Journal of Computer Engineering, Electronics and Information Technology (COELITE)

Journal homepage: https://ejournal.upi.edu/index.php/COELITE

A Comparative Evaluation of State-of-the-Art Deep Face Recognition Models on the Indonesian Muslim Student Face Dataset

Mokhamad Arfan Wicaksono[1], Ibrahim Faizal Burhan[2], Hadi Permana[3], Muhung Anggarawan[4], Bakhriyah Firdausi[5], Muhammad Rafid Amrullah[6], Ratnasari Nur Rohmah[7]*

[1,2,3,4,5] Department of Software Engineering Technology, Akademi Digital Bandung, Indonesia
[6] Software Engineer, Avey, Qatar
[7] Faculty of Engineering, Universitas Muhammadiyah Surakarta, Indonesia
Correspondence E-mail: marfanw@gmail.com

ABSTRACT

Face recognition systems perform well on Western-centric benchmarks, but their reliability on diverse populations is less studied. This paper evaluates six state-of-the-art models on the Indonesian Muslim Student Face Dataset (IMSFD), containing 26,760 images of 68 individuals under varied conditions, including hijab use. Models tested include FaceNet, AdaFace, MagFace, and ElasticFace-Arc using identification and verification protocols, with metrics such as CMC, ROC/DET, EER, and decidability index, along with computational profiling. Results show AdaFace-R50 achieves the best rank-1 accuracy at 91.70% (95% CI: [90.98%, 92.38%]), significantly outperforming others. Performance varies across subsets, with A2 reaching ≥99.40% accuracy, while B1 is most challenging (≤86.05%). These findings stress the importance of diverse datasets and highlight the effectiveness of modern margin-based loss functions.

ARTICLE INFO

Article History:
Submitted/Received 16 Jan 2026
First Revised 23 Feb 2026
Accepted 09 Mar 2026
First Available online 01 Apr 2026
Publication Date 01 Apr 2026

Keyword:
face recognition, deep learning, comparative evaluation, Indonesian faces, angular margin loss, biometric performance.

© 2026 Universitas Pendidikan Indonesia

1. INTRODUCTION

Face recognition has become one of the most widely deployed biometric technologies, with applications spanning security, surveillance, mobile device authentication, and identity verification [1]. Recent advances in deep learning, particularly the development of discriminative loss functions such as ArcFace [2], CosFace [3], and their variants, have pushed face recognition accuracy to near-perfect levels on established benchmarks like Labeled Faces in the Wild [4] and MegaFace [5].

However, multiple studies have demonstrated that face recognition performance varies significantly across demographic groups defined by race, ethnicity, and gender [6]. Models trained predominantly on Western faces often exhibit degraded performance on underrepresented populations. This disparity is particularly acute for individuals who wear religious head coverings, such as the hijab, which partially occludes the face boundary and alters the overall facial appearance context [7].

Indonesia, the world's most populous Muslim-majority country, presents a unique evaluation context. The Indonesian Muslim Student Face Dataset [8] was specifically curated to address the scarcity of face datasets representing Indonesian individuals, many of whom wear hijab. Despite the availability of IMSFD, no comprehensive comparative evaluation of modern face recognition models has been conducted on this dataset.

This paper addresses this gap by presenting a rigorous, reproducible evaluation of six state-of-the-art face recognition models on IMSFD. Our contributions are as follows:

1. We conduct the first comprehensive comparative evaluation of six modern deep face recognition models—FaceNet [9] (two variants), AdaFace [10] (two architectures), MagFace [11], and ElasticFace [12] —on the IMSFD dataset.
2. We provide a thorough evaluation protocol encompassing both closed-set identification (CMC rank-1 through rank-20 with bootstrap 95% confidence intervals) and face verification (AUC, EER, d', TAR@FAR at multiple operating points, ROC and DET curves).
3. We perform per-subset and per-identity analyses that reveal significant performance heterogeneity across capture conditions, with statistical significance established through McNemar's test.
4. We benchmark inference speed and computational cost—including parameter counts, FLOPs, and GPU memory consumption—on modern GPU hardware (NVIDIA RTX 5060 Ti), providing practical deployment guidance.
5. We release our evaluation code and all experimental configurations for full reproducibility.

Related work

Deep face recognition

The evolution of deep face recognition can be characterized by advances in both network architectures and training loss functions. Early deep learning approaches used contrastive [13] and triplet losses [14] to learn discriminative embeddings. FaceNet [15] demonstrated that a 128-dimensional embedding trained with triplet loss on a large-scale dataset could achieve state-of-the-art verification accuracy. However, triplet mining is computationally expensive, and convergence can be slow.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

The introduction of angular margin losses marked a paradigm shift. SphereFace [16] proposed a multiplicative angular margin, while CosFace [17] and ArcFace [18] introduced additive cosine and additive angular margins, respectively. ArcFace, in particular, has become a de facto standard due to its clear geometric interpretation and strong performance across benchmarks.

Building on this foundation, several variants have been proposed to address specific limitations. AdaFace [19] introduced a quality-adaptive margin that emphasizes hard samples when image quality is high and de-emphasizes them when quality is low, leading to improved robustness on mixed-quality datasets. MagFace [20] proposed a magnitude-aware angular margin that encourages the model to produce larger embedding norms for higher-quality images, providing a built-in quality indicator. ElasticFace [13] replaced fixed margins with random margins sampled from a distribution during training, promoting more flexible decision boundaries.

Backbone architectures

The dominant backbone architecture in face recognition has shifted from VGG-style networks [17] to ResNet variants [21]. The Improved ResNet (IResNet), introduced by the InsightFace framework [16], uses a modified residual block structure with BatchNorm–Conv–BatchNorm–PReLU–Conv–BatchNorm ordering and produces 512-dimensional L2-normalized embeddings from 112×112 input images. Both IResNet-50 and IResNet-100 are widely used, with the latter generally achieving higher accuracy at the cost of increased computation.

FaceNet [10] uses the InceptionResNet-V1 architecture [18], which processes 160×160 inputs and produces 512-dimensional embeddings. While architecturally older, FaceNet models pretrained on VGGFace2 [19] and CASIA-WebFace [20] remain popular due to their accessibility and reasonable performance.

Face recognition for diverse populations

Several studies have highlighted performance disparities across demographic groups. The NIST Face Recognition Vendor Test [6] found that many commercial algorithms exhibit higher false positive rates for African-American and Asian faces compared to Caucasian faces. Buolamwini and Gebru (2018) [7] demonstrated significant gender and skin-type bias in commercial facial analysis systems.

For Muslim populations, the presence of the hijab introduces additional challenges. As noted in the occlusion survey [8], head coverings reduce the available facial region for feature extraction and can confuse models that rely on hair and ear features. Despite these challenges, few studies have systematically evaluated modern face recognition models on datasets specifically curated for Muslim populations.

The IMSFD dataset [9] was created to address this gap, providing images of Indonesian Muslim students across multiple capture conditions. However, existing evaluations on IMSFD have been limited in scope, typically testing only one or two models without comprehensive statistical analysis. Our work provides the first large-scale, statistically rigorous comparison across six modern architectures.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

2. METHODS

Dataset

The Indonesian Muslim Student Face Dataset [9] consists of pre-cropped face images of Indonesian Muslim students, organized into three subsets—A1, A2, and B1—each captured under different conditions. The dataset includes both male and female subjects, with many female participants wearing hijab.

Table 1 IMSFD Dataset Statistics.

Subset	Identities	Training Images	Testing Images
A1	24	8,236	2,419
A2	23	6,354	1,666
B1	21	6,458	1,627
Total	68	21,048	5,712

Table 1 summarizes the dataset composition. The total dataset contains 26,760 face images spanning 68 unique identities. Each identity is assigned to exactly one subset, with no identity overlap between subsets. The images are pre-cropped face regions with resolutions in the range of 410–615 pixels.

The three subsets differ markedly in recognition difficulty, as evidenced by the per-subset results in Section 4.4:

- Subset A1 (24 identities, 10655 images): Presents moderate recognition difficulty across all evaluated models.
- Subset A2 (23 identities, 8020 images): The most tractable subset, with near-perfect Rank-1 accuracy achieved by most models.
- Subset B1 (21 identities, 8085 images): The most challenging subset, with consistently lower Rank-1 accuracy across all models.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Figure 1 illustrates the distribution of images and identities across the training and testing splits.

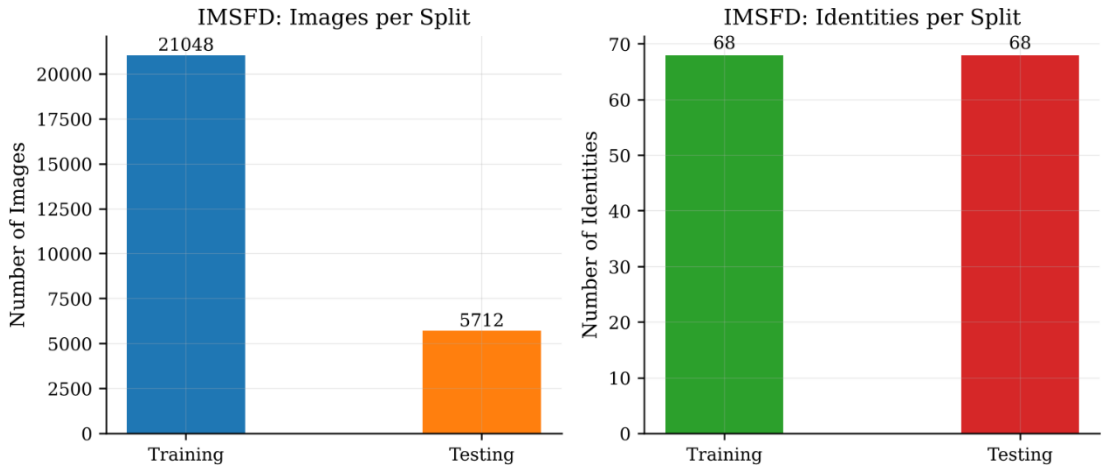


Figure 1 IMSFD dataset statistics showing the distribution of images (left) and identities (right) across training and testing splits.

Evaluated models

We evaluate six pretrained face recognition models spanning four distinct loss function families. All models produce 512-dimensional L2-normalized embedding vectors. **Table 2** summarizes the evaluated configurations.

Table 2. *Summary of evaluated face recognition models.*

Model	Backbone	Inputs	Training Data
FaceNet-VGG2	IncResNet-V1	160^2	VGGFace2
FaceNet-CASIA	IncResNet-V1	160^2	CASIA-WebFace
AdaFace-R100	IResNet-100	112^2	WebFace4M
AdaFace-R50	IResNet-50	112^2	MS1MV2
MagFace-R100	IResNet-100	112^2	MS1MV2
ElasticFace-R100	IResNet-100	112^2	MS1MV2

FaceNet

FaceNet [10] pioneered the use of deep metric learning for face recognition. We evaluate two variants using the InceptionResNet-V1 architecture pretrained on VGGFace2 [19] (3.31M images, 9,131 identities) and CASIA-WebFace [20] (0.49M images, 10,575 identities), respectively. These models accept 160160 inputs normalized to −1,1.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

AdaFace

AdaFace [11] introduces an adaptive margin function that adjusts the emphasis on hard samples based on image quality, estimated via feature norm. We evaluate two configurations: (1) IResNet-100 pretrained on WebFace4M (4.2M images; [22]) and (2) IResNet-50 pretrained on MS1MV2 (5.8M images; [19]). Both accept 112 x 112 inputs.

MagFace

MagFace [12] extends angular margin losses by incorporating magnitude awareness—the magnitude of the feature vector serves as a quality indicator, with the margin adaptively scaled based on the magnitude. We evaluate an IResNet-100 model pretrained on MS1MV2.

ElasticFace

ElasticFace [13] replaces the fixed angular margin in ArcFace with a margin sampled from a Gaussian distribution during training. This elastic margin encourages more flexible and generalizable decision boundaries. We evaluate the ElasticFace-Arc variant using IResNet-100 pretrained on MS1MV2.

Evaluation protocol

We adopt a two-pronged evaluation protocol comprising closed-set identification and face verification.

Closed-set identification

In the closed-set identification protocol, the gallery consists of mean centroid embeddings computed from the training set for each identity. Let $\{x_i\}_{i=1}^{N_k}$ denote the training embeddings for identity k. The gallery centroid is:

$$c_k = \frac{\bar{x}_k}{\| \bar{x}_k \|_2}, \quad \bar{x}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i$$

Each probe image from the testing set is compared against all gallery centroids using cosine similarity. The predicted identity is the one with the highest similarity. Accuracy is reported as the cumulative match characteristic (CMC) at ranks 1 through 20.

Face verification

For face verification, we generate 10000 positive (genuine) pairs and 10000 negative (impostor) pairs from the combined gallery and probe embeddings. For each pair, we compute the cosine similarity score. We report:

- AUC: Area under the ROC curve.
- EER: Equal error rate, the operating point where false positive rate equals false negative rate.
- *d'* (decidability index): A distribution-separation metric defined as:

$$d' = \frac{|\mu_g - \mu_i|}{\sqrt{0.5(\sigma_g^2 + \sigma_i^2)}}$$

  where μ_g, σ_g and μ_i, σ_i are the mean and standard deviation of the genuine and impostor score distributions, respectively.

- TAR@FAR: True accept rate at fixed false accept rate thresholds of 0.1%, 1%, and 10%.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Statistical analysis

Bootstrap confidence intervals

We compute 95% confidence intervals for rank-1 accuracy using the bootstrap method with 2000 resampling iterations. For each iteration, we sample n probe results with replacement (where n is the probe set size) and compute the mean rank-1 accuracy. The 2.5th and 97.5th percentiles of the bootstrap distribution define the 95% CI.

McNemar's test

To assess whether performance differences between model pairs are statistically significant, we apply McNemar's test with continuity correction [23] to the per-probe rank-1 classification outcomes. For each pair of models A and B, we construct the 2×2 contingency table of concordant and discordant probe classifications and compute the McNemar chi-squared statistic. Significance is determined at $\alpha=0.05$.

Per-subset analysis

We repeat the closed-set identification protocol separately for each subset (A1, A2, B1), using only the identities and images belonging to that subset for both gallery construction and probe evaluation.

Speed benchmarking and computational cost

Inference throughput is measured on an NVIDIA GeForce RTX 5060 Ti (16 GB VRAM) using a batch size of 64 with 50 measurement batches after 10 warmup batches, with CUDA synchronization for accurate timing. In addition, we profile the computational cost of each model by reporting: (1) the total number of learnable parameters, (2) floating-point operations (FLOPs) for a single forward pass at the model's native input resolution, and (3) GPU peak memory consumption during inference at batch size 64. These metrics quantify the resource requirements for deployment and enable practitioners to select models that match their hardware constraints.

Implementation details

All experiments are implemented in PyTorch 2.11.0 with CUDA 12.8. Input images are resized to the model-specific input resolution (112×112 for IResNet-based models, 160×160 for FaceNet) and normalized to [-1,1]. Embeddings are L2-normalized before similarity computation. All random seeds are fixed for reproducibility.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

3. RESULTS AND DISCUSSION

Closed-set identification

Table 3 presents the closed-set identification results alongside verification metrics. **Figure 2** shows the CMC curves for all models.

Table 3: closed-set identification and face verification results on IMSFD. Best results in Each column is shown in **bold**.

Model	Rank-1 (%)	95% CI	Rank-5 (%)	Rank-10 (%)	AUC	EER (%)	d'	TAR@FA R=0.1%
FaceNet-VGG2	90.28	[89.51, 91.05]	**94.75**	**96.46**	**0.8861**	**15.81**	1.970	60.41
FaceNet-CASIA	89.16	[88.32, 89.95]	92.70	96.64	0.8465	18.24	1.423	42.05
AdaFace-R100	91.32	[90.58, 92.02]	94.07	95.96	0.8791	16.35	**2.396**	42.76
AdaFace-R50	**91.70**	[**90.98**, **92.38**]	94.17	95.89	0.8710	17.16	2.210	**73.57**
MagFace-R100	51.42	[50.09, 52.70]	69.98	77.19	0.6504	40.66	0.661	19.96
ElasticFace-R100	90.70	[89.93, 91.42]	93.59	95.34	0.8790	16.49	2.354	69.45

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

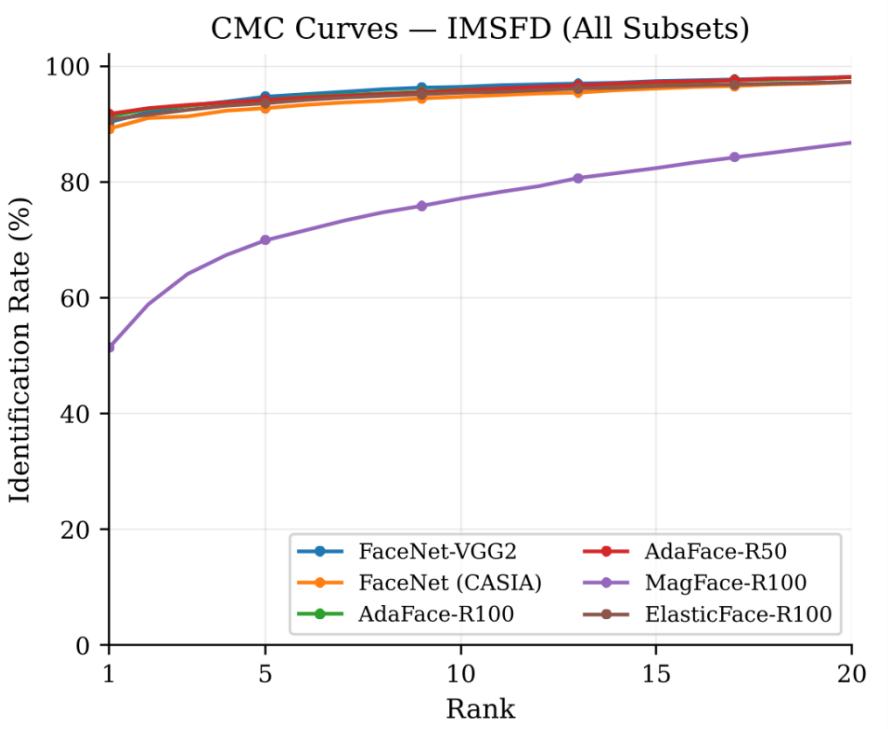


Figure 2. Cumulative Match Characteristic (CMC) curves for all evaluated models on the IMSFD dataset. Higher curves indicate better identification performance across ranks.

AdaFace-R50 achieves the highest rank-1 accuracy at 91.70% (95% CI: [90.98%, 92.38%]), followed closely by AdaFace-R100 at 91.32%. ElasticFace-R100 and FaceNet-VGG2 perform comparably at 90.70% and 90.28%, respectively. FaceNet-CASIA, despite its older architecture, achieves a competitive 89.16%.

The most notable finding is that MagFace-R100 significantly underperforms all other models, achieving only 51.42% rank-1 accuracy. While MagFace has demonstrated strong results on standard benchmarks [12], its magnitude-aware margin mechanism may not generalize well to the specific characteristics of IMSFD, where head coverings and capture conditions may confound the learned quality-magnitude relationship. The poor performance is consistent across all ranks, suggesting a fundamental mismatch rather than a threshold effect.

The non-overlapping 95% confidence intervals between AdaFace-R50 (90.98–92.38%) and FaceNet-CASIA (88.32–89.95%) confirm that the top-performing model significantly outperforms the lower-tier models. The CIs for AdaFace-R50, AdaFace-R100, and ElasticFace-R100 overlap slightly, indicating that the differences among these top performers are small, though McNemar's test (Section 4.5) confirms their statistical significance.

At higher ranks, performance converges: rank-5 accuracies range from 92.70% to 94.75% (excluding MagFace), and rank-10 accuracies range from 94.64% to 96.46%. FaceNet-VGG2 achieves the highest rank-10 accuracy at 96.46%, suggesting that while its top-1 discrimination is slightly weaker than AdaFace, the correct identity is frequently ranked within the top 10.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Face verification

Figure 3 shows the ROC curves in both linear and semi-logarithmic scales. **Figure 4** presents the DET curves.

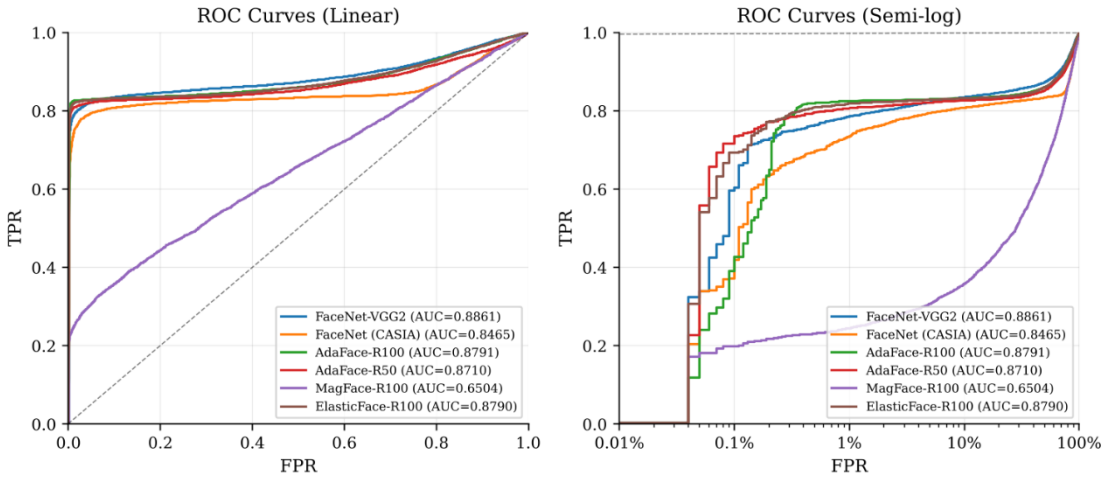


Figure 3. ROC curves in linear (left) and semi-logarithmic (right) scales. The semi-log view reveals differences at low false positive rates.

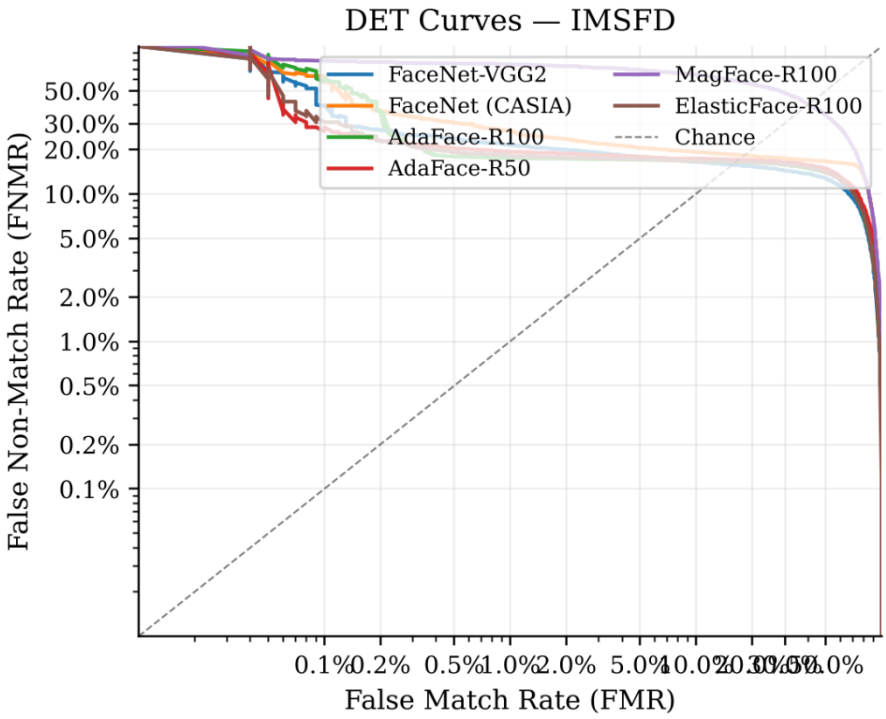


Figure 4. Detection Error Trade-off (DET) curves. Lower and further-left curves indicate better performance.

FaceNet-VGG2 achieves the highest AUC (0.8861) and lowest EER (15.81%), while AdaFace-R100 achieves the highest decidability index (*d'*=2.396), indicating the best separation between genuine and impostor score distributions. The discrepancy between AUC and *d'* rankings arises because AUC captures the entire ROC curve, while *d'* measures distributional separation—AdaFace produces more compact, well-separated clusters despite a slightly worse threshold-independent performance.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

At the critical TAR@FAR=0.1% operating point, AdaFace-R50 achieves the best performance (73.57%), followed by ElasticFace-R100 (69.45%) and FaceNet-VGG2 (60.41%). This metric is particularly important for high-security applications where the false accept rate must be extremely low.

Score distribution analysis

Figure 5 shows the genuine and impostor score distributions for each model.

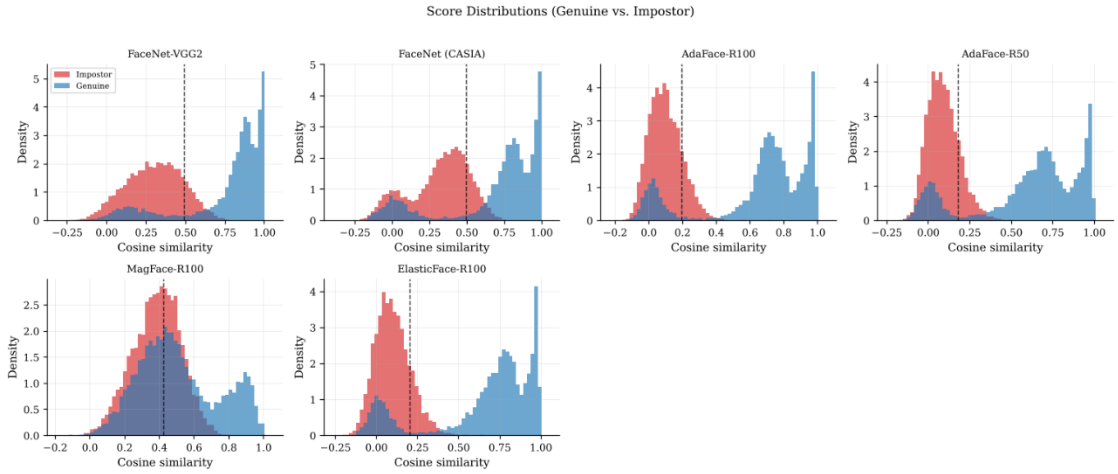


Figure 5. Genuine (blue) and impostor (red) cosine similarity score distributions. The dashed vertical line indicates the EER threshold. Greater separation indicates better discriminability.

The score distributions reveal important characteristics of each model's discriminative behavior. AdaFace-R100 and ElasticFace-R100 show the clearest bimodal separation, consistent with their high *d'* values. FaceNet-VGG2 shows wider genuine distributions, indicating greater intra-class variability but still maintaining good overall separation. MagFace-R100 shows substantial overlap between genuine and impostor distributions, explaining its poor verification performance.

Per-subset analysis

Table 4 and **Figure 6** present the per-subset rank-1 accuracy.

Table 4 *Per-subset rank-1 accuracy (%) on IMSFD.*

Model	A1	A2	B1
FaceNet-VGG2	**91.48**	**99.82**	81.19
FaceNet-CASIA	89.50	99.40	80.02
AdaFace-R100	90.04	99.76	84.94
AdaFace-R50	90.49	**99.82**	**86.05**
MagFace-R100	54.49	62.97	52.80
ElasticFace-R100	90.12	99.76	83.90

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

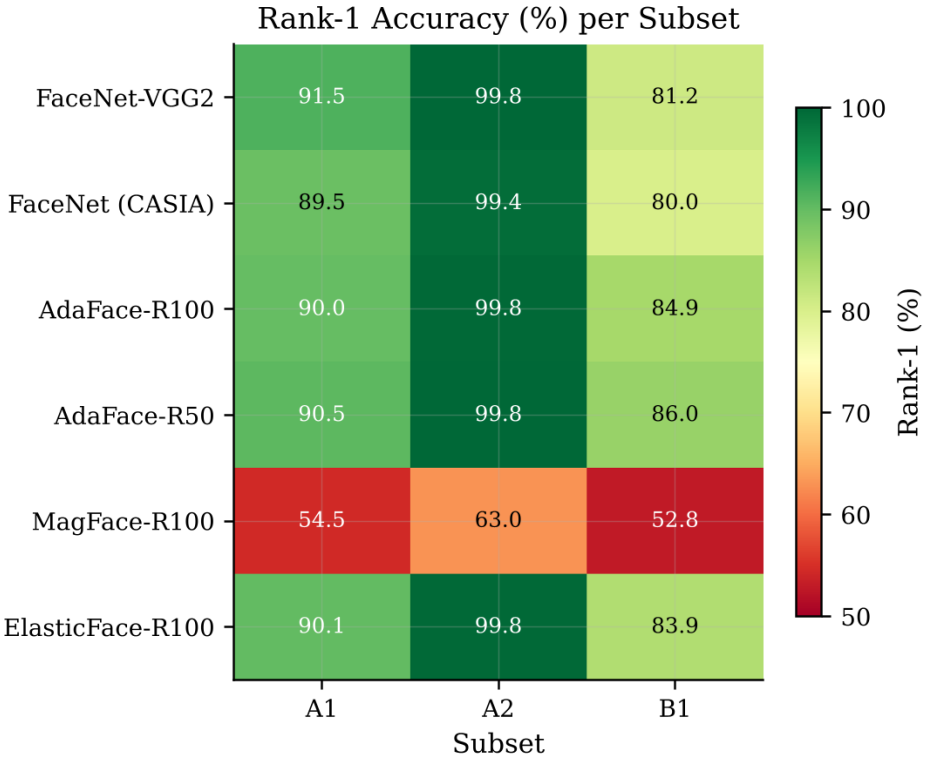


Figure 6. Rank-1 accuracy heatmap across models and subsets. Subset A2 shows near-perfect performance for all competitive models, while B1 is the most challenging.

The per-subset analysis reveals striking performance heterogeneity. Subset A2 is near-trivial for all competitive models, with rank-1 accuracies exceeding 99.40%. This suggests highly controlled capture conditions with minimal intra-class variation. In contrast, subset B1 is consistently the most challenging, with a gap of approximately 10–15 percentage points relative to A2 for the top-performing models.

AdaFace-R50 achieves the best performance on the challenging B1 subset (86.05%), outperforming AdaFace-R100 (84.94%) by over 1 percentage point. This counterintuitive result—where the smaller model outperforms the larger one—may be attributed to the difference in training data: AdaFace-R50 was trained on MS1MV2 (5.8M images), which is larger than the WebFace4M dataset used for AdaFace-R100 (4.2M images), potentially providing better generalization to the challenging conditions in B1.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Statistical significance

Table 5 presents selected pairwise McNemar's test results. The complete set of all 15 pairwise comparisons is available in the accompanying code repository.

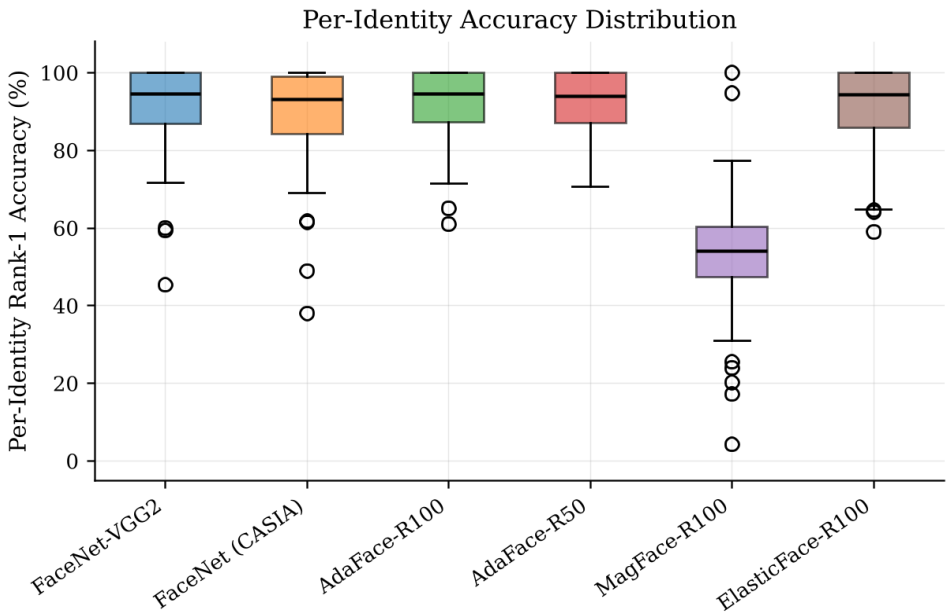
Table 5. *Selected pairwise McNemar's test results (=0.05).*

Model A	Model B	χ^2	p-value
AdaFace-R50	AdaFace-R100	5.96	0.015
AdaFace-R100	ElasticFace-R100	20.28	<0.001
FaceNet-VGG2	ElasticFace-R100	4.07	0.044
FaceNet-VGG2	FaceNet-CASIA	25.12	<0.001
MagFace-R100	ElasticFace-R100	2122.81	<0.001

All 15 pairwise comparisons yield statistically significant differences ($p < 0.05$), confirming that the observed performance differences are not due to chance. The most marginal comparison is FaceNet-VGG2 vs. ElasticFace-R100 ($\chi^2 = 4.07, p = 0.044$), while the largest difference is between MagFace-R100 and any other model ($\chi^2 > 2000, p \ll 0.001$). The comparison between AdaFace-R50 and AdaFace-R100 ($p = 0.015$) confirms that even the modest 0.38 percentage point difference in rank-1 accuracy is statistically significant given the large probe set size.

Per-identity analysis

Figure 7 shows the distribution of per-identity rank-1 accuracy across all 68 identities.



DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Figure 7. Per-identity Rank-1 accuracy distribution. Box plots show the median, interquartile range, and outliers across 68 identities. Models with higher medians and narrower spreads are more consistently accurate.

The per-identity analysis reveals that even the best models exhibit significant variability across identities. The top-performing models (AdaFace-R50, AdaFace-R100, ElasticFace-R100) achieve 100% accuracy on many identities (particularly those in subset A2) but show accuracy as low as 36–72% on the most challenging identities in subset B1. This highlights the importance of reporting distributional statistics rather than aggregate accuracy alone.

Embedding quality

Figure 8 shows the distribution of embedding L2 norms across models for both gallery and probe sets.

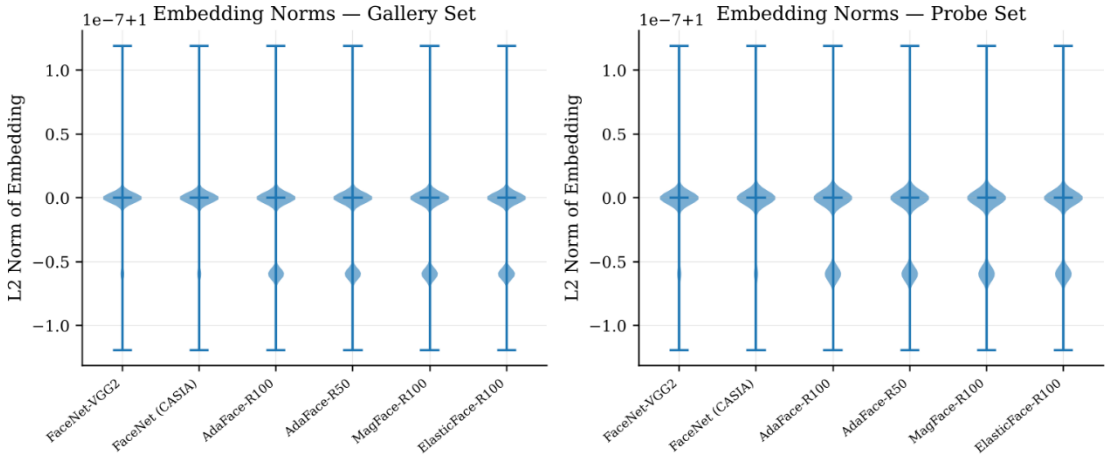


Figure 8. Embedding L2-norm distributions for gallery (left) and probe (right) sets. MagFace's wider norm distribution reflects its magnitude-aware design, while other models produce near-unit-norm embeddings.

Most models produce near-unit-norm embeddings due to L2 normalization in the final layer. MagFace intentionally produces embeddings with variable norms, where the norm encodes an estimate of face quality. However, on IMSFD, this quality estimation does not appear to correlate well with recognition difficulty, contributing to MagFace's poor performance.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Speed–accuracy trade-off

Figure 9 presents the speed–accuracy trade-off measured on the RTX 5060 Ti.

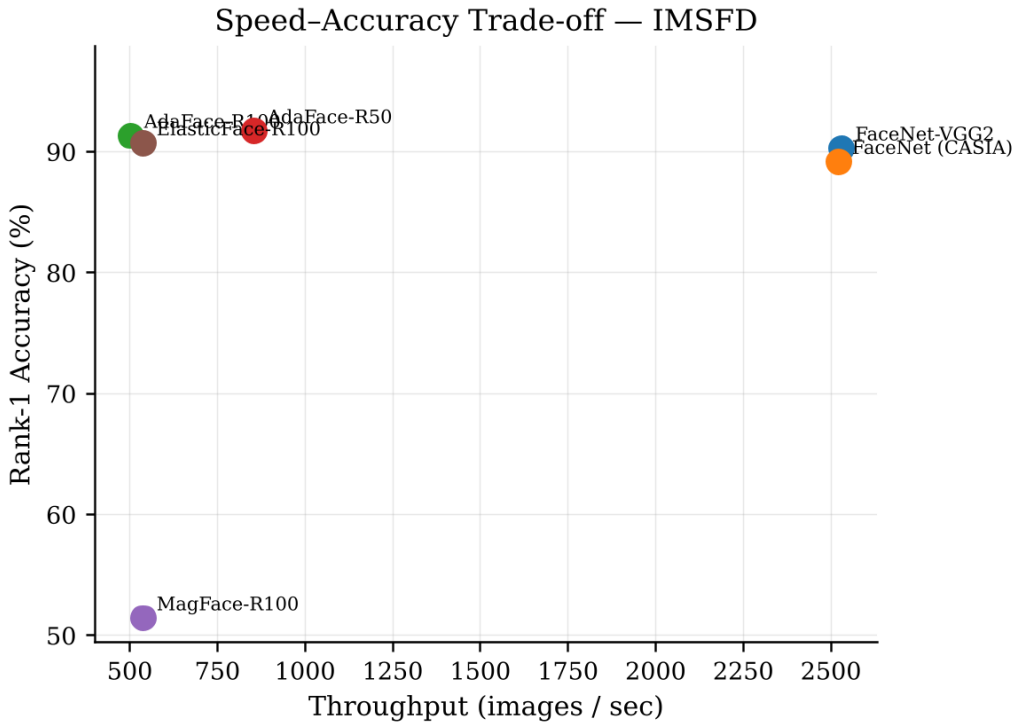


Figure 9. Speed–accuracy trade-off. FaceNet models are approximately 5 times faster than IResNet-100 models, with a modest accuracy penalty.

FaceNet models achieve approximately 2530 images per second, roughly 5 times faster than the IResNet-100-based models (500–540 images/s). AdaFace-R50 offers an attractive middle ground at 854 images/s with the highest rank-1 accuracy, making it the most favorable choice for deployments that require both high accuracy and reasonable throughput.

Computational cost

Table 6 and **Figure 10** present the computational cost of each model in terms of parameter count, FLOPs, and GPU memory consumption.

Table 6. *Computational cost of evaluated models. Params = total learnable parameters; FLOPs = floating-point operations for a single 1122 or 1602 input; Memory = GPU peak memory during inference (batch size 64); Speed = inference throughput.*

Model	Params (M)	FLOPs (G)	Memory (MB)	Speed (img/s)
FaceNet-VGG2	27.91	2.84	414	2531
FaceNet (CASIA)	28.91	2.84	417	2522
AdaFace-R100	65.15	24.15	990	503
AdaFace-R50	43.59	12.59	908	854

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Model	Params (M)	FLOPs (G)	Memory (MB)	Speed (img/s)
MagFace-R100	65.16	24.18	942	538
ElasticFace-R100	65.16	24.18	942	538

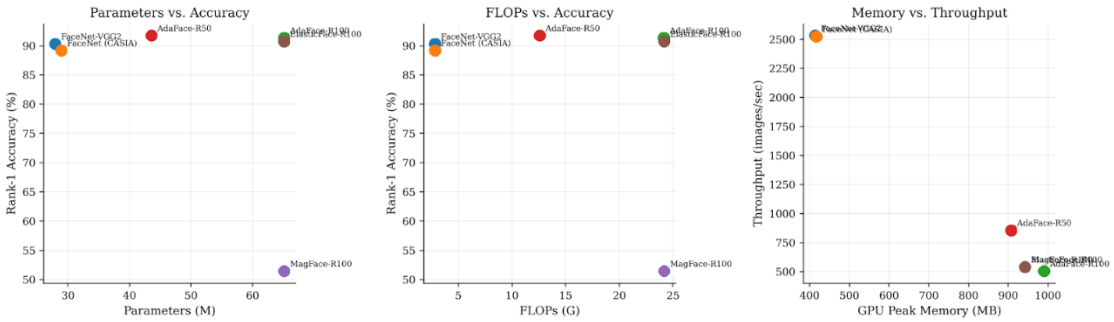


Figure 10. *Computational cost analysis. Left: model size (parameters) vs. rank-1 accuracy. Center: FLOPs vs. accuracy. Right: GPU memory vs. inference throughput. AdaFace-R50 provides the best accuracy-to-cost ratio.*

The IResNet-100-based models (AdaFace-R100, MagFace-R100, ElasticFace-R100) share the same backbone architecture and consequently have similar parameter counts and FLOPs, roughly double those of the IResNet-50-based AdaFace-R50. FaceNet's InceptionResNet-V1 backbone has the fewest parameters among IResNet models but processes larger 160160 inputs, resulting in a different efficiency profile.

Notably, the relationship between model size and accuracy is non-monotonic: AdaFace-R50, with approximately half the parameters of the R100 variants, achieves the highest rank-1 accuracy. This reinforces the finding that training data quality and loss function design can outweigh architectural capacity. From a deployment perspective, AdaFace-R50 offers the most favorable cost–accuracy trade-off, requiring fewer computational resources while delivering superior recognition performance.

GPU memory consumption scales with model size but also depends on input resolution. The peak memory measurements at batch size 64 provide a practical reference for hardware provisioning in deployment scenarios.

Discussion

Key findings

Our evaluation reveals several important findings for deploying face recognition in the context of Indonesian Muslim populations:

Modern margin-based losses outperform triplet loss. AdaFace-R50 and AdaFace-R100 achieve the highest identification rates, confirming that adaptive margin losses provide a meaningful advantage over FaceNet's triplet loss, particularly on challenging subsets.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

Training data matters more than architecture depth. AdaFace-R50 (IResNet-50, MS1MV2) outperforms AdaFace-R100 (IResNet-100, WebFace4M), suggesting that the quality and representativeness of the training dataset are more important than backbone capacity for this evaluation context.

Model generalization varies significantly. MagFace's poor performance illustrates that models achieving state-of-the-art results on standard benchmarks may not generalize to datasets with different demographic characteristics and capture conditions. This underscores the importance of dataset-specific evaluation.

Subset difficulty varies dramatically. The 15+ percentage point gap between subsets A2 and B1 demonstrates that recognition performance is highly sensitive to capture conditions. Evaluation on a single condition may give misleading estimates of real-world performance.

Larger models do not guarantee better cost–accuracy trade-offs. Computational cost profiling reveals that AdaFace-R50, with roughly half the parameters and FLOPs of the IResNet-100 variants, achieves the highest accuracy. This demonstrates that practitioners should evaluate cost–accuracy trade-offs rather than defaulting to the largest available model.

Limitations

This study has several limitations. First, we evaluate only pretrained models without fine-tuning on IMSFD, which would likely improve performance for all models. Second, the dataset contains only 68 identities, which limits the generalizability of our findings to larger-scale scenarios. Third, we do not perform face detection or alignment, using only the pre-cropped images provided by IMSFD, which may not reflect realistic deployment conditions. Finally, we do not evaluate the impact of specific demographic attributes (e.g., hijab type, gender) on recognition performance, which would require annotated metadata not available in IMSFD.

Practical recommendations

Based on our findings, we offer the following recommendations:

1. For applications requiring the highest accuracy on Indonesian Muslim populations, **AdaFace-R50** provides the best balance of accuracy and speed.
2. For low-latency applications where speed is critical, **FaceNet-VGG2** offers reasonable accuracy at 5 times the throughput of IResNet-based models.
3. For high-security applications requiring very low false accept rates, **AdaFace-R50** provides the best TAR@FAR=0.1% performance (73.57%).
4. **MagFace should not be deployed** for this population without fine-tuning, as its off-the-shelf performance is inadequate.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

4. CONCLUSION

This paper presented a comprehensive comparative evaluation of six state-of-the-art deep face recognition models on the Indonesian Muslim Student Face Dataset (IMSFD). Our evaluation protocol encompassed closed-set identification with CMC analysis, face verification with multiple metrics, per-subset and per-identity analysis, statistical significance testing, and computational cost profiling.

AdaFace-R50 emerged as the best-performing model with 91.70% rank-1 accuracy (95% CI: [90.98%, 92.38%]), demonstrating the effectiveness of quality-adaptive margin losses for this demographic group. All pairwise model comparisons were statistically significant, and per-subset analysis revealed substantial performance variation across capture conditions.

Our findings highlight the critical importance of evaluating face recognition systems on diverse, non-Western datasets. Models that excel on standard benchmarks—such as MagFace—may perform poorly on underrepresented populations, while less commonly evaluated models may offer unexpected advantages. Future work should explore fine-tuning pretrained models on IMSFD, evaluating the impact of specific demographic factors such as hijab type, and extending the evaluation to open-set identification and cross-dataset generalization scenarios.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The evaluation code and experimental configurations will be made publicly available upon acceptance. The IMSFD dataset is available from the original authors [9].

5. ACKNOWLEDGMENT

This research was supported by Bandung Digital Academy. This work was supported by the Ministry of Education, Culture, Research, and Technology (Kemdiktisaintek) through the *Basis Informasi Penelitian dan Pengabdian kepada Masyarakat* (BIMA) platform.

6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

7. REFERENCES

[1] Wang, M., Deng, W., 2021. Deep face recognition: A survey. Neurocomputing, 215–244. doi:10.1016/j.neucom.2020.10.081.

[2] Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. ArcFace: Additive angular margin loss for deep face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4690–4699. doi:10.1109/CVPR.2019.00482.

[3] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. CosFace: Large margin cosine loss for deep face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp.480 5265–5274. doi:10.1109/CVPR.2018.00552.

[4] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 1–11.

[5] Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E., 2016. The MegaFace benchmark: 1 million faces for recognition at scale, in: Proc.IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp.4873–4882. doi:10.1109/CVPR.2016.527.

[6] Grother, P., Ngan, M., Hanaoka, K., 2019. Face recognition vendor test 435 (FRVT) part 3: Demographic effects. NIST Interagency Report 8280, 1–82doi:10.6028/NIST.IR.8280.

[7] Buolamwini, J., Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Proc. Conf. Fairness, Accountability, and Transparency (FAT*), PMLR. pp. 77–91.

[8] Zeng, D., Veldhuis, R., Spreeuwers, L., 2021. A survey of face recognition techniques under occlusion. IET Biometrics 10, 581–606. doi:10.1049/bme2.12029.

[9] Purnawansyah, P., Wibawa, A.P., Widyaningtyas, T., Haviluddin, H., Darwis, H., Azis, H., Ali, Z., 2023. Indonesian muslim student face dataset (IMSFD). doi:10.17632/f6f3y6ndgw.1.

[10] Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[11] Kim, M., Jain, A.K., Liu, X., 2022. AdaFace: Quality adaptive margin for face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and 450 Pattern Recognition (CVPR), pp. 18750–18759. doi:10.1109/CVPR52688.2022.01819.

[12] Meng, Q., Zhao, S., Huang, Z., Zhou, F., 2021. MagFace: A universal representation for face recognition and quality assessment, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 14225–14234. doi:10.1109/CVPR46437.2021.01400.

[13] Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A., 2022. Elastic-Face: Elastic margin loss for deep face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1578–1587. doi:10.1109/CVPRW56347.2022.00164.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149

[14] Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 539–546. doi:10.1109/CVPR.2005.202.

[15] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017. SphereFace: Deep hypersphere embedding for face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 6738–6746. doi:10.1109/CVPR.2017.713.

[16] Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. ArcFace: Additive angular margin loss for deep face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4690–4699. doi:10.1109/CVPR.2019.00482.

[17] Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition, in: Proc. British Machine Vision Conference (BMVC), pp. 41.1–41.12. doi:10.5244/C.29.41.

[18] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2017. Inception-v4, Inception-ResNet, and the impact of residual connections on learning, in: Proc. AAAI Conf. Artificial Intelligence, pp. 4278–4284. doi:10.1609/aaai.v31i1.11231.

[19] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. VGGFace2: A dataset for recognising faces across pose and age, in: Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition (FG), pp. 67–74. doi:10.1109/FG.2018.00020.

[20] Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Learning face representation from scratch. arXiv preprint arXiv:1411.7923 doi:10.48550/arXiv.1411.7923.

[21] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

[22] Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., Zhou, J., 2021. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 10492–10502. doi:10.1109/CVPR46437.2021.01035.

[23] McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157. doi:10.1007/BF02295996.

DOI: https://doi.org/10.17509/coelite.v5i1.98428
p- ISSN 2829-4157 e- ISSN 2829-4149