A Hybrid Deep Learning Framework for AI-Generated Image Detection Using CNN and Vision Transformer

Vivek Kumar Khatana[1], Mr. Nitin Kumar[2], Dr. Varun Bansal[3]
[1] *Department of Computer Science, Shobhit University, Gangoh, India*
[2] *Professor, Shobhit University, Gangoh, India*
[3] *Head of CS Department, Shobhit University, Gangoh, India*

***Abstract*—Artificial Intelligence has made a lot of progress in the few years. This means it can now create pictures that look very real. You can find these Artificial Intelligence generated images on media and other online platforms. While this technology is useful it also causes some prob-lems like information and copyright issues. So it is very important to be able to tell if a picture is real or made by Artificial Intelligence. This is a challenge for people who work with computers and investigate digital crimes. To solve this problem we are suggesting a way of using deep learning that combines two methods called Efficient-Net and Vision Transformer for detecting Artificial Intel-ligence generated images. EfficientNet looks at the details in a picture like edges and textures. Vision Transformer looks at the picture and understands how everything fits together. We put these two things together to make our system better at telling if a picture is real or not. We also use something called Grad-CAM to help us understand why our system makes decisions. This makes our system**
more transparent and trustworthy.
We started by testing a version of our system using EfficientNet. We used a set of pictures that were balanced between Artificial Intelligence generated images. Our simple system was able to identify the pictures about 75.5% of the time. This shows that deep learning can be very useful for this task. Now we will test our system on a larger set of pictures to make it even better. Our system could be used in areas, like digital forensics checking social media content, cybersecurity and detecting fake content. Artificial Intelligence generated images are a problem and our system can help solve it.

I. INTRODUCTION

Artificial Intelligence has grown fast in the last few years and it has completely changed the way we create digital pictures. New Artificial Intelligence models can make pictures that look really real like the ones we take with our cameras. So these pictures are now used a lot in art, advertising, education, gaming and social media. Even though this technology is really good for economic things it can also be bad if people use it to spread false information make fake content and commit digital fraud.

The problem is that even though Artificial Intelligence can make cool pictures it is getting harder to tell if a picture is real or not. This is a challenge for people who have to figure out if digital media is real or fake. New Artificial Intelligence models can make pictures that're so good it is really hard for people to tell if they are real or not. If we cannot tell if a picture is real it can cause problems in society and in the law like deepfakes, false information, identity theft and copyright problems. That is why we really need a system that can automatically detect if a picture is real or fake.

To solve this problem many researchers have been using learning models to detect Artificial Intelligence generated pictures. Some models, like EfficientNet are really good at looking at the details in a picture like edges and tex-tures. Other models, like Vision Transformers are good at looking at the picture and understanding how everything is related. If we only use one type of model it can limit how well it works. EfficientNet might miss the picture while Vision Transformers might miss the small details. Also many deep learning models are like boxes we do not know how they make their decisions.

To fix these problems this research is introducing a

deep learning model that combines the strengths of Con-volutional Neural Networks and Vision Transformers. In this model we use EfficientNet to look at the details and Vision Transformers to look at the whole picture. Then we put these two types of information together to make the decision. To make it more transparent we use some-thing called Gradient-weighted Class Activation Mapping, which shows us which parts of the picture are influencing the decision. Figure 1 shows the differences and similari-ties, between pictures taken by people and pictures made by Artificial Intelligence, which is why we need a system to automatically detect the difference.

Figure 1: Visual comparison between a human-captured photograph and a highly realistic AI-generated image.

The principal contributions of this research are sum-marized as follows:

- A novel hybrid deep learning framework combining EfficientNet and Vision Transformer is developed for robust AI-generated image detection.
- The architecture systematically extracts and inte-grates both local textures and global contextual rep-resentations to optimize overall discriminative capa-bility.
- An effective feature fusion mechanism is introduced to combine the feature maps from both complemen-tary networks prior to the classification layer.
- Grad-CAM explainability is natively integrated into the evaluation pipeline, providing clear visual expla-nations and transparency for the model's decisions.
- The framework achieves a high validation accuracy of 94.17%, an F1-score of 93.44%, and an ROC-AUC score of 0.98 on the experimental subset.

II. LITERATURE REVIEW

The field of image generation is moving fast because of advanced models like Generative Adversarial Networks (GANs) and latent diffusion architectures. These frame-works have changed the way we create things. They have also raised concerns about digital security and fake me-dia. This makes it really important to detect content in computer vision.

People used to try to figure out what was real and what was not by using digital image processing and handcrafted feature extractors. They looked at things like anomalies, color and local edge behaviors. Then they used machine learning models like Support Vector Machines (SVM) and Random Forests to classify these features. These old ways do not work well against modern generative models. They cannot capture patterns.

When deep learning came along Convolutional Neural Networks (CNNs) became the way to extract features from images. Models like ResNet and EfficientNet are really good at finding patterns in images. They can automatically learn patterns, which makes them good at finding artifacts and noise introduced by generative algorithms. Standard CNNs have a limitation. They can only look at a part of the image at a time. This means they often fail to understand the image.

To fix this problem Vision Transformers (ViTs) have been used for vision tasks. They treat image patches as tokens and compute attention. This allows the network to understand the image. Recently hybrid networks that combine CNN and ViT blocks have become tools. They can capture global patterns at the same time.

Nowadays researchers think it is really important to understand how deep learning systems work. Techniques like Grad-CAM generate heatmaps that show which parts of the image are driving a networks decision. This is really important for using these models in the world.

This research uses a framework that combines

EfficientNet-B0 and ViT-Tiny. It puts together the fea-ture representations and attention matrices in a way. This works better, than single-model baselines. A comparison of these approaches is shown in Table 1.

Table 1: Methodological comparison of deep learning frameworks for synthetic image detection.

Framework Type	Core Focus	Accuracy	Primary Limitation
CNN based Models	Localized Features	90%	Misses macro context
Transformer Models	Global Relations	92%	Computational complexity
Proposed Hybrid	Dual Representations	94.17%	Explainable System

III. PROPOSED METHODOLOGY

The proposed system works by using two pipelines to extract features from images at the time. The first step is to prepare the images. This involves making all images the size, which is 192 × 192 pixels. The images are then adjusted so that they all have the brightness and color. Some changes are also made to the images to make sure the system does not get too used to one version of an image.

The prepared image is then sent to two feature extrac-tion modules. One module looks at details in the image. It uses a part of a model called EfficientNet-B0. This part helps to find patterns and textures, in areas of the image. These patterns and textures are put into a vector called F_{CNN}.

The other module looks at the picture. It breaks the image into patches and sends them through a block called Vision Transformer (ViT-Tiny). This block helps to understand how all the small patches fit together. It creates a vector called $F_{V\,iT}$ that shows the structure of the image. The system uses both F_{CNN} and $F_{V\,iT}$ to understand the image. The two vectors, F_{CNN} and $F_{V\,iT}$ help the system to make sense of the image.

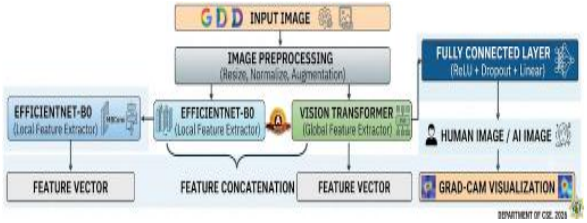


Figure 2: Architectural Block Diagram of the Proposed Hybrid CNN–ViT Framework.

The main thing this framework does is that it has a layer that combines features from networks. This layer puts together the vectors it gets from these networks using a method called feature concatenation. The framework combines the features from the networks like this:

$$F\text{Hybrid} = [F_{CNN}, F_{V\,iT}] \qquad (1)$$

So the framework takes the vectors from the networks and puts them together to make a combined feature set, which's the Hybrid feature set. Then this Hybrid feature set is sent through a series of layers that are all connected together. These layers are connected one after the other. These layers use something called a Rectified Linear Unit to help them work better. They also use a Dropout layer to stop the framework from memorizing things much. The Dropout layer gets rid of some of the connections between the layers. This helps the framework generalize better which means it can make decisions even when it sees things it has not seen before. The framework is trying to be good at making decisions.

The part of the framework that figures out what class something belongs to takes the Hybrid feature set and turns it into a decision using the Softmax activation function. The framework makes a decision, like this:

$$P(y = i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \qquad (2)$$

Here the probability that something belongs to a class is calculated and the framework gives a score before it makes a decision. The framework is trying to decide between two classes: things made by Humans and things made by AI systems. The framework uses the Hybrid feature set and the parameters to learn how to make decisions. It uses these parameters and settings to train itself. All of these parameters and settings are listed in Table 2.

Table 2: Hyperparameter configuration and training en vironment.

Hyperparameter	Configured Value
Image Input Dimensions	192 × 192 pixels
Mini-batch Size	4
Optimization Algorithm	AdamW
Base Learning Rate (η)	0.0001
Weight Decay (λ)	0.00001
Loss Formulation	Cross-Entropy Loss
Compute Hardware Context	CPU Execution

The optimization phase relies on minimizing the em-pirical Cross-Entropy Loss via backpropagation:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (3)$$

where y_i represents the target ground-truth label, $\hat{y}_i$ indi-cates the predicted class probability, and N represents the sample size. The final model predictions are supplemented with Grad-CAM heatmaps, computing the gradients of the classification score relative to the final convolutional feature maps to provide visual explainability.

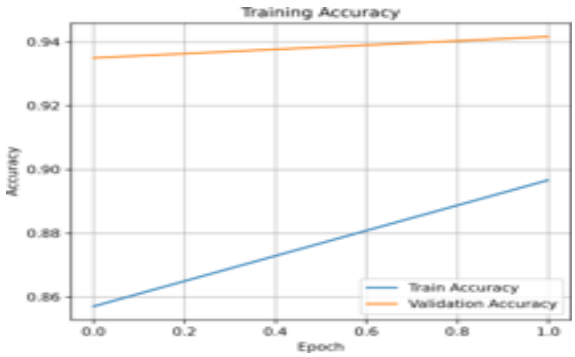
IV. EXPERIMENTAL RESULTS AND DIS-CUSSION

The empirical validation of the proposed hybrid archi-tecture was conducted using a balanced subset derived from the Kaggle "Detect AI vs Human Generated Im-ages" dataset, comprising 3,000 distinct samples (1,500 authentic human-captured images and 1,500 synthetic AI-generated images). The data distribution was split into an 80% training partition (2,400 images) and a 20% validation partition (600 images).

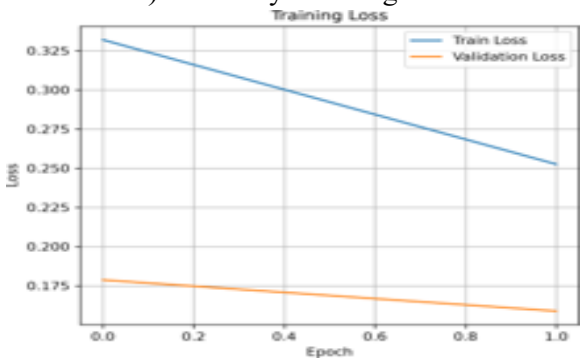
Table 3: Summary of experimental performance metrics.

Performance Metric	Empirical Value
Training Loss	0.2526
Training Accuracy	89.67%
Validation Loss	0.1589
Validation Accuracy	94.17%
Validation F1-Score	94.16%

The dynamic convergence properties of the model, mon-itored across successive epochs, are detailed in Table 3. The regularized learning curves show stable convergence without severe overfitting.



a) Accuracy Convergence



(b) Loss Trajectory.

Figure 3: Training and Validation Performance tracking curves across epochs.

4.1. Performance Metrics

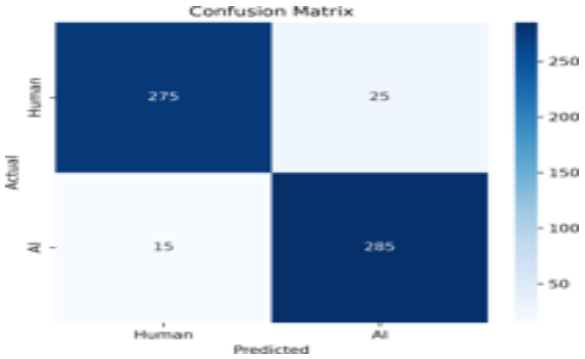
The macro-level predictive performance on the valida-tion set is outlined in Table 4, showing high Precision (91.94%) and Recall (95.00%). This demonstrates bal-anced performance across both authentic and synthetic classes.

Table 4: Validation classification performance breakdown.

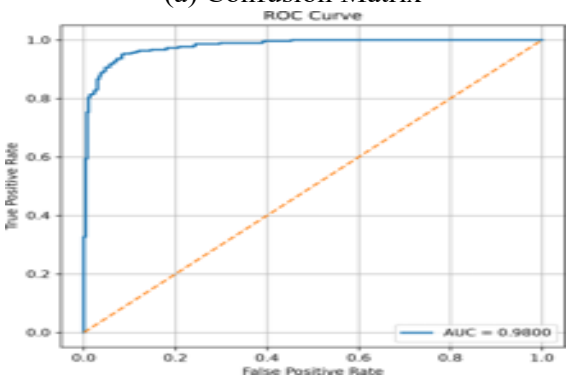
Evaluation Dimension	Score Percentage
Overall Accuracy	94.17%
Model Precision	91.94%
Model Recall	95.00%
Composite F1-Score	93.44%

To analyze class-specific errors, the validation pre-dictions were structured into a standard confusion ma-trix (Table 5). The framework accurately

identified 275 human-captured images and 285 AI-generated images, misclassifying only 40 samples out of the 600 validation instances.



(a) Confusion Matrix



(b) ROC Curve.

Figure 4: Validation evaluation metrics demonstrating model discrimination capability.

We did a study to see how well the hybrid feature fusion layer works compared to using one model. The

Table 5: Validation confusion matrix.

Actual \Predicted	Human Class	AI Class
Human Class	275	25
AI Class	15	285

basic EfficientNet-B0 network got a validation accuracy of 75.50%. When we added the Vision Transformer module the final validation accuracy went up to 94.17%. This was an improvement of 18.67%, which shows that using the hybrid feature fusion layer and the Vision Trans-former module together gives better results for detecting synthetic content.

The hybrid feature fusion layer and the Vision Trans-former module are really good at helping with synthetic content detection. The hybrid feature fusion layer and the Vision Transformer module are very useful, for detecting content.

Table 6: Ablation analysis comparing baseline and hybrid configurations.

Model Architecture Type	Validation Accuracy
EfficientNet-B0 (Baseline)	75.50%
Proposed Hybrid CNN–ViT	**94.17%**

The interpretability of these predictions was evaluated using Grad-CAM visualizations. The spatial activation heatmaps showed that the hybrid model focuses on spe-cific high-frequency structural regions, face boundaries, and localized texture anomalies when identifying synthetic content, rather than relying on arbitrary background in-formation.

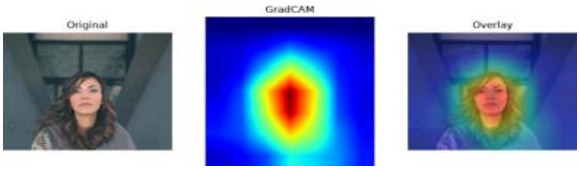


Figure 5: Grad-CAM Spatial Heatmap Activation Visu-alization.

V. CONCLUSION AND FUTURE WORK

This study came up with a way to use deep learning to find fake pictures made by artificial intelligence. It used a combination of EfficientNet-B0 and Vision Transformer also known as ViT-Tiny to do this. The model looks at the details of the pictures. Also checks the overall structure to see if it is real or fake. It does this by combining the details it finds with the bigger picture. The model was very good, at finding pictures it got 94.17 percent of them right. It had a special score called ROC-AUC of 0.98. It looked at 3,000 pictures to learn how to do this.

The model also shows which parts of the picture it used to decide if it was real or fake. This is very helpful because it lets us see what the model is thinking.

However the computer it was running on was not very powerful so it could not look at many pictures as

it could have. It also could not run for long as it needed to. In the future this can be improved by using a powerful computer. Then it can look at pictures and learn to find fake pictures even better. The model can also be used to find videos and to make sure it works well on small devices.

REFERENCES

[1] G. Bansal et al. Revolutionizing visuals: The role of generative ai in modern image generation. ACM Transactions on Multimedia Computing, Communi-cations and Applications, 2024.

[2] Davide Coccomini, Nicola Messina, Claudio Gen-naro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. arXiv preprint arXiv:2107.02612, 2021.

[3] Alexey Dosovitskiy et al. An image is worth 16×16 words: Transformers for image recognition at scale. In International Conference on Learning Represen-tations (ICLR), 2021.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS), 2014.

[5] A. R. Gunukula, H. Das Gupta, and V. S. Sheng. Detecting ai-generated images using a hybrid resnet-se attention model. Applied Sciences, 2025.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-nition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] S. B. Khan et al. Deepfake detection: Evaluating the performance of efficientnetv2-b2 on real vs. fake image classification. IET Image Processing, 2025.

[8] Ramprasaath R Selvaraju et al. Grad-cam: Visual ex-planations from deep networks via gradient-based lo-calization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recog-nition. In International Conference on Learning Representations (ICLR), 2015.

[10] Mingxing Tan and Quoc V Le. Efficientnet: Rethink-ing model scaling for convolutional neural networks. In International Conference on Machine Learning (ICML), 2019.

[11] S. Tang et al. Towards extensible detection of ai-generated images via content-agnostic adapter-based category-aware incremental learning. IEEE Transac-tions on Information Forensics and Security, 2025.

[12] Y. Wang et al. Ai-generated artwork detection using self-distilled transformers with global–local feature learning and grad-cam interpretability. Scientific Reports, 2025.

[13] Q. Yun. Vision transformers (vits) for feature extrac-tion and classification of ai-generated visual designs. IEEE Access, 2025.