

Deepfake Image Detection Using EfficientNetB4

Divyanshi Rathor¹, Prof. Kamlesh Patidar²

M Tech Researcher¹, HOD²

Department of Computer Science and Engineering^{1,2}

Jawaharlal Institute of Technology, Borawan, Khargone, Madhya Pradesh, India^{1,2}

Abstract—

The rapid development of generative artificial intelligence has made synthetic facial images increasingly difficult to distinguish from authentic photographs. This paper presents a hybrid deepfake image detection framework that combines EfficientNetB4 with a Self-Attention module. EfficientNetB4 is used as the principal feature extractor because its compound-scaled architecture provides rich spatial representations with comparatively efficient computation. The Self-Attention block is added to model long-range relationships among facial regions and to emphasize manipulation-sensitive features such as texture discontinuities, inconsistent illumination, blending boundaries, and abnormal facial details. The framework uses preprocessing, augmentation, transfer learning, feature fusion, batch normalization, dropout, and binary classification. Experiments are conducted on the OpenForensics dataset, which contains diverse real and manipulated facial images captured under unconstrained conditions. According to the thesis experiments, the fine-tuned model achieved 95.09% validation accuracy and the independent test evaluation produced 84.33% accuracy, 92.96% precision, 74.04% recall, 82.33% F1-score, 0.9253 ROC-AUC, and 0.9367 PR-AUC. The results indicate that attention-guided feature refinement can complement convolutional representations for deepfake detection, although the difference between validation and test performance also highlights the need for cross-dataset and real-world evaluation.

Keywords—deepfake detection, EfficientNetB4, self-attention, OpenForensics, transfer learning, facial forgery, digital forensics, computer vision.

I. INTRODUCTION

Digital images are now central to news, communication, identity verification, entertainment, and online social interaction. At the same time, generative models can synthesize or manipulate faces with a level of realism that challenges manual inspection. Deepfake images may be used for benign content creation, but they can also support misinformation, impersonation, fraud,

reputational damage, and the production of fabricated evidence. Automated image authentication has therefore become an important problem in computer vision and multimedia forensics.

Conventional forensic approaches often rely on manually designed cues, including compression statistics, noise inconsistencies, color artifacts, and boundary irregularities. Such cues can be useful, but modern generation pipelines may remove or alter them. Convolutional neural networks (CNNs) learn hierarchical representations directly from data and have consequently become common in image forgery detection. However, convolution is inherently local: subtle evidence may be distributed across several facial regions, while background context and illumination can influence the interpretation of a local texture.

This work investigates a hybrid architecture based on EfficientNetB4 and Self-Attention. EfficientNetB4 supplies multi-level spatial features through MBConv blocks and compound scaling, whereas Self-Attention enables interactions among spatial locations in the extracted feature map. The central hypothesis is that combining local texture extraction with global contextual reasoning will improve the discrimination of authentic and manipulated faces.

The main contributions are: (1) a transfer-learning-based EfficientNetB4 detector adapted for binary real/fake classification; (2) an attention module for contextual feature refinement; (3) an end-to-end preprocessing and regularization pipeline; and (4) an evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC on OpenForensics.

II. RELATED WORK

Deepfake detection research has developed from handcrafted image-forensic features to deep CNNs, frequency-domain methods, transformers, and multimodal systems. Early benchmarks such as FaceForensics, Celeb-DF, and the DeepFake Detection Challenge dataset enabled systematic evaluation, but many reported results were obtained under controlled or intra-dataset conditions. Performance can decline when images are compressed, generated by unseen methods, or collected from real social-media environments.

CNN-based detectors are effective because their successive layers learn edges, textures, facial parts, and higher-level artifacts. Transfer learning is particularly useful because models pretrained on large image corpora already contain general visual features. EfficientNet, introduced by Tan and Le, improves the accuracy–efficiency trade-off by scaling depth, width, and input resolution together. EfficientNetB4 processes 380×380 inputs and is larger than the baseline B0 model, making it suitable when fine facial details are important.

Attention mechanisms provide a complementary strategy. In the scaled dot-product formulation, feature vectors are projected to queries, keys, and values, and the attention output is computed as $A(Q,K,V) = \text{softmax}(QK^T/\sqrt{d})V$. This operation permits one region to use information from other regions. For

A. Dataset and preprocessing. The thesis uses the OpenForensics dataset, a multiface forgery benchmark containing real and manipulated facial images with varied lighting, occlusions, viewpoints, backgrounds, and face-wise annotations. The reported class distribution is approximately balanced: 95,201 real images and 94,134 fake images. The stated split contains 140,002 training images, 39,428 validation images, and 10,905 test images. Images are converted to RGB, resized to the model input size, normalized, and augmented using operations such as horizontal flipping, rotation, zoom, and brightness variation.

B. EfficientNetB4 backbone. The input image is forwarded through the ImageNet-pretrained EfficientNetB4 backbone. Its MBConv blocks use expansion, depthwise convolution, squeeze-and-excitation recalibration, projection, and residual connections where applicable. Lower layers learn generic edges and textures, while deeper layers encode facial structure, skin patterns, illumination, and possible manipulation traces. During initial training, the backbone is largely frozen so that the newly added classification layers can adapt without destroying pretrained representations.

C. Self-Attention feature refinement. Let $F \in \mathbb{R}^{(N \times C)}$ denote the flattened feature map, where N is the number of spatial locations and C is the channel dimension. Learnable projections produce $Q = FW_Q$, $K = FW_K$, and $V = FW_V$. The attention-enhanced representation is $H = \text{softmax}(QK^T/\sqrt{d})V$. The module is intended to assign greater weight to informative relationships among facial regions and to reduce the influence of irrelevant background content.

D. Feature fusion and classification. The attention representation is combined with global pooling features

facial forensics, such contextual interaction can help relate a local mouth or eye artifact to skin texture, lighting, and surrounding facial structure.

Recent literature also emphasizes generalization. Standardized frameworks such as DeepfakeBench have been proposed to improve reproducibility, while real-world benchmarks show that laboratory accuracy does not necessarily translate to deployment. These observations motivate the present study and inform its limitations: the reported model should be viewed as a dataset-based detector rather than a universal authenticity guarantee.

III. PROPOSED METHODOLOGY

from the EfficientNetB4 output. The fused vector is passed through batch normalization and dropout, followed by dense layers and a sigmoid output for binary classification. A probability above the selected decision threshold is interpreted as one class and a probability below it as the other. The binary cross-entropy objective is $L = -[y \log(p) + (1-y)\log(1-p)]$, where y is the ground-truth label and p is the predicted fake probability.

E. Training strategy. The model is optimized with Adam. Model checkpoints are selected using validation AUC, and early stopping or restoration of the best weights is used to limit overfitting. In the fine-tuning stage, upper EfficientNetB4 layers are unfrozen and trained with a small learning rate so that high-level features can adapt to deepfake-specific patterns while generic low-level features remain stable.

IV. EXPERIMENTAL SETUP AND METRICS

The evaluation uses the independent test partition after training and fine-tuning. Accuracy measures the fraction of correctly classified samples. Precision measures the proportion of predicted fake samples that are actually fake, while recall measures the proportion of fake samples detected by the model. The F1-score is the harmonic mean of precision and recall. ROC-AUC summarizes discrimination across thresholds, and PR-AUC is especially informative when precision–recall behavior is important for forensic screening.

The reported fine-tuning trace shows improvement from epoch 7 through epoch 16. Training accuracy increased from 87.14% to 97.19%, validation accuracy increased from 90.87% to 95.09%, and validation AUC reached 0.9892. The best checkpoint was saved for final testing. These validation values should not be confused with the independent test values, because validation performance

is used during model selection whereas test performance is intended to estimate generalization.

V. RESULTS AND DISCUSSION

Table I presents the principal test results transcribed from the thesis. The model obtains high precision and specificity, indicating that its positive predictions are comparatively reliable and that false alarms for the real class are limited. The recall value is lower than precision, which means that a non-trivial fraction of manipulated images remains undetected at the selected threshold. In a forensic workflow, this trade-off should be considered explicitly; a screening system may choose a threshold that favors recall over raw accuracy.

Table I: Principal test performance.

Metric	Value
Accuracy	84.33%
Precision	92.96%
Recall / Sensitivity	74.04%
Specificity	94.25%
F1-score	82.33%
ROC-AUC	0.9253
PR-AUC	0.9367

The per-class report in the thesis further indicates that the fake class achieved 78.65% precision, 94.25% recall, and 85.8% F1-score, whereas the real class achieved 92.69% precision, 74.04% recall, and 82.4% F1-score. This asymmetry is consistent with a detector that identifies most manipulated examples but occasionally labels authentic images as fake or misses difficult real/fake cases depending on the operating threshold.

Table II: Reported dataset split.

Split	Images	Purpose
Training	140,002	Parameter learning
Validation	39,428	Model selection and tuning
Testing	10,905	Independent evaluation
Total	190,335	Reported split total

Table II summarizes the reported data split. Because the source thesis also describes the dataset as containing approximately 190,000 images, the counts should be checked against the final dataset manifest before publication. This paper preserves the reported experimental values rather than silently changing them.

VI. LIMITATIONS AND FUTURE WORK

The validation–test gap is an important limitation. A validation accuracy of 95.09% and a test accuracy of 84.33% suggest that the model may be sensitive to distribution differences, image sources, compression, or

the particular split used in experimentation. In addition, the current system performs image-level binary classification and does not by itself provide a calibrated explanation, provenance record, or legally sufficient authenticity decision. Attention weights should not automatically be interpreted as definitive manipulation masks.

Future work should evaluate cross-dataset generalization, unseen generators, social-media compression, adversarial perturbations, and low-resolution inputs. Other useful directions include explicit forgery-region segmentation, frequency-domain branches, uncertainty calibration, threshold selection for forensic risk, model compression for edge deployment, and multimodal analysis of image, video, and audio evidence. Repeated experiments with confidence intervals and subject-disjoint splits would also improve reproducibility.

VII. CONCLUSION

This paper presented a hybrid EfficientNetB4 with Self-Attention framework for deepfake facial-image detection. The convolutional backbone captures hierarchical local evidence, while attention-based refinement models relationships among spatial regions. On the OpenForensics evaluation reported in the source thesis, the method achieved 84.33% test accuracy, 92.96% precision, 74.04% recall, 82.33% F1-score, 0.9253 ROC-AUC, and 0.9367 R-AUC. The results support the usefulness of combining efficient CNN features with contextual attention, while the validation–test difference emphasizes that robust deployment requires broader, cross-domain evaluation.

REFERENCES

- [1] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2019.
- [2] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
- [3] B. Dolhansky et al., “The DeepFake Detection Challenge (DFDC) dataset,” arXiv:2006.07397, 2020.
- [4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A new dataset for deepfake forensics,” in Proc. CVPR, 2020.
- [5] A. Rossler et al., “FaceForensics++: Learning to detect manipulated facial images,” in Proc. ICCV, 2019.

- [6] Z. Yan et al., “DeepfakeBench: A comprehensive benchmark of deepfake detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2023.
- [7] OpenForensics dataset and experimental configuration, as described in the source dissertation, “Deepfake Image Detection Using EfficientNetB4,” Jawaharlal Institute of Technology, Borawan, 2023–2026.
- [8] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, 2016.