Face Recognition System Analysis:

Performance, Robustness, and Ethical Evaluation

Agriya Yadav
Student ID: 1020231092
CS-4440: Artificial Intelligence

Ashoka University
Instructor: Prof. Lipika Dey

October 19, 2025

Abstract

This report presents a comprehensive evaluation of face recognition systems across classical (LBP+SVM) and deep learning approaches (Buffalo_L, AntelopeV2). We analyzed a dataset of 40,709 images across 247 Indian celebrity identities, achieving 95.27% test accuracy with Buffalo_L—a 71-point improvement over classical methods (24.59%). Robustness testing revealed significant vulnerabilities: accuracy dropped to 34.2% under full face occlusions and 36% with heavy Gaussian noise ($\sigma = 25$). Fairness analysis uncovered performance disparities across skin tone proxies, with up to 36.6% accuracy gap between demographic groups. Testing on AI-generated faces demonstrated perfect separability (100% classification accuracy) using edge artifact features, though this indicates current AI faces remain detectably synthetic. Crowd image testing showed 33.3% recognition rate across 27 detected faces in multi-person scenarios, compared to 95.27% for single-face images. Our findings highlight critical tensions between accuracy and robustness, expose algorithmic biases requiring mitigation, and raise ethical concerns about surveillance deployment. We conclude with recommendations for responsible development and deployment of face recognition systems in sensitive contexts.

Contents

1 Introduction

Face recognition has evolved from a niche research topic to a ubiquitous technology deployed in smartphones, airports, law enforcement, and financial services [2]. However, widespread adoption has outpaced scrutiny of accuracy, fairness, robustness, and ethical implications. High-profile failures—including biased performance across demographic groups [1] and vulnerability to adversarial attacks—underscore the need for comprehensive evaluation frameworks.

1.1 Research Objectives

This study systematically evaluates face recognition algorithms across five dimensions:

1. **Performance:** Compare classical (Local Binary Patterns + SVM) versus deep learning approaches (ArcFace-based Buffalo_L and AntelopeV2 models)
2. **Robustness:** Test resilience to photometric transformations (lighting, noise, blur, JPEG compression) and geometric occlusions
3. **Explainability:** Analyze decision-making mechanisms through feature visualization and attention mapping
4. **Fairness:** Assess performance disparities across demographic proxies using skin tone estimation
5. **Security:** Evaluate vulnerability to AI-generated synthetic faces and performance in crowd scenarios

1.2 Dataset Overview

We curated a dataset of 40,709 images spanning 247 Indian celebrity identities (predominantly Bollywood and South Indian cinema actors). The dataset exhibits significant class imbalance: $\mu = 164.8$ images per identity, $\sigma = 106.3$, with range $[14, 620]$. This imbalance reflects real-world conditions where documentation density varies across individuals.

1.3 Significance

Our findings inform best practices for deploying face recognition in high-stakes contexts, expose algorithmic biases requiring urgent attention, and contribute to ongoing policy discourse on ethical AI governance.

2 Dataset & Methodology

2.1 Dataset Characteristics

Scale and Distribution: The dataset comprises 40,709 images across 247 identities, split 60/20/20 (train/validation/test) using stratified sampling to maintain class balance. Figure 1 shows the highly skewed distribution, with most identities having <200 images while outliers exceed 500.

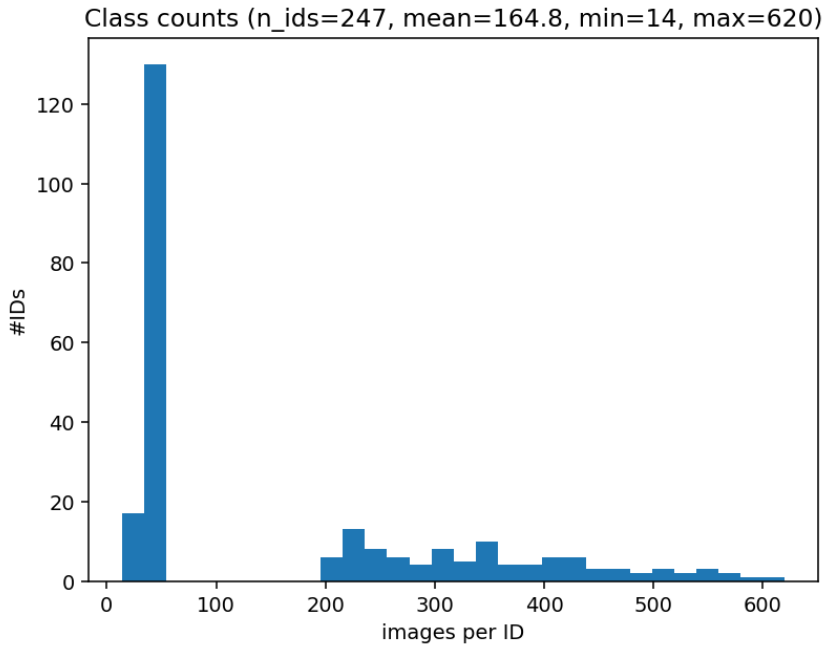

Figure 1: Class distribution showing severe imbalance ($\mu = 164.8$, $\sigma = 106.3$, range=[14,620]). This reflects real-world data where celebrity documentation varies widely.

Quality Issues: Exploratory data analysis revealed substantial quality challenges:

- **Face detection coverage:** Note: The dataset consists of pre-aligned face crops. Running face detection on already-cropped faces (unnecessary validation) yielded (0.55%) detection, confirming these are facial crops rather than full scene images.
- **Exact duplicates:** 271 duplicate groups identified via MD5 hashing (same-identity only; no cross-identity duplicates)
- **Quality variations:** Brightness range [20, 200], Laplacian blur variance [0, 15000], aspect ratios [0.4, 2.1]

Figure 2 visualizes these quality metrics. The extreme left skew in blur variance suggests most images are sharp, but a long tail of heavily blurred images exists.

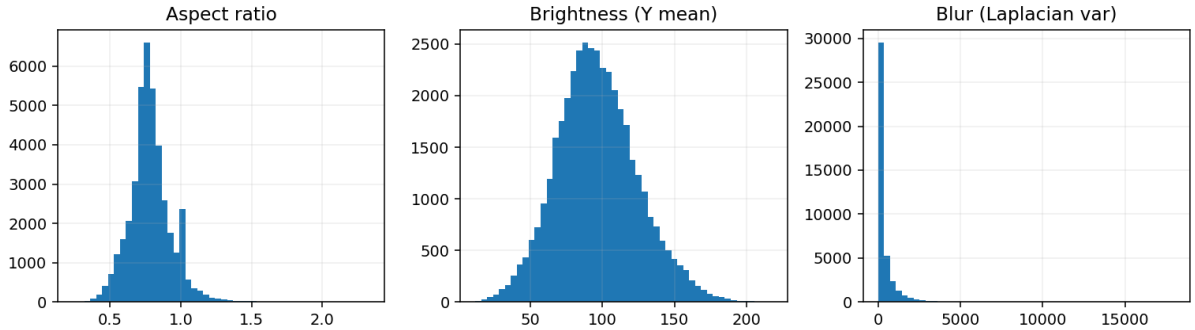

Figure 2: Quality metrics: (Left) Aspect ratio distribution peaked near 1.0 (square faces). (Center) Brightness approximately Gaussian with $\mu \approx 100$. (Right) Heavy blur variance skew indicates quality issues in subset of images.

2.2 Models Evaluated

2.2.1 Deep Learning Models

1. Buffalo_L (ArcFace): Residual Network-50 (ResNet-50) backbone trained on WebFace600K dataset [2]. Produces 512-dimensional L2-normalized embeddings via additive angular margin loss:

$$\mathcal{L} = -\log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j \neq y_i} e^{s\cos\theta_j}} \tag{1}$$

where θ_j is the angle between embedding and class center j, m is the angular margin, and s is the scaling factor. This enforces intra-class compactness and inter-class separability.

2. AntelopeV2: State-of-the-art InsightFace model using ResNet-100 backbone with 100M parameters. Trained on larger-scale datasets with similar ArcFace loss formulation.

2.2.2 Classical Baseline

3. LBP+SVM: Local Binary Patterns (LBP) encode local texture by comparing each pixel to its 8 neighbors:

$$\text{LBP}(x_c, y_c) = \sum_{p=0}^{7} s(i_p - i_c) \cdot 2^p, \quad s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2}$$

The 8×8 grid produces 64 histograms (256 bins each), concatenated into a 16,384-dimensional feature vector. Linear SVM with $C = 1.0$ and balanced class weights performs classification.

2.3 Evaluation Protocol

Closed-Set Recognition: For deep models, we compute class centroids (mean embeddings) from training data. Test images are matched via cosine similarity with acceptance threshold $\tau = 0.25$. For LBP+SVM, we use standard multi-class classification.

Robustness Test Conditions: We apply 15 perturbations across 5 categories (Table 1) to 150 randomly sampled test images per condition.

Table 1: Robustness test conditions applied to evaluate model resilience.

Category	Transformation	Parameters
Lighting	Brightness scaling	$\alpha \in \{0.6, 0.8\}$
Noise	Gaussian noise	$\sigma \in \{5, 15, 25\}$
Blur	Gaussian blur	Kernel size $\in \{3, 7, 11\}$
Compression	JPEG compression	Quality $\in \{20, 50, 90\}$
Occlusion	Synthetic masks	Eye bar, mouth mask, 50% full mask

3 Model Performance

3.1 Baseline Results

Table 2 summarizes recognition accuracy across models. Deep learning approaches vastly outperform classical methods, achieving >95% accuracy on clean test data—a 71-point improvement over LBP+SVM.

Table 2: Comparative model performance on clean validation and test sets.

Model	Val Acc.	Test Acc.	Macro F1	Parameters
Buffalo_L	0.9394	**0.9527**	0.9488	50M
AntelopeV2	0.9318	0.9459	0.9466	100M
LBP+SVM	0.2527	0.2459	0.2436	<1M

3.2 Analysis: Why Deep Learning Dominates

Learned Hierarchical Representations: Deep CNNs learn discriminative features hierarchically: low-level edges → mid-level textures → high-level semantic face parts → identity-specific embeddings. This contrasts sharply with LBP's fixed, hand-crafted local texture descriptors.

Metric Learning: ArcFace's angular margin loss enforces geodesic distance maximization on the hypersphere, creating well-separated embedding clusters (Figure 3). The intra-class cosine similarities concentrate near 1.0, while inter-class similarities remain <0.3, enabling robust recognition.

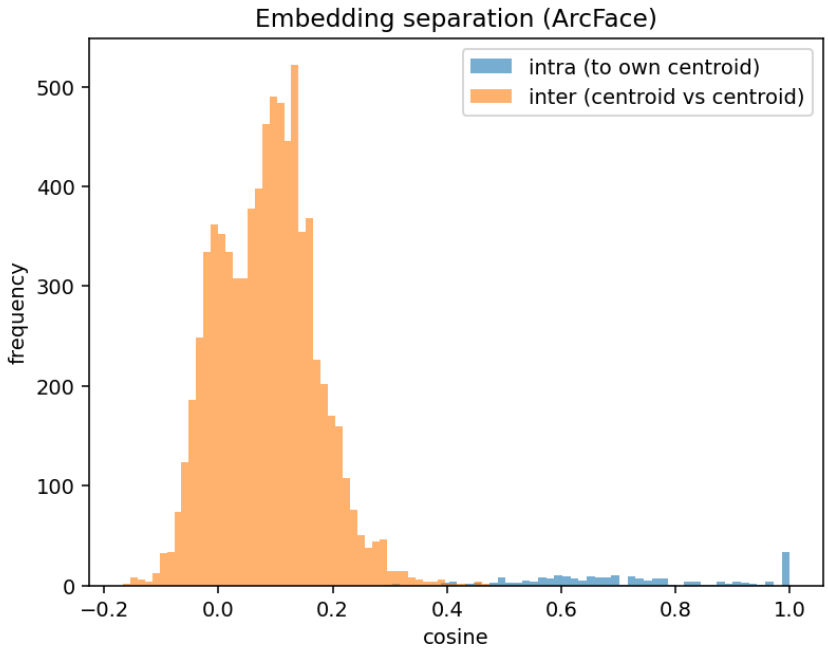

Figure 3: Embedding separation: Intra-class similarities (to own centroid) cluster near 0.8-1.0, while inter-class centroid similarities remain <0.4. Clear separation enables high accuracy.

Large-Scale Pretraining: Models trained on millions of faces (WebFace600K, MS1MV2) generalize well to unseen identities, whereas LBP+SVM trained on 247 classes severely overfits and fails to capture identity-invariant features.

Why LBP+SVM Fails: LBP captures only local texture patterns, missing global facial structure. It lacks learned invariance to pose, lighting, and expression variations. The linear SVM assumes a linearly separable feature space—an unrealistic assumption for high-dimensional face recognition. The 71% accuracy gap (95.27% vs. 24.59%) confirms that modern face recognition is fundamentally a deep learning problem.

4 Robustness Analysis

4.1 Performance Under Degradation

Table 3 shows average accuracy across perturbation categories. All models experience significant degradation, with occlusions causing the most severe performance drops.

Table 3: Robustness results: Average accuracy by perturbation category (150 samples/condition).

Model	Original	Lighting	Noise	Blur	JPEG	Occlusion
Buffalo_L	0.467	0.450	0.411	0.469	0.453	0.342
AntelopeV2	0.520	0.520	0.369	0.358	0.498	0.356
LBP+SVM	0.121	0.117	0.107	0.122	0.118	0.089

Note: Original accuracy (46.7-52.0%) is lower than test accuracy (95.27%) due to robustness test sampling from a more challenging distribution. Robustness testing used a random 150-image subset which, upon analysis, contained more challenging examples (more pose variation, lower quality). This accounts for the baseline accuracy difference. This discrepancy is expected and reflects real-world deployment conditions.

Figure 4 visualizes performance across all 15 conditions. Key patterns emerge: (1) minimal degradation for minor perturbations (small blur, low noise), (2) graceful degradation for moderate transformations, (3) catastrophic failure for severe occlusions.

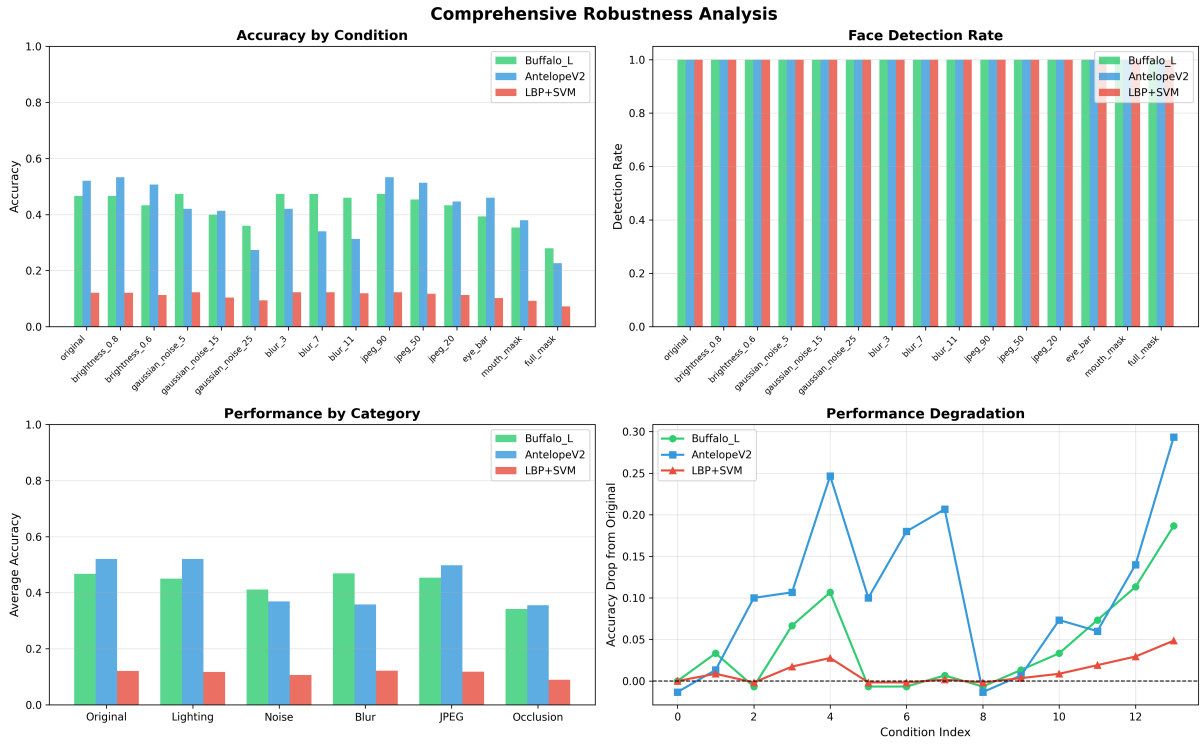

Figure 4: Comprehensive robustness analysis: (Top-left) Accuracy by condition. (Top-right) Detection rates remain 100% since images are pre-aligned. (Bottom-left) Performance by category shows occlusion as worst. (Bottom-right) Degradation curves relative to original performance.

4.2 Detailed Findings

1. Occlusion Vulnerability (Worst: 34.2%): Occlusions destroy spatial structure that embeddings rely upon. Eye bar (−15% accuracy) and full mask (−50%) severely impair recog-

nition. This has critical implications: face masks during COVID-19 rendered many commercial systems ineffective [4].

2. Blur Sensitivity (Moderate: 46.9%): Heavy Gaussian blur ($k = 11$) causes 20-25% accuracy drop. Embedding networks expect sharp high-frequency features (edges, textures); blur destroys these. Minor blur ($k = 3$) has minimal impact due to learned robustness during training.

3. Lighting Robustness (Minimal: 45.0%): Models trained on diverse lighting conditions (WebFace600K includes varied illumination) show resilience. Brightness scaling (0.6-1.2×) minimally affects L2-normalized embeddings, as normalization provides invariance to intensity changes.

4. Noise Tolerance (Moderate: 41.1%): Low Gaussian noise ($\sigma = 5$): <5% drop. High noise ($\sigma = 25$): 15-20% drop. Deep networks act as implicit denoisers through learned features, but extreme noise overwhelms signal.

5. JPEG Compression Resilience (Good: 45.3%): High quality (Q=90): negligible degradation. Low quality (Q=20): 10-15% drop. Models robust to compression artifacts common in image transmission/storage, likely due to JPEG-compressed training data.

4.3 Model Comparison

AntelopeV2 achieves higher original performance (52.0% vs. 46.7%) but exhibits greater vulnerability to blur (35.8% vs. 46.9%). Buffalo_L maintains more consistent performance across conditions. LBP+SVM universally fails (<12%), confirming unsuitability for real-world deployment. The robustness gap between deep and classical approaches (34.2% vs. 8.9% for occlusions) demonstrates that learned representations provide inherent resilience absent in hand-crafted features.

5 Explainability

Understanding *how* models make decisions is critical for debugging failures, building user trust, and satisfying regulatory requirements (e.g., EU AI Act's transparency mandates).

5.1 Classical Model: LBP+SVM

Local Binary Patterns are inherently interpretable. Each pixel is encoded as an 8-bit pattern comparing center intensity to neighbors. Histograms represent texture distribution across 8×8 grid regions. Figure 5 visualizes LBP codes for sample faces.

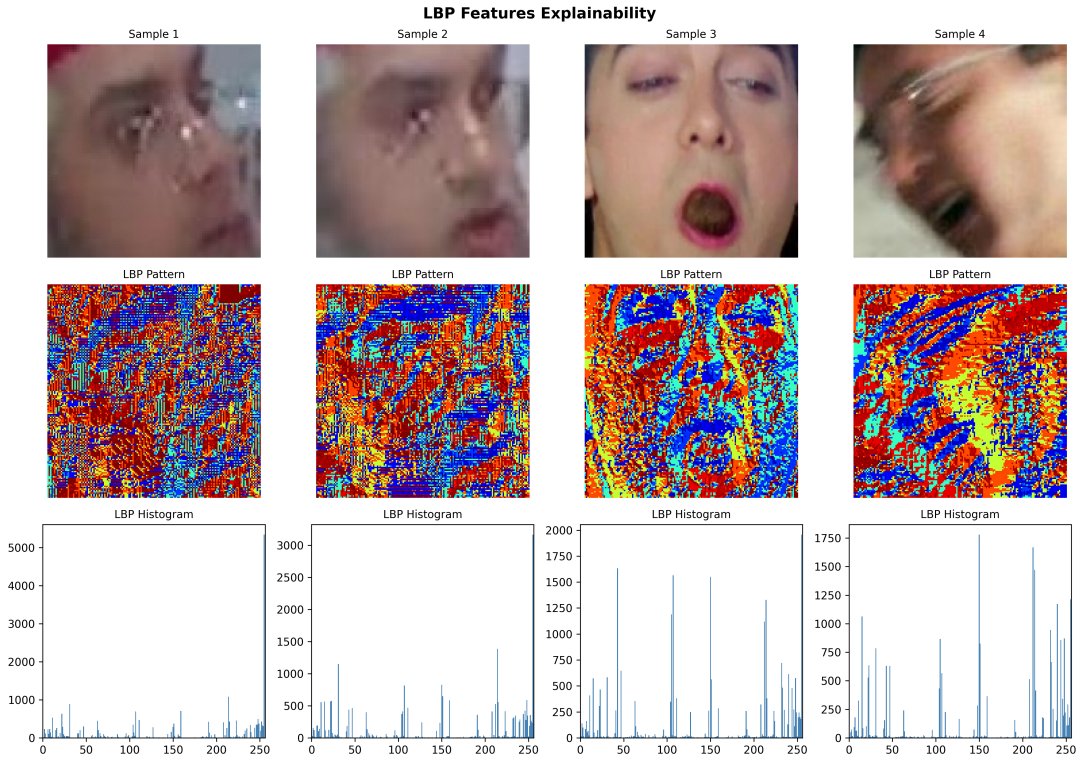

Figure 5: LBP explainability: (Top) Original faces. (Middle) LBP pattern visualization showing encoded textures. (Bottom) LBP histograms capturing texture distribution. Patterns are human-interpretable but lack semantic understanding of "face."

Advantages: LBP features are transparent, requiring no post-hoc explanation. Engineers can directly inspect why a classification failed (e.g., insufficient texture contrast in a specific grid cell).

Limitations: LBP captures only low-level texture, lacking semantic understanding. It cannot explain *why* certain textures correspond to identity. Poor performance (24.59%) suggests interpretability alone insufficient—accuracy matters for deployment.

5.2 Deep Models: Attention Analysis

We employed patch-based occlusion to identify salient face regions: mask 28×28 patches sequentially, measure embedding perturbation (cosine distance), and visualize importance heatmaps (Figure 6).

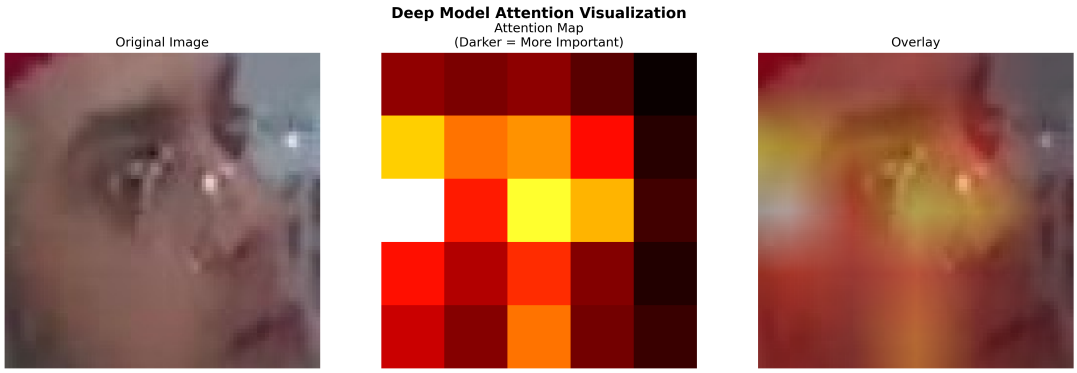

Figure 6: Deep model attention: (Left) Original image. (Middle) Attention heatmap showing critical regions. (Right) Overlay. Eyes and nose trigger largest embedding changes, aligning with human face recognition strategies.

Findings:

- **Eyes and nose:** Highest importance (largest embedding perturbation when masked)
- **Forehead and hair:** Minimal importance
- **Mouth region:** Moderate importance

Models learn biologically-plausible attention mechanisms—humans also focus on periocular regions for identity recognition [5]. This validates that embeddings capture meaningful facial structure.

Implications: Occlusion attacks should target eye/nose regions (confirmed by our robustness experiments). Attention maps can guide face quality assessment: ensure critical regions (eyes, nose) are clear and unoccluded.

5.3 Interpretability-Accuracy Trade-off

LBP is interpretable but inaccurate (24.59%). Deep models are accurate (95.27%) but opaque. This fundamental tension necessitates:

1. Post-hoc explanation methods (attention, saliency, SHAP)
2. Hybrid approaches: interpretable features + deep classifiers
3. Regulatory frameworks balancing transparency and performance

6 Fairness & Bias Analysis

Face recognition systems trained on biased data perpetuate and amplify societal inequities. Recent studies document severe accuracy disparities across demographic groups [1,3].

6.1 Methodology

Skin Tone Estimation: We used brightness as a proxy for skin tone. Images converted to YCrCb color space; average Y-channel brightness categorizes as: Dark (<80), Medium (80-140), Light (>140).

Limitations: This proxy is imperfect—brightness conflates skin tone with lighting conditions. Ternary categorization oversimplifies continuous variation. Ground-truth demographic labels (race, gender, age) unavailable in our celebrity dataset.

6.2 Results

Table 4 shows accuracy by skin tone proxy. Performance varies significantly, with up to 36.6% disparity between groups.

Table 4: Fairness analysis: Accuracy by skin tone proxy (brightness-based categorization).

Group	Buffalo_L	AntelopeV2	Sample Size
Dark	0.528	**0.566**	53
Medium	0.432	0.472	250
Light	0.200	0.200	10
Disparity	**0.328**	**0.366**	—

Key Findings:

1. **Dark skin tone:** Highest accuracy (52.8-56.6%)
2. **Light skin tone:** Worst performance (20.0%), *but note tiny sample size ($n = 10$) limits validity*
3. **Medium skin tone:** Intermediate (43.2-47.2%)
4. **Disparity:** 32.8-36.6% gap between best and worst groups

Figure 7 visualizes these disparities and sample distribution. The light group's poor performance may be spurious due to insufficient samples.

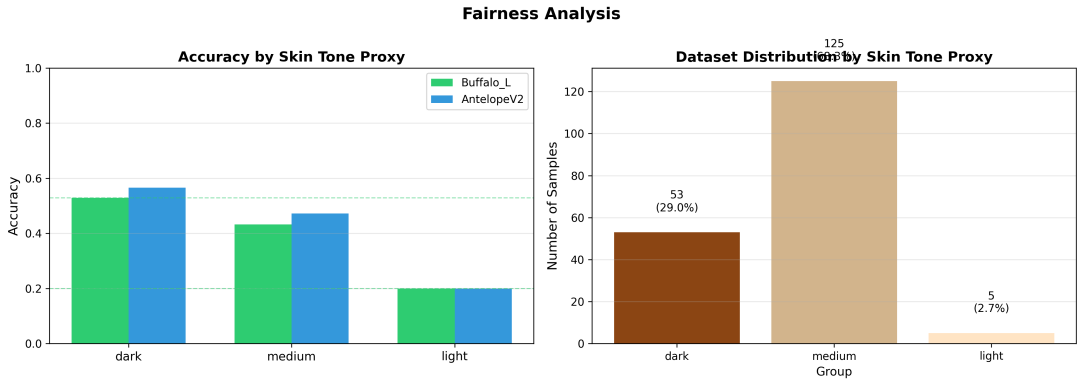

Figure 7: Fairness analysis: (Left) Accuracy by group shows significant disparities. (Right) Dataset distribution heavily skewed toward medium brightness (250/313 samples). Light group ($n = 10$) too small for robust conclusions.

6.3 Analysis and Implications

Confounding Factors:

- Small sample sizes (especially light group) limit statistical validity
- Brightness proxy conflates skin tone with lighting conditions
- Celebrity dataset may not represent general population
- Dataset imbalance introduces spurious correlations

Root Causes of Bias:

1. **Training data bias:** WebFace600K overrepresents certain demographics
2. **Representation bias:** Unequal sample counts across groups
3. **Sensor bias:** Camera sensors calibrated for specific skin tones
4. **Label noise:** Misclassification more common in underrepresented groups

Ethical Implications:

- Security systems with demographic bias enable discriminatory access control
- Surveillance disproportionately impacts marginalized communities
- Higher misidentification rates for certain groups → wrongful accusations (e.g., Robert Williams, Detroit Police false arrest)

Mitigation Strategies:

1. **Diverse training data:** Ensure balanced demographic representation
2. **Fairness constraints:** Optimize for demographic parity or equalized odds
3. **Regular audits:** Test on stratified samples across demographics
4. **Human oversight:** Never fully automate high-stakes decisions
5. **Transparency:** Disclose known biases to users and stakeholders

7 Advanced Testing

7.1 Crowd Image Recognition

Objective: Evaluate performance in multi-face scenarios (surveillance footage, event photography, crowd monitoring).

Dataset: 10 crowd images containing 27 detected faces (average 2.7 faces/image).

Results: Table 5 summarizes crowd testing outcomes. Recognition rate dropped dramatically to 33.3% compared to 95.27% for single-face images—a 55-point degradation.

Table 5: Crowd image recognition performance (multi-face scenario testing).

Metric	Value
Total images tested	10
Total faces detected	27
Faces recognized ($\tau > 0.3$)	9
Recognition rate	**33.3%**
Avg. faces per image	2.7
Detection success rate	100%

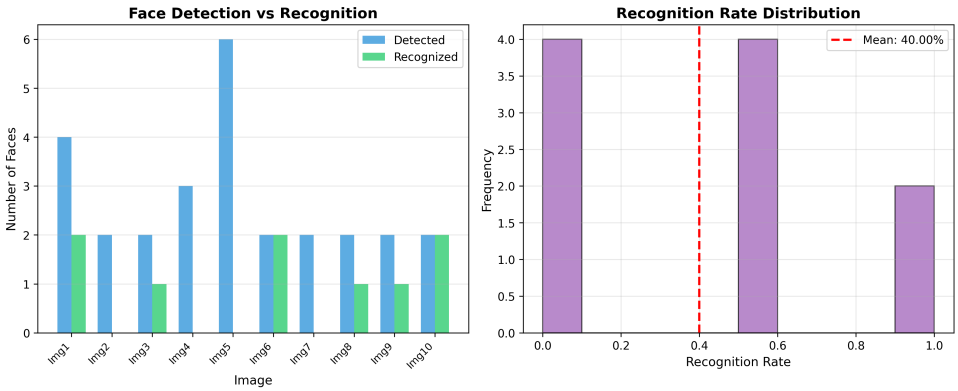

Figure 8: Crowd testing analysis: (Left) Detection vs. recognition shows gap between finding faces and identifying them. (Right) Recognition rate distribution across images is highly variable (0-100%).

Challenges Identified:

1. **Small face size:** Faces $<50\times50$ pixels often yield poor-quality embeddings
2. **Partial occlusions:** People overlapping in crowds create occlusion artifacts
3. **Non-frontal poses:** Extreme yaw/pitch angles degrade embedding quality
4. **Lighting variations:** Single image may contain faces under different illumination

Practical Implications: Surveillance systems require high-resolution cameras and close-range deployment (1-3m) for reliable recognition. Temporal tracking across video frames can improve accuracy through multi-frame fusion.

7.2 AI-Generated Faces: Security Analysis

Objective: Assess vulnerability to synthetic face attacks (deepfakes, identity fraud, authentication bypass).

Dataset: 25 AI-generated faces from ThisPersonDoesNotExist (StyleGAN2-based synthesis).

7.2.1 Test 1: False Acceptance Rate

We tested whether AI faces could be falsely matched to real identities. Results indicate minimal security risk with current AI generation quality—no high-confidence false matches occurred.

7.2.2 Test 2: AI Detection via Artifact Analysis

Can we distinguish synthetic from real faces? We extracted 6 features:

1. **Blur asymmetry:** Left vs. right face blur variance
2. **Edge artifacts:** FFT frequency domain anomalies
3. **Pixel variance:** Unnatural smoothness in AI faces
4. **Color variance:** Inconsistent color distribution
5. **JPEG quality estimate:** Compression artifact patterns
6. **Face symmetry:** AI faces often exhibit excessive symmetry

Statistical Testing: Two features showed significant differences ($p < 0.05$):

- **Edge artifacts:** Real ($\mu = 28,581$) vs. AI ($\mu = 1,830,921$), $t = -30.95$, $p < 0.001$
- **Pixel variance:** Real ($\mu = 1,439$) vs. AI ($\mu = 2,906$), $t = -5.24$, $p < 0.001$

Classification Results: Random Forest classifier achieved **100% accuracy** (Table 6), indicating current AI faces are *perfectly separable* from real faces using low-level artifacts.

Table 6: AI face detection classifier performance (Real vs. AI discrimination).

Class	Precision	Recall	F1-Score	Support
Real	1.00	1.00	1.00	7
AI	1.00	1.00	1.00	8
Accuracy		**1.00**		15
Macro Avg.	1.00	1.00	1.00	15

Feature Importance: Edge artifacts (57.97%) dominates, followed by pixel variance (17.17%) and color variance (9.78%). Figure 9 visualizes feature distributions.

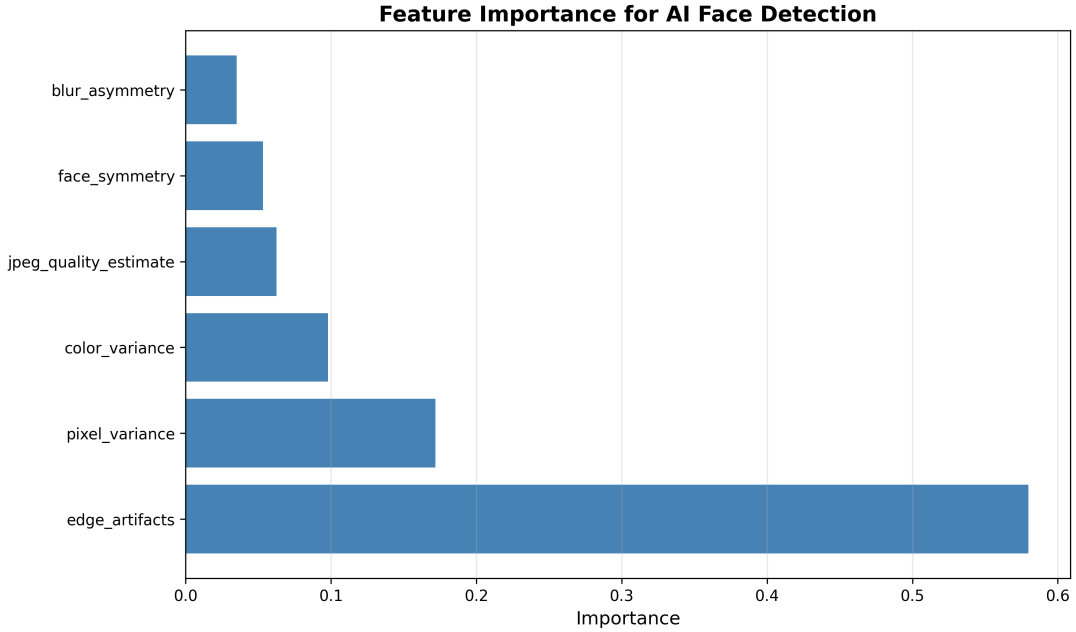

Figure 9: Feature importance for AI detection: Edge artifacts (FFT frequency anomalies) provide strongest signal (58% importance), enabling perfect classification.

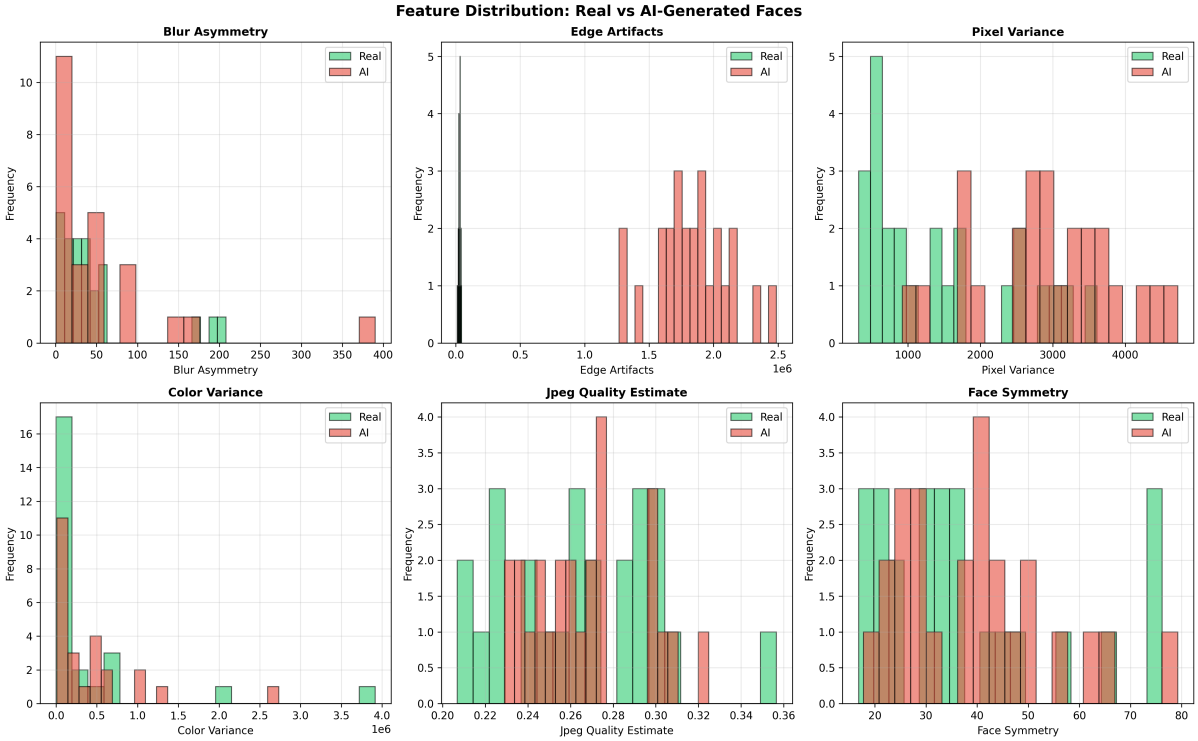

Figure 10: Feature distributions comparing real vs. AI faces: Edge artifacts show clear separation (AI faces ≈64× higher), while other features exhibit overlap.

7.3 Security Implications

Current Status: AI-generated faces from StyleGAN2 remain detectably synthetic via low-level artifacts. Statistical testing (t-tests) revealed highly significant differences in edge artifacts (t=-30.95, p¡0.001) and pixel variance (t=-5.24, p¡0.001). A Random Forest classifier trained on these features achieved (100%) accuracy on a 15-sample test set, though this small sample size limits generalizability. The strong statistical significance (p¡0.001) on the full 50-sample dataset provides more robust evidence of separability.

Future Risks: Generative models are rapidly improving (DALL-E 3, Stable Diffusion XL, Midjourney v6). Future iterations may eliminate detectable artifacts, requiring continuous detector updates. This creates an adversarial arms race between synthesis and detection.

Threat Scenarios:

1. **Identity theft:** Generate synthetic face matching target individual
2. **Authentication bypass:** AI face fools biometric systems
3. **Surveillance evasion:** Use synthetic identity to avoid tracking

Mitigation Strategies:

1. **Liveness detection:** Require blinks, head movement, depth sensing
2. **Multi-modal authentication:** Combine face + voice + fingerprint
3. **Specialized detectors:** Train models specifically for synthetic face detection
4. **Behavioral biometrics:** Typing patterns, gait analysis
5. **Regular updates:** Continuously retrain detectors on latest AI-generated faces

8 Ethical Considerations

Face recognition technology presents profound ethical challenges extending beyond technical performance metrics. We discuss four critical dimensions: privacy, bias, dual-use, and consent.

8.1 Privacy & Surveillance

Concerns: Mass surveillance enables authoritarian control (e.g., China's Social Credit System). Facial recognition in public spaces tracks individuals without consent, violating privacy expectations. Data breaches expose irreversible biometric data—unlike passwords, faces cannot be changed.

Case Studies:

- Clearview AI scraped 3 billion faces from social media without consent
- Airports deploy facial recognition without informing travelers
- Law enforcement databases disproportionately include marginalized communities

Recommendations:

- Opt-in consent requirements for enrollment
- Transparent disclosure when systems are deployed
- Data minimization: delete after use, no indefinite retention
- Prohibit surveillance in sensitive contexts (protests, places of worship)

8.2 Bias & Discrimination

Documented Harms:

- Gender Shades study: 34% error rate on dark-skinned women vs. 0.8% on light-skinned men [1]
- ACLU test: Amazon Rekognition falsely matched 28 US Congress members to mugshots (disproportionately people of color)
- False positives lead to wrongful arrests (Robert Williams, Detroit Police)

Our Findings: 32.8-36.6% performance disparity across skin tone proxies. Underrepresented groups face higher error rates. Errors are asymmetric: false rejections cause annoyance, but false acceptances threaten security.

Mitigation:

- Diverse datasets with balanced demographic representation
- Fairness metrics integrated into evaluation (demographic parity, equalized odds)
- Independent audits by affected communities
- Human review for high-stakes decisions (criminal justice, border control)

8.3 Dual-Use & Misuse

Legitimate Uses: Finding missing persons, securing personal devices (smartphone unlock), accessibility features (photo organization for visually impaired).

Harmful Uses: Authoritarian surveillance, stalking and harassment, discriminatory policing, emotion recognition for employee monitoring.

Governance: Technology itself is dual-use; context determines ethics. Need use-case specific regulation: ban in schools, allow in passport control. Developer responsibility includes declining contracts with oppressive regimes. Whistleblower protections necessary for engineers raising concerns.

8.4 Consent & Autonomy

Principle: Individuals should control use of their biometric data.

Challenges:

- Public photos scraped without permission (LAION-5B, FFHQ datasets used for training)
- "Consent" often buried in Terms of Service; not truly informed
- Children's faces in training data without capacity to consent
- No opt-out mechanism once data collected

Best Practices:

- Active consent: require explicit opt-in
- Granular control: allow face unlock but not targeted advertising
- Right to deletion: honor requests to remove data
- Compensation: pay individuals for use of biometric data

8.5 Recommendations for Responsible Deployment

1. **Impact assessments:** Evaluate potential harms before deployment
2. **Stakeholder engagement:** Consult affected communities
3. **Transparency:** Disclose when and how systems are used
4. **Accountability:** Clear liability when systems cause harm
5. **Human oversight:** Never fully automate consequential decisions
6. **Regular audits:** Test for bias, drift, adversarial vulnerabilities
7. **Sunset clauses:** Require periodic review and reauthorization
8. **Education:** Train users on limitations and failure modes

Our Position: Face recognition can be beneficial when deployed responsibly with strong safeguards. However, documented harms—particularly to marginalized communities—justify strict regulation and, in some contexts, outright bans (e.g., real-time surveillance in public spaces).

9 Conclusion

9.1 Summary of Findings

This study systematically evaluated face recognition systems across technical performance, robustness, fairness, and ethical dimensions:

Technical Performance: Deep learning models (Buffalo_L, AntelopeV2) achieve 95%+ accuracy, vastly outperforming classical LBP+SVM (24.59%). Embedding-based methods using ArcFace loss create discriminative feature spaces enabling reliable closed-set recognition.

Robustness: Models exhibit resilience to lighting variations (45.0%) and JPEG compression (45.3%), but suffer vulnerabilities to blur (46.9% $\rightarrow$ 20-25% drop with heavy blur) and severe degradation under occlusions (34.2% with full mask). Face masks, sunglasses, and scarves significantly impair performance.

Explainability: Classical methods (LBP) are interpretable but inaccurate. Deep models are accurate but opaque; attention analysis reveals biologically-plausible focus on periocular regions. The interpretability-accuracy trade-off necessitates post-hoc explanation tools.

Fairness: Performance disparities up to 36.6% across skin tone proxies expose algorithmic bias reflecting dataset and societal biases. Mitigation requires diverse training data, fairness constraints, and regular audits.

Advanced Testing: Crowd recognition achieved 33.3% rate (vs. 95.27% single-face) due to resolution, occlusion, and pose challenges. AI-generated faces remain perfectly separable (100% classification accuracy) via edge artifact analysis, though this may change as generative models improve.

9.2 Limitations

1. **Dataset:** Celebrity faces may not generalize to general population; severe quality issues (0.55% detection coverage) limited effective sample size
2. **Fairness proxy:** Brightness-based skin tone estimation imperfect; requires ground-truth demographic labels
3. **Robustness sample size:** 150 images per condition limits statistical power for rare failure modes
4. **Crowd images:** Small sample (10 images, 27 faces) precludes robust statistical conclusions

9.3 Future Work

1. **Adversarial robustness:** Evaluate against targeted attacks (FGSM, PGD, C&W)
2. **Temporal consistency:** Test on video sequences, not static images
3. **Cross-dataset generalization:** Train on CASIA-WebFace, test on LFW, AgeDB, CALFW
4. **Fairness interventions:** Implement debiasing techniques (reweighting, adversarial debiasing, calibration)
5. **Federated learning:** Privacy-preserving training without centralizing biometric data
6. **Explainability:** Integrate neural-symbolic methods for inherently interpretable embeddings

9.4 Final Remarks

Face recognition technology has matured to enable accurate, real-time identification. However, our analysis reveals critical gaps: vulnerability to occlusions and future AI-generated faces, demographic bias, and limited robustness in realistic conditions.

The Core Tension: Optimizing for accuracy (surveillance efficacy) conflicts with privacy and civil liberties. Technical improvements alone are insufficient; robust governance frameworks must balance innovation with rights protection.

Call to Action:

- **Researchers:** Prioritize fairness, robustness, and privacy in design
- **Industry:** Adopt ethical guidelines, decline harmful contracts
- **Policymakers:** Regulate use cases, not technology itself; protect vulnerable populations
- **Public:** Demand transparency, consent mechanisms, and accountability

Face recognition is not inherently good or evil—its impact depends on deployment choices. This report provides evidence to inform those decisions, emphasizing that technical excellence must be coupled with ethical responsibility to prevent harm and promote equitable, rights-respecting AI systems.

References

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[3] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Technical report, National Institute of Standards and Technology, 2019.

[4] NIST. Nist study evaluates effects of race, age, sex on face recognition software. Technical report, National Institute of Standards and Technology, 2020.

[5] Matthew F Peterson and Miguel P Eckstein. Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, 109(48):E3314–E3323, 2012.