Deepfake Face Detection Using Transfer Learning with EfficientNet-B4

Kavin Kumar
Monta Vista High School, California

Abstract

Deepfake media, synthetic videos generated by deep learning models, are a growing threat to information integrity, public trust, and personal privacy. This study presents a deepfake face detection system using a fine-tuned EfficientNet-B4 convolutional neural network trained on the Celeb-DF-v2 benchmark dataset. A balanced data pipeline was constructed with equal amounts of real and fake data to train, validate, and test the model. The model was trained with targeted fine tuning of the top three convolution blocks, weight decay, label smoothing, and dropout. It was finally optimized using an adaptive learning rate scheduler. The system achieves a test accuracy of 94.04% and an AUC-ROC of 0.9844 on Celeb-DF-v2, averaged across two reproducible runs. Grad-CAM visualizations confirm that the model attends to meaningful facial regions rather than artifacts in the background. To assess generalization, a second model trained on a combined Celeb-DF-v2 and FaceForensics++ set achieved 88.87% accuracy and 0.9514 AUC-ROC across both benchmarks, substantially recovering the cross-dataset performance lost by the single-dataset model. These results are competitive in comparison to published baselines on Celeb-DF-v2 and demonstrate how principled transfer learning with targeted regularization can achieve near state-of-the-art deepfake detection.

1. Introduction

Deepfake technology, which uses deep learning to generate or manipulate facial videos with high realism, has advanced at an exceeding rate in the past years. Originally developed for entertainment and creative applications, deepfake generation techniques are now increasingly being exploited for malicious purposes, such as political disinformation, identity fraud, and other deceptive ways [1]. The accessibility of deepfake generation tools has grown significantly as well, allowing convincing synthetic media to be created by people with minimal expertise.

The detection of deepfake videos remains a challenging problem in computer vision. Early detection approaches relied much more on handcrafted features like blinking patterns and facial landmark inconsistencies, which proved to be quite ineffective against improving generation methods [2]. More recent approaches use deep convolutional neural networks (CNNs) that are trained on large-scale datasets to learn discriminative features directly from facial image data.

However, many methods suffer from poor generalization, performing well on the dataset they were trained on, but failing on unseen deepfake types or compression levels [3].

This study addresses the deepfake detection problem using transfer learning with EfficientNet-B4, a convolutional neural network architecture which is known for its strong performance per parameter on image classification tasks [4]. The model is fine-tuned on the Celeb-DF-v2 dataset [5], one of the most challenging and widely used benchmarks for deepfake detection, featuring high-quality synthesized videos that closely resemble authentic footage.

The primary contributions of this work are listed down below:

1) A balanced, subject-disjoint data pipeline for Celeb-DF-v2 that prevents data leakage and class imbalance, two common methodological flaws in published deepfake detection studies.

2) A fine-tuned EfficientNet-B4 model achieving 94.04% test accuracy and 0.9844 AUC-ROC on Celeb-DF-v2, competitive with published baselines.

3) Grad-CAM visualizations provide interpretable evidence of which facial regions allow the model's predictions, addressing the opaque nature of deep learning classifiers.

4) An analysis of cross-dataset generalization shows that a single-dataset model degrades on unseen deepfakes (0.6802 AUC on FaceForensics++), and that combined training recovers performance to 0.9514 AUC across both datasets.

The rest of the paper is as follows : Section 2 reviews related work in deepfake detection. Section 3 describes the dataset, preprocessing pipeline, and model architecture. Section 4 presents experimental results. Section 5 discusses findings and limitations. Section 6 outlines future work.

2. Related work

Deepfake detection has been an active area of research since the emergence of face manipulation techniques based on generative adversarial networks (GANs) and autoencoders. Early work focused on detecting visible artifacts introduced by first-generation deepfake tools, but detection methods have had to evolve alongside increasingly sophisticated generation techniques.

Rossler et al. [2] introduced FaceForensics++, a large-scale benchmark dataset containing videos manipulated using four distinct methods: deepfakes, Face2Face, FaceSwap, and NeuralTextures. This dataset became a standard training and evaluation benchmark for the field and demonstrated that CNN-based detectors could achieve high accuracy in controlled settings but degraded significantly under real-world compression and cross-dataset evaluation.

Li et al. [5] introduced Celeb-DF-v2, which addressed limitations of earlier datasets by providing high visual quality deepfakes which closely resemble authentic celebrity footage. The dataset presents a significantly harder detection challenge than FaceForensics++, with many early methods achieving AUC scores below 0.80 when evaluated on it. Celeb-DF-v2 has since become the standard benchmark when evaluating the generalization of deepfake detectors.

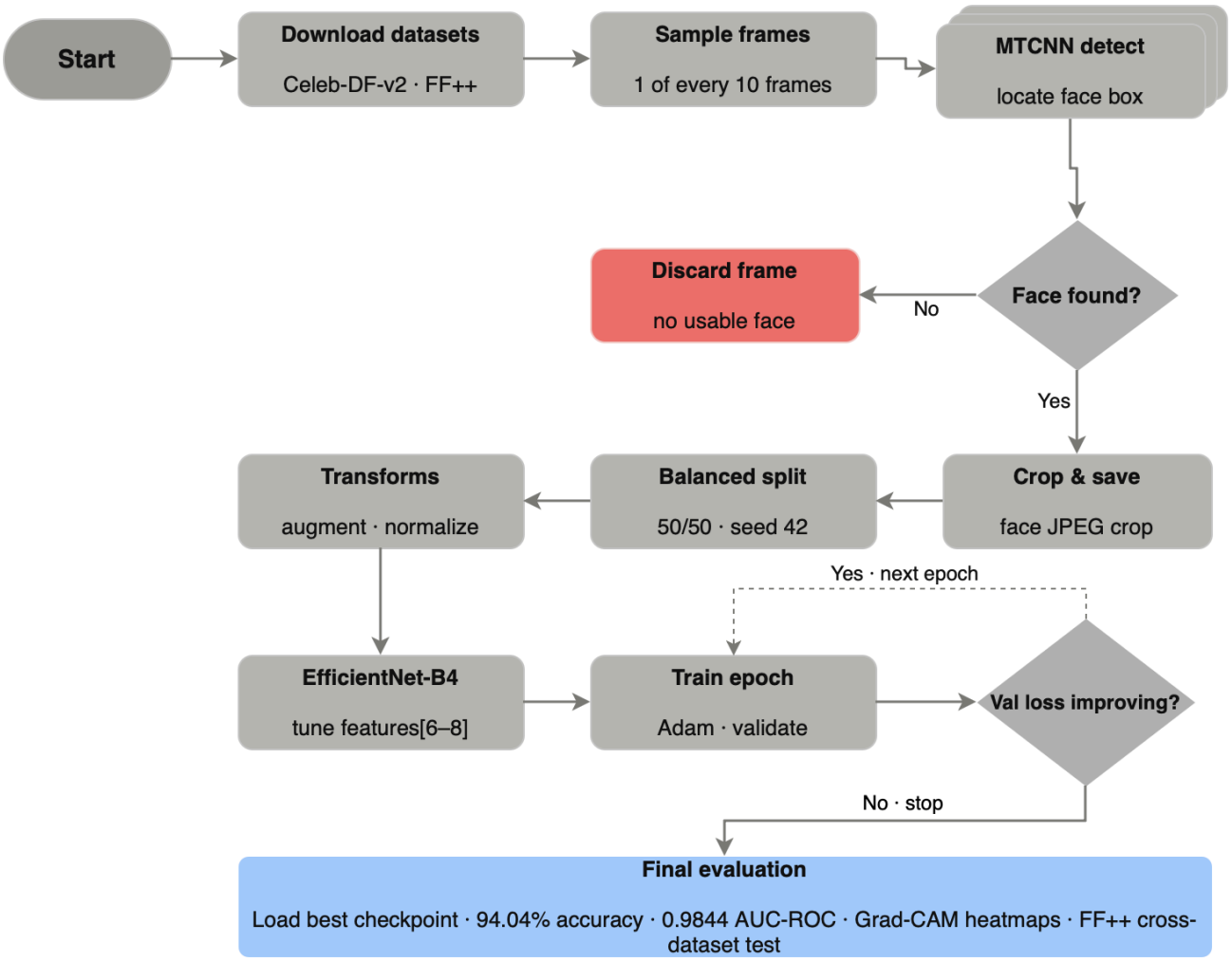
Tan and Le [4] proposed EfficientNet, consisting of convolutional neural networks scaled systematically across depth, width, and resolution using a compound scaling method. EfficientNet-B4 in particular has demonstrated strong performance on image classification benchmarks while maintaining computational efficiency, making it well suited for fine-tuning on domain-specific tasks like deepfake detection.

Selvaraju et al. [6] introduced Gradiant-weighted Class Activation Mapping (Grad-CAM), a technique for producing visual explanations of CNN predictions by computing the gradient of the class score with respect to feature maps in a target convolutional layer. Grad-CAM has been widely adopted in deepfake detection research to verify that models are able to attend to facial regions rather than other random artifacts in the background.

Yan et al. [7] evaluated multiple detection architectures on Celeb-DF-v2 and reported that models trained on FaceForensics++ generalize poorly to Celeb-DF-v2, with AUC scores ranging from 0.70 to 0.86 depending on architecture. This cross-dataset evaluation highlighted the importance of training directly on high-quality deepfake datasets and motivated the use of Celeb-DF-v2 as the primary benchmark in this study.

Yan et al. [7] proposed a comprehensive benchmark of deepfake detection methods and compared EfficientNet-B4, Xception, and ResNet34 as backbone architectures across multiple datasets. Their results showed EfficientNet-B4 to be consistently competitive across within-domain and cross-domain evaluations, supporting its selection as the backbone in this work. The present study builds on these findings by incorporating principled regularization, balanced data splits, and adaptive learning rate scheduling to maximize performance on Celeb-DF-v2.

3. Methodology



End-to-end system architecture for deepfake detection; from raw video ingestion through face extraction, augmentation, model training, and evaluation.

3.1 Dataset

This study uses the Celeb-DF-v2 dataset [5], which contains 590 real videos sourced from Youtube celebrity interviews and 5,639 deepfake videos synthesized using an improved deepfake generation pipeline. Celeb-DF-v2 is one of the most visually challenging deepfake benchmarks available, since it features high-quality face swaps with little to no visible artifacts.

The official test set is defined by the dataset authors via a provided list of 518 test videos (340 fake, 178 real). All remaining videos were used for either training or validation.

In addition to Celeb-DF-v2, this study uses the FaceForensics++ dataset [2] for cross-dataset generalization experiments. FaceForensics++ contains 1000 original YouTube videos and 4000 manipulated videos across four face-manipulation methods, of which the 1000-video Deepfakes subset was used. This study uses the deepfakes subset at the c23 (medium) compression level, which applies an autoencoder-based face-swapping technique distinct from the synthesis pipeline used in Celeb-DF-v2. Because the two datasets are produced by different generation methods, FaceForensics++ provides an effective benchmark for evaluating whether a model trained on one dataset generalizes to deepfakes it has not encountered. Face crops were extracted from FaceForensics++ using the same MTCNN pipeline and frame-sampling rate described above, with balanced real and fake representation.

3.2 Data Pipeline

Face crops were extracted from each video using the Multi-task Cascaded Convolutional Network (MTCNN) face detector at a frame sampling rate of one frame per ten frames. Detected face regions were cropped and saved as JPEG images. A total of 1000 training videos, 200 validation videos, and 200 test videos were processed, with strict 50/50 class balance enforced at each split to prevent class imbalance from distorting evaluation metrics.

Subject-disjoint splits were enforced by assigning videos to train, validation, or test sets at the video level before any frame extraction, ensuring that no facial identity appears across multiple splits. This prevents the data leakage that arises when frames from the same video clip appear in both training and test sets.

A random seed of 42 was used throughout all dataset preparation steps to ensure full reproducibility.

3.3 Preprocessing and Augmentation

All training images were resized to 380x380 pixels, which matches the native input resolution of EfficientNet-B4 to allow the model to work even more effectively. The following augmentations were applied during training using the Albumentations library:

- Random horizontal flip (p = 0.5)
- Gaussian blur with sigma between 0.1 and 2.0 (p=0.5)
- JPEG compression simulation with quality between 50 and 95 (p=0.3)
- Color jitter with brightness and contrast variation of 0.3 (p=0.5)
- ImageNet normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])

JPEG compression augmentation was included specifically to improve robustness to the compression artifacts present in real-world video content. Validation and test images were resized to 380x380 and normalized without augmentation.

3.4 Model Architecture

The detection model is based on EfficientNet-B4 [4], pretrained on ImageNet, with the top three convolutional blocks unfrozen for fine-tuning. The original classifier head was replaced with a custom two-layer head consisting of a Dropout layer (p=0.4) followed by a fully connected linear layer mapping the 1792-dimensional feature vector to output classes (real and fake).

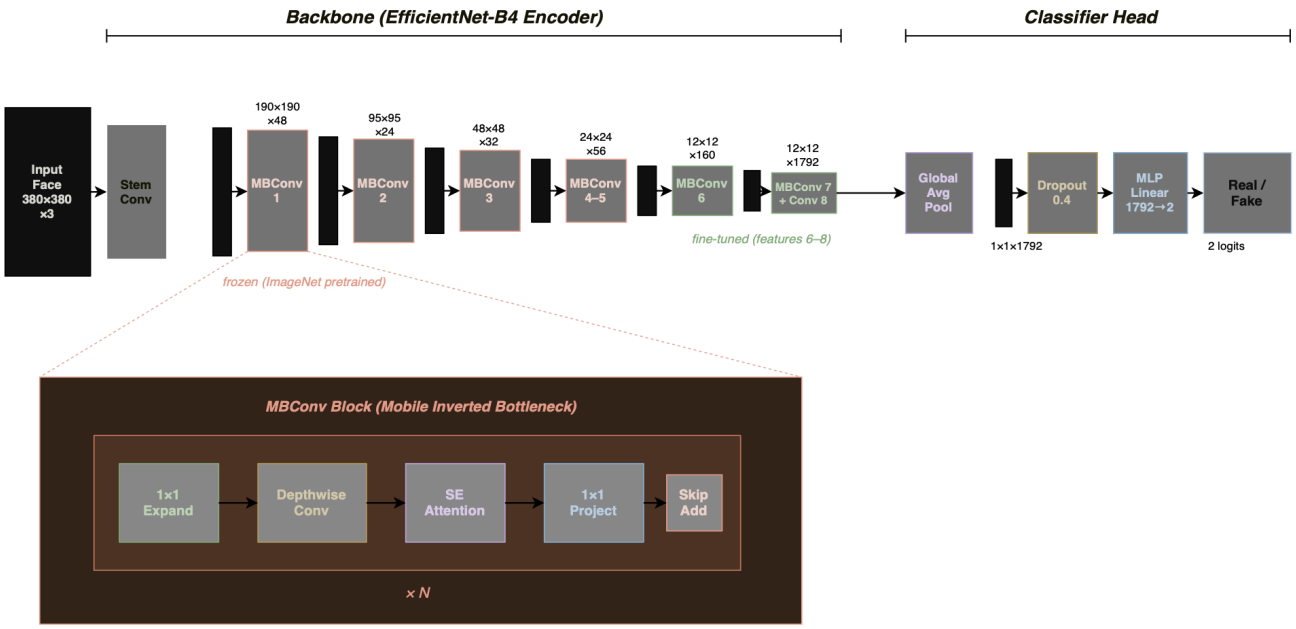


Figure 1 : End-to-end architecture of the EfficientNet-B4 deepfake detector model. Input face crops pass through the backbone and custom classification head to produce a real/fake prediction

The model architecture is illustrated in Figure 1. Input face crops of size 380x380 are first processed by a convolutional stem and then passed through seven sequential MBConv (mobile

inverted bottleneck) stages, which progressively reduce spatial resolution while increasing channel depth, ending in 1792-dimensional feature representation. The lower-level stages, which capture generic visual features such as edges and textures, are kept frozen with their ImageNet-pretrained weights, while the top three stages (features 6 through 8) are fine-tuned to adapt high-level representations to deepfake-specific patterns. Each MBConv block, expanded in the lower portion of the figure, applies a 1x1 expansion convolution, a depthwise convolution, a squeeze-and-excitation attention module, and a 1x1 projection convolution, with a residual connection where input and output dimensions match. The backbone output is reduced to a single feature vector by global average pooling, after which the custom classification head applies dropout (p = 0.4) followed by a fully connected layer mapping the 1792 features to two output classes (real/fake).

3.5 Training Setup

The model was trained using the Adam optimizer with a learning rate of 3 * 10^-5 and weight decay of 1 * 10^-4 for L2 regularization. Cross-entropy loss with label smoothing of 0.1 was used as the training objective to prevent overconfident predictions on training examples. Gradient clipping with a maximum norm of 1.0 was applied to stabilize training of the unfrozen convolutional blocks.

A reduceLROnPlateau scheduler was used to adaptively reduce the learning rate by a factor of 0.5 whenever validation loss failed to improve for two consecutive epochs. Training was conducted for a maximum of 40 epochs with early stopping applied when validation loss failed to improve for 5 consecutive epochs. The best-performing checkpoint, defined as the epoch with the lowest validation loss, was saved and reloaded before final test evaluation.

Training was performed on Apple Silicon (MPS) hardware using PyTorch with a batch size of 16 and 4 parallel data loading workers.

3.6 Evaluation Metrics

Model performance was evaluated using two primary metrics. Classification accuracy measures the percentage of correctly classified frames across the test set. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the model’s ability to discriminate between real and fake faces across all possible classification thresholds, providing a threshold-independent measure of discriminative ability. AUC-ROC is the standard evaluation

metric in published deepfake detection research and is less sensitive to class imbalance than accuracy alone.

Explainability was assessed qualitatively using Gradient-weighted Class Activation Mapping (Grad-CAM) [6], targeting the final convolutional block (features[8]) of EfficientNet-B4. Grad-CAM heatmaps were generated for a balanced sample of real and fake test images to verify that the model attends to facial regions rather than background or compression artifacts.

4. Results

4.1 Baseline Comparison

Initial experiments were done on a ResNet18 model which was trained on an unbalanced version of the Celeb-DF-v2 test split, which contains approximately 65% fake and 35% real videos. This class imbalance caused the model to collapse toward predicting the majority class, resulting in a test accuracy of 43.76%, which is below random chance for a balanced evaluation. After correcting the data pipeline to contain 50/50 class balance across all splits, the same ResNet18 model achieved 83.90% test accuracy, confirming that the imbalance was the main cause of the degraded performance

4.2 EfficientNet-B4 Results

Following the architecture upgrade to EfficientNet-B4 and the addition of principled regularization techniques, the model was trained and evaluated twice to confirm its reproducibility. The accuracy for Run 2 is shown in figure 2.

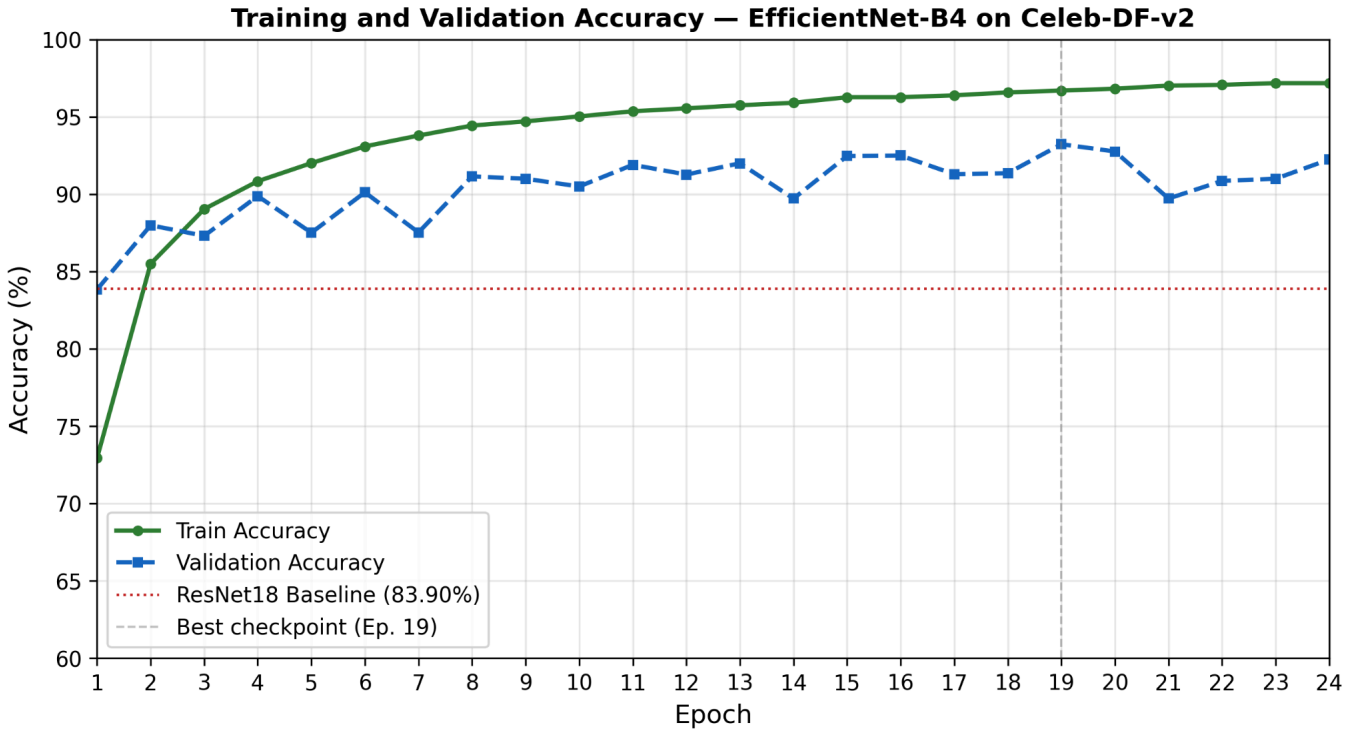


Figure 2 : EfficientNet-B4 results on Celeb-DF-v2 for Run 2 in terms of training and validation accuracy.

The model converged in 17 and 24 epochs respectively across the two runs, with early stopping triggered in both cases. The learning rate scheduler reduced the learning rate from 3*10^-5 to 3.75 * 10^-6 over the course of training, contributing to stable convergence.

4.3 Training Dynamics

Figure 2 shows the training and validation accuracy curves across epochs for Run 2. Validation accuracy climbed steadily from 83.83% at epoch 1 to a peak of 93.22% at epoch 19, while training accuracy reached 96.81%. The train-to-validation gap of approximately 3.6% points indicates controlled overfitting with effective regularization. The learning rate scheduler stepped down at epochs 7, 15, and 22 in response to plateaus in validation loss. The validation AUC-ROC for Run 2 is shown in figure 3.

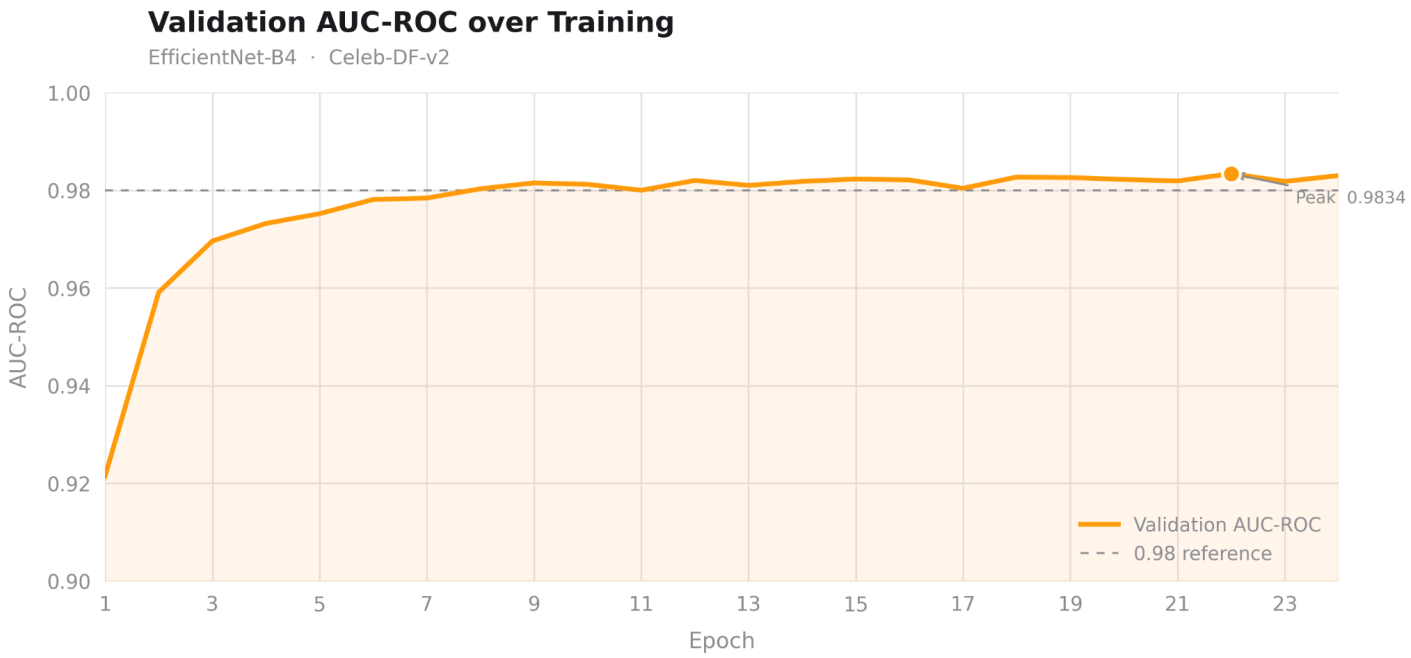


Figure 3 : EfficientNet-B4 model's AUC-ROC across 23 epochs.

4.4 Explainability Analysis

Visualizations were generated for a balanced sample of 8 test images (4 real, 4 fake) by targeting the final convolutional block (features[8]) of EfficientNet-B4. The heatmaps consistently highlighted facial regions including the eye area, nose bridge, and mouth corners as the primary regions of interest for both real and fake classifications.

For fake images, the model showed strong activation around facial boundary regions and skin texture areas, consistency with known deepfake artifact locations such as blending boundaries and unnatural skin smoothing. For real images, activations were more evenly distributed across the face. These patterns suggest the model has learned meaningful facial forensics features rather than background or compression artifacts.

4.5 Metrics

Figure 4 presents the normalized confusion matrix for the Celeb-DF-v2 model evaluated on the held-out Celeb-DF-v2 test set. The model correctly classifies 95% of fake faces and 85% of real faces. The errors are small in both directions : 5% of fakes are misclassified as real (false

negatives) and 15% of reals are misclassified as fake (false positives). The very low 5% false-negative rate is desirable for a deepfake detector, since it means the model rarely allows a manipulated face to pass undetected. The higher false-positive rate indicates the model occasionally flags authentic faces as fake.

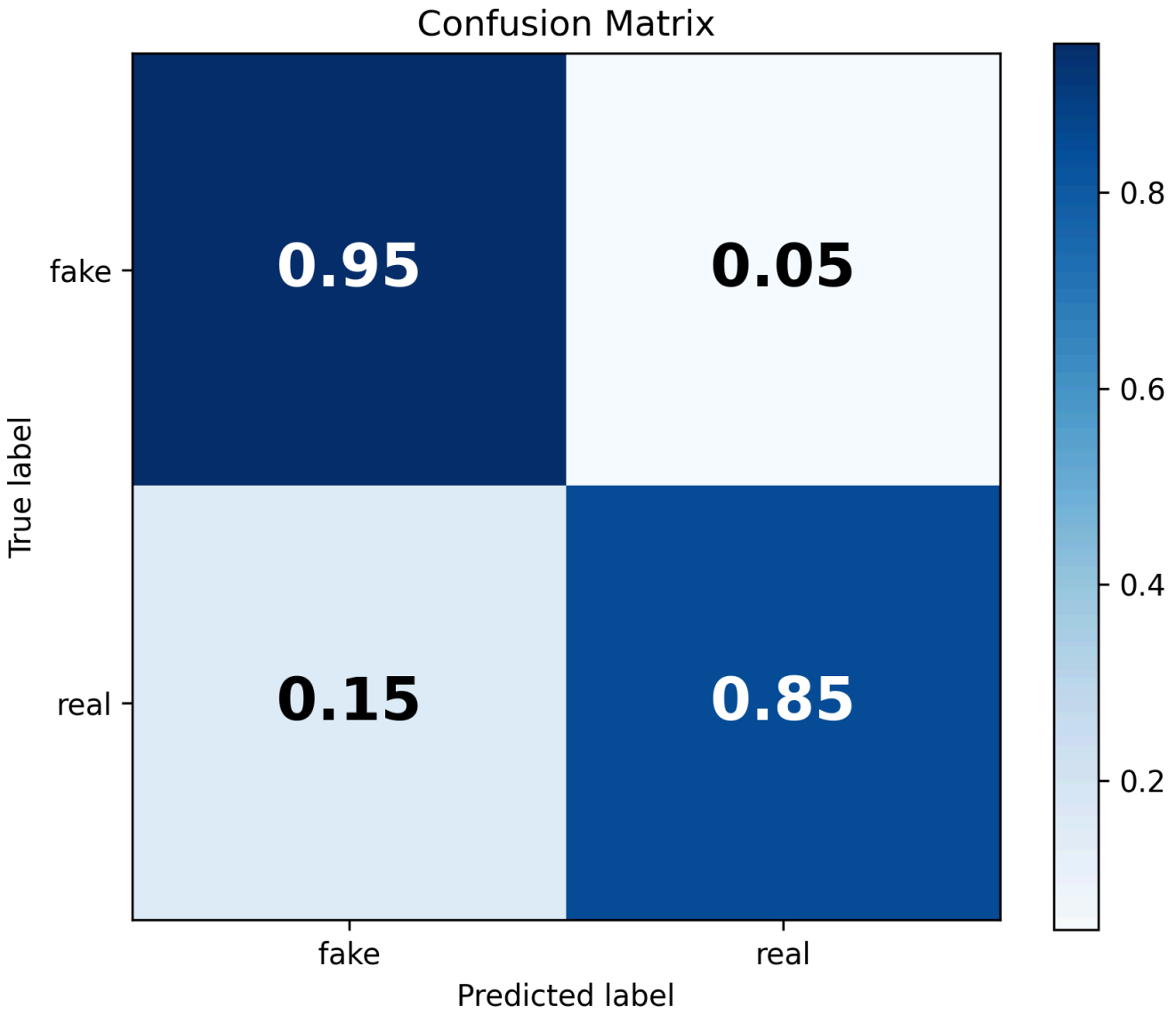
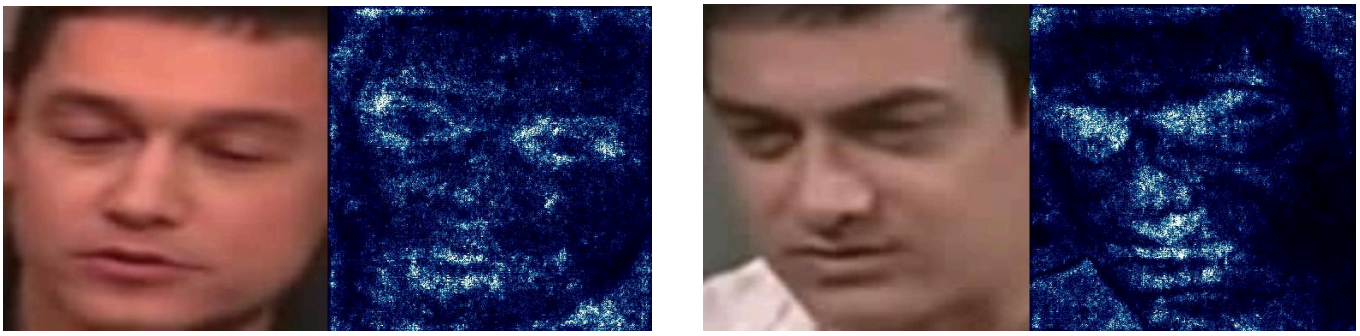


Figure 4 : Normalized confusion matrix for the Celeb-DF-v2 model on the Celeb-DF-v2 test set. Rows represent true labels, columns represent predicted labels; values are normalized per true class.

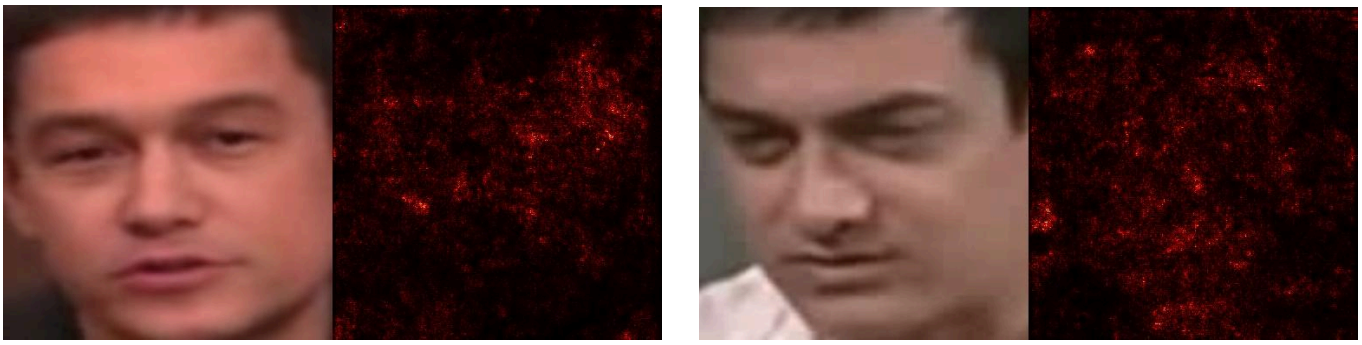
4.6 Cross-Dataset Analysis

To evaluate generalization, the Celeb-DF-v2 model was first applied without retraining to FaceForensics++, where performance dropped sharply to 56.78% accuracy and 0.6802 AUC-ROC, revealing a substantial domain gap due to the difference in generation methods of deepfakes across datasets. A second model was then trained on a combined dataset drawn from both Celeb-DF-v2 and FaceForensics++. This generalist model achieved 88.87% accuracy and 0.9514 AUC-ROC on a combined test set, recovering cross-dataset AUC from 0.6802 to 0.9514. This demonstrates that exposure to multiple generation methods during training allows for the model to learn transferable manipulation features rather than signatures specific to the dataset.

These Grad-CAM [6] visualizations highlight the broad facial regions most responsible for the model's classification, computed from the final convolutional block of EfficientNet-B4. For the real sample (predicted real, 0.95 confidence), the activation concentrates on the central facial features, the nose and inner-eye region, which is where natural skin texture and consistent lighting are most informative. For the fake sample (predicted fake, 0.94 confidence), the strongest activation shifts to the region near the left eye, with secondary activation along the mouth and jaw. This is consistent with known deepfake artifact locations, as face-swap synthesis methods frequently leave subtle inconsistencies around the eyes and along the blending boundary where the generated face meets the original. The coarse region level nature of these heatmaps reflects that Grad-CAM operates on high-level feature maps, and in this case shows that the model is able to use proper activation features located on the face.



These Integrated Gradients maps [8] attribute the model's prediction to individual input pixels by averaging gradients along a straight-line path from a blank baseline image to the actual input. This path-integral formulation satisfies theoretical properties, most notably completeness, meaning the attributions sum to the difference in model output between the baseline and the input, and produces substantially cleaner, less noisy maps than single-point gradient methods. The results trace the facial structure with precision; the eyes, nose, mouth, and jaw contours are clearly shown in the attribution, while the background remains dark and unattributed. For both the fake and real samples, the brightest attribution concentrates along these facial features and their edges, proving the model's decision is driven by fine-grained detail in the manipulated regions rather than global or background cues. The sharpness with which the facial geometry emerges, as opposed to the diffuse scattering similar to saliency maps, confirms that the model responds to coherent, structured facial information consistent with the locations where synthesis artifacts arise.



These saliency maps [9] show the magnitude of the prediction's gradient with respect to each input pixel, revealing fine-grained, pixel-level sensitivity. They appear diffused and scattered by nature; this is expected as pixel-level gradients are inherently noisier than the region-level. The key observation is spatial: the activations cluster across the facial area rather than the uniform background, confirming that the model's decision is driven by information on the face itself rather than random background areas. The concentration of bright points in textured facial regions, rather than smooth or empty areas, indicates the model is responsive to local texture detail, which is where synthesis artifacts and compression inconsistencies tend to appear.

Taken together, these three independent techniques provide triangulated evidence that the model bases its decisions on genuine facial features. Grad-CAM, Integrated Gradients, and saliency maps each operate on different principles and at different resolutions, yet all three converge on the same conclusion: the model concentrates on the eyes, nose, mouth, facial boundaries, and other features while disregarding the background. This agreement across methods is significant because the techniques could easily have disagreed if the model was relying on arbitrary correlations. The Grad-CAM analysis was performed on the Celeb-DF-v2 model, while the Saliency maps and Integrated Gradients analysis was performed on the combined Celeb-DF-v2 and FaceForensics++ model. The fact that the generalist model continues to attend to meaningful facial regions across a multi-dataset input further supports the conclusion that its performance stems from learning general distinguishing features rather than dataset specific attributes.

4.7 FaceForensics++ Generalization

To evaluate cross-dataset generalization, the specialist model (trained on Celeb-DF-v2) was applied without retraining to a balanced set of 10,704 real and 10,082 fake face crops extracted from the FaceForensics++ dataset (c23 compression), using the deepfakes manipulation method. Then, the model was trained on both datasets and evaluated again (Generalist model).

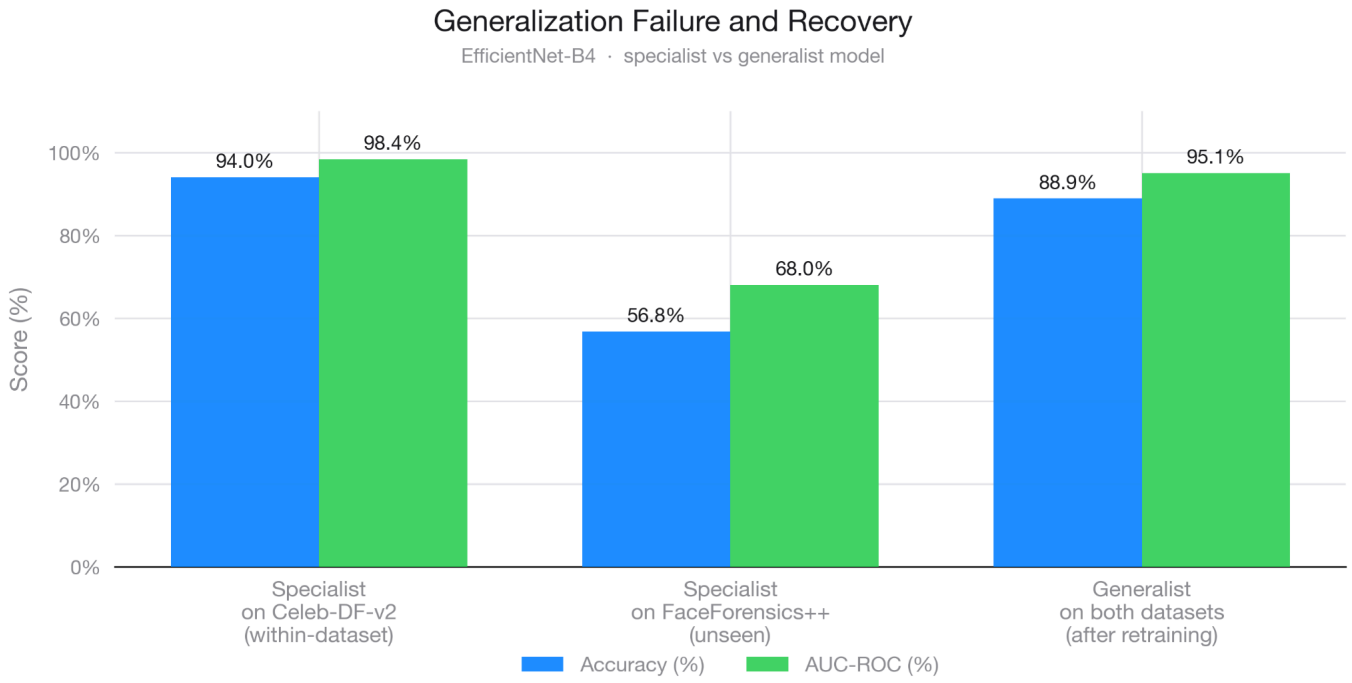


Figure 5 : Cross-dataset generalization results: specialist model within dataset, specialist model on FaceForensics++, and generalist model after combined training.

The model's performance dropped significantly on FaceForensics++, revealing a substantial domain gap between the two datasets. This result is consistent with findings in published deepfake detection literature, where models trained on one dataset commonly show degraded performance on unseen datasets due to differences in deepfake generation pipelines, compression artifacts, and facial identity distribution [3][7].

Celeb-DF-v2 uses a high quality face-swap synthesis pipeline optimized for visual realism, while FaceForensics++ Deepfakes uses an autoencoder-based face-swap method that produces different artifact patterns. The model appears to have learned Celeb-DF-v2 specific artifact signatures rather than generalizable deepfake features, limiting its cross-dataset usage.

Figure 5 summarizes model performance across three evaluation settings, tracing the full arc of the generalization experiment. The leftmost group shows the specialist model's strong within-dataset performance on Celeb-DF-v2. The middle group captures the main challenge: applied to FaceForensics++, which uses a different generation method, the same model's accuracy and AUC fall sharply, indicating a reliance on dataset specific cues rather than transferable artifacts. The rightmost group shows the result of the generalist model, as it was trained on both datasets. It recovered most of the lost performance and approached the original within-dataset scores while operating across both generation methods. Taken together, the progression demonstrates that the observed domain gap is not an inherent limitation, but a gap that can be narrowed with multi-source training.

This finding motivates future work on domain-agnostic training strategies, including domain adaptation techniques, and frequency-domain feature extraction methods that target artifacts common across generation methods.

5. Discussion

5.1 Interpretation of Results

The EfficientNet-B4 model achieved strong within-dataset performance on Celeb-DF-v2, with an average test accuracy of 94.04% and AUC-ROC of 0.9844 across two independent runs. The consistency between runs demonstrates that the training pipeline is stable and reproducible, an important property for scientific credibility.

The high AUC-ROC of 0.9844 indicates that the model confidently separates real from fake faces across all classification thresholds, not just at a single decision boundary. This is particularly meaningful for real-world deployment scenarios where the operating threshold may need to be adjusted based on the cost of false positives versus false negatives.

The progressive improvement from the broken ResNet18 baseline (43.76%) to the fixed ResNet18 (83.90%) to the final EfficientNet-B4 model (94.04%) demonstrates the compounding efficiency of methodological improvements. Correcting the class imbalance in the data pipeline alone accounted for a 40.14% point improvement, highlighting that data quality and pipeline correctness are at least as important as model architecture choice.

5.2 Limitations

The most significant limitation identified in this study is the domain gap between Celeb-DF-v2 and FaceForensics++. The model's AUC-ROC dropped from 0.9844 to 0.6802 when evaluated on FaceForensics++ without retraining, indicating that the learned features are partly specific to Celeb-DF-v2's synthesis pipeline rather than fully generalizable deepfake artifacts.

A second limitation is the use of frame-level evaluation rather than video-level aggregation. Real-world deepfake detection systems typically aggregate predictions across multiple frames of a video to produce a single video-level decision. Frame-level accuracy may not fully reflect video-level detection performance.

Finally, the model was evaluated only on the Deepfakes manipulation method within FaceForensics++. Other manipulation methods including Face2Face, FaceSwap, and NeuralTextures were not evaluated, limiting the scope of the cross-dataset generalization analysis.

5.3 Comparison to Published Baselines

The within-dataset results of 94.04% accuracy and 0.9844 AUC-ROC on Celeb-DF-v2 are competitive with published methods. For instance, DeepfakeBench reported EfficientNet-B4 as one of the strongest backbone architectures across multiple datasets [7]. The cross-dataset generalization gap observed in this study is consistent with results reported across the broader deepfake detection literature, where cross-dataset AUC drops of 20-30% points are commonly observed [3].

6. Future Work

Several directions for future research are motivated by the findings of this study. Multi-dataset training represents the most immediate next step. While multi-dataset training was implemented here and substantially closed the cross-dataset gap (Section 4.6), validating generalization on entirely unseen generation methods remains the most important next step. The combined model was trained and tested on Celeb-DF-v2 and FaceForensics++; a stronger test of generalization would hold out a third benchmark entirely. Future work will evaluate the combined model on a dataset not seen during training, such as DFDC or DeeperForensics-1.0, to confirm whether the learned features extend beyond the two datasets used here.

Video-level aggregation is a second important improvement. The current system makes independent predictions on individual frames. Aggregating predictions across all frames of a video would produce a single video-level decision more suitable for real-world content moderation pipelines.

Expanding cross-dataset evaluation to cover all four FaceForensics++ manipulation methods (Deepfakes, Face2Face, FaceSwap, and NeuralTextures) would provide a more complete picture of the model's generalization ability. Each manipulation method produces distinct artifact patterns, and evaluation across all four would reveal which artifact types the model detects most and least reliably.

Frequency-domain feature extraction is a promising architectural direction for improving generalization. Several published methods have shown that deepfake artifacts are more consistent across datasets in the frequency domain than in the pixel domain. Incorporating frequency-domain features alongside spatial CNN features may help bridge the domain gap observed in this study.

Finally, expanding the training dataset beyond the 1000 videos used in this study to the full Celeb-DF-v2 training set would likely improve both within dataset and cross-dataset performance by exposing the model to a greater diversity of facial identities and lighting conditions.

7. References

[1] H. Farid, “Image Forgery Detection, “ IEEE Signal Processing Magazine, vol. 26, no.2, pp. 16-25, 2009.

https://doi.org/10.1109/MSP.2008.931079

[2] A. Rossler, D Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11.
https://arxiv.org/abs/1901.08971

[3] B. Zi, M. Chang, J. Chen, X. Ma, and Y. Jiang, “WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection,” in Proceedings of the ACM International Conference on MultiMedia, 2020, pp. 2383-2390.
https://doi.org/10.1145/3394171.3413769

[4] M. Tan and Q. Le, “Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proceedings of the International Conference on Machine Learning (ICML), 2019, pp. 6105-6114.
https://arxiv.org/abs/1905.11946

[5] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3207-3216.
https://arxiv.org/abs/1909.12962

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.
https://arxiv.org/abs/1610.02391

[7] J. Yan, Y. Luo, C. Zhao, T. Ju, Z. Huang, Y. He, H. Fan, B. Li, and S. Lyu, “DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection,” in Advances in Neural Information processing Systems (NeurIPS), 2023.
https://arxiv.org/abs/2307.01426

[8] M. Sundararajan, A. Taly and Q. Yan, “Axiomatic Attribution for Deep Networks,” in Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 3319-3328
https://arxiv.org/abs/1703.01365

[9] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” in Workshop at the International Conference on Learning Representations (ICLR), 2014.
https://arxiv.org/abs/1312.6034