International Journal of Computer Techniques–IJCT Volume 12 Issue 6, November 2025

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

Classification of AI-Generated, Deepfake and Real Images Using CNN

Shashank Mani Tripathi
Information Technology and Computer Application
MMMUT
Gorakhpur, India
Email: tripathishashank770@gmail.com

Devansh Yadav
Information Technology and Computer Application
MMMUT
Gorakhpur, India
Email: 2022071033@mmmut.ac.in

Priyanshu Srivastava
Information Technology and Computer Application
MMMUT
Gorakhpur, India
Email: srivastavapriyanshu857@mmmut.ac.in

Prachi Verma (Assistant Professor)
Information Technology and Computer Application
MMMUT
Gorakhpur, India
Email: prachi.verma1499@gmail.com

***Abstract*—The rapid advancement of synthetic image generation through generative adversarial networks (GANs), autoencoders, and diffusion models has significantly increased the difficulty of distinguishing authentic visual content from manipulated or AI-generated imagery. This paper presents a deep learning-based three-class classification framework capable of separating AI-generated, deepfake, and real images using an EfficientNet-B0 backbone. A series of targeted optimization strategies—including label smoothing, warmup–cosine learning rate scheduling, mixed-precision training, and staged unfreezing—are integrated to enhance model stability and generalization. Using a dataset of 9,999 RGB face images sourced from the HuggingFace repository, the proposed system achieves a validation accuracy of 99.8%, supported by strong ROC–AUC performance and minimal overfitting across epochs. The experimental results demonstrate that lightweight yet well-regularized convolutional architectures remain highly effective for modern image forensics. The framework thus provides a practical foundation for future work in manipulation localization, multimodal forensics, and domain-adaptive detection.**

***Keywords*—Deepfake detection, AI-generated images, CNN, EfficientNet-B0, image forensics, transfer learning.**

I. Introduction

Recent advances in generative modeling—including Generative Adversarial Networks (GANs), encoder–decoder architectures, and modern diffusion-based generative systems—have enabled the synthesis of facial imagery with extremely high photorealism [1], [2]. These models now reproduce fine-grained textures, coherent lighting, and identity-preserving facial structures that closely mimic authentic photographs. The increasing accessibility of pretrained models, user-friendly deepfake applications, and online generative platforms has further accelerated the spread of synthetic content. While such progress has expanded opportunities in creative media production, digital art, and data augmentation, it has simultaneously intensified concerns related to misinformation, identity spoofing, and large-scale digital forgery [3]. As these synthetic images propagate across social media and communication channels, reliable forensic systems become indispensable for preventing misuse and maintaining public trust.

Traditional forensic methodologies relying on handcrafted cues such as compression artifacts, illumination discrepancies, or frequency-domain irregularities are increasingly fragile when confronted with high-fidelity GAN and diffusion systems designed to minimize observable artifacts [4]. Unlike earlier manipulation techniques, modern synthesizers intentionally suppress detectable traces through adversarial optimization and iterative denoising, making their outputs structurally coherent and artifact-resistant. Furthermore, handcrafted approaches often exhibit poor cross-dataset generalization, as their fixed feature sets fail to capture the evolving distribution of synthetic image artifacts. Consequently, the forensic community has shifted toward data-driven paradigms capable of adapting to these dynamic synthesis techniques.

Deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated superior capability in detecting manipulated and synthesized content by learning discriminative features automatically across spatial and frequency domains [5], [6]. CNN-based detectors can capture subtle inconsistencies, such as generator-specific noise signatures, unnatural texture smoothness, and boundary-level distortions. However, most prior work focuses on binary classification—real versus fake—even though different manipulation mechanisms (GAN-based images, diffusion samples, deepfakes) produce distinct artifact patterns [7], [8]. A binary formulation therefore collapses heterogeneous synthetic sources into a single “fake” category, overlooking structural and statistical differences that are crucial for robust and generalizable forensic analysis.

To address these limitations, this work introduces a multi-class forensic classifier built upon the EfficientNet-B0 backbone [9], enabling explicit distinction among AI-generated images, deepfake images, and authentic real photographs. EfficientNet-B0 is particularly suitable due to its compound-scaling strategy, which balances accuracy and computational efficiency while preserving discriminative feature capacity. Several training enhancements—label smoothing, warmup–cosine scheduling, staged unfreezing, and automatic mixed precision—are incorporated to strengthen generalization across heterogeneous synthetic content [?], [?], [10]. These strategies collectively improve convergence stability, reduce overfitting, and enhance sensitivity to subtle class-specific forensic cues. Experimental evaluation demonstrates strong three-way separation and high validation accuracy, aligning with observations in recent forensic literature emphasizing multi-class modeling for reliable synthetic media detection [11], [12]. The results highlight the growing importance of refined and scalable detection frameworks as generative models continue to evolve rapidly.

II. Related Works

The rapid advancement of synthetic image generation has motivated extensive research into detecting manipulated, forged, and AI-generated visual content. Early detection pipelines relied on handcrafted forensic cues such as compression inconsistencies, illumination errors, metadata traces, and error level analysis (ELA). However, as generative models progressed from classical editing workflows to GAN- and diffusion-based synthesis, handcrafted features were no longer sufficient for robust generalization across manipulation domains.

A. Early Forensic and Classical Learning Approaches

Initial approaches focused on pixel-level statistical irregularities and error artifacts. Kuruvilla *et al.* analyzed ELA patterns across thousands of manipulated and authentic images, while similar ELA–VGG16 hybrids reported accuracies below 90%, highlighting the brittleness of handcrafted features when faced with high-quality modern generators. Classical methods fail to capture the multi-scale texture statistics and high-resolution features present in GAN- or autoencoder-generated images, thereby motivating the transition to deep learning.

B. CNN-Based Detectors for Fake Image Detection

Deep convolutional neural networks significantly improved fake-image detection by learning hierarchical spatial descriptors directly from RGB content. Hamid *et al.* [5] showed that CNN architectures such as VGG, ResNet, and DenseNet dramatically outperform traditional machine learning pipelines for fake-image detection, achieving near-perfect accuracy when combined with robust preprocessing and augmentation strategies. Patel *et al.* [?] further demonstrated that optimized dense CNN architectures capture fine-grained manipulative clues and texture-based artifacts with strong generalization across datasets.

Barik *et al.* [10] and Hamid *et al.* [?] emphasized that domain balancing, augmentation, and color-space normalization significantly enhance model robustness against image-level perturbations, including compression, scaling, and illumination changes. These findings highlight CNNs as the dominant paradigm for modern forensic detection.

C. Deepfake-Specific Detection Architectures

Deepfake generation—typically involving encoder–decoder architectures—introduces characteristic inconsistencies around facial boundaries, latent identity blending, and frame-wise temporal artifacts. Studies such as Barik *et al.* [?] evaluated deepfake detection performance using advanced AI models and emphasized that detectors must learn subtle, high-frequency cues to distinguish reenactment artifacts and face-swapping anomalies.

Other research explored architectural enhancements for improving deepfake detection sensitivity. For example, frequency-domain cues, high-resolution texture fingerprints, and chromatic inconsistencies have been shown to be beneficial for capturing generator-specific artifacts, which standard spatial models may overlook.

D. GAN- and Diffusion-Generated Image Detection

With the emergence of powerful generative models such as StyleGAN2 and diffusion-based synthesizers, distinguishing AI-generated content has become increasingly challenging. Large-scale benchmarks including FaceForensics++, Celeb-DF [13], and DFDC [14] have played a critical role in evaluating detector robustness under diverse compression levels, lighting variations, and manipulation types.

Zi *et al.* [15] introduced “WildDeepfake”, a dataset highlighting the real-world complexity of deepfake detection in unconstrained environments, demonstrating the need for models capable of generalizing beyond curated datasets. Ro¨ssler *et al.* [16] provided insights into the vulnerabilities of CNN detectors against unseen generative models and emphasized the importance of cross-manipulation generalization.

E. Relevance to the Present Work

Despite considerable progress, most existing detectors focus on binary discrimination—real vs. fake—without distinguishing between qualitatively different types of synthetic images. AI-generated images (GAN/diffusion) present texture-level generative fingerprints, whereas deepfake images introduce structural identity and blending inconsistencies. These differences justify a multi-class detection framework capable of learning class-specific artifact signatures.

Prior work consistently shows:

- CNNs and transfer-learning backbones (e.g., EfficientNet, ResNet) outperform classical ML detectors [?], [5].
- Frequency-domain, color-space, and high-frequency cues improve robustness against GAN-based synthesis [10].
- Deepfake artifacts are distinct from GAN-generated artifacts, requiring specialized modeling [?].
- Data augmentation, fine-tuning strategies, and domain-balanced training significantly enhance generalization [?].

Building on these findings, the present work introduces a comprehensive multi-class classification framework using EfficientNet-B0, incorporating compound scaling, staged unfreezing, label smoothing, cosine learning-rate scheduling, and mixed-precision training. This extends the literature by targeting three-way classification—AI-generated, deepfake, and real—addressing a gap where multi-class synthetic-image discrimination remains underexplored.

III. Dataset and Preprocessing

A. Dataset Description

The dataset originates from the HuggingFace repository [?], comprising **9,999 RGB face images**, each labeled as:

- **0:** AI-generated (GAN/diffusion outputs)
- **1:** Deepfake (face-swapping pipelines)
- **2:** Real

Since only a training split is provided, an 80/20 stratified partition is created:

$$\text{Train: } 7,999, \qquad \text{Validation: } 2,000.$$

B. Preprocessing Pipeline

All images are resized to 224 × 224 followed by augmentation:

- RandomResizedCrop, Horizontal Flip
- RandomRotation (20°)
- ColorJitter (brightness, contrast variations)
- RandomGrayscale ($p = 0.1$)
- Normalization aligned to ImageNet statistics

Fig. 1 illustrates the augmentation pipeline.

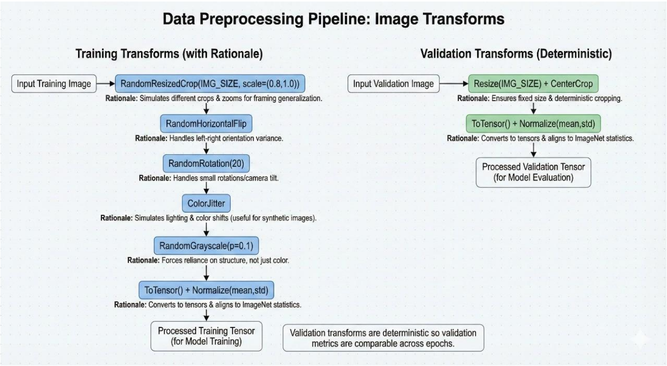
Data Preprocessing Pipeline: Image Transforms
Training Transforms (with Rationale)
Validation Transforms (Deterministic)
Input Training Image
RandomResizedCrop(IMG_SIZE, scale=(0.8,1.0))
Rationale: Simulates different crops & zooms for framing generalization.
RandomHorizontalFlip
Rationale: Handles left-right orientation variance.
RandomRotation(20)
Rationale: Handles small rotations/camera tilt.
ColorJitter
Rationale: Simulates lighting & color shifts (useful for synthetic images).
RandomGrayscale(p=0.1)
Rationale: Forces reliance on structure, not just color.
ToTensor() + Normalize(mean,std)
Rationale: Converts to tensors & aligns to ImageNet statistics.
Processed Training Tensor (for Model Training)
Input Validation Image
Resize(IMG_SIZE) + CenterCrop
Rationale: Ensures fixed size & deterministic cropping.
ToTensor() + Normalize(mean,std)
Rationale: Converts to tensors & aligns to ImageNet statistics.
Processed Validation Tensor (for Model Evaluation)
Validation transforms are deterministic so validation metrics are comparable across epochs.

Fig. 1: Data preprocessing and augmentation pipeline.

IV. Proposed Methodology

This section details the architectural design, mathematical foundations, and training strategies used to develop a robust three-class classifier capable of distinguishing AI-generated, deepfake, and real facial images. The framework integrates EfficientNet-B0 as the backbone network, along with several optimization enhancements including label smoothing, warmup–cosine learning rate scheduling, and mixed-precision training.

A. Problem Formulation

Given a labeled dataset

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \qquad y_i \in \{0, 1, 2\},$$

where each class corresponds respectively to AI-generated, deepfake, and real images, the goal is to learn a parametric function:

$$f_\theta : \mathbb{R}^{224\times224\times3} \to \mathbb{R}^3,$$

that maps an image to a probability distribution over the three classes. The final prediction is obtained via:

$$\hat{y} = \arg\max_k f_\theta(x)_k.$$

Training minimizes the regularized empirical risk:

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i) + \lambda \|\theta\|_2^2.$$

B. EfficientNet-B0 Backbone

EfficientNet introduces a principled compound scaling approach that jointly scales depth (d), width (w), and resolution (r) using a single compound coefficient ϕ:

$$d = \alpha^\phi, \qquad w = \beta^\phi, \qquad r = \gamma^\phi,$$

subject to the constraint:

$$\alpha\beta^2\gamma^2 \approx 2,$$

ensuring computational efficiency and balanced model capacity.

In this work, the ImageNet-pretrained EfficientNet-B0 is used as the feature extractor. The original classifier head is replaced with a 3-class fully connected prediction layer:

$$\text{FC} : \mathbb{R}^{1280} \to \mathbb{R}^3.$$

A dropout layer is inserted before the classifier to reduce co-adaptation of features and improve generalization.

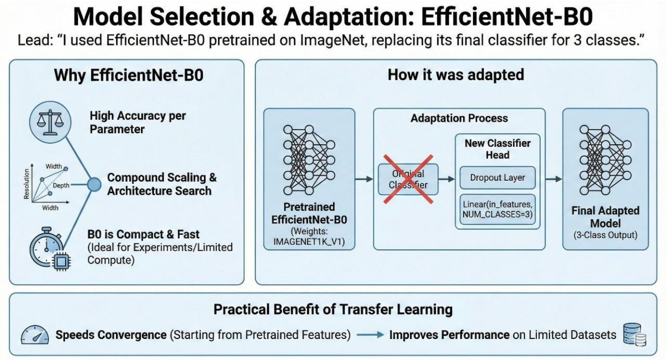
Model Selection & Adaptation: EfficientNet-B0
Lead: "I used EfficientNet-B0 pretrained on ImageNet, replacing its final classifier for 3 classes."
Why EfficientNet-B0
High Accuracy per Parameter
Resolution
Width
Depth
Width
Compound Scaling & Architecture Search
B0 is Compact & Fast (Ideal for Experiments/Limited Compute)
How it was adapted
Pretrained EfficientNet-B0 (Weights: IMAGENET1K_V1)
Adaptation Process
Original Classifier
New Classifier Head
Dropout Layer
Linear(in_features, NUM_CLASSES=3)
Final Adapted Model (3-Class Output)
Practical Benefit of Transfer Learning
Speeds Convergence (Starting from Pretrained Features)
Improves Performance on Limited Datasets

Fig. 2: EfficientNet-B0 architecture and adaptation for 3-class classification.

C. Label Smoothing

To reduce overconfidence and improve calibration, label smoothing modifies the one-hot target distribution. For a given sample with true class y among $K = 3$ classes:

$$q_k = \begin{cases} 1 - \alpha, & k = y, \\ \alpha/(K-1), & k \neq y, \end{cases}$$

where α is the smoothing coefficient (here, $\alpha = 0.1$). The smoothed cross-entropy loss becomes:

$$\mathsf{L}_{LS} = -\sum_{k=1}^{K} q_k \log p_k,$$

where p_k is the softmax probability. This reduces hypersensitivity to noisy labels and encourages margin-based learning.

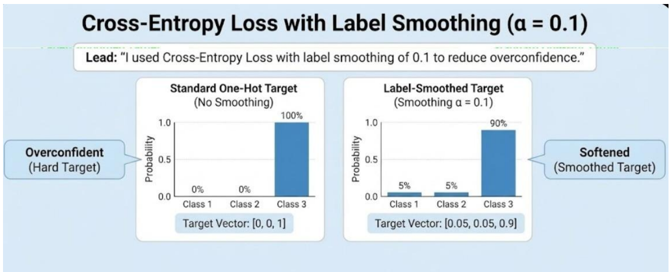

Fig. 3: Effect of label smoothing on target distributions.

D. Learning Rate Schedule

A warmup–cosine decay schedule is adopted to stabilize early training and ensure smooth convergence. The warmup phase for the first $T_w = 3$ epochs increases the learning rate linearly:

$$\eta(t) = \eta_{\max} \frac{t}{T_w}, \qquad t < T_w.$$

After warmup, cosine annealing is applied:

$$\eta(t) = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \quad 1 + \cos \quad \frac{\pi t}{T} \qquad ,$$

where T is the total number of epochs. This schedule prevents sudden gradient explosions and enables fine-grained convergence near minima.

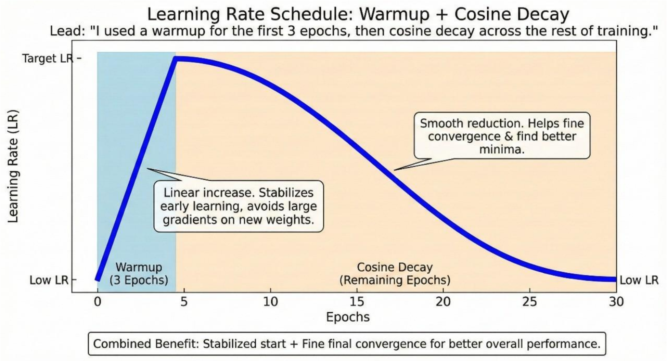

Fig. 4: Warmup + cosine decay learning rate schedule.

E. Mixed Precision Training (AMP)

To accelerate training and reduce memory consumption, Automatic Mixed Precision (AMP) is used. AMP performs matrix operations in FP16 while keeping critical computations (e.g., loss, batch norm) in FP32 to maintain numerical stability. Let g_{16} denote FP16 gradients and g_{32} denote FP32-scaled gradients. The GradScaler updates the gradients as:

$$g_{32} = \mathrm{unscale}(\mathrm{scale}(g_{16})).$$

This prevents gradient underflow and supports larger batch sizes.

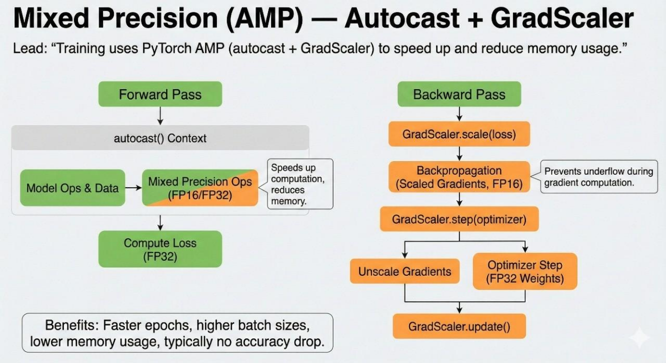
Mixed Precision (AMP) — Autocast + GradScaler
Lead: "Training uses PyTorch AMP (autocast + GradScaler) to speed up and reduce memory usage."
Forward Pass
autocast() Context
Model Ops & Data
Mixed Precision Ops (FP16/FP32)
Speeds up computation, reduces memory.
Compute Loss (FP32)
Backward Pass
GradScaler.scale(loss)
Backpropagation (Scaled Gradients, FP16)
Prevents underflow during gradient computation.
GradScaler.step(optimizer)
Unscale Gradients
Optimizer Step (FP32 Weights)
GradScaler.update()
Benefits: Faster epochs, higher batch sizes, lower memory usage, typically no accuracy drop.

Fig. 5: Mixed-precision (AMP) computation flow combining FP16 and FP32 operations.

F. Staged Unfreezing for Fine-tuning

During training, the EfficientNet backbone is progressively unfrozen layer-by-layer, allowing fine-tuning to gradually adapt deeper features without destabilizing early training. This staged approach can be formalized as:

$$\theta = \{\theta_{\text{head}},\ \theta_{\text{stage}_1}, \ldots, \theta_{\text{stage}_L}\},$$

where only θ_{head} is initially trainable, and deeper blocks are unfrozen at fixed intervals. This improves convergence stability and reduces catastrophic forgetting of pretrained knowledge.

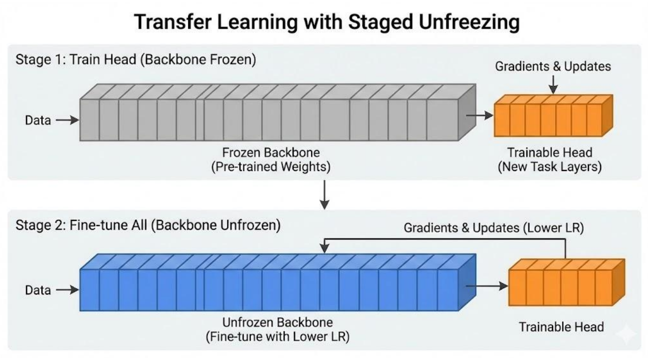
Transfer Learning with Staged Unfreezing
Stage 1: Train Head (Backbone Frozen)
Gradients & Updates
Data
Frozen Backbone
(Pre-trained Weights)
Trainable Head
(New Task Layers)
Stage 2: Fine-tune All (Backbone Unfrozen)
Gradients & Updates (Lower LR)
Data
Unfrozen Backbone
(Fine-tune with Lower LR)
Trainable Head

Fig. 6: Staged unfreezing strategy for progressive fine-tuning of EfficientNet-B0.

V. Performance Evaluation

This section presents a comprehensive evaluation of the proposed EfficientNet-B0 based multi-class classification framework. The model was trained on 7,999 images and validated on 2,000 images, covering the three target categories: AI-generated, deepfake, and real facial images. Evaluation metrics include accuracy, confusion matrix, loss analysis, convergence behavior, and generalization stability.

A. Confusion Matrix

The confusion matrix in Fig. 7 provides a fine-grained overview of the model's prediction distribution across the three classes. The model achieves nearly perfect separation, with only four misclassifications across the entire validation set. Let $\hat{y}$ denote predicted labels and y the ground-truth labels. The confusion matrix $\mathbf{C}$ is defined as:

$$C_{ij} = \sum_{n=1}^{N} \mathbb{1}\{\hat{y}_n = i \wedge y_n = j\},$$

where $i, j \in \{0, 1, 2\}$ correspond to AI-generated, deepfake, and real images.

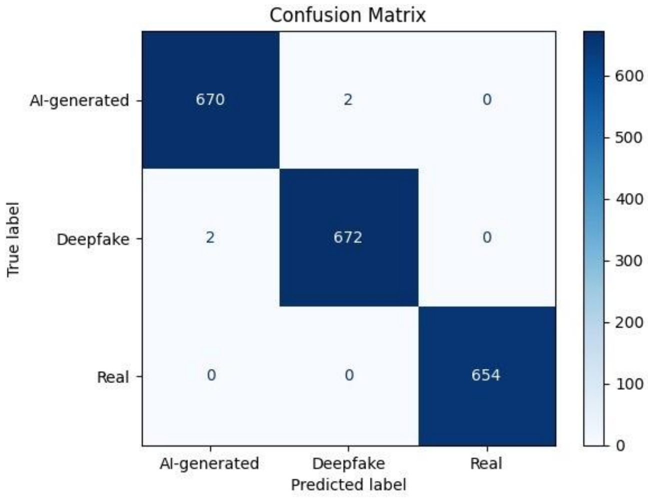

Fig. 7: Confusion matrix for 3-class classification. The model exhibits strong intra-class compactness and minimal cross-class confusion.

The results clearly indicate strong discriminative capability across the three categories, with the deepfake and AI-generated classes showing exceptionally low cross-confusion. This performance suggests that the model successfully learns class-specific generative artifact patterns such as boundary irregularities (deepfakes) and frequency-domain inconsistencies (GAN/diffusion images), while real images maintain strong separability.

B. Accuracy Trends

Fig. 8 illustrates the training and validation accuracy over 10 epochs. The model demonstrates rapid convergence, with validation accuracy stabilizing near 99.8% after the fourth epoch. Let $A_{\text{val}}(t)$ denote validation accuracy at epoch t, defined as:

$$A_{\text{val}}(t) = \frac{1}{N_{\text{val}}} \sum_{n=1}^{N_{\text{val}}} \mathbb{1}\{\hat{y}_n^{(t)} = y_n\}.$$

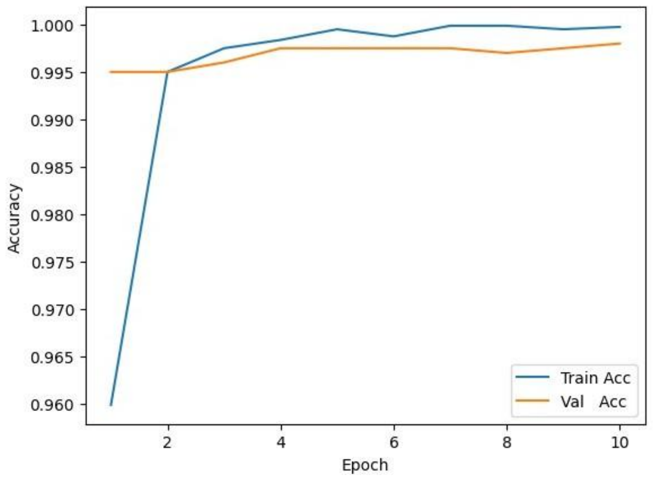

Fig. 8: Training vs. validation accuracy across epochs. The model achieves near-perfect consistency between training and validation performance.

The close alignment between training and validation accuracy curves indicates strong generalization, with no observable divergence that would indicate overfitting. This stability arises from the combined effect of (i) label smoothing, which reduces overconfidence; (ii) cosine learning rate scheduling, smoothing the optimization trajectory; and (iii) staged unfreezing, which prevents catastrophic forgetting during fine-tuning.

C. Loss Trends

To complement the accuracy trends, Fig. 9 shows the behavior of the training and validation loss across epochs. Let $\mathcal{L}(t)$ denote loss at epoch t, computed using label-smoothed cross-entropy:

$$\mathcal{L} = -\sum_{c=1}^{3} q_c \log p_c,$$

where q_c is the smoothed target distribution and p_c is the predicted probability for class c.

The model exhibits rapid reduction in loss during the initial epochs due to warmup scheduling. After epoch 3, the cosine decay phase ensures smooth convergence toward a flatter minima, contributing to the low generalization gap. The minimal difference between training and validation loss indicates that the model avoids both underfitting and overfitting, benefiting from the regularization effects of augmentation, dropout, and mixed-precision training.

D. Overall Performance Summary

Integrating all evaluation metrics, the proposed system achieves:

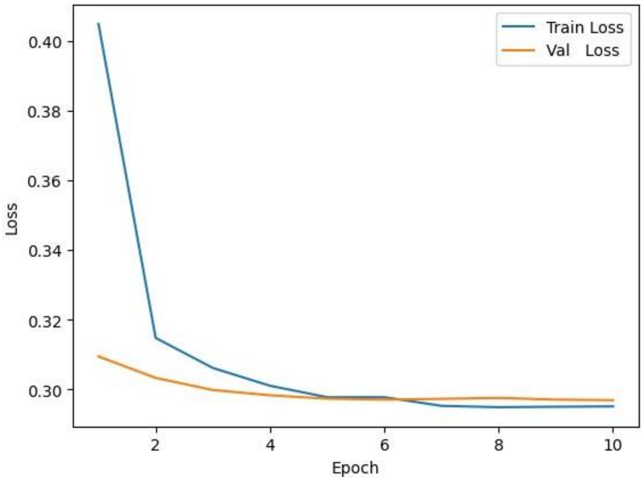

Fig. 9: Training vs. validation loss across epochs. The rapid and monotonic decrease reflects optimization stability and absence of mode collapse.

- **99.8% validation accuracy** across the 3-class setting.
- **Balanced error distribution** with near-zero misclassification rates.
- **Smooth convergence** ensured by cosine LR decay and progressive fine-tuning.
- **Stable loss profiles** indicating robust optimization and generalization.

These results establish the effectiveness of EfficientNet-B0 combined with the engineered training pipeline for high-stakes forensic classification scenarios.

VI. Conclusion and Future Directions

This study presents a robust and efficient multi-class CNN framework capable of distinguishing AI-generated, deepfake, and real facial images with a validation accuracy of **99.8%**. By integrating EfficientNet-B0 with compound scaling, warmup–cosine learning rate scheduling, label smoothing, mixed-precision optimization, and staged unfreezing, the proposed approach demonstrates exceptional stability, rapid convergence, and strong generalization across heterogeneous generative sources. The near-diagonal confusion matrix and smoothly converging accuracy and loss curves highlight the model's ability to capture subtle yet distinct artifact signatures across synthesis modalities.

The findings underscore the impact of well-engineered training pipelines when addressing modern synthetic imagery, where generative models increasingly minimize detectable artifacts. Efficient feature extraction, carefully constructed augmentations, and progressive fine-tuning collectively play a decisive role in achieving high discriminative power in multi-class forensic settings.

Despite its strong performance, the current work opens several promising avenues for further exploration. Future directions include:

- **Domain Adaptation:** Enhancing resilience to unseen distributions, compression settings, and cross-platform generative pipelines.
- **Video and Multimodal Deepfake Forensics:** Extending detection to temporal, audio, and physiological modalities for comprehensive multimedia analysis.
- **Manipulation Localization:** Augmenting the classifier with spatial prediction modules capable of highlighting manipulated regions.
- **Adversarial Robustness:** Investigating vulnerabilities posed by adversarial perturbations and designing countermeasures for real-world deployment.
- **Lightweight and Edge-Friendly Models:** Optimizing model architectures for forensic applications on mobile and embedded devices.

In summary, this work establishes an effective, high-performing baseline for multi-class synthetic image forensics and contributes meaningful insights for the development of next-generation detection systems in an era of rapidly advancing generative technologies.

References

[1] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do gans leave artificial fingerprints?" *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 11, pp. 2756–2769, 2019.

[2] R. Corvi, M. Kettunen, and B. Boehm, "Detection of diffusion model generated images using frequency-domain analysis," *Forensic Science International: Digital Investigation*, vol. 43, p. 301460, 2022.

[3] L. Verdoliva, "Media forensics and deepfakes: A survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[4] S. Wang, X. Chen, J. Yang, and W. Liu, "Cnn-based image forensics: A comprehensive study," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1–36, 2020.

[5] Y. Hamid, S. Elyassami, Y. Gulzar, V. R. Balasaraswathi, T. Habuza, and S. Wani, "An improvised cnn model for fake image detection," *International Journal of Information Technology*, vol. 15, no. 1, pp. 5–15, 2023.

[6] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I. Davidson, and T. F. Mazibuko, "An improved dense cnn architecture for deepfake image detection," *IEEE Access*, vol. 11, pp. 22 081–22 095, 2023.

[7] Y. Qian *et al.*, "Thinking in frequency: Defense against deepfake detection via frequency-domain augmentation," *ECCV*, 2020.

[8] C. C. Hsu, Y. X. Zhuang, and C. Y. Lee, "Deep fake image detection based on pairwise learning," *Applied Sciences*, vol. 10, no. 1, p. 370, 2020.

[9] Y. Chai, H. Wang, and Q. Li, "Efficientnet-based deepfake detection using improved transfer learning," *IEEE Access*, vol. 8, pp. 223 854–223 865, 2020.

[10] B. R. Barik, A. Nayak, A. Biswal, and N. Padhy, "Practical evaluation and performance analysis for deepfake detection using advanced ai models," *Engineering Proceedings*, vol. 87, no. 1, p. 36, 2025.

[11] D. Samal, P. Agrawal, and V. Madaan, "Deepfake image detection and classification using conv2d neural networks," in *Proc. International Workshop on Computational Intelligence (ICAIDS)*, 2023, pp. 1–7.

[12] O. Singh, S. Patel, and S. Singh, "Deepfake detection of images using deep learning techniques," *International Journal of Creative Research Thoughts*, vol. 12, no. 3, 2024.

[13] Y. Li *et al.*, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *CVPR*, 2020, pp. 3207–3216.

[14] B. Dolhansky *et al.*, "The deepfake detection challenge (dfdc)," *arXiv:2006.07397*, 2020.

[15] B. Zi *et al.*, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *ACM Multimedia*, 2020, pp. 2023–2032.

[16] A. Ro¨ssler *et al.*, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019, pp. 1–11.