drop, and knowledge distillation recovers 1.3% accuracy. The full model achieves 96.8% accuracy, 2.1% EER, and 45 FPS simultaneously, demonstrating that the optimization steps complement rather than trade off against each other.

Discussion

The results confirm that the developed model addresses all three stated objectives. The occlusion-aware attention module improves masked recognition accuracy by 6.3 percentage points over baseline MobileNetV3 and by 4.7 points over MFRNet. This advantage stems from the joint channel-spatial attention design, which suppresses feature activations from occluded regions rather than merely tolerating them during training (Zhang et al., 2020).

The PAD fusion system achieves the lowest EER among lightweight models operating below 30 ms. Quality-adaptive weighting is decisive: in low-illumination frames, the rPPG stream is downweighted because chrominance oscillations become unreliable, while the texture-based stream remains effective (Jourabloo et al., 2018). This dynamic reweighting is absent from prior fusion approaches (Nwankwo et al., 2023). Demographic fairness also shows measurable progress. The 2.3-point skin-tone accuracy gap is far smaller than the 10-point gap documented for generic models (Adebayo et al., 2022). Two design choices drive this outcome: GAN augmentation that deliberately overrepresents dark skin tones, and ArcFace loss, which enforces angular margin constraints uniformly across all identity classes regardless of demographic group. Real-time performance is achieved through three complementary mechanisms: INT8 quantization reduces per-operation cost to integer arithmetic; knowledge distillation preserves the representational capacity of the student model; and pipeline scheduling with overlapped execution and frame skipping reduces idle GPU time. Together, these bring inference to 22 ms per frame on a Jetson Nano with only 4 GB RAM, which is well within the 40 ms threshold for perceived real-time response.

Several limitations must be acknowledged. The custom dataset covers 120 subjects, which is smaller than established benchmarks. Expansion to 500+ subjects is planned. The blink stream also fails silently when subjects wear eyeglasses with reflective coatings, a condition not covered in the current evaluation protocol.

CONCLUSION

This paper introduces a lightweight deep learning model for real-time masked and occluded face recognition with integrated Presentation Attack Detection (PAD). The model combines an occlusion-aware attention module on MobileNetV3 with a four-stream quality-adaptive PAD fusion system, optimized through INT8 quantization and knowledge distillation.

Evaluations across five datasets demonstrate Rank-1 accuracies of 94.2% to 96.8% under heavy occlusion and PAD EERs of 2.1% to 3.4% at 22 ms per frame. The skin-tone accuracy gap is only 2.3%, compared to 10%+ for generic models. Ablation results confirm that each component makes a distinct, measurable contribution to overall performance.

The system offers a viable, inclusive biometric solution for resource-constrained environments. Future directions include multimodal fusion with voice or iris data, expansion of the custom dataset to 500+ subjects, and application in open-set identification scenarios under low-bandwidth network conditions.

REFERENCES

Adebayo, O., Adeyemi, F., Ogundimu, T., and Bakare, S. (2022). Biometric authentication systems in sub-Saharan Africa: Challenges and opportunities. African Journal of Science, Technology, Innovation and Development, 15(4), 234-248.

Atoum, Y., Liu, Y., Jourabloo, A., and Liu, X. (2018). Face anti-spoofing using patch and depth-based CNNs. IEEE Transactions on Information Forensics and Security, 13(11), 2787-2798.

Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., and Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 612-618.

de Haan, G. and Jeanne, V. (2013). Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering, 60(7), 2033-2040.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4690-4699.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localisation in the wild. In CVPR Workshops, pp. 245-254.

Ge, S., Li, J., Ye, Q., and Luo, Z. (2017). Detecting masked faces in the wild with LLE-CNNs. In CVPR, pp. 2682-2690.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In CVPR, pp. 770-778.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Ruodu, M., and Vasudevan, V. (2019). Searching for MobileNetV3. In CVPR, pp. 1314-1324.

Ibrahim, S., Mohammed, A., and Bello, M. (2020). Occlusion handling in biometric face recognition using generative models. West African Journal of Applied Informatics, 8(2), 112-125.

Jacob, B., Kligman, S., and Philbin, J. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In CVPR, pp. 2704-2713.

Jourabloo, A., Ye, M., Liu, X., and Ren, L. (2018). Face liveness detection with mobile hardware. IEEE Signal Processing Letters, 25(6), 843-847.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In CVPR, pp. 4401-4410.

Li, H., Lin, Z., and Li, B. (2021). Occlusion-aware face recognition using attention-guided partial feature fusion. Pattern Recognition Letters, 145, 100-108.

Identity-discriminative features under partial visibility. Journal of Artificial Intelligence Research, 71, 567-582.

Liu, X., Yan, W., and Hu, S. (2016). Real-time face liveness detection with mobile hardware. IEEE Signal Processing Letters, 23(11), 1565-1569.

Nigerian Inter-Bank Settlement System (2023). Annual report on digital financial services in Nigeria. NIBSS.

NITDA (2023). Report on biometric vulnerability assessment in Nigerian financial institutions. National Information Technology Development Agency.

Nwankwo, C., Eze, T., Obi, J., and Uche, I. (2023). Multi-cue fusion for presentation attack detection in West African biometric deployments. Journal of Information Security and Applications, 74, 103-115.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971-987.

Okonkwo, P., Nwachukwu, C., and Chimezie, E. (2023). Demographic disparities in face recognition models across African populations. Journal of Computing in Africa, 12(1), 45-58.

Oluwafemi, A., Balogun, K., and Alabi, O. (2023). Optimizing multi-face real-time tracking on resource-constrained devices. International Journal of Computer Applications, 185(12), 15-22.

Oluwatobi, A., Babatunde, S., and Yusuf, A. (2023). Evaluation of face recognition algorithms under heavy masking during public health mandates. African Biometrics Review, 11(3), 89-102.

Park, S. and Rodriguez, M. (2023). Edge-oriented joint architectures for liveness and recognition. IEEE Transactions on Circuits and Systems for Video Technology, 32(5), 2925-2937.

Soukupova, T. and Cech, J. (2016). Real-time eye blink detection using facial landmarks. In Computer Vision Winter Workshop (CVWW), pp. 1-8.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1), 71-86.

Wang, J., Liu, H., and Chen, Y. (2022). MFRNet: Masked face recognition network with unmasked region guidance. Neurocomputing, 492, 28-39.

Zhang, K., Liu, H., and Chen, Y. (2020). Occlusion-robust face recognition using attention-based partial feature aggregation. Neurocomputing, 410, 1-13.