IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151159

Training Data Diversity and Distribution Shift in AI-Generated Image Detection: An Empirical Study

E Niranjan[1], Arushi[2], Deesha S Chajer[3], Niveditha S[4]

Associate Professor, ECE, Bangalore Institute of Technology, Bengaluru, Karnataka, India[1]

Student, 4th Yr ECE, Bangalore Institute of Technology, Bengaluru, Karnataka, India[2]

Student, 4th Yr ECE, Bangalore Institute of Technology, Bengaluru, Karnataka, India[3]

Student, 4th Yr ECE, Bangalore Institute of Technology, Bengaluru, Karnataka, India[4]

Abstract: The proliferation of high-fidelity generative models such as Stable Diffusion, DALL-E, MidJourney, and Gemini has made automated detection of AI-generated images an increasingly important research problem. While many published detectors report high in-domain accuracy, their performance under distribution shift is often overlooked. This paper presents an empirical study of binary real-versus-AI image classifiers, examining how training data composition and architecture choice affect accuracy on images from sources not seen during training. Models trained only on public data (ResNet18 and EfficientNet-B3) reached 53.33% and 61.67% on a 60-image held-out set of images from web-based generators and personal photographs, whereas a ResNet18 trained with an additional 360 self-collected images reached 75.00%, rising to 81.67% after threshold calibration. On unseen Tiny-GenImage samples (n=800) the three models scored between 72.12% and 74.75%, and the data-augmented model scored lower than the public-only ResNet18 on a 2,000-image slice (68.85% vs. 72.60%), indicating a trade-off between the added sources and the original public distribution. Because the two architectures have near-identical parameter counts (11.18M and 10.70M) and the runs differ in training volume and schedule, we draw no conclusion about model capacity. The results suggest that adding even a small amount of data from the target sources can matter more than the choice between these two architectures.

Keywords: AI-generated image detection, deep learning, convolutional neural networks, dataset diversity, transfer learning.

I. INTRODUCTION

Recent advances in diffusion-based and adversarial generative models have made synthetic images increasingly difficult to distinguish from authentic photographs. This capability raises concerns for misinformation, digital forensics, journalism, and public trust in visual media [1]. Consequently, automated detection of AI-generated images has emerged as an active research area.

A recurring limitation in this literature is that detectors are frequently evaluated only on held-out data drawn from the same distribution as their training set — that is, the same generators and similar real-image sources. Such evaluation can substantially overstate real-world performance, since detectors may learn generator-specific artifacts (e.g., frequency-domain signatures or compression patterns) rather than generalizable indicators of synthetic content [2], [3].

This paper investigates that gap directly. We train and evaluate several model configurations for binary real-versus-AI image classification, with particular emphasis on performance under distribution shift: test images drawn from sources that were not part of the public training datasets, namely images generated with Google Gemini, ChatGPT, and other web-based tools, and personal photographs. Our research question is: how does adding a small amount of source-diverse data change detection accuracy under distribution shift, compared with changing the architecture?

Our key contributions are:

- A comparison of ResNet18 and EfficientNet-B3, which have near-identical parameter counts with a binary head (11.18M and 10.70M), showing mixed results: EfficientNet-B3 scored higher on the 60-image shifted set (61.67% vs. 53.33%) but lower on unseen Tiny-GenImage samples (72.12% vs. 74.75%, n=800), with the runs differing in training volume and schedule.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151159

- An empirical observation that adding a modest number (360) of self-collected images from additional sources accompanied an increase from 53.33% to 75.00% on the shifted set, together with a small decrease on Tiny-GenImage, indicating a trade-off between the added sources and the original public distribution.
- Evaluation of the resulting model on larger unseen Tiny-GenImage slices (n=800 and n=2000) and an independent 30-image set, to examine how accuracy and the calibrated threshold transfer beyond the 60-image set.
- A decision-threshold calibration procedure that further corrects a directional prediction bias observed in the base model.

II. RELATED WORK

Several large-scale benchmarks have been proposed for AI-generated image detection. GenImage [1] provides over one million real and synthetic image pairs spanning multiple diffusion models and GANs, using shared ImageNet [4] classes across generators. The Artifact dataset [2] extends coverage to 25 distinct generative models, including both older GAN-based architectures (StyleGAN, ProGAN) and modern diffusion models, explicitly targeting generalizable detection. SynthBuster [3] provides 1,000 images from each of nine diffusion models, including commercial, closed-source generators (DALL-E 2/3, MidJourney, Adobe Firefly) that were underrepresented in earlier open datasets, alongside several Stable Diffusion versions and GLIDE. More recently, the Defactify/MS-COCOAI dataset [5] paired MS COCO [6] real images with captions used as generation seeds across five modern diffusion and commercial models, enabling caption-controlled comparison between real and synthetic images.

A consistent theme across this literature is the existence of a generalization gap: detectors trained on one set of generators frequently show substantially reduced accuracy when evaluated on unseen generators [1], [5]. This has been attributed to reliance on generator-specific artifacts rather than universal cues of synthetic content. Our results are consistent with this pattern: the public-only models scored markedly lower on a held-out set of images from other sources than on unseen images from their own training datasets, and adding a small amount of source-diverse data was associated with a large gain on that held-out set.

III. DATASETS

Three data sources were combined for this study:

- Defactify (MS-COCOAI) [5]: MS COCO-derived real images paired with AI images from Stable Diffusion 2.1, Stable Diffusion 3, SDXL, DALL-E 3, and MidJourney v6.
- Tiny-GenImage: A reduced-scale derivative of GenImage [1], providing ImageNet-derived real images and AI images from Stable Diffusion 1.4/1.5, GLIDE, ADM, VQDM, Wukong, and BigGAN.
- Self-collected data (n=360; 180 real, 180 AI): Images generated using Google Gemini and OpenAI ChatGPT's image generation feature, plus personal photographs captured on team members' smartphone cameras — sources not included in the two public datasets above, and specifically targeting generators and image styles found to be poorly handled by the initial model.

The final training pool contained 7,860 images (7,500 public and 360 self-collected), split 85/15 into training and validation sets. A held-out evaluation set of 60 images (30 real photographs and 30 AI-generated images from Gemini, ChatGPT, and other web-based generators) was collected separately by the authors and excluded entirely from all training runs, to serve as a consistent shifted-distribution test set across experiments. Because this set was collected with tools similar to those used for the self-collected training data, it measures performance on those kinds of sources rather than on fully independent generators, and some of the web-based tools may use generator families that also appear in the public training data.

IV. METHODOLOGY

A. Model Architectures

Two convolutional architectures, both initialized with ImageNet-pretrained weights, were evaluated: ResNet18 [7], a residual network, and EfficientNet-B3 [8], a compound-scaled network. With a binary classification head replacing the original 1000-class ImageNet output, the two models have similar parameter counts (11.18M and 10.70M, respectively), so the comparison contrasts architecture families rather than model size.

B. Training Configuration

All models were fine-tuned using the AdamW optimizer [9] at an input resolution of 224×224. The initial public-only ResNet18 was trained on 3,000 public images for 4 epochs at a learning rate of 1×10^{-4}. The final ResNet18 was trained for 8 epochs on 7,860 images (7,500 public and 360 self-collected) with a learning rate of 1×10^{-4} and weight decay of

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151159

1×10^{-4}. The public-only EfficientNet-B3 was trained on 9,000 public images (3,000 Defactify real, 3,000 Defactify AI, and 3,000 Tiny-GenImage). An initial configuration (learning rate 3×10^{-4}, 12 epochs, label smoothing 0.1) overfit; a subsequently tuned configuration (learning rate 1×10^{-4}, 6 epochs, label smoothing 0.1, one warm-up epoch followed by cosine decay) was used for all reported EfficientNet-B3 results. Standard augmentation (random horizontal flip, rotation, color jitter) was applied during training, with additional random translation for EfficientNet-B3. Consequently, the runs differ in training volume, epochs, and optimization details, and the architecture comparison is not a controlled ablation.

C. Evaluation Protocol

To prevent data leakage, Tiny-GenImage test images were always drawn from a portion of the dataset excluded from the corresponding training runs (verified by index offset when streaming from Hugging Face Hub); for the comparison in Table I, the first 800 samples from stream position 3,000 onward were used, which lies beyond every sample used in training. The 60-image set was kept in a separate folder never used for training. Unless stated otherwise, models were evaluated at the default decision threshold of 0.5. We report accuracy, precision, recall, and F1 score, computed with the AI-generated class as positive.

D. Threshold Calibration

The default classification rule assigns the AI label when the model's predicted AI-probability exceeds 0.5. For the final ResNet18 we additionally performed a post-hoc threshold sweep (0.30 to 0.70 in increments of 0.02) on the 60-image set, selecting the threshold that maximized accuracy (0.68; accuracy was tied at 0.70, so the optimum lies at the edge of the searched range), to correct for an observed directional bias toward predicting the AI class.

V. EXPERIMENTS AND RESULTS

A. Architecture Comparison

Table I compares ResNet18 and EfficientNet-B3 on the 60-image shifted set and on 800 unseen Tiny-GenImage samples, all at the default threshold of 0.5. With a binary head the two architectures have similar parameter counts (11.18M vs. 10.70M). The results do not favor either architecture consistently: the public-only EfficientNet-B3 (trained on 9,000 images) scored higher than the public-only ResNet18 (3,000 images) on the 60-image set (61.67% vs. 53.33%), while ResNet18 scored 2.6 points higher on the Tiny-GenImage samples (74.75% vs. 72.12%), a difference within the sampling error expected at n=800. Both public-only models scored markedly lower on the shifted 60-image set than on Tiny-GenImage, consistent with the generalization gap described in Section II. Because the models differ in training volume, epochs, and optimization settings, Table I should be read as a descriptive comparison rather than a controlled test of model capacity.

TABLE I COMPARISON OF ARCHITECTURES AND TRAINING DATA ON THE 60-IMAGE SHIFTED SET AND UNSEEN TINY-GENIMAGE SAMPLES (THRESHOLD = 0.50)

Model	Training images	Parameters	60-image set	Tiny-GenImage (n=800)
ResNet18, public only	3,000	11.18M	53.33%	74.75%
EfficientNet-B3, public only	9,000	10.70M	61.67%	72.12%
ResNet18, public + self-collected	7,860	11.18M	75.00%	73.75%

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151159

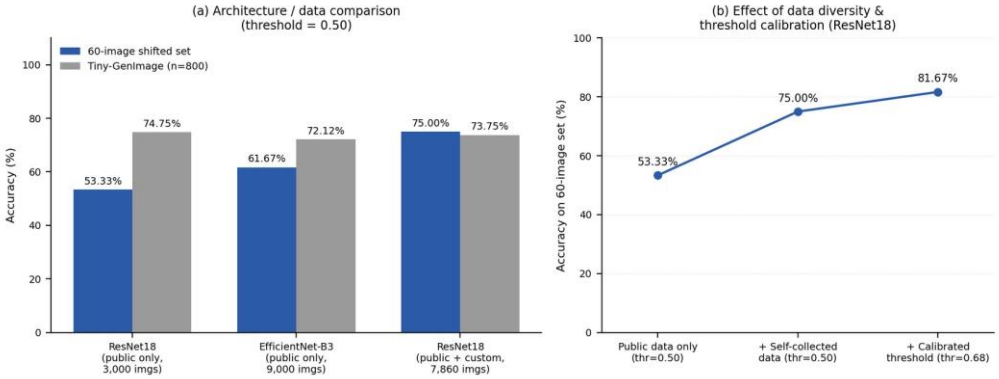


Fig. 1 (a) Accuracy of the three models on the 60-image shifted set and on 800 unseen Tiny-GenImage samples (Table I). (b) Effect of adding self-collected data and threshold calibration on the ResNet18 model (Table II).

A. Effect of Training Data Diversity

Table II summarizes the ResNet18 results as self-collected data and threshold calibration were added. The public-only ResNet18 reached 53.33% on the 60-image set; the ResNet18 trained with 360 self-collected images (4.6% of its 7,860 training images) reached 75.00%, an increase of 21.67 percentage points, and also exceeded the public-only EfficientNet-B3 (61.67%), which was trained on more images (9,000). This gain should be interpreted with caution: the public-only baseline was trained on fewer images (3,000 vs. 7,860) and for fewer epochs (4 vs. 8), and the 60-image set shares source types (Gemini, ChatGPT, and personal photographs) with the self-collected training data, so the result reflects improvement on those sources and does not demonstrate generalization to entirely unseen generators. Over the same change, accuracy on unseen Tiny-GenImage samples did not improve (Table I: 74.75% to 73.75% at n=800). Threshold calibration on the same 60-image set provided a further apparent 6.67-point improvement (81.67%); as discussed in Section V-C, part of this specific gain reflects threshold selection on the evaluation set itself and should be interpreted alongside the leakage-free transfer result.

TABLE II EFFECT OF DATA DIVERSITY AND THRESHOLD CALIBRATION

Configuration	Threshold	Accuracy	Test Set	N
Public data only (3,000 images)	0.50	53.33%	60-image set	60
+ Self-collected data (7,860 images)	0.50	75.00%	60-image set	60
+ Calibrated threshold	0.68	81.67%	60-image set	60
Large-scale, default threshold	0.50	72.55%	Tiny-GenImage	2000
Large-scale, transferred threshold	0.68	76.45%	Tiny-GenImage	2000

B. Large-Scale Validation and Threshold Transfer

To examine how the final model behaves beyond the 60-image set, it was additionally evaluated on 2,000 images independently sampled from Tiny-GenImage (excluding all images used in training). At the default threshold (0.5), this yielded 72.55% accuracy — a number free of any threshold-selection leakage, since no calibration was performed on this data. Critically, we then transferred the threshold value calibrated on the separate 60-image set (0.68) to this 2,000-image set without any recalibration, obtaining 76.45% accuracy. This value is directly comparable to, and modestly higher than, the 72.55% default-threshold result, and is nearly identical to the 76.75% obtained when the threshold was instead calibrated directly on this same 2,000-image set — indicating that threshold overfitting had a negligible practical effect in this instance, though we report the leakage-free 76.45% figure as the methodologically correct estimate. On a different 2,000-image slice (stream positions 3,000–5,000), the same model scored 68.85% at the default threshold (precision 62.41%, recall 94.80%), which indicates that accuracy on Tiny-GenImage varies by several points between slices.

Using a normal approximation to the binomial distribution, the 95% confidence interval for the 60-image result is approximately ±9.8 percentage points (71.9–91.5%), and for the 2,000-image transferred-threshold result approximately ±1.9 percentage points (74.6–78.3%); the overlap between these intervals supports the conclusion that the smaller-sample result reflects genuine model behavior rather than sampling noise. An additional independent check on a freshly collected 30-image set (15 real, 15 AI-generated; not used in training or in any threshold selection) yielded 80.00% accuracy at the transferred threshold, providing a further leakage-free confirmation.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151159

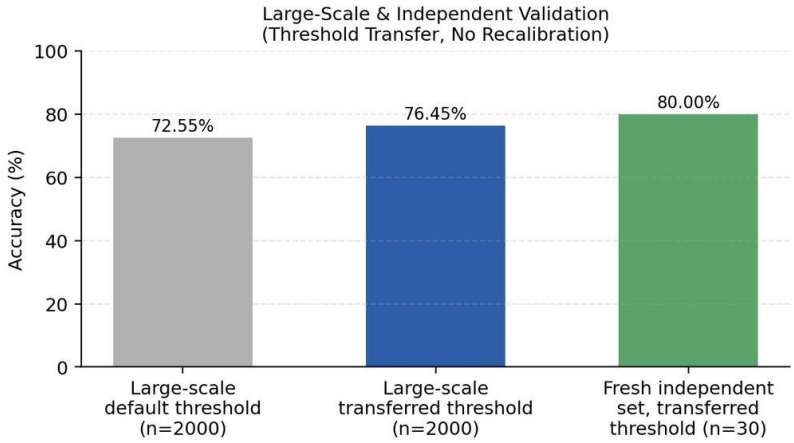


Fig. 2 Accuracy at default vs. transferred threshold on the 2,000-image large-scale set, and on the independent 30-image freshly collected set.

C. Error Analysis

In nearly all evaluated settings, misclassifications were asymmetric: the models predicted the AI class for real images more often than the reverse. For the final ResNet18 on the 60-image set at threshold 0.5, 13 real images were classified as AI versus 2 AI images classified as real; the only exception was EfficientNet-B3 on the 60-image set (12 versus 11). This pattern may reflect real photographs whose compression characteristics, resolution, or capture conditions differ from the curated real-image distributions in the public training datasets (e.g., MS COCO [6], ImageNet [4]), although we did not test this explanation directly. Threshold calibration on the 60-image set reduced the bias (13 to 8 false positives, with false negatives rising from 2 to 3) but did not remove it.

VI. DISCUSSION

The main findings of this study are as follows. First, models trained only on public data performed markedly worse on the shifted 60-image set (53.33% and 61.67%) than on unseen samples from their own training datasets (72.12%–74.75%), consistent with the generalization gap reported in prior work [1], [5]. Second, adding 360 self-collected images from additional sources was associated with a large gain on the 60-image set (to 75.00%), larger than the difference between the two architectures (8.3 points), even though the augmented model used fewer training images than the public-only EfficientNet-B3. Third, this gain did not carry over to Tiny-GenImage, where the augmented model scored slightly lower than the public-only ResNet18 (73.75% vs. 74.75% at n=800; 68.85% vs. 72.60% at n=2000), which suggests a trade-off between accuracy on the added sources and accuracy on the original public distribution. Because the two architectures have nearly identical parameter counts and the runs differ in training volume and schedule, the results do not isolate the effect of model capacity, and we make no claim that larger models generalize worse.

A further methodological observation is that validation accuracy computed on a random split of the training data pool did not reliably predict performance on held-out data: the tuned EfficientNet-B3 achieved 85.48% validation accuracy on a random split of the pool used for training, yet only 72.12% on unseen Tiny-GenImage samples (a 13-point gap) and 61.67% on the 60-image shifted set. This underscores the importance of source-level (not merely sample-level) held-out evaluation when assessing generalization claims.

VII. CHALLENGES AND FUTURE WORK

- The shifted test set contains only 60 images (a 95% confidence interval of roughly ±10 percentage points) and shares source types with the self-collected training data; larger test sets from independent generators and capture devices are needed.
- The runs differ in training volume (3,000, 7,860, and 9,000 images), epochs, and optimization settings, so neither the architecture comparison nor the effect of the 360 self-collected images is a controlled ablation; a public-only ResNet18 trained on the same 7,500 public images with identical settings is needed to isolate the latter. Only two architectures were compared; broader sweeps (e.g., Vision Transformers, CLIP-based detectors) were outside the scope of this study.
- Self-collected data did not include GAN-based generators; future work could incorporate datasets such as Artifact [2] to test generalization to adversarially-trained generators.



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151159

- The calibrated decision threshold was tuned on a specific held-out set; its optimality may not transfer exactly to other deployment distributions. The best threshold (0.68) lay at the edge of the searched range (0.30–0.70), and accuracy on Tiny-GenImage differed by several points between slices (72.55% and 68.85% at the default threshold).

VIII. CONCLUSION

This paper presented an empirical investigation of how training data composition and architecture choice affect AI-generated image detection under distribution shift. Models trained only on public data scored 53.33% (ResNet18) and 61.67% (EfficientNet-B3) on a held-out set of images from other sources, whereas a ResNet18 trained with 360 additional self-collected images scored 75.00% (81.67% after threshold calibration), at the cost of slightly lower accuracy on unseen Tiny-GenImage samples. Since the two architectures have near-identical parameter counts and the runs differ in training volume and schedule, the study does not establish an effect of model capacity. The results suggest that practitioners with constrained data-collection budgets may obtain larger gains on a target distribution by adding even a small amount of data from that distribution than by switching between these two architectures; a controlled ablation and larger, independent test sets are needed to confirm this.

IX. CODE AND DATA AVAILABILITY

Source code, training notebooks, and the evaluation scripts used in this study are available at https://github.com/Arusheee/ai-vs-real-image-detector. A live demonstration of the trained model is deployed at https://ai-vs-real-image-detector-arushee.streamlit.app/, where users can upload an image to obtain a Real / AI-generated / Uncertain prediction with confidence score.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the Department of Electronics and Communication Engineering, Bangalore Institute of Technology, Bengaluru, for providing the facilities, guidance, and encouragement that made this work possible. The authors also thank the creators of the GenImage and Defactify (MS-COCOAI) datasets for making their data publicly available, and the Hugging Face platform for hosting the datasets used in this study.

REFERENCES

[1]. M. Zhu et al., “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image,” arXiv:2306.08571, 2023.
[2]. M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, and S. A. Fattah, “ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection,” in Proc. IEEE Int. Conf. Image Process. (ICIP), 2023, pp. 2200–2204.
[3]. Q. Bammey, “Synthbuster: Towards Detection of Diffusion Model Generated Images,” IEEE Open J. Signal Process., vol. 5, pp. 1–9, 2024, doi: 10.1109/OJSP.2023.3337714.
[4]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in Proc. IEEE CVPR, 2009, pp. 248–255.
[5]. R. Roy et al., “A Comprehensive Dataset for Human vs. AI Generated Image Detection,” arXiv:2601.00553, 2026.
[6]. T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” in Proc. ECCV, 2014, pp. 740–755.
[7]. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.
[8]. M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proc. ICML, 2019.
[9]. I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in Proc. ICLR, 2019.

 This work is licensed under a Creative Commons Attribution 4.0 International License