# Face Recognition Model Comparison Results

Date: January 26, 2026

Test Environment: Windows 11, Python 3.11, ONNX Runtime

Database: PostgreSQL (Azure-hosted)

Total Indexed Visitors: ~71,500-71,700

## Executive Summary

A comprehensive comparison was conducted between SFace (128-dimensional) and ArcFace (512-dimensional) face recognition models, including threshold optimization to find the best configuration for each. SFace demonstrated superior performance across all metrics.

### Complete Comparison Summary (Single View)

```
+============================================================================+
|          FACE RECOGNITION MODEL COMPARISON: SFace vs ArcFace               |
|                        Test Date: January 26, 2026                         |
+============================================================================+
|  TEST CONFIGURATION                                                        |
|  ------------------------------------------------------------------------  |
|  Database: PostgreSQL (Azure)     |  Indexed Visitors: ~71,500             |
|  Test Subjects: 7 persons         |  Total Queries: 350                    |
|  Queries per Person: 50           |  Min Entries/Person: 100               |
+============================================================================+
|  MODEL SPECIFICATIONS                                                      |
|  ------------------------------------------------------------------------  |
|                          |  ArcFace          |  SFace            |  Winner  |
|  Feature Dimensions      |  512-dim          |  128-dim          |  SFace   |
|  Index Size (vectors)    |  71,526           |  71,677           |  -       |
|  Index File Size         |  ~146 MB          |  ~37 MB           |  SFace   |
|  Best F1 Threshold       |  0.90             |  0.70             |  -       |
+============================================================================+
|  ACCURACY RESULTS (at Best F1 Thresholds)                                  |
|  ------------------------------------------------------------------------  |
|                          |  ArcFace @ 0.90   |  SFace @ 0.70     |  Winner  |
|  ------------------------------------------------------------------------  |
|  * TOP-1 ACCURACY        |  73.71%           |  81.71%           |  SFace   |
|    Top-5 Accuracy        |  82.57%           |  92.86%           |  SFace   |
|    Top-10 Accuracy       |  86.57%           |  93.14%           |  SFace   |
|  ------------------------------------------------------------------------  |
|    Precision             |  73.71%           |  82.99%           |  SFace   |
|    Recall                |  100.00%          |  98.95%           |  ArcFace |
|    F1 Score              |  84.87%           |  90.27%           |  SFace   |
+============================================================================+
|  SCORE ANALYSIS                                                            |
|  ------------------------------------------------------------------------  |
|  Avg Same-Person Score   |  0.9946           |  0.9337           |  ArcFace |
|  Avg Diff-Person Score   |  0.9848           |  0.7899           |  SFace   |
|  * SCORE SEPARABILITY    |  0.0098           |  0.1438           |  SFace   |
|                          |                   |  (14.7x better)   |          |
+============================================================================+
|  PERFORMANCE                                                               |
|  ------------------------------------------------------------------------  |
|  Processing Time (350q)  |  31.6 sec         |  14.9 sec         |  SFace   |
|  Queries/Second          |  11.1 q/s         |  23.5 q/s         |  SFace   |
|  Relative Speed          |  1.0x             |  2.12x faster     |  SFace   |
+============================================================================+
|  PER-PERSON BREAKDOWN                                                      |
|  ------------------------------------------------------------------------  |
|  Person               |  ArcFace    |  SFace      |  Winner                |
|  ------------------------------------------------------------------------  |
|  Joana Celoso         |  90.0%      |  96.0%      |  SFace (+6.0%)         |
|  Dannie rose Estomo   |  98.0%      |  100.0%     |  SFace (+2.0%)         |
|  Marygrace Gabatas    |  78.0%      |  80.0%      |  SFace (+2.0%)         |
|  Janelle Vallido      |  86.0%      |  84.0%      |  ArcFace (+2.0%)       |
|  Emarjoy Maramba      |  58.0%      |  70.0%      |  SFace (+12.0%)        |
|  Regi Fugoso          |  34.0%      |  62.0%      |  SFace (+28.0%)        |
```

```
|  Sid Villacorta       |  72.0%      |  80.0%      |  SFace (+8.0%)          |
|  ---------------------------------------------------------------------------  |
|  TOTAL WINS           |  1/7        |  6/7        |  SFace                  |
+=============================================================================+
|  FINAL VERDICT                                                              |
|  ========================================================================   |
|                                                                             |
|   *** RECOMMENDED MODEL: SFace ***                                          |
|                                                                             |
|   ? +8.0% better Top-1 Accuracy (81.71% vs 73.71%)                         |
|   ? +5.4% better F1 Score (90.27% vs 84.87%)                               |
|   ? 14.7x better score separability for reliable matching                  |
|   ? 2.12x faster processing speed                                          |
|   ? 4x smaller storage (128-dim vs 512-dim)                                |
|   ? Wins 6 out of 7 test subjects                                          |
|                                                                             |
|   Recommended Threshold: 0.70 (best F1) or 0.45 (max recall)               |
|                                                                             |
+=============================================================================+
```

### Quick Reference Table

| Metric | ArcFace @ 0.90 | SFace @ 0.70 | Winner |
|---|---|---|---|
| Top-1 Accuracy | 73.71% | 81.71% | SFace (+8.0%) |
| Best F1 Score | 84.87% | 90.27% | SFace (+5.4%) |
| Precision | 73.71% | 82.99% | SFace (+9.3%) |
| Recall | 100.00% | 98.95% | ArcFace (+1.0%) |
| Score Separability | 0.0098 | 0.1438 | SFace (14.7x better) |
| Processing Speed | 31.6s | 14.9s | SFace (2.12x faster) |

### All Threshold Tests Summary

| Test | ArcFace Threshold | SFace Threshold | ArcFace Top-1 | SFace Top-1 | Winner |
|---|---|---|---|---|---|
| Best F1 | 0.90 | 0.70 | 73.71% | 81.71% | SFace |
| Max Recall | 0.40 | 0.45 | 73.71% | 81.71% | SFace |
| Same Threshold | 0.55 | 0.55 | 73.71% | 81.71% | SFace |

Note: Conclusion: SFace consistently outperforms ArcFace by +8% regardless of threshold configuration.

### Recommended Thresholds

| Model | Use Case | Threshold | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| SFace | Best F1 (Recommend | 0.70 | 82.99% | 98.95% | 90.27% |
| SFace | Max Recall | 0.45 | 81.71% | 100.00% | 89.94% |
| ArcFace | Best F1 | 0.90 | 73.71% | 100.00% | 84.87% |
| ArcFace | Max Recall | 0.40 | 73.71% | 100.00% | 84.87% |

Note: Note: SFace @ 0.70 is the recommended configuration, offering the best balance of precision (82.99%) and recall (98.95%) with the highest F1 score (90.27%).

## 1. Test Methodology

### 1.1 Cross-Image Recognition Test

Unlike simple self-recognition tests, the cross-image test evaluates real-world performance:

1. Find repeat visitors: Query database for individuals with ?100 separate entries

2. For each person: Take one image as query, search for OTHER images of the SAME person

3. Success criteria: Another image of the same person appears in top-1 results

### 1.2 Test Dataset

| Parameter | Value |
|---|---|
| Persons tested | 7 |
| Total entries | 1,472 |
| Queries per person | 50 |
| Total queries | 350 |

### 1.3 Test Subjects

- Joana Celoso
- Dannie rose Estomo
- Marygrace Gabatas
- Janelle Vallido
- Emarjoy Maramba
- Regi Fugoso
- Sid Villacorta

## 2. Threshold Optimization Results

### 2.1 SFace Threshold Analysis

| Threshold | Top-1 Acc | Precision | Recall | F1 Score | Notes |
|---|---|---|---|---|---|
| 0.30-0.60 | 81.71% | 81.71% | 100.00% | 89.94% | Max recall |
| 0.65 | 81.71% | 81.66% | 99.65% | 89.76% | |
| 0.70 | 81.71% | 82.99% | 98.95% | 90.27% | Best F1 |
| 0.75 | 81.71% | 85.58% | 95.45% | 90.25% | Higher precision |
| 0.80 | 81.71% | 90.44% | 86.01% | 88.17% | Max precision |

SFace Optimal: 0.70 - Achieves the best balance with 90.27% F1 score while maintaining 98.95% recall.

### 2.2 ArcFace Threshold Analysis

| Threshold | Top-1 Acc | Precision | Recall | F1 Score | Notes |
|---|---|---|---|---|---|
| 0.30-0.92 | 73.71% | 73.71% | 100.00% | 84.87% | Max recall |
| 0.94 | 73.71% | 73.64% | 99.61% | 84.68% | |
| 0.96 | 73.71% | 74.27% | 98.45% | 84.67% | |
| 0.97 | 73.71% | 75.38% | 96.12% | 84.50% | |
| 0.98 | 73.71% | 76.39% | 90.31% | 82.77% | |
| 0.99 | 73.71% | 84.98% | 76.74% | 80.65% | Max precision |

ArcFace Optimal: 0.90 - All scores are so high that thresholds 0.30-0.92 give identical results (100% recall). This indicates poor score separability - the model struggles to distinguish same-person from different-person matches.

### 2.3 Key Insight: Score Distribution

The threshold optimization reveals a critical difference:

- ArcFace: Scores are clustered very close together (0.98-0.99+), making threshold selection nearly irrelevant until very high values. The narrow score gap (0.0098) means the model cannot reliably distinguish between matches.
- SFace: Scores are better distributed with a clear gap (0.1438) between same-person (~0.93) and different-person (~0.79) matches. This allows meaningful threshold tuning.

## 3. Cross-Image Accuracy Results

### 3.1 Comparison at Optimal Thresholds

| Metric | ArcFace @ 0.90 | SFace @ 0.70 | Winner |
|---|---|---|---|
| Top-1 Accuracy | 73.71% | 81.71% | SFace (+8.0%) |
| Top-5 Accuracy | 82.57% | 92.86% | SFace (+10.3%) |
| Top-10 Accuracy | 86.57% | 93.14% | SFace (+6.6%) |
| Precision | 73.71% | 82.99% | SFace (+9.3%) |
| Recall | 100.00% | 98.95% | ArcFace (+1.0%) |
| F1 Score | 84.87% | 90.27% | SFace (+5.4%) |

### 3.2 Score Analysis

| Metric | ArcFace | SFace | Analysis |
|---|---|---|---|
| Avg Same-Person Score | 0.9946 | 0.9337 | ArcFace higher |
| Avg Different-Person Score | 0.9848 | 0.7899 | SFace rejects better |
| Score Gap (Separability) | 0.0098 | 0.1438 | SFace 14.7x better |

### 3.3 Per-Person Breakdown

| Person | ArcFace | SFace | Winner |
|---|---|---|---|
| Joana Celoso | 90.0% | 96.0% | SFace |
| Dannie rose Estomo | 98.0% | 100.0% | SFace |
| Marygrace Gabatas | 78.0% | 80.0% | SFace |
| Janelle Vallido | 86.0% | 84.0% | ArcFace |
| Emarjoy Maramba | 58.0% | 70.0% | SFace |
| Regi Fugoso | 34.0% | 62.0% | SFace |
| Sid Villacorta | 72.0% | 80.0% | SFace |
| Total Wins | 1 | 6 | SFace |

### 3.4 Performance

| Metric | ArcFace | SFace |
|---|---|---|
| Test Duration | 23.7s | 12.3s |
| Queries/Second | 14.8 | 28.5 |
| Relative Speed | 1.0x | 1.93x |

### 3.5 Comparison at Alternative Thresholds (0.40 vs 0.45)

An additional test was run using more aggressive thresholds to verify consistency:

| Metric | ArcFace @ 0.40 | SFace @ 0.45 | Winner |
|---|---|---|---|
| Top-1 Accuracy | 73.71% | 81.71% | SFace (+8.0%) |
| Top-5 Accuracy | 82.57% | 92.86% | SFace (+10.3%) |
| Top-10 Accuracy | 86.57% | 93.14% | SFace (+6.6%) |
| Score Gap | 0.0098 | 0.1438 | SFace (14.7x better) |
| Test Duration | 24.7s | 13.4s | SFace (1.84x faster) |

Per-Person Results:

| Person | ArcFace @ 0.40 | SFace @ 0.45 | Winner |
|---|---|---|---|
| Joana Celoso | 90.0% | 96.0% | SFace |
| Dannie rose Estomo | 98.0% | 100.0% | SFace |

| | | | |
|---|---|---|---|
| Marygrace Gabatas | 78.0% | 80.0% | SFace |
| Janelle Vallido | 86.0% | 84.0% | ArcFace |
| Emarjoy Maramba | 58.0% | 70.0% | SFace |
| Regi Fugoso | 34.0% | 62.0% | SFace |
| Sid Villacorta | 72.0% | 80.0% | SFace |
| Total Wins | 1 | 6 | SFace |

Key Finding: SFace consistently outperforms ArcFace regardless of threshold selection, confirming the robustness of the SFace advantage for this dataset.

### 3.6 Comparison at Best F1 Thresholds (0.70 vs 0.90)

This test uses the thresholds identified as optimal during threshold optimization testing:

- SFace @ 0.70: Best F1 score (90.27%) with balanced precision/recall
- ArcFace @ 0.90: Best achievable F1 score (84.87%) while maintaining 100% recall

| Metric | | | |
|---|---|---|---|
| Top-1 Accuracy | 73.71% | 81.71% | SFace (+8.0%) |
| Top-5 Accuracy | 82.57% | 92.86% | SFace (+10.3%) |
| Top-10 Accuracy | 86.57% | 93.14% | SFace (+6.6%) |
| Score Gap | 0.0098 | 0.1438 | SFace (14.7x better) |
| Test Duration | 31.6s | 14.9s | SFace (2.12x faster) |

Per-Person Results at Best F1 Thresholds:

| Person | ArcFace @ 0.90 | SFace @ 0.70 | Winner |
|---|---|---|---|
| Joana Celoso | 90.0% | 96.0% | SFace |
| Dannie rose Estomo | 98.0% | 100.0% | SFace |
| Marygrace Gabatas | 78.0% | 80.0% | SFace |
| Janelle Vallido | 86.0% | 84.0% | ArcFace |
| Emarjoy Maramba | 58.0% | 70.0% | SFace |
| Regi Fugoso | 34.0% | 62.0% | SFace |
| Sid Villacorta | 72.0% | 80.0% | SFace |
| Total Wins | 1 | 6 | SFace |

Conclusion: Even at their individually optimized thresholds for best F1 score, SFace maintains its 8% accuracy advantage over ArcFace while being over 2x faster.

## 4. Index Verification

### 4.1 Index Statistics

| Model | Vectors Indexed | Dimensions | Index File |
|---|---|---|---|
| SFace | 71,677 | 128 | `hnsw_sface_128.bin` |
| ArcFace | 71,526 | 512 | `hnsw_arcface_512.bin` |

### 4.2 Self-Recognition Test

Both models pass self-recognition (score = 1.0):

```
SFace:   cm15scvft000lll0ccwg8r4me: 1.0000 ?
ArcFace: cm15scvft000lll0ccwg8r4me: 1.0000 ?
```

## 5. Technical Configuration

### 5.1 Model Details

| Model | File | Dimensions | Optimal Threshold |
|---|---|---|---|
| SFace | `face_recognition_sface_2021de | 128 | 0.70 |
| ArcFace | `arcface.onnx` | 512 | 0.90 |

### 5.2 Face Detection

| Component | Model |
|---|---|
| Face Detector | YuNet (`face_detection_yunet_2 |
| Face Alignment | 5-point landmark alignment |

### 5.3 HNSW Index Parameters

| Parameter | Value |
|---|---|
| M (connections per layer) | 16 |
| ef_construction | 200 |
| ef_search | 50 |
| Metric | Cosine similarity |

## 6. Recommendations

### 6.1 Model Selection

Recommended: SFace for the following reasons:

| Factor | SFace Advantage |
|---|---|
| Accuracy | +8% Top-1, +10% Top-5 |
| F1 Score | +5.4% (90.27% vs 84.87%) |
| Separability | 14.7x better score gap |
| Speed | 1.93x faster |
| Storage | 4x smaller (128-dim vs 512-dim |

### 6.2 Threshold Configuration

| Use Case | SFace Threshold | ArcFace Threshold |
|---|---|---|
| Balanced (Recommended) | 0.70 | 0.90 |
| High Recall (fewer misses) | 0.55 | 0.90 |
| High Precision (fewer false ma | 0.80 | 0.99 |

### 6.3 Why ArcFace Underperforms Here

ArcFace's poor score separability (0.0098) suggests potential issues:

1. Model-data mismatch: ArcFace may be optimized for different face characteristics

2. Feature distribution: All embeddings cluster too closely in 512-dim space

3. Index compatibility: Consider re-evaluating the indexing approach

### 6.4 Future Improvements

1. Expand test dataset: Include more persons with varied demographics

2. Test edge cases: Poor lighting, extreme angles, masks

3. ROC curve analysis: Full precision-recall curves

4. Ensemble approach: Consider combining models for critical applications

# 7. Appendix

## 7.1 Test Scripts

| Script | Purpose |
|---|---|
| `tests/optimize_thresholds.py` | Threshold optimization across |
| `tests/compare_recognizers.py` | Cross-image comparison test |
| `tests/verify_indexes.py` | Index loading and self-recogni |
| `tests/test_accuracy.py` | Basic accuracy testing |
| `tests/test_cross_image_accura | Cross-image accuracy testing |
| `scripts/rebuild_for_recognize | Rebuild HNSW index for specifi |

## 7.2 Commands Used

```
# Threshold optimization (both models)
python tests/optimize_thresholds.py
# Comparison at single threshold (same for both)
python tests/compare_recognizers.py --threshold 0.55
# Comparison at lower thresholds (max recall)
python tests/compare_recognizers.py --arcface-threshold 0.40 --sface-threshold 0.45
# Comparison at Best F1 thresholds (optimal balance)
python tests/compare_recognizers.py --arcface-threshold 0.90 --sface-threshold 0.70
# Index verification
python tests/verify_indexes.py
```

## 7.3 Raw Threshold Optimization Output

```
 Threshold    Top-1 Acc    Precision    Recall      F1 Score
------------------------------------------------------------
      0.30       81.71%       81.71%    100.00%       89.94%
      0.35       81.71%       81.71%    100.00%       89.94%
      0.40       81.71%       81.71%    100.00%       89.94%
      0.45       81.71%       81.71%    100.00%       89.94%
      0.50       81.71%       81.71%    100.00%       89.94%
      0.55       81.71%       81.71%    100.00%       89.94%
      0.60       81.71%       81.71%    100.00%       89.94%
      0.65       81.71%       81.66%     99.65%       89.76%
      0.70       81.71%       82.99%     98.95%       90.27% << BEST F1
      0.75       81.71%       85.58%     95.45%       90.25%
      0.80       81.71%       90.44%     86.01%       88.17%
```

```
 Threshold    Top-1 Acc    Precision    Recall      F1 Score
------------------------------------------------------------
      0.30       73.71%       73.71%    100.00%       84.87%
      ...        (same results through 0.92)
      0.94       73.71%       73.64%     99.61%       84.68%
      0.96       73.71%       74.27%     98.45%       84.67%
      0.97       73.71%       75.38%     96.12%       84.50%
      0.98       73.71%       76.39%     90.31%       82.77%
      0.99       73.71%       84.98%     76.74%       80.65%
```

Document Version: 2.3

Last Updated: January 26, 2026

Recommended Configuration: SFace @ 0.70 (Best F1: 90.27%, Top-1: 81.71%)