

Comparison of Deep Learning Approaches for Fake Image Classification

Fatih BAYRAM^{1*}

¹*Mechatronic Engineering, Faculty of Technology, Afyon Kocatepe University, Türkiye*

*(fatihbayram@aku.edu.tr)

(Received: 08 February 2026, Accepted: 15 February 2026)

(7th International Conference on Engineering, Natural and Social Sciences ICENSOS 2026, February 06-07, 2026)

ATIF/REFERENCE: Bayram, F. (2026). Comparison of Deep Learning Approaches for Fake Image Classification, *International Journal of Advanced Natural Sciences and Engineering Researches*, 10(2), 134-138.

Abstract – Rapid advancements in generative artificial intelligence models have increasingly made it difficult for humans to distinguish fake images from real ones, giving rise to serious risks in terms of security, ethics, and information reliability. In this study, deep learning-based approaches for the automatic detection of AI-generated images are systematically investigated. Experiments are conducted on the CIFAKE dataset, a balanced, publicly available benchmark comprising 60,000 real and 60,000 synthetic images. EfficientNet-B0, EfficientNet-B3, and EfficientNet-B6 architectures with varying depths and capacities are evaluated using three different loss function variants: standard cross-entropy (CE), attention-enhanced cross-entropy (Attn+CE), and an attention-based composite loss function (Attn+Composite).

To analyze the effect of random initialization, all models are trained with five different random seeds. Performance is evaluated using Accuracy, Precision, Recall, F1-score, PR-AUC, and Expected Calibration Error (ECE) to assess prediction reliability. In addition, the statistical significance of performance differences between variants is examined using the McNemar test. The results demonstrate that the mid-depth EfficientNet-B3 architecture provides a more balanced trade-off between performance and stability. While the Attn+CE loss function yields meaningful improvements for specific architectures, the Attn+Composite variant does not consistently outperform simpler alternatives. Overall, this study presents a comprehensive evaluation that jointly considers architectural depth, loss function design, classification performance, and calibration reliability in fake image detection.

Keywords – Fake image detection, generative artificial intelligence, deep learning, loss functions, attention mechanisms

I. INTRODUCTION

With the rapid development of artificial intelligence, distinguishing synthetically generated images from real ones has become increasingly challenging for human perception. This difficulty facilitates the widespread dissemination of manipulation, security threats, ethical violations, and related adverse consequences. Consequently, the development of automated systems capable of detecting AI-generated fake content has become necessary to ensure information reliability. Deep learning techniques have demonstrated strong capabilities in automatically identifying synthetic images that are difficult for humans to recognize. In particular, deep learning-based methods are highly effective at capturing subtle artifacts and inconsistencies present in artificial images.

Publicly available datasets play a crucial role in enabling fair comparisons and performance evaluation of such methods [1]–[8]. The CIFAKE dataset has emerged as a prominent benchmark for real-versus-fake image classification [9]. CIFAKE is a balanced dataset composed of real images derived from CIFAR-10 and synthetic images generated using modern generative models [9], [10]. Such large-scale, open-access datasets facilitate systematic evaluations of different network architectures under identical experimental conditions.

Loss functions also play a critical role in determining the performance of deep learning models. However, relying solely on Accuracy for performance evaluation can lead to misleading conclusions [11], [12]. Therefore, a comprehensive analysis using multiple evaluation metrics is required. Statistical tests help determine whether observed performance differences are due to random chance or represent genuine improvements. Paired statistical methods such as the McNemar test are widely used to compare models trained and evaluated on the same dataset [13], [14].

II. MATERIALS AND METHODS

In this study, the CIFAKE dataset, consisting of a balanced total of 120,000 images (60,000 real and 60,000 synthetic), is utilized. The dataset is constructed using the CIFAR-10 dataset and contains 32×32 pixel images across 10 categories (e.g., airplane, automobile, cat). These characteristics make CIFAKE a suitable benchmark for comparing different deep learning architectures.

Three EfficientNet architectures with varying depths and capacities—EfficientNet-B0, EfficientNet-B3, and EfficientNet-B6—are examined to analyze their classification performance [15]. One of the most influential factors affecting classification performance is the choice of loss function. Accordingly, three different loss function variants are evaluated under identical architectures: standard cross-entropy (CE), attention-enhanced cross-entropy (Attn+CE), and an attention-based composite loss function (Attn+Composite).

All experiments are conducted under identical training conditions. To assess the effect of random initialization, each model is trained with five different random seeds. Model performance is evaluated using Accuracy, Precision, Recall, F1-score, and PR-AUC metrics. In addition, prediction reliability is assessed using the Expected Calibration Error (ECE) [16]. The statistical significance of performance differences is assessed using paired McNemar tests.

III. RESULTS

This section presents experimental results obtained on the CIFAKE dataset and compares the performance of different EfficientNet architectures and loss function variants. All results are derived from five different random seeds and evaluated using multiple performance metrics. Table 1 reports the best classification performance achieved across all seeds for each architecture and loss function variant.

Table 1. Best performance values across all random seeds

Architecture	Variant	Best Accuracy	Precision	Recall	F1-score	PR-AUC
EfficientNet-B0	CE	97.56	96.71	98.46	97.58	99.74
EfficientNet-B0	Attn+CE	97.34	96.88	97.82	97.35	99.70
EfficientNet-B0	Attn+Composite	97.29	97.32	97.26	97.29	99.63
EfficientNet-B3	CE	97.55	97.00	98.00	97.49	99.70
EfficientNet-B3	Attn+CE	97.58	96.56	98.67	97.61	99.73
EfficientNet-B3	Attn+Composite	97.50	98.00	97.00	97.49	99.70
EfficientNet-B6	CE	97.48	98.00	97.00	97.49	99.70
EfficientNet-B6	Attn+CE	97.53	98.11	96.92	97.51	99.67
EfficientNet-B6	Attn+Composite	97.58	97.00	98.00	97.49	99.70

The results indicate that all models achieve high classification Accuracy and F1-scores on the CIFAKE dataset. Variations in Precision and Recall among models with similar Accuracy values highlight the limitations of relying on a single metric. The consistently high PR-AUC values (>99%) demonstrate strong discriminative capability across all experiments. Figure 1 illustrates the multi-seed Accuracy results (mean \pm standard deviation) for different architectures and loss function variants.

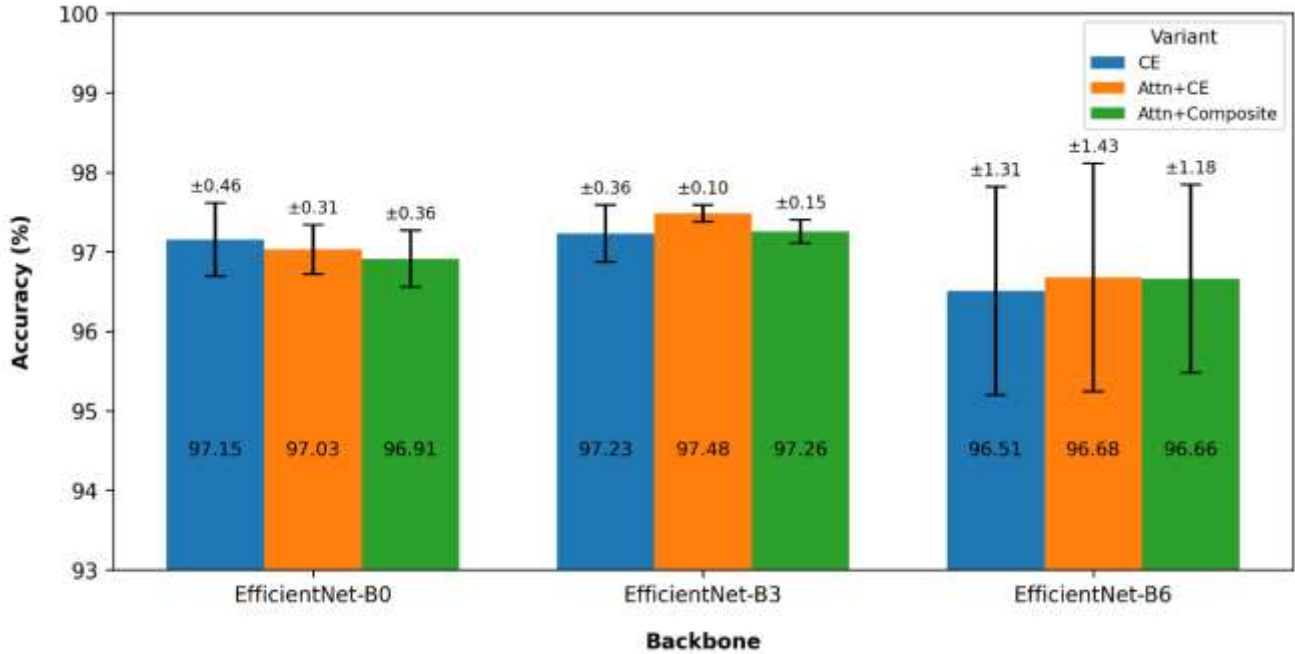


Fig. 1 Accuracy results (mean \pm std) obtained from multiple random seed initializations

The multi-seed analysis reveals notable differences in both average classification Accuracy and stability across architectures and loss function variants. The EfficientNet-B3 architecture with the Attn+CE loss achieves the highest Accuracy with the lowest standard deviation, indicating a favorable balance between performance and reproducibility. Although the deeper EfficientNet-B6 architecture achieves high performance for some seeds, it shows higher overall variance. These findings suggest that the impact of loss function design is architecture-dependent. While the CE loss yields more stable results for EfficientNet-B0, Attn+CE provides the best performance for EfficientNet-B3. In contrast, the Attn+Composite variant does not consistently offer additional benefits. These observations highlight the importance of jointly considering architectural depth and loss function design, as well as the critical role of multi-seed evaluation in assessing model reliability.

In addition to classification performance, model reliability is evaluated using the ECE metric. Table 2 presents the ECE values obtained for the best-performing seed of each variant.

Table 2. ECE values for different architectures and variants

Architecture	Variant	ECE (%)
EfficientNet-B0	CE	0.42
EfficientNet-B0	Attn+CE	0.50
EfficientNet-B0	Attn+Composite	0.65
EfficientNet-B3	CE	0.38
EfficientNet-B3	Attn+CE	0.52
EfficientNet-B3	Attn+Composite	0.73
EfficientNet-B6	CE	0.45
EfficientNet-B6	Attn+CE	0.55
EfficientNet-B6	Attn+Composite	0.78

The results demonstrate that high classification Accuracy does not necessarily correspond to low calibration error. In particular, the Attn+Composite variant produces higher ECE values for specific architectures, underscoring the importance of calibration analysis alongside performance metrics.

To assess the statistical significance of performance differences, McNemar tests are conducted. Table 3 reports the chi-square (χ^2) values obtained from pairwise comparisons.

Table 3. McNemar test results for different architectures and loss function variants

Architecture	Comparison	χ^2
EfficientNet-B0	CE vs. Attn+CE	10.33
EfficientNet-B0	CE vs. Attn+Composite	43.94
EfficientNet-B0	Attn+CE vs. Attn+Composite	14.13
EfficientNet-B3	CE vs. Attn+CE	2.71
EfficientNet-B3	CE vs. Attn+Composite	2.78
EfficientNet-B3	Attn+CE vs. Attn+Composite	11.48
EfficientNet-B6	CE vs. Attn+CE	2.58
EfficientNet-B6	CE vs. Attn+Composite	5.03
EfficientNet-B6	Attn+CE vs. Attn+Composite	11.19

The McNemar test results indicate statistically significant differences between specific architecture-variant pairs, particularly between Attn+CE and Attn+Composite for EfficientNet-B0 and EfficientNet-B6. Following standard practice, χ^2 values exceeding 3.84 indicate statistical significance at the 0.05 significance level. These findings statistically validate the influence of loss function and model selection on classification performance.

IV. DISCUSSION

The findings of this study demonstrate that deep learning-based methods exhibit strong discriminative capability in fake image detection and that loss function design plays a critical role in performance. Experiments conducted with multiple random seeds reveal that increased architectural depth does not necessarily guarantee improved performance. Notably, EfficientNet-B3 achieves higher average Accuracy and lower variance compared to deeper architectures, providing a more balanced trade-off between performance and stability. This suggests that overly deep models may not always be advantageous, particularly for low-resolution datasets such as CIFAKE.

Analysis of loss function variants indicates that the attention-enhanced Attn+CE loss yields meaningful performance improvements for specific architectures. In contrast, the more complex Attn+Composite variant fails to demonstrate consistent superiority, suggesting that increased complexity alone does not ensure better performance. Furthermore, ECE analysis reveals that some models with high classification Accuracy may exhibit inferior calibration reliability. The McNemar test results confirm that the observed improvements are not due to random chance, supporting the experimental and statistical robustness of the proposed evaluation.

V. CONCLUSION

In this study, real-versus-synthetic image classification is investigated on the CIFAKE dataset using different EfficientNet architectures and loss function variants. Experiments conducted with multiple random seed initializations enable a comprehensive comparison not only in terms of peak performance but also in terms of model stability and reproducibility. The results indicate that mid-depth architectures offer a more balanced trade-off between performance and stability, whereas attention-based loss-function variants exhibit architecture-dependent effects. Calibration analysis highlights that high Accuracy does not necessarily imply reliable predictions, emphasizing the importance of considering reliability in fake image

detection. Statistical analyses further confirm the significance of performance differences across variants. Overall, this study presents a holistic evaluation that jointly considers architecture, loss function, performance, and calibration, providing valuable insights for future research on fake image detection.

REFERENCES

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [2] Q. Xu, X. Jiang, T. Sun, H. Wang, L. Meng, and H. Yan, "Detecting artificial intelligence-generated images via deep trace representations and interactive feature fusion," *Information Fusion*, vol. 112, Art. no. 102578, 2024.
- [3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.
- [4] C. Zhu *et al.*, "The evolution and future perspectives of artificial intelligence generated content," *arXiv preprint*, arXiv:2412.01948, 2024.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [7] C. Akdoğan, T. Özer, and Y. Oğuz, "PP-YOLO: Deep learning-based detection model to detect apple and cherry trees in orchard based on histogram and wavelet preprocessing techniques," *Computers and Electronics in Agriculture*, vol. 232, Art. no. 110052, 2025.
- [8] N. E. Bengi, B. Orhan, and İ. Koyuncu, "Deep learning-based classification of skin lesions," in *Proc. 5th Int. Boğaziçi Scientific Research and Innovation Congress*, Türkiye, 2024.
- [9] J. J. Bird and A. Lotfi, "CIFAKE: Image classification and explainable identification of AI-generated synthetic images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024.
- [10] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Computer Science, University of Toronto, Toronto, ON, Canada, 2009.
- [11] C. Li, K. Liu, and S. Liu, "A survey of loss functions in deep learning," *Mathematics*, vol. 13, no. 15, Art. no. 2417, Jul. 2025.
- [12] D. Powers, "Evaluation: From precision, recall and F-measure to ROC and informedness," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [13] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947.
- [14] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional," *BMC Medical Research Methodology*, vol. 13, Art. no. 91, 2013.
- [15] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.
- [16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.