# Multi-Phased Approach for the Detection of AI-Generated Images using Customizable Neural Networks

**Team:**

Bonala Shanmukesh  
23BCE20070  
[Shanmukesh.23bce20070@vitapstudent.ac.in](mailto:Shanmukesh.23bce20070@vitapstudent.ac.in)

Amith Gerri

23BCE20008

[amith.23bce20008@vitapstudent.ac.in](mailto:amith.23bce20008@vitapstudent.ac.in)

Kaustuv Gupta

23BCE7232  
[kaustuv.23bce7232@vitapstudent.ac.in](mailto:kaustuv.23bce7232@vitapstudent.ac.in)

## Abstract

The rapid advancement of generative AI models has enabled the creation of highly realistic synthetic images, making it increasingly challenging to distinguish them from authentic photographs. This proliferation of "deepfakes" and other AI-generated content poses significant risks, including the spread of misinformation and the erosion of digital trust. This project proposes a multi-phased approach to develop a robust system for detecting AI-generated images. **Phase 1** will involve implementing a **Convolutional Neural Network (CNN)** using transfer learning to establish a baseline for classification. **Phase 2** will explore the capabilities of a **Region-based Convolutional Neural Network (R-CNN)** to identify localized artifacts within images. Finally, **Phase 3** will focus on the development of a novel, custom-designed model architecture and a specialized dataset to push the boundaries of detection accuracy and generalizability. The ultimate goal is to create a comprehensive and effective tool to verify the authenticity of digital visual media.

## Introduction

In recent years, the line between real and artificially generated visual content has blurred considerably. Generative models, such as Generative Adversarial Networks (GANs) and Diffusion Models, can now produce images of such high fidelity that they are often indistinguishable from real photographs to the naked eye. While this technology has incredible applications in art, design, and entertainment, it also opens the door for malicious use. The ability to create convincing but fake images can be exploited to create propaganda, defame individuals, or generate fraudulent evidence. Therefore, the development of reliable automated tools to detect these forgeries is not just a technical challenge, but a critical necessity for maintaining the integrity of our digital information ecosystem. This project aims to tackle this challenge through a structured, phased implementation of increasingly sophisticated deep learning models.

### **Literature Review**

The detection of AI-generated images is a rapidly evolving field of research that has progressed in lockstep with the advancements in generative models themselves. The core challenge lies in identifying the subtle, often imperceptible, statistical traces left behind by the generation process.

Initial research efforts were largely focused on detecting images from **Generative Adversarial Networks (GANs)**. A foundational study by Wang et al. (2020) in **"CNN-generated images are surprisingly easy to spot... for now"** revealed that the up-sampling component common in many CNN-based generators, including GANs, introduces distinct artifacts in the frequency domain. These "GAN fingerprints" manifest as a spectral peak that can be reliably identified, allowing for high-accuracy detection. This approach, centered on frequency analysis, became a cornerstone of early detection methods.

However, the landscape shifted dramatically with the advent of **Diffusion Models**. As explored by Lorenz, P., et al. (2023) in their ICCV workshop paper **"Detecting Images Generated by Deep Diffusion Models Using Their Local Intrinsic Dimensionality,"** these newer models do not exhibit the same predictable frequency artifacts as GANs. Diffusion models construct images in a fundamentally different way, resulting in forgeries that are more globally coherent and lack the tell-tale signs that detectors were trained to find. This necessitated a move away from simple frequency analysis towards more sophisticated and robust feature extraction methods. The work by Lorenz et al. proposes using Local Intrinsic Dimensionality (LID) to capture subtle differences in the feature space, demonstrating a more effective approach for these modern generators.

This need for more robust features led researchers to leverage powerful, pre-trained architectures through **transfer learning**. As detailed by Guarnera, L., et al. (2023) in **"Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks,"** large models like ResNet, VGG, and EfficientNet, which are already experts at identifying complex patterns and textures, can be fine-tuned to become highly effective forgery detectors. This methodology forms the basis of our **Phase 1** investigation.

Furthermore, research has shown that while AI-generated images are becoming more realistic overall, they often fail in specific, localized areas. Artifacts frequently appear in complex details like hands, eyes, or the consistency of textures and shadows. This observation supports a move towards models that can analyze specific image regions, which is the core idea behind our **Phase 2** plan to use a **Region-based CNN (R-CNN)**. By isolating and classifying different parts of an image, an R-CNN can potentially identify localized inconsistencies that a global, full-image classifier might average out and miss.

Finally, the most significant challenge in the field today is **generalization**. As highlighted in numerous surveys, including the work by Yao, T., et al. (2023) in **"Towards Understanding the Generalization of Deepfake Detectors from a Game-Theoretical View"** (ICCV 2023), detectors trained on one type of generative model (e.g., Stable Diffusion) often perform poorly on images from another (e.g., Midjourney). To combat this, state-of-the-art approaches are exploring **ensemble and hybrid models**, which combine the strengths of different architectures (e.g., CNNs and Vision Transformers) to create a more resilient detector. This pursuit of a generalizable, robust solution is the primary motivation for our **Phase 3**, where we will design a novel model architecture.

## Objectives

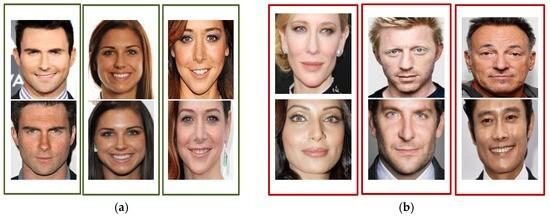
This project has two primary objectives:

1. **To systematically evaluate the effectiveness of different CNN-based architectures for AI-generated image detection.** This will be achieved by implementing and comparing a standard CNN (using transfer learning), an R-CNN, and a custom-designed model. The performance of each will be rigorously measured using metrics such as accuracy, precision, recall, and F1-score.
2. **To develop a novel model and a corresponding dataset that improves upon existing detection methods.** This involves not only designing a unique model architecture in Phase 3 but also curating a challenging and diverse dataset of real and AI-generated images from various sources. The goal is to create a solution that is more robust and can generalize better to new, unseen types of AI-generated content.

### **Preparing Datasets**

The project will utilize different datasets across its three phases:

* **Phase 1 & 2:** We will begin with a well-established, publicly available dataset. The **CIFAKE dataset** is an excellent choice. It contains 120,000 images (60,000 real images from CIFAR-10 and 60,000 AI-generated images created with Stable Diffusion). This provides a large, balanced dataset for our initial model training and evaluation.
  + **Data Source:** Kaggle - <https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images>
  + **Preprocessing:** The images will be resized to a standard input size (e.g., 224x224 pixels), and pixel values will be normalized. Data augmentation techniques (e.g., rotation, flipping, brightness adjustments) will be applied to the training set to improve model robustness.

  
 Fig-1 : Images showing both real and AI-generated images

* **Phase 3:** We will create a **custom dataset**. This dataset will be more challenging and diverse than CIFAKE. It will include:
  + High-resolution real photographs from various sources.
  + AI-generated images from a wide range of modern models (e.g., Midjourney, DALL-E 3, newer versions of Stable Diffusion).
  + Images that have undergone common post-processing steps like compression, resizing, and social media filtering, to simulate real-world conditions.

The project will be implemented in three distinct phases, each with a different model architecture:

#### **Phase 1: Baseline Model (CNN with Transfer Learning)**

* **Model:** We will use a pre-trained **ResNet50** model as our base.
* **Method:** We will employ **transfer learning**. The convolutional base of ResNet50 (pre-trained on the ImageNet dataset) will be used as a feature extractor. We will freeze these layers initially to retain their learned knowledge.
* **Custom Head:** We will add a custom classification head on top of the ResNet50 base. This will consist of:
  1. A GlobalAveragePooling2D layer.
  2. A Dense layer with a ReLU activation function.
  3. A Dropout layer to prevent overfitting.
  4. A final Dense output layer with a sigmoid activation for binary classification (Real vs. Fake).

#### **Phase 2: Localized Feature Detection (R-CNN)**

* **Model:** A **Region-based Convolutional Neural Network (R-CNN)** or one of its more efficient variants like **Faster R-CNN**.
* **Method:** Unlike a standard CNN that classifies the entire image, an R-CNN first identifies regions of interest (RoIs) within the image and then classifies each region. This approach is powerful for our task because AI-generated images often contain *localized* artifacts (e.g., distorted hands, strange textures in a specific area). By analyzing specific regions, the R-CNN may be able to detect subtle inconsistencies that a global classifier might miss.

#### **Phase 3: Proposed Custom Model**

* **Model:** The architecture for this phase will be designed based on the findings from Phases 1 and 2.
* **Potential Methods:** We will explore several advanced concepts:
  + **Attention Mechanisms:** To allow the model to focus on the most informative parts of an image.
  + **Ensemble Learning:** Combining the predictions of multiple different models (e.g., our CNN from Phase 1 and a Vision Transformer) to create a more robust final prediction.
  + **Frequency Domain Analysis:** Incorporating layers that analyze the frequency spectrum of the image, as this is often where AI-generation artifacts reside.

This phased approach will allow us to build a deep and comprehensive understanding of the problem and systematically work towards a state-of-the-art solution.