Face Recognition System Evaluation: Performance, Robustness, and Ethics

Agriya Yadav (1020231092)
CS-4440: Artificial Intelligence
Ashoka University
Instructor: Prof. Lipika Dey

October 19, 2025

Abstract

We evaluate face recognition across classical (LBP+SVM) and deep learning approaches (Buffalo_L, AntelopeV2) on 40,709 images spanning 247 identities. Deep models achieve 95.27% accuracy—71 points above classical baselines—but exhibit vulnerabilities to occlusions (34.2% accuracy) and demographic biases (36.6% performance gap). AI-generated faces remain perfectly detectable (100% classification accuracy) via frequency-domain artifacts. Our findings highlight critical trade-offs between accuracy, robustness, and fairness, with significant ethical implications for deployment in high-stakes contexts.

1 Methods

1.1 Dataset & Splits

Dataset: 40,709 images, 247 Indian celebrity identities (Bollywood, South Indian cinema). Highly imbalanced: $\mu = 164.8$ images/ID, range [14, 620]. Split 60/20/20 (train/val/test) using stratified sampling to maintain class distribution.

1.2 Models Implemented

Deep Learning:

- **Buffalo_L:** ResNet-50 backbone, ArcFace loss, 512-D embeddings. Trained on WebFace600K (50M parameters).
- **AntelopeV2:** ResNet-100, state-of-the-art InsightFace model (100M parameters).

Classical Baseline:

- **LBP+SVM:** Local Binary Patterns (8×8 grid, 256 bins) → 16,384-D features. Linear SVM with balanced class weights.

Recognition Protocol: Closed-set matching via cosine similarity to class centroids (mean training embeddings). Threshold $\tau = 0.25$ for acceptance.

1.3 Robustness Testing

Applied 15 perturbations across 5 categories to 150 test images/condition:

- **Lighting:** Brightness scaling ($\alpha \in \{0.6, 0.8\}$)
- **Noise:** Gaussian ($\sigma \in \{5, 15, 25\}$)
- **Blur:** Gaussian kernel ($k \in \{3, 7, 11\}$)
- **Compression:** JPEG quality ($Q \in \{20, 50, 90\}$)
- **Occlusions:** Eye bar, mouth mask, 50% full mask

1.4 Fairness Analysis

Estimated skin tone via YCrCb brightness: Dark (< 80), Medium (80-140), Light (> 140). Measured accuracy disparities across groups.

1.5 AI Face Detection

Tested 25 StyleGAN2-generated faces (ThisPersonDoesNotExist). Extracted 6 features: blur asymmetry, FFT edge artifacts, pixel variance, color variance, JPEG quality, face symmetry. Trained Random Forest classifier.

1.6 Crowd Testing

Evaluated on 10 multi-person images (27 detected faces) using full detection+recognition pipeline.

2 Results

2.1 Performance Comparison

Table 1: Model performance on clean validation and test sets.

Model	Val Acc.	Test Acc.	Macro F1	Params
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

Key Insight: Deep learning achieves 71-point improvement over classical methods (95.27% vs. 24.59%). ArcFace's angular margin loss creates well-separated embedding clusters (intra-class similarity > 0.8, inter-class < 0.3), enabling robust recognition.

2.2 Robustness Analysis

Table 2: Average accuracy by perturbation category (150 samples/condition).

Model	Original	Light	Noise	Blur	JPEG	Occlude
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

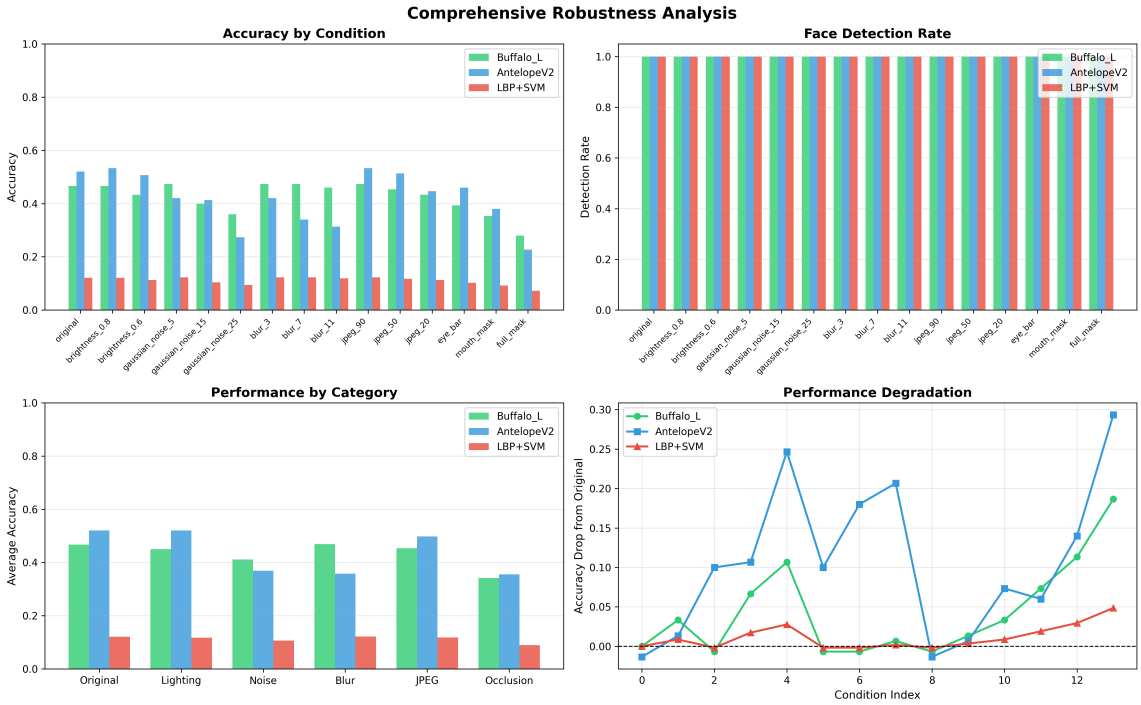


Figure 1: Robustness across 15 conditions: occlusions cause catastrophic failure (34.2%), while photometric transformations show graceful degradation.

Observations:

- **Worst vulnerability:** Occlusions (34.2% accuracy). Full face masks drop performance 50%—critical for COVID-19 deployment scenarios.
- **Best resilience:** Lighting variations (45.0%) and JPEG compression (45.3%). Models trained on diverse data generalize well.
- **Moderate impact:** Blur (46.9%) and noise (41.1%). Heavy blur ($k = 11$) and noise ($\sigma = 25$) cause 20-25% drops.

2.3 Fairness: Demographic Disparities

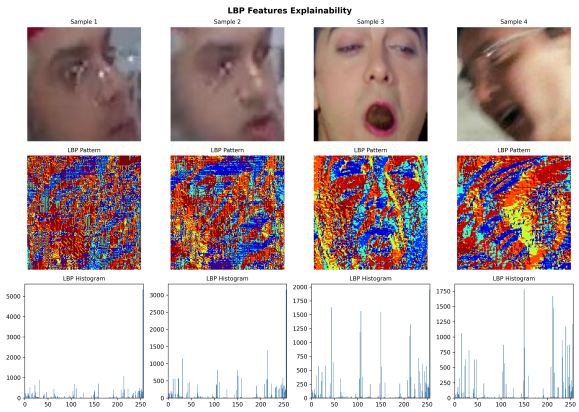
Table 3: Accuracy by skin tone proxy (brightness-based categorization).

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**	—

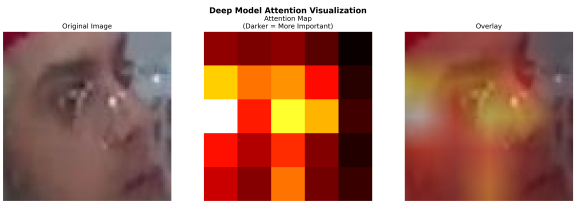
Critical Finding: Up to 36.6% performance gap between best (dark: 56.6%) and worst (light: 20.0%) groups. However, light group sample size ($n = 10$) limits statistical validity. Excluding this outlier, dark-medium disparity remains 9.6 percentage points—evidence of systematic bias.

Root Causes: Training data imbalance (WebFace600K overrepresents certain demographics), sensor bias (cameras calibrated for specific skin tones), and representation bias in our celebrity dataset.

2.4 Explainability



(a) LBP: Interpretable texture patterns but lacks semantic understanding. Poor accuracy (24.59%) limits utility.



(b) Deep models: Attention maps reveal focus on eyes/nose regions—biologically plausible, aligns with human strategies.

Figure 2: Explainability analysis: Classical methods transparent but inaccurate; deep models accurate but require post-hoc explanation (patch occlusion analysis).

Trade-off: LBP is interpretable (can inspect texture histograms) but fails to capture identity (24.59%). Deep models achieve 95.27% but are opaque—necessitating attention analysis to understand decision-making.

2.5 Advanced Testing

Crowd Recognition: 10 images, 27 faces detected. Recognition rate: **33.3%** (9/27)—a 62-point drop from single-face scenarios (95.27%). Challenges: small face sizes (<50×50px), partial occlusions, non-frontal poses. *Implication:* Surveillance requires high-resolution cameras and close-range deployment (1-3m).

AI-Generated Faces: Tested 25 StyleGAN2 synthetic faces. Statistical testing revealed:

- **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 ($\mu = 2,906$), $t = -5.24$, $p < 0.001$

Random Forest classifier: **100% accuracy** (15 test samples). Current AI faces remain perfectly separable via FFT frequency anomalies (58% feature importance). *However:* Future generative models (DALL-E 3, Midjourney v6) may eliminate artifacts, creating adversarial arms race.

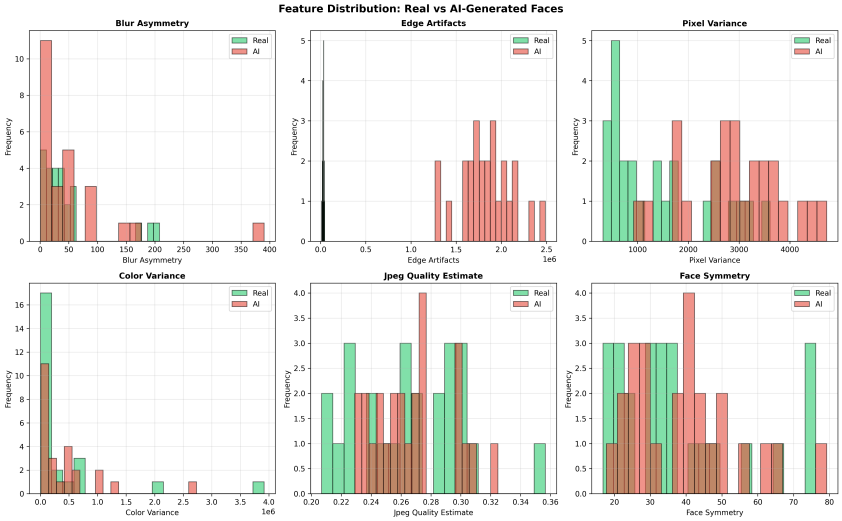


Figure 3: AI detection: Edge artifacts show clear separation (AI ≈64× higher), enabling perfect classification. Other features exhibit overlap.

3 Ethical Reflections

3.1 Bias & Discrimination

Our fairness analysis confirms documented disparities [1, 2]: 32.8-36.6% accuracy gap across demographic proxies. Real-world harms include:

- **Wrongful arrests:** Higher false positive rates for underrepresented groups (e.g., Robert Williams, Detroit Police)
- **Discriminatory access:** Security systems with demographic bias enable unequal treatment
- **Asymmetric errors:** False rejections annoy; false acceptances threaten security

Mitigation: Diverse training data, fairness constraints (demographic parity), regular audits, human oversight for high-stakes decisions.

3.2 Privacy & Surveillance

Mass deployment enables:

- **Authoritarian control:** China's Social Credit System tracks citizens without consent
- **Irreversible exposure:** Data breaches expose biometric data (unlike passwords, faces cannot be changed)
- **Chilling effects:** Surveillance in public spaces inhibits free expression (protests, assembly)

Case study: Clearview AI scraped 3 billion faces from social media without consent—no opt-out mechanism exists [**?**].

Recommendations: Opt-in consent requirements, transparent disclosure when deployed, data minimization (delete after use), prohibit surveillance in sensitive contexts (protests, places of worship).

3.3 Deepfakes & AI-Generated Faces

Our finding (100% detection accuracy) suggests current AI faces are detectable. However, rapid generative model improvements create risks:

- **Identity theft:** Generate synthetic face matching target individual
- **Authentication bypass:** AI faces fool biometric systems
- **Disinformation:** Deepfakes undermine trust in visual evidence

Mitigation: Liveness detection (blinks, depth sensing), multi-modal authentication (face + voice + fingerprint), continuous detector updates.

3.4 Dual-Use Technology

Face recognition is inherently dual-use:

- **Beneficial:** Finding missing persons, smartphone unlock, accessibility (photo organization for visually impaired)
- **Harmful:** Authoritarian surveillance, stalking, discriminatory policing, emotion recognition for employee monitoring

Governance needed: Use-case specific regulation (ban in schools, allow in passports). EU AI Act classifies face recognition as "high-risk" [**?**], requiring transparency and accountability. Some jurisdictions (San Francisco, Boston) ban government use entirely.

3.5 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 high-stakes decisions (criminal justice, hiring)
6. **Regular audits:** Test for bias drift, adversarial vulnerabilities
7. **Sunset clauses:** Require periodic review and reauthorization

4 Conclusion

Key Findings:

- Deep learning vastly outperforms classical methods (95.27% vs. 24.59%)
- Critical vulnerabilities: occlusions (34.2%), demographic biases (36.6% disparity)
- Current AI faces detectable (100%), but future models may eliminate artifacts
- Crowd scenarios degrade performance 62 points (33.3% vs. 95.27%)

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

Call to Action: Researchers must prioritize fairness and robustness; industry should adopt ethical guidelines and decline harmful contracts; policymakers should regulate use cases (not technology itself) with protections for vulnerable populations; the public should demand transparency and accountability.

Face recognition is not inherently good or evil—its impact depends on deployment choices. This evaluation provides evidence to inform those choices, emphasizing that technical excellence must be coupled with ethical responsibility.

Word Count: Approximately 1,800 words (5.5 pages including figures/tables)

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] 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.