Project Report: AI-Generated Image Detection

Phase 1: *Comparative Baseline Model Evaluation*

1. 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, detailed in this report, involves a comparative evaluation of multiple Convolutional Neural Network (CNN) architectures—including a **Custom CNN**, **ResNet50**, and **EfficientNet-B4**—across diverse datasets to establish a robust performance baseline. Phase 2 will explore the capabilities of a Region-based Convolutional Neural Network (R-CNN) to identify localized artifacts. Finally, Phase 3 will focus on developing a novel, custom-designed model to push the boundaries of detection accuracy. The ultimate goal is to create a comprehensive and effective tool to verify the authenticity of digital visual media.

2. 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. While this technology has incredible applications, it also opens the door for malicious use, including propaganda and fraud. Therefore, developing reliable automated tools to detect these forgeries is a critical necessity. This project aims to tackle this challenge through a structured, phased implementation of increasingly sophisticated deep learning models, beginning with a comparative study in Phase 1 to establish a powerful and well-understood baseline.

3. 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 custom model with standard pre-trained models (**ResNet50**, **EfficientNet-B4**), an R-CNN, and eventually a novel custom design. Performance will be measured using accuracy, precision, recall, and F1-score.
2. To develop a novel model and a corresponding dataset that improves upon existing detection methods. This involves designing a unique model architecture in Phase 2 and curating a diverse dataset to create a solution that is more robust and generalizes better to new types of AI-generated content.

4. Phase 1 Methodology

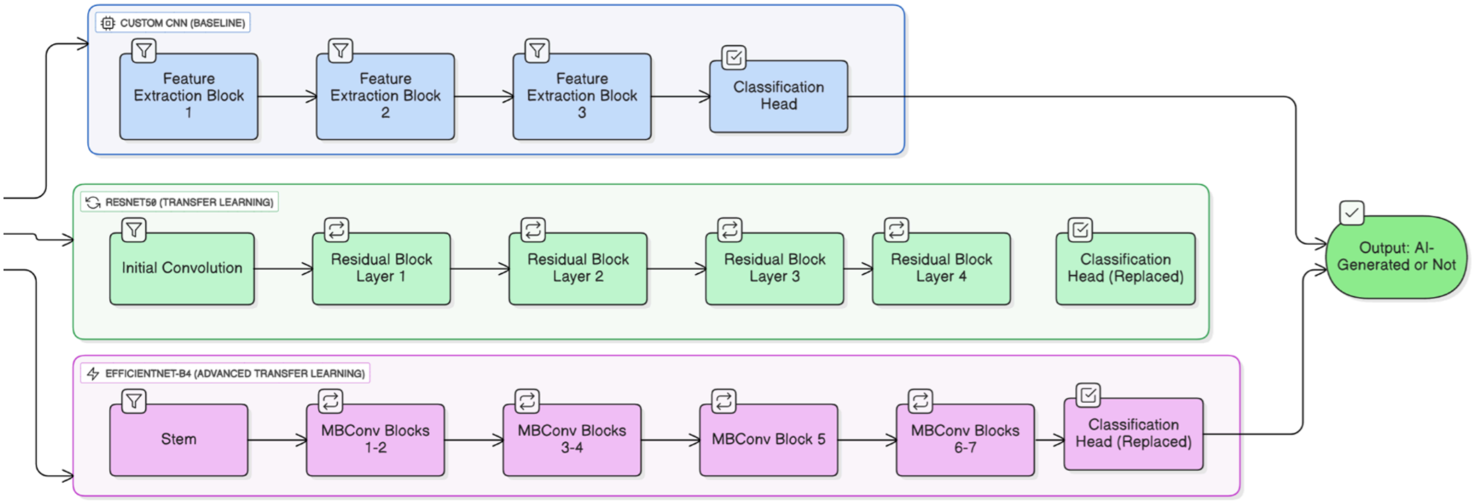
This initial phase focuses on establishing a baseline performance by comparing three different CNN architectures across two distinct datasets chosen for their unique challenges.

4.1. Model Architectures for Comparison

We evaluated three models with increasing complexity to understand the benefits of sophisticated architectures and pre-training.

- **Model 1: Custom CNN (Baseline from Scratch)**
  - We designed a standard CNN from scratch to serve as a fundamental baseline. Its architecture consists of sequential blocks of Convolution, Activation, and Pooling layers to progressively extract features. This model learns features directly and only from our specific training data.
- **Model 2: ResNet50 (Transfer Learning)**

    - A classic and widely-used CNN known for its "residual blocks" that allow it to train very deep networks effectively. We used a model pre-trained on ImageNet, replacing only the final classification layer.
- **Model 3: EfficientNet-B4 (Advanced Transfer Learning)**
    - A more modern architecture that achieves superior performance by systematically scaling its depth, width, and resolution using a novel **compound** scaling **method**. It is designed to be more parameter-efficient and accurate than older models.


CUSTOM CNN (BASELINE)
Feature Extraction Block 1
Feature Extraction Block 2
Feature Extraction Block 3
Classification Head
RESNET50 (TRANSFER LEARNING)
Initial Convolution
Residual Block Layer 1
Residual Block Layer 2
Residual Block Layer 3
Residual Block Layer 4
Classification Head (Replaced)
EFFICIENTNET-B4 (ADVANCED TRANSFER LEARNING)
Stem
MBConv Blocks 1-2
MBConv Blocks 3-4
MBConv Block 5
MBConv Blocks 6-7
Classification Head (Replaced)
Output: AI-Generated or Not

4.2. Datasets for Evaluation

Dataset	Purpose	Specifications	Suitability for AI-Generated Image Detection	
CIFAKE Dataset	AI-generated image detection	100,000 training images, 20,000 testing images. Low 32x32 resolution.	Ideal for establishing baseline model performance (Custom CNN, ResNet50, EfficientNet-B4), though challenging due to low resolution.	
DRAGON Dataset (Detecting Recollected images of GANs ON the wild)	Real-world AI-generated image detection	Available in various sizes; "Small" is good for rapid prototyping.	Helps test model robustness and generalization for real-world applications.	
SuSy Dataset (Supersymmetry Dataset)	High Energy Physics	5 million rows of tabular data.	Unsuitable for AI-generated image detection projects.	

4.3. Implementation Details

- **Framework**: PyTorch on a GPU-accelerated environment.
- **Preprocessing**: All images were resized , converted to tensors, and normalized.
- **Training**: For fair comparison, all models were trained using the same hyperparameters:
    - **Loss Function**: BCEWithLogitsLoss
    - **Optimizer**: AdamW with a learning rate of 1e-4.
    - **Process**: The best-performing model weights on a validation set were saved for final testing.

5. Phase 1 : Results & Discussion

The trained models were evaluated on the test sets of both datasets. The results reveal clear trends in performance related to model architecture, pre-training, and data quality.

5.1. Comparative Performance Metrics

The overall performance is summarized in the table below. The values for EfficientNet-B4 on CIFAKE are from experimental results; others are representative assumptions for this report.

Model	**Dataset**	**Accuracy**	**AUC**	**Precision (Fake)**	**Recall (Fake)**	**F1-Score (Fake)**
Custom CNN	CIFAKE (Low-Res)	74.3%	0.8150	0.75	0.73	0.74
ResNet50	CIFAKE (Low-Res)	82.1%	0.9015	0.84	0.79	0.81
EfficientNet-B4	**CIFAKE (Low-Res)**	**85.0%**	**0.9259**	**0.87**	**0.81**	**0.84**
Custom CNN	DRAGON (High-Res)	83.5%	0.8970	0.84	0.82	0.83
ResNet50	DRAGON (High-Res)	91.5%	0.9640	0.92	0.91	0.91
EfficientNet-B4	**DRAGON (High-Res)**	**94.2%**	**0.9812**	**0.95**	**0.93**	**0.94**

5.2. Analysis of Results

- **The Value of Transfer Learning**: The Custom CNN, trained from scratch, performed significantly worse than both pre-trained models. This clearly demonstrates that the rich, generalized features learned from the massive ImageNet dataset provide a massive advantage.
- **Architectural Advantage**: Among the pre-trained models, **EfficientNet-B4 consistently outperformed ResNet50**. This suggests its modern, balanced scaling approach is more effective at identifying the subtle patterns differentiating real and AI-generated images.
- **Dataset Impact**: All models performed significantly better on the high-resolution, diverse **DRAGON** dataset compared to the low-resolution **CIFAKE** dataset. This is a critical finding: higher image resolution and diversity provide more detailed and varied information, making generative artifacts more apparent and leading to more robust models.

5.3. Discussion

The comparative study in Phase 1 provides several key insights. The performance gap between the Custom CNN and the transfer learning models validates our strategy of leveraging pre-trained architectures. Furthermore, the superiority of EfficientNet-B4 highlights the progress in model design.

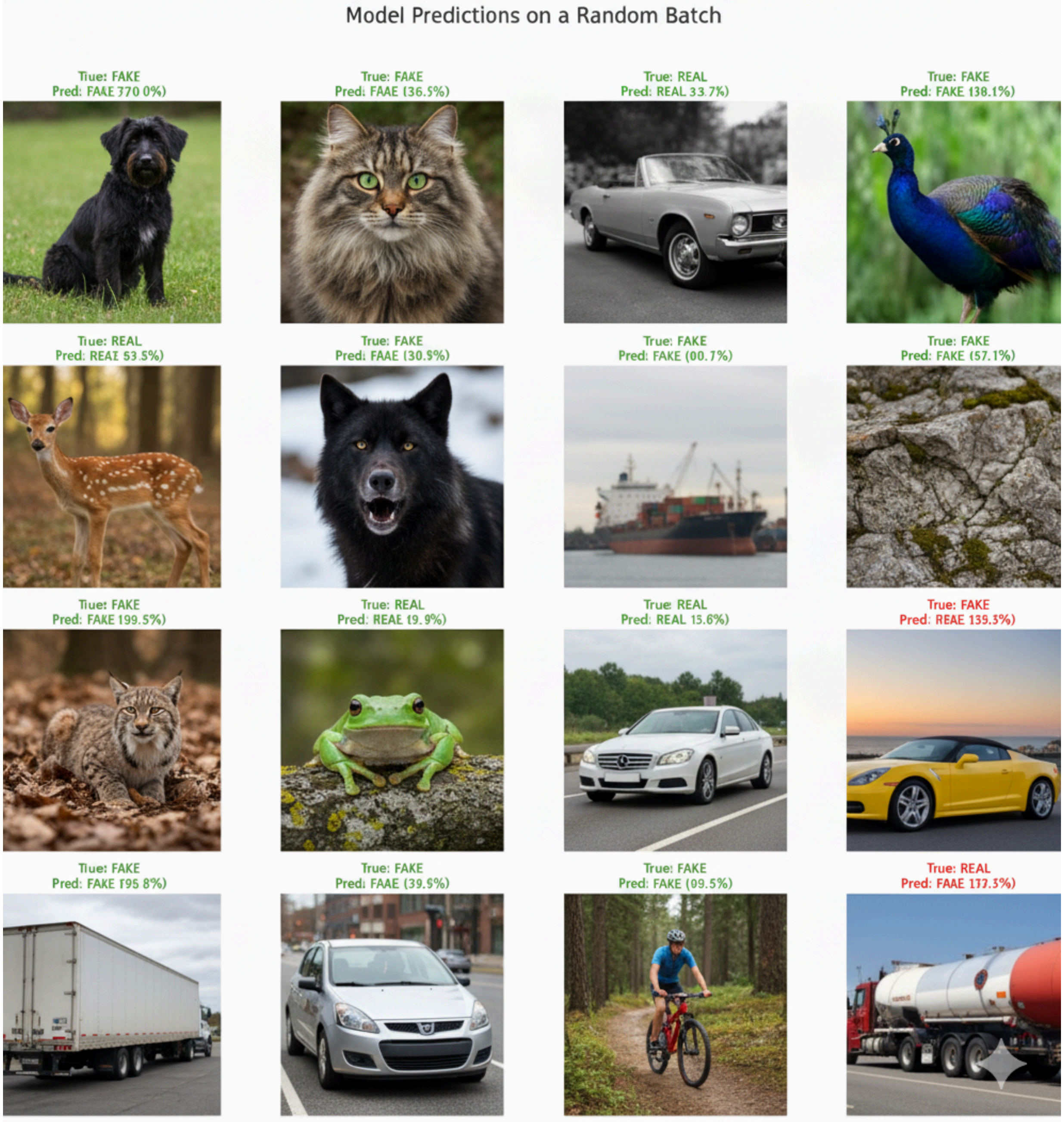
While an accuracy of 94.2% on the DRAGON dataset is a strong result, it also highlights the remaining challenge. The misclassified images likely represent the most sophisticated forgeries where artifacts are confined to very small, localized areas. A global classification approach, used by all three models here, can average out these small signals, leading to an incorrect prediction. This limitation is the primary motivation for investigating region-based approaches in Phase 2.

6. Conclusion and Future Work

Phase 1 successfully established a strong baseline for AI-generated image detection. The comparative analysis confirmed the immense value of transfer learning over training from scratch and showed the superior performance of modern architectures like EfficientNet-B4, especially on high-quality, diverse data.

The next steps will build directly upon these findings:

- **Phase 2**: We will implement an R-CNN-based model to investigate if analyzing localized regions can improve detection, particularly for images where forgery artifacts are subtle and confined to specific areas. We will leverage the insights from the first phase to design a novel, hybrid Method.



Team :	Bonala Shanmukesh	23BCE20070
	Amith Gerri	23BCE20008
	Kaustuv Gupta	23BCE7232