Hierarchical Multi-Class Deepfake Detection Using EfficientNetB4 for Source Attribution

Mukesh Kumar[1], *Raviraj Choudhary*[1], and *Gaurav Meena*[1,*]
[1]Department of Computer Science, Central University of Rajasthan, Ajmer-305817, Rajasthan, India

Abstract. The rising of deepfake media (images, videos and audio etc.) has become a major issue in the digital era. Deepfake generation techniques such as Generative Adversarial Networks and Diffusion Models, shows major challenge to individual's privacy, democracy and national security. While the previously deepfake detection techniques deals with binary classification such as real and fake, it remains explored to detect deepfakes with multiple generation techniques concurrently. Despite the recent improvements, no existing work has addressed the problem of mapping all three major generation families (GAN-based, diffusion-based, and face manipulation techniques) into a single hierarchical pipeline, which is a crucial research gap. In order to overcome these limitations this paper we present a multi- class deepfake detection framework capable of distinguishing between real images and nine distinct synthesis methods: DALL-E, Face2Face, FaceSwap, StyleGAN, NeuralTextures, Stable Diffusion, DeepFaceLab, FaceShifter, Midjourney. We used transfer learning with EfficientNetB4 as backbone and proposed hierarchy of three levels and achieves 98.89 % and 98.52% accuracy for binary and multiclass attribution, respectively. The framework further achieves 99.31% precision for binary classification and a macro-average AUC of 0.9995 for multiclass attribution. Our confusion matrix analysis shows that the model is able to differentiate between the various generation strategies and that it is particularly successful in finding diffusion-based approaches and traditional face manipulation techniques. The hierarchical structure offers actionable attribution for forensic use, enabling applications in content moderation, forensic investigations, and media verification.

Keywords: Deepfake, Generative Adversarial Networks, Diffusion Models, Deep Learning, Transfer Learning

1. Introduction:

Forgery of digital images is one of the critical problem in this digital world. The fast improvement in generative AI models has promoted the generation of synthetic media, enabling both creative purposes and for misuse. But the recent advancements in diffusion models (DALL-E [1], Stable Diffusion [2], Midjourney) and advanced face manipulation techniques (Face2Face, DeepFakeLab) have made it increasingly difficult to differentiate authentic images from fake images. Current Deepfake detection studies are primarily concern with a binary question of whether an image is real or fake. It has major shortcomings, but, as it is not an effective solution for both forensic analysis and platform dependent, where knowledge of the generation method is necessary.

While deepfake detection has made substantial advancements, a key research gap remains: no prior work considers the identification of three types of synthetic image generation methods (GAN-based, diffusion-based, and other traditional face manipulation approaches) in a unified framework. Moreover, existing methods that aim for multiclass attribution are restricted to a single generation model family (e.g., only diffusion models or only GANs) with at most 88% multiclass accuracy [7]. In order to recognising deepfakes, we propose a hierarchical classification model that work on nine different synthesis processes and one another class i.e., real. The transfer learning choice of our approach is paired with EfficientNetB4, a well-known convolutional architecture, which is efficient and accurate due to its compound scaling feature.

Contributions of this study can be summarised as follows:

- Hierarchical three level framework for multi-class deepfake detection across diverse generation methods including both GAN-based and diffusion-based approaches.

*Corresponding author: gaurav.meena@curaj.ac.in

- Demonstration that transfer-learning with EfficientNetB4 achieves over 99% accuracy in attributing images to specific synthesis methods.
- An in-depth confusion matrix analysis revealed similarities among different generations methods.

2. Related works

Deepfake detection is advanced from the handmade artifacts analysis to deep learning. Early approaches contain low-level visual artifacts in GAN generated images such as unnatural textures, color inconsistencies, and pixel-level anomalies due to adversarial training. Frequency domain analysis was an important technique in which GAN images exhibit abnormal high frequency patterns relative to the natural decay of real images [3]. Another important field was the facial and biological inconsistencies. Imperfect face-swapping often caused messed-up facial landmarks (eyes, nose, mouth etc). Other biological cues like abnormal eye blinking and pupil behaviour disturbances, corneal reflections, were also useful.

More recently, detection has turned to deep learning classifiers trained on large datasets such as FaceForensics++ [4], where they employ CNNs to learn discriminative features end-to-end. They can improve on known manipulations, but are unable to generalise across datasets. Modern approaches increasingly emphasize hybrid and multimodal methods that provide spatial, temporal cues. This is driven by diffusion models which produce fewer perceptible artifacts than GANs and require robust, adaptive and multimodal detection strategies.

The majority of deepfake detection studies view the task as a binary classification process, requiring separation between real and fake content. This formulation simplifies learning since forgeries are more general than artifacts specific to a generator. Most commonly, CNN-based models extract spatial features from images or frames [5]. Binary detectors are still dominant because of their ease of use and their ability to perform on known data but have several major limitations; poor generalization in the results among generation methods (e.g., GAN-based trained models failing in diffusion-based fakes).

Rossler et al. (2019) [4] address the growing concern of identifying realistic facial manipulations with a standard of large-scale benchmark for facial counterfeiting detection. They stimulate over 1.8 million real-world video frames which are altered with four manipulation techniques: Deepfakes, FaceSwap, face2face and neural textures, and examined under realistic compression by both human observer and automated detector. There results shows that deep learning, mainly XceptionNet with face-region cropping, is far superior than human and hand-made approaches. In the perspective of manipulation-wise precision, highest detection precision is 96.36 for DeepFakes, 90.29 for FaceSwap; due to artifacts of identity replacement, 86.86 for Face2Face, which modifies only facial expressions; Neural textures had 80.67 precision value, which uses GANs to generate images that can be difficult to detect, and pristine images scores the lowest precision value i.e., 52.40. These results indicate that detection is of a high degree, dependent on the type of manipulation and GAN-based expression manipulations being the most challenging.

Gragnaniello et al. (2021) [6] tested the assumption that the classification of synthetic image from real one is easy in binary classification process. They train the detectors on the set of images which are generated by ProGAN and test on eight unseen GAN architectures StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, RelGAN, and GauGAN. Several important factors requiring robustness are identified in this study, but only in case of Gan generated images, and improving the architectural design from 90-95% from previous studies. Despite these gains, they conclude that real-time detection remains far from addressing as GAN evolution grows rapidly, forensic evidence disappears from post-processing and drastic drop under down-sampling. One of the limitation of this work is that it entirely focused on GAN-based generation methods since other new and widely used generative techniques, such as face manipulation techniques and diffusion-based models, are not included in the analysis which does not extend its applicability to modern generative scenarios.

Sha et al. (2023) [7] introduced DE-FAKE, a methodical framework for identifying and classifying tampered images produced from text-to-image models. They test Stable Diffusion, Latent Diffusion, GLIDE, and DALLE 2 on the MSCOCO and Flickr30k datasets by using either image only CNN detectors or by using a hybrid approach by embedding images with prompts and prompts via CLIP, with BLIP-generated captions when prompts are unavailable. Empirical outcomes shows that the hybrid classifier significantly exceeds image-only detectors in measuring up to 93% detection accuracy across model and that frequency domain analysis indicate common artifacts for diffusion models. For attribution, a multiclass-classifier categories the source generation models of images at up to 88% accuracy, indicating the presence of unique visual fingerprints for each model. But, since the all-generative models are Diffusion-based in this dataset, the efficiency of the framework has been experimentally validated merely for diffusion models and thus it could be generalised to GAN-based or other image manipulation techniques as an open limitation.

3. Methodology

3.1 Dataset

The labelled deepfake image collection [8] is a curated dataset for deepfake detection and attribution, which comprises 18,124 images, 5,890 of which are pristine and 12,234 are fake such as: 2000 images from DALL-E, 1000 images from Face2Face, 1000 images from FaceShifter [10], 1600 images from FaceSwap, 930 images from Midjourney, 1000 images from NeuralTextures [11], 1606 images from DeepFaceLab [12], 2098 images from Stable Diffusion, and 1000 images from StyleGAN [13]. This benchmark is then used to assess the performance of a detector, classifying between real and fake images and if deteted image is fake then what kind of fake? Such methods are accomplished for generating realistic images, making it exceptionally hard to identify original from tampered images.

3.2 Proposed Model

3.2.1. Backbone Architecture

We use EfficientNet-B4 [9] as a backbone architecture for feature extraction from input images, is published by Tan and Mingxing in 2019. As shown in figure 1, the architecture of this neural network is based on the compound scaling principle, that scales the depth, width and resolution of networks systematically using a compound coefficient. Instead of choosing one or two dimensions at random, EfficientNet uses a procedure of compound scaling that scales all three dimensions with a fixed proportion to achieve a much better performance with usually a smaller number of parameters and FLOPS (Floating Point Operations Per Second).

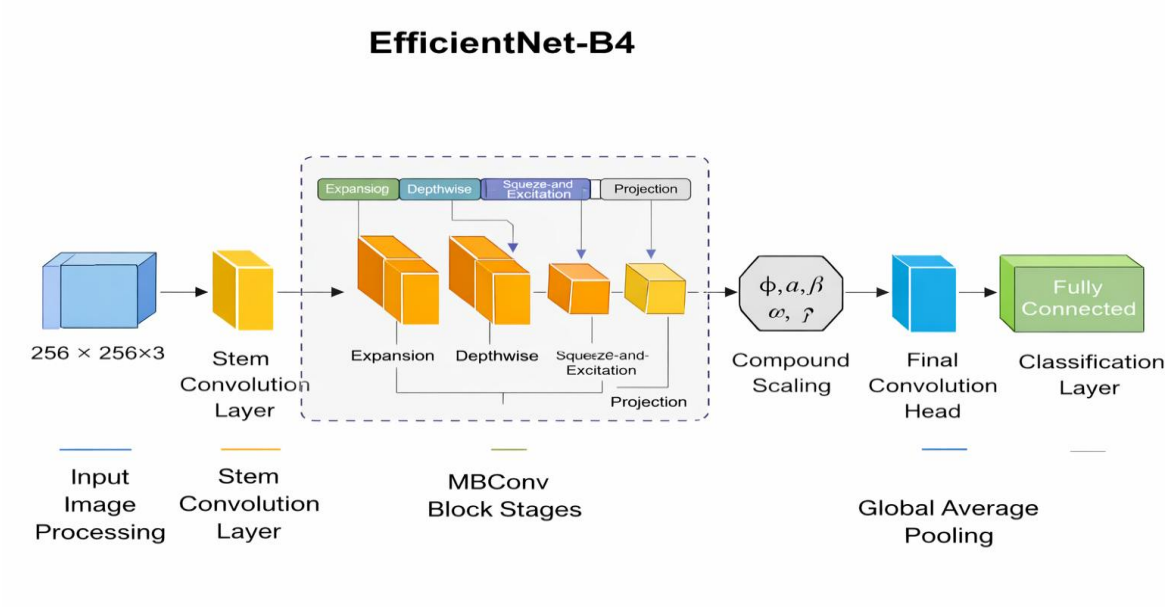


Fig. 1. EfficientNetB4

Compound Scaling Method: A compound coefficient φ used by compound scaling approach which uniformly scales the dimensions of neural network according to:

- Depth: d = α^φ
- Width: w = β^φ
- Resolution: r = γ^φ

where α, β, and γ are constants obtained from grid search, and $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$. Total FLOPS increase by approximately 2^φ for each scaling step. For EfficientNet-B4, φ = 3, resulting in a network with 17 million parameters and an input resolution of 256*256 pixels we adapt it for our detection framework.

3.2.2. Proposed Three-Level Architecture Hierarchy

Our proposed deepfake detection framework employs a hierarchy of three level architecture, with the intension to systematically identity and classify altered facial images. This approach uses Efficientnet-B4 as the main neural

network for feature extraction, which uses specialized detection modules at different layers to address multiple features of different deepfake generation techniques. This framework then processes the input images through a sequence of increasingly specialized steps in analysis until it finally determines whether it is real image or fake and if fake then which manipulation technique it uses.
Let

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

Equation (1) denote the training dataset, where $x_i \in R^{H\times W\times 3}$ represents an input face image resized to 256×256, and y_i denotes its ground-truth label. Each fake sample is additionally associated with a coarse category label $c_i \in \{GAN, DM, OTH\}$ and a fine-grained generation method label m_i.

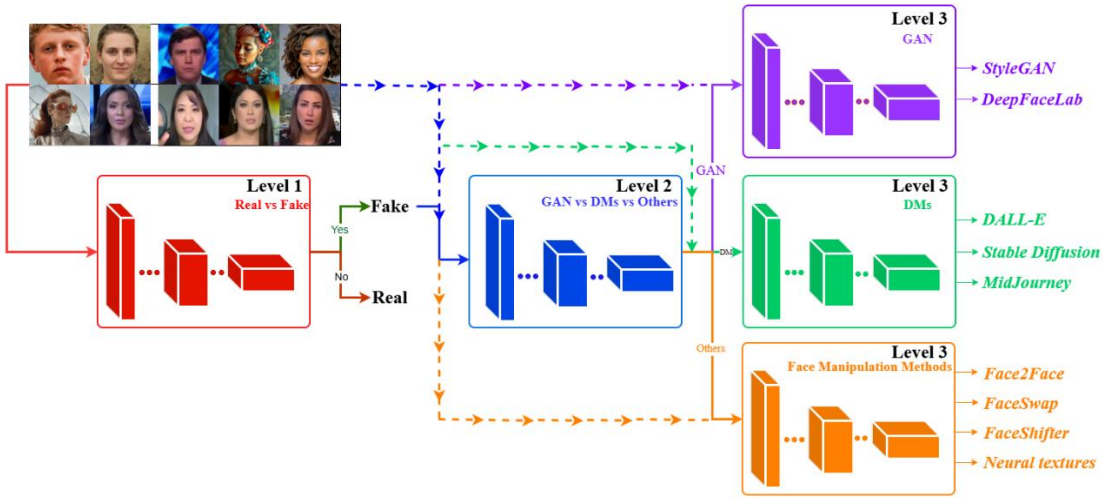


Fig.2. Hierarchy of three level architecture

Level 1: Real vs Fake Classification
The first level of the architecture as shown in figure 2, functions as a binary classifier that distinguishes authentic (real) facial images from manipulated (fake) ones. Input images are resized to 256 × 256 × 3 and processed using the EfficientNet-B4 backbone, which provides high accuracy with low computational cost through compound scaling. The network extracts discriminative features using MBConv blocks and aggregates them via global average pooling, followed by a fully connected layer that outputs the final real/fake decision. Images predicted as fake are forwarded to Level 2 for further analysis, while those classified as real exit the pipeline.

Let $f_1(x; \theta_1)$ denote the Level-1 EfficientNet-B4 classifier parameterized by θ_1. Given an input image x, the network **produces** posterior probabilities: $p^{(1)} = f_1(x; \theta_1) = [p_R, p_F]$, where p_R and p_F correspond to the probabilities of real and fake classes, respectively. The predicted label is: $\hat{y}^{(1)} = \arg\max p^{(1)}$.

The loss of training level-1 is categorial cross-entropy is given in equation 2.

$$\mathcal{L}_1 = -\sum_{k\in\{R,F\}} y_k^{(1)} \log p_k^{(1)} \quad (2)$$

where $y_k^{(1)}$ denotes the one hot encoded ground truth label and only samples classified as fake ($\hat{y}^{(1)}$) are forwarded to level-2.

Level 2: GAN vs DM vs Others Classification
On the second level of proposed architecture, multiclass classification is applied to identify alteration technique of images detected as fake. We use the same feature extractor named as EfficientNet-B4, the identification head is reformatted to differentiate between GAN-based, DM-based and other face manipulation methods. This level detects techniques-specific artifacts like frequency anomalies in GANs and noise patterns in DMs and other cues in case of other face manipulation techniques. After completion the task of predicted class, images are passed to the corresponding level 3 for further deepfake generation attribution.

For fake samples, level-2 performs gross categorisation into GANs, DMs and other face manipulation techniques. Let f2(x, θ2) represents the level-2 EfficientNet-B4 classifier. The output probabilities are:

$$p^{(2)} = f_2(x; \theta_2) = [p_{GAN}, p_{DM}, p_{OTH}]$$

And the predicted gross category is:

$$\hat{c} = \arg\max p^{(2)}$$

The loss of level-2 is defined as:

$$\mathcal{L}_2 = -\sum_{j\in\{GAN,DM,OTH\}} y_j^{(2)} \log p_j^{(2)} \quad (3)$$

Level 3: Technique-Specific Classification

Third level of proposed methodology is comprises of three specialized modules along with an identification for specific manipulators and methods in their respective categories. First module concentrates on identifying specific GAN-based generation methods, including StyleGAN and DeepFaceLab. The extracted features are processed through a specialized classification head trained to determine pattern specific to each manipulation tool, for the system to recognise the origin of the altered image. Second module is directed towards DALL-E, Stable Diffusion, and MidJourney. The detector then analyzes high frequency features and temporal consistency markers for each diffusion-based tool to achieve fine-grained identification. Last module of our proposed methodology explores traditional facial manipulation techniques including such as Face2Face, FaceSwap, FaceShifter, and Neural-Texture. The classifier employs robust feature extraction from EfficientNet-b4 to identify signatures of alteration, such as boundary discontinuities, lighting inconsistencies, anatomical illegibility.

Depending up on the predicted coarse category ĉ, the input is routed to one of three specialized modules of level-3. Let f_3^{GAN}, f_3^{DM}, and f_3^{OTH}, denotes the EfficientNet-B4 models for GAN-based, DM-fusion, and other face manipulation techniques, respectively. For s selected branch b∈ {GAN, DM, OTH}, the prediction is:

$$p^{(3)} = f_3^b(x, \theta_3^b)$$

And the final method label is obtained as:

$$\hat{m} = arg\ max\ p^{(3)}.$$

Each level-3 classifier is optimised using the cross-entropy loss is shown in equation (4):

$$\mathcal{L}_3^b = -\sum_{l\in\mathcal{M}_b} y_l^{(3)} \log p_l^{(3)} \quad (4)$$

Where $\mathcal{M}_b$ denoted the set of generation methods within branch b.

All levels are trained independently. The total optimized objective is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \sum_b \lambda_3^b \mathcal{L}_3^b, \quad (5)$$

Where λ_2=0.5, and $\lambda_3^b = 0.3$ are weighting coefficient handling the contribution of each level for all branches b∈ {GAN, DM, OTH} over the range [0.1,1.0].

3.3 Performance Metrics

The proposed approach's performance will be evaluated against previously created algorithms on the basis of various robust parameters like Precision, Recall, and F1 score etc.

Table 1. Performance Evaluation Metrics

Sr No.	Metric	Description	Mathematical Formula (%)
1	Precision	Accuracy of optimistic predictions.	$Precision = \frac{T_p}{T_p + F_p} \times 100$
2	Recall	Accurate predictions over all accurate predictions	$Recall = \frac{T_p}{T_p + F_n} \times 100$
3	F1 Score	Harmonic mean of accuracy and recall	$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \times 100$
4	Accuracy	Percentage of accurate forecasts among all predictions	$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100$
5	Specificity	Percentage of non-fake were accurately identified	$Specificity = \frac{T_n}{T_n + F_p} \times 100$
6	Geometric Mean	Balance between sensitivity and specificity.	$Geometic\ Mean = \sqrt{Senstivity \times Specificity}$

3.4 Implementation Details

The EfficientNet-B4 model is loaded with pre-trained ImageNet weights, with the final classification layers trained separately for each level. Images are resized to 256×256×3 to fit the EfficientNet-B4 input size. Images of 18,124 are divided into 80% training and 20% validation using image_dataset_from_directory with a random seed of 42, to ensure repeatability and stratified sampling to maintain class balance across the training and validation sets. Each stage is trained separately using the Adam optimizer with initial learning rate 1×10^{-4}. For all levels, the batch size is set to 32. The model is trained for up to 10 epochs. Regularization (dropout with a rate of 0.5) is used before the final classification layer to reduce overfitting.

4. Result and Discussions

This portion is concerned with narrating research outcomes. In this section, we explore the assessment methodologies and programming tools used in the analysis, and evaluating the neural network models.

4.1 Experimental Setup

The programming processes are performed in Python on a X86-64 Ubuntu 18.04.4 LTS. The system has 16 GB of RAM and an intel core i7-8550U processor that performing at the speed of 1.80GHz.

4.2 Experimental Outcomes

4.2.1 Accuracy and loss

4.2.1.1 Accuracy and Loss Curve of Binary Classification

The accuracy and loss curve of training and validation continuous and successful learning for the binary classification task. Curve shows that training accuracy increases quickly, and with respect to validation accuracy, stays above 99% as shown in figure 3 (a), and shows a strong discriminative performance. As expected, both training and validation loss decreases significantly as shown in figure 3 (b), there are small fluctuations in validation loss after convergence. The near correspondence between training and validation curves shows that there is efficient optimization, controlled model complexity, and no overfitting, which validates the robustness of level 1 for binary classification of proposed approach.

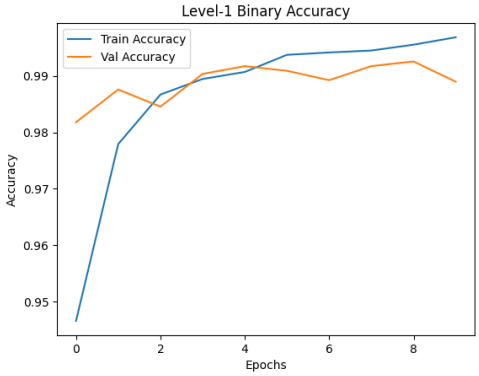


Fig. 3. (a) Accuracy Curve

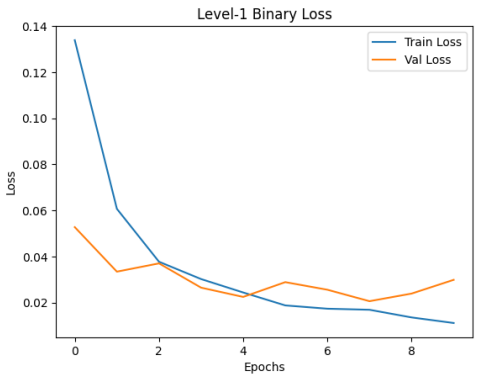


Fig. 3. (b) Loss Curve

4.2.1.2 Accuracy and Loss Curve of Multiclass Attribution

The accuracy and loss curves of training and validation for the multiclass attribution task remains stable and effective learning behaviors. As shown in figure 4 (a), training accuracy improves consistently from early epoch, along with validation accuracy, reached about 98.52% uniform with effective class separation and reliable generalisation across multiple classes. Simultaneously, training and validation loss as shown in figure 4 (b), decrease sharply during early epochs, and slowly balanced at low values. The small difference between training and validation curves throughout training shows the controlled model complexity and the absence of overfitting at level 2 of proposed approach for multiclass attribution. The fast convergence and steady validation accuracy directly validates the benefit of EfficientNetB4's compound scaling that simultaneously scales depth, width and

resolution to adapt to both low-level forgery patterns and high-level semantic information. The fact that there is no overfitting at both Level-1 and Level-2 directly validates that independent training of each level with dropout rate 0.5 prevents over-fitting.

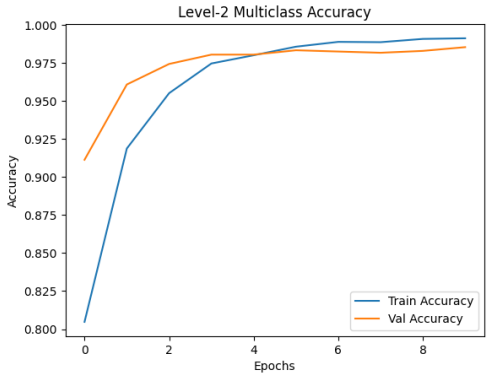


Fig. 4. (a) Accuracy Curve

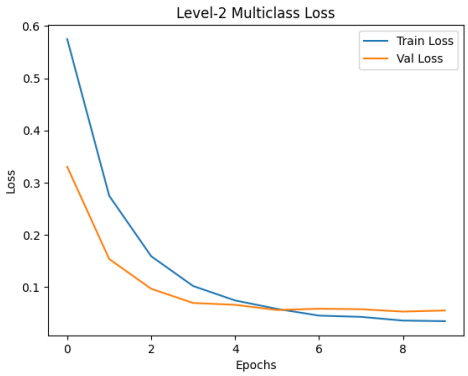


Fig. 4. (b) Loss Curve

4.2.2 Confusion Matrix

The confusion matrix for model performance in binary and multiclass attribution is very strong. Figure 5 (a) shown that the model correctly identifies 2428 fake images and 1156 real images in the binary classification task i.e., real VS Fake, whereas only 8 fake images are mistakenly classified as real and 32 real images are classified as fake, which is a very small number of misclassifications that indicating great discriminative ability and sensitivity to fake detection. Similarly, the multiclass confusion matrix as shown in figure 5 (b), illustrates that the maximum images are correctly identified through the main diagonal for all deepfake categories including DALL-E, Stable Diffusion, StyleGAN, FaceSwap, FaceShifter, DeepFaceLab, Face2Face, Midjourney, and NeuralTextures. There are few misclassifications among similar image manipulation techniques. Overall, the supremacy of all diagonal values and minimal diagonal-off errors in both metrices reveals high accuracy, class-wise discrimination robustness for the proposed framework for binary classification as well as multiclass attribution. The low number of misclassifications between visually similar approaches (FaceSwap/FaceShifter, DALL-E/Stable Diffusion) directly demonstrates the benefits of three-level hierarchical routing. The separation of GAN-based, diffusion-based and face manipulation methods at Level-2 before further attribution avoids the confusion between families that a single-level classifier would experience.

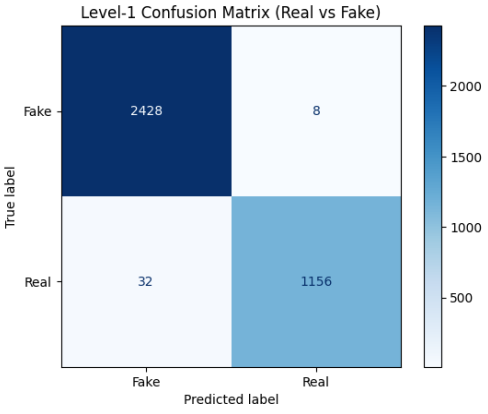


Fig. 5. (a) Confusion Matrix (Binary)

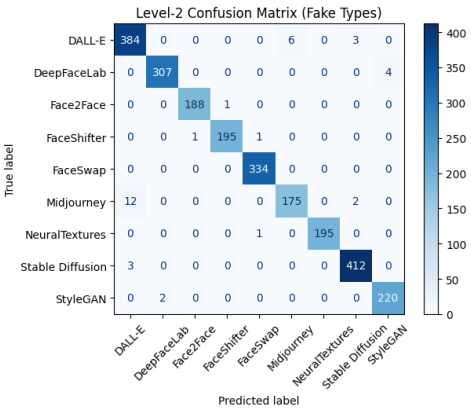


Fig. 5. (b) Confusion Matrix (Multiclass)

4.2.3 Evaluation Metrices

The evaluation metrics for Level-1 (Binary) and Level-2 (Multiclass) classification tasks shows in table 2 (a), (b), also show that the proposed model is strong and reliable. The model is fairly accurate in the binary classification task, which has a 98.89 % accuracy, indicating effective separation between real and fake images. 99.31 % precision for binary classification reflects a very low false-positive rate, while recall of level-1 of proposed approach 97.31% which confirms the model's strong ability to detect fake images. F1-score of binary classification is 98.30% that emphasis a balanced trade-off between precision and recall, and the high specificity 99.67% demonstrates an accurate determination of real images. G-Mean of 98.48% also verifies balanced

performance both classes. Similarly, model has best accuracy of 98.535 for multiclass attribution as shown in figure 5 (b), and illustrating effective discrimination among the various fake image categories. Precision of 98.63% and recall of 98.37%, shows that model has consistent and reliable predictions across different classes, and the F1-score of 98.49% is a strong indicator of string identification. The very high specificity of 99.81% and G-Mean of 99.08% also point to minimal class confusion and good generalizations among different fake generation techniques.

Table 2 (a): Evaluation Metrics (Binary)

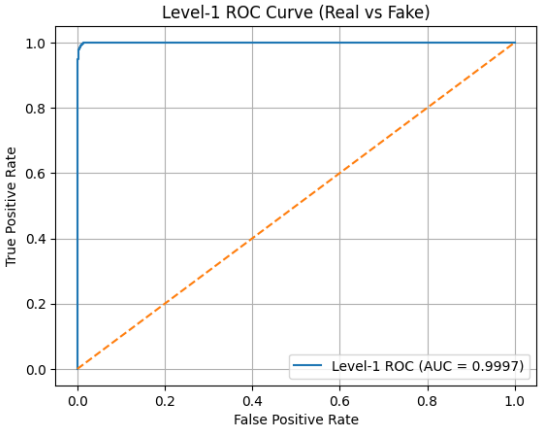
LEVEL-1 (Binary) METRICS	
Accuracy	: 0.9889624724061811
Precision	: 0.993127147766323
Recall	: 0.9730639730639731
F1-score	: 0.9829931972789115
Sensitivity	: 0.9730639730639731
Specificity	: 0.9967159277504105
G-Mean	: 0.9848189481691334

Table 2 (b): Evaluation Metrics (Multiclass)

LEVEL-2 (Multiclass) METRICS	
Accuracy	: 0.9852820932134096
Precision	: 0.9862879304740859
Recall	: 0.9837089079099768
F1-score	: 0.9849366649348136
Sensitivity	: 0.9837089079099768
Specificity	: 0.9981114860508524
G-Mean	: 0.9908840294986533

4.2.4 AUROC Curve

The ROC curve analysis for both Level-1 (Binary) and Level-2 (Multiclass) classification as shown in figure 6 (a), and (b), further confirms the outstanding discriminative capability of the proposed model. For Level-1 binary classification (Real vs Fake), the ROC curve lies very close to the top-left corner, achieving an exceptionally high AUC of 0.9997, which indicates near-perfect separation between real and fake images with an extremely low false positive rate and very high true positive rate across thresholds. Similarly, the AUROC curve of level-2 illustrates strong functioning across all fake image classes, with individual class AUC values reaching 1.000 for seven types manipulation out of 9, and a macro-average AUC of approximately 0.9995. This proves the model is very sensitive to distinguish between different fake generation methods. Entirely, the ROC analysis on both level shows the model's reliability and best classification performance for both binary and multiclass deepfake detection tasks. The 100% AUC (1.000) for seven of nine manipulation types, especially for all three diffusion-based methods (DALL-E, Stable Diffusion, Midjourney), directly validates that the dedicated diffusion branch at Level-3 effectively identifies unique noise and frequency artifacts of diffusion models. The macro-average AUC of 0.9995 also directly validates the hierarchical decomposition that offers better discrimination for all nine synthesis methods.



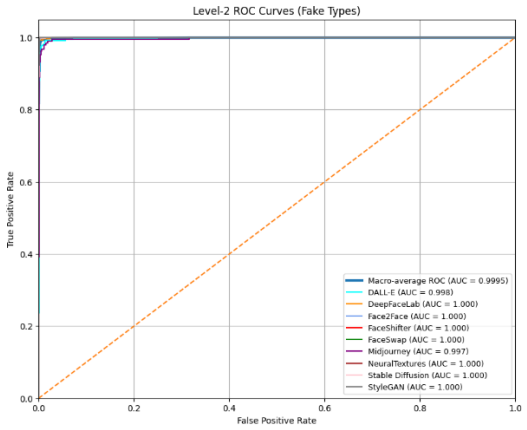


Fig. 6. (a) AUROC Curve (Binary) **Fig. 6. (b)** AUROC Curve (Multiclass)

4.3 Comparison Table: Proposed Model vs. State-of-the-Art Architectures

The current state of art comparison as shown in table 3, displays that earlier deepfake detection approaches performed well but were limited in particular type of manipulation, or had slightly lower multiclass accuracy i.e., approximately 88%. Earlier results reported accuracies of between 90-96% accuracy and lower multiclass attribution ability but the proposed approach far exceeds these. The Level-1 binary classifier is 98.89% accurate with 99.31% precision, and is highly accurate when detecting real and fake images. Also, the Level-2 which is multiclass attribution model, achieves 98.53% accuracy and 98.63% precision, prove that it is easy to distinguish between multiple GAN-based, DM-based and other face manipulation methods for tampered image generation

techniques. In general, the proposed three level hierarchical framework is more robust, generalise and precise to attribution than other any.

Table3: The comparative performance analysis of proposed approach with the other state of arts

References	Year	Generation / Manipulation type	Detection Type	Accuracy / Precision
Rossler et al. [4]	2019	Face manipulation Techniques	Binary	Precision: DeepFakes 96.36%, FaceSwap 90.29%, Face2Face 86.86%, Neural Textures 80.67%, Pristine 52.40%
Gragnaniello et al. [6]	2021	GAN Generated Images	Binary	Accuracy: 90–95%
Sha et al. [7]	2023	Diffusion Model	Binary + Multiclass Attribution	Binary: 93% Multiclass: 88%
Proposed Method (Level-1)	**2026**	**GAN + DMs + Other Methods**	**Binary classification**	**Accuracy 98.89%, Precision 99.31%,**
Proposed Method (Level - 2)	**2026**	**GAN + DMs + Other Methods**	**Multiclass Attribution**	**Accuracy 98.53%, Precision 98.63%,**

4.4 Discussion

The proposed hierarchical three-level deepfake detection framework is very performant at 98.89% for binary classification i.e., Real VS Fake, and attaining 98.52% accuracy for multiclass attribution. Using the EfficientNet-B4 backbone, the model is able to capture both low-level artifacts and high-level semantic cues across multiple fake image generation methods simultaneously. It achieves almost perfect precision to detect all diffusion-based alterations with AUC =1.000 (DALL-E, Stable Diffusion, Midjourney), while also showing high performance with traditional manipulation methods. Level-1 has robust classification performance with 99.31% precision and Level-2 provides fine-grained attribution with a macro average AUC of 0.9995, substantially outperforming among prior works which is limited to binary or single-method detection. The proposed hierarchical approach decreases the complexity of classification and allows strong generalization across all nine manipulation methods.

5. Conclusion and Future Scope

This paper presents a hierarchical three level deepfake detection and attribution framework that achieves 98.89% accuracy in binary real–fake classification and 98.52% accuracy in multiclass attribution across nine synthesis techniques. In combination with transfer learning, the EfficientNet-B4 framework is better than others and consistently highly performant over GAN-based techniques, diffusion models, and the traditional facial manipulation techniques with near-perfect 1.000 AUC scores for diffusion-based deepfakes. The hierarchical layout of this model offers high-resolution screening with low false positives, followed by coarse-grained forensic attribution with balance of computational efficiency and analytical depth. We also validate our hierarchical framework by comparing it with a flat baseline: an end-to-end direct 10-class classifier using a single EfficientNet-B4. After 10 epochs, the flat baseline achieves 97.30% validation accuracy (with overfitting as indicated by 99.39% training accuracy), while the proposed framework reaches 98.89% accuracy and 99.31% precision at Level-1 and 98.52% accuracy and 98.63% precision at Level-2, with consistently high specificity of 99.67% and 99.81% respectively, showing that the coarse-to-fine hierarchical decomposition reduces class confusion and performs better across all nine manipulation classes. The highest rate of misclassification is between visually similar methods - FaceSwap and FaceShifter, and DALL-E and Stable Diffusion - owing to similar boundary artifacts and noise patterns respectively. The Level-2 step substantially curtails these confusions by routing GAN-based, diffusion-based and face manipulation techniques before conducting fine-grained attribution. StyleGAN and DeepFaceLab have the lowest number of misclassifications due to their unique frequency artifacts that are captured by EfficientNet-B4's compound-scaled convolution. Beyond the detection, the framework contains forensic information that can be used for forensic analysis by identifying the generated data. This makes it very versatile for a real-world application for content moderation, digital forensics and media authentication, yet extensible to other deepfake generation processes.

6. References

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, Zero-shot text-to-image generation, *Proc. Mach. Learn. Res.* **139**, 8821–8831 (2021), https://proceedings.mlr.press/v139/ramesh21a.html.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2022**, 10684–10695 (2022), https://doi.org/10.1109/CVPR52688.2022.01042.

[3] Y. Zhang, Z. Pang, S. Huang, et al., Unmasking AI-created visual content: A review of generated images and deepfake detection technologies, *J. King Saud Univ. Comput. Inf. Sci.* **37**, 148 (2025), https://doi.org/10.1007/s44443-025-00154-8.

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, FaceForensics++: Learning to detect manipulated facial images, *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)* **2019**, 1–11 (2019), https://doi.org/10.1109/ICCV.2019.00009.

[5] J. C. Ng, M. B. Jasser, B. Issa, S.-S. M. Ajibade, and H. N. Chua, A review on deepfake image detection approaches, techniques and methods, *Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (I2CACIS)* **2025**, 512–517 (2025), https://doi.org/10.1109/I2CACIS65476.2025.11100461.

[6] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, Are GAN generated images easy to detect? A critical analysis of the state-of-the-art, *Proc. IEEE Int. Conf. Multimed. Expo (ICME)* **2021**, 1–6 (2021), https://doi.org/10.1109/ICME51207.2021.9428429.

[7] Z. Sha, Z. Li, N. Yu, and Y. Zhang, DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models, *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)* **2023**, 3417–3431 (2023), https://doi.org/10.1145/3576915.3616588.

[8] J. Bottu, Labeled deepfake image collection, *Kaggle Dataset* (n.d.), https://www.kaggle.com/datasets/jayanthbottu/labeled-deepfake-image-collection.

[9] M. Tan and Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, *Proc. Mach. Learn. Res.* **97**, 6105–6114 (2019), https://proceedings.mlr.press/v97/tan19a.html.

[10] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, FaceShifter: Towards high fidelity and occlusion-aware face swapping, *arXiv preprint* **2019**, arXiv:1912.13457 (2019), https://arxiv.org/abs/1912.13457.

[11] P. Henzler, N. J. Mitra, and T. Ritschel, Learning a neural 3D texture space from 2D exemplars, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2020**, 8157–8166 (2020), https://doi.org/10.1109/CVPR42600.2020.00769.

[12] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, and W. Zhang, DeepFaceLab: Integrated, flexible and extensible face-swapping framework, *Pattern Recognit.* **141**, 109628 (2023), https://doi.org/10.1016/j.patcog.2023.109628.

[13] T. Karras, S. Laine, and T. Aila, A style-based generator architecture for generative adversarial networks, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2019**, 4401–4410 (2019), https://doi.org/10.1109/CVPR.2019.00453.