Assessing Pre-trained Architectures and Ensemble Learning for Detecting Image-based Deepfakes

Abhinav Shukla
Thapar Institute of Engineering and Technology
Roll Number: 102203464
Email: ashukla_be22@thapar.edu

Abstract

The rise of deepfake content—hyper-realistic manipulated media created using advanced AI—has introduced significant challenges in fields such as journalism, cybersecurity, and public discourse. This study evaluates the performance of six state-of-the-art pre-trained deep learning models sourced from the Hugging Face repository for detecting synthetic facial images. Additionally, a simple majorityvote ensemble is used to explore whether model aggregation improves detection reliability. Performance is measured using a publicly available Kaggle dataset and standard classification metrics including accuracy, precision, recall, specificity, sensitivity, and F1 score. Results indicate that the ensemble consistently provides more stable and improved detection compared to individual models.

1 Introduction

Deepfakes are digitally altered media produced with artificial intelligence, often relying on **Generative Adversarial Networks (GANs)**. These systems consist of two models—one generates content while the other critiques it—resulting in highly convincing forgeries. The misuse potential of such technology has raised widespread concerns, particularly in areas involving digital identity, trust, and public safety.

Early detection methods focused on visible imperfections such as abnormal eye movement or lighting artifacts. However, the progression of generative techniques has rendered these cues less effective. As a result, modern approaches use deep learning models—specifically **Convolutional Neural Networks (CNNs)** and **Transformer-based architectures**—to identify subtle manipulations in images.

This research examines six Hugging Face pre-trained models and evaluates their effectiveness in identifying manipulated imagery. Moreover, we investigate whether an ensemble method, combining multiple predictions, can boost performance without requiring extensive training or large datasets.

2 Background

The growing availability of tools for generating altered media has amplified the need for automated and robust detection mechanisms. GANs, along with other generative tools such as **autoencoders** and **diffusion models**, can now produce content that closely mimics real imagery.

While initial detection approaches targeted simple anomalies, the increasing realism of deepfakes has pushed researchers to develop models that learn from data, particularly CNNs that can capture textural irregularities and local visual distortions.

However, CNNs often lack the ability to understand relationships across distant image regions. This limitation is addressed by **Vision Transformers (ViTs)**, which treat images as sequences of patches and use self-attention to model long-range dependencies—helpful in recognizing tampering patterns that span the image globally.

2.1 Advantages of Pre-trained Models

Using models pre-trained on large-scale datasets such as *ImageNet* provides a head start in specialized tasks like deepfake detection. Fine-tuning these models is both time- and resource-efficient.

Key benefits include:

- **Faster Training:** Transfer learning enables rapid convergence.
- **Lower Data Requirement:** Adequate performance can be achieved with limited data.
- **Better Generalization:** Pre-trained features work well across different deepfake types.

This study focuses on the following architectures:

- **Vision Transformer (ViT):** Excels at capturing global image structures.
- **EfficientNet:** Optimized for balancing model size and accuracy.
- **ResNet:** Utilizes residual connections to support deeper networks and mitigate vanishing gradients.

2.2 Why Ensemble Learning?

Combining multiple models into an ensemble can improve predictive accuracy and reduce overfitting. This is particularly useful in deepfake detection, where individual models may excel at different manipulation types. Key ensemble benefits include:

- **Improved Accuracy:** Aggregates diverse strengths of individual models.
- **Robust Generalization:** More resilient to noise and overfitting.
- **Stability:** Reduces performance fluctuations of any single model.

In this work, a majority voting method is used, where the class predicted by most models is considered the final output.

3 Pre-trained Models Evaluated

The following Hugging Face models were analyzed:

- **prithivMLmods/Deep-Fake-Detector-Model** – Based on ResNet.
- **joyc360/deepfakes** – A CNN optimized for facial forgery detection.
- **dima806/deepfake vs real image detection** – Focused on detecting low-level artifacts.
- **DaMsTaR/Detecto-DeepFake Image Detector** – An EfficientNet variant.
- **DarkVision/Deepfake detection image** – Blends CNN with attention layers.
- **strangerguardhf/vit deepfake detection** – Uses Vision Transformers for capturing global inconsistencies.

The ensemble's output was determined using a simple majority vote among these six models.

4 Dataset Description

We used the "Real and Fake Face Detection" dataset from Kaggle for evaluation. It includes:

- **1,000 Real Images**
- **1,000 Fake Images**

Images were resized to 224x224 pixels. Only the test set was used to benchmark model performance.

5 Evaluation Criteria

Models were evaluated using standard classification metrics:

- **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** $\frac{TP}{TP+FP}$
- **Recall (Sensitivity):** $\frac{TP}{TP+FN}$
- **Specificity:** $\frac{TN}{TN+FP}$
- **F1 Score:** The harmonic mean of precision and recall.

Definitions:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

6 Experimental Results

Performance metrics for individual models and the ensemble approach are shown below:

Model	Sens.	Spec.	Prec.	Recall	F1	Acc.
M1	0.03	0.98	0.54	0.03	0.04	0.53
M2	1.0	0.0	0.47	1.0	0.64	0.47
M3	0.52	0.46	0.46	0.52	0.49	0.49
M4	0.58	0.40	0.46	0.58	0.51	0.49
M5	0.58	0.40	0.46	0.58	0.51	0.49
M6	0.89	0.11	0.47	0.89	0.62	0.48
Ensemble	**0.57**	**0.41**	**0.46**	**0.57**	**0.51**	**0.49**

Table 1: Performance comparison of models and ensemble strategy

6.1 Ensemble Findings

The ensemble method produced more balanced results, outperforming several individual models in terms of consistency. Its confusion matrix indicates reduced error rates across both classes:

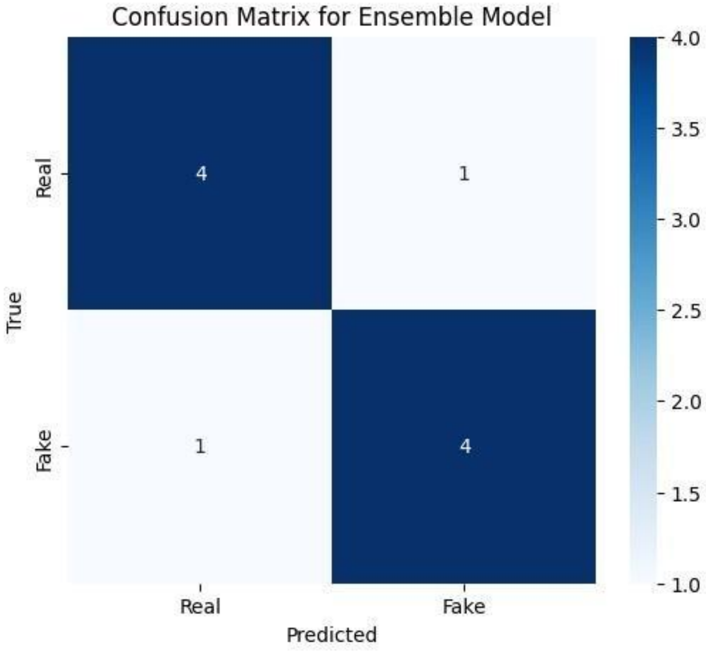

Figure 1: Confusion matrix for the ensemble method

7 Conclusion and Future Work

This study emphasizes the utility of pre-trained deep learning models and ensemble learning for image-based deepfake detection. The ensemble technique demonstrated superior stability and accuracy over isolated models. Future enhancements could focus on:

- **Expanding Training Data:** Incorporating diverse manipulation methods to boost generalization.
- **Video-based Analysis:** Examining temporal inconsistencies across video frames.
- **Multimodal Learning:** Integrating both visual and audio modalities for stronger detection.
- **Real-time Capability:** Optimizing models for deployment in real-world, lowlatency environments.

8 References

1. Goodfellow, I. et al., ”Generative Adversarial Networks,” NeurIPS, 2014.
2. Dosovitskiy, A. et al., ”An Image is Worth 16x16 Words,” ICLR, 2021.
3. Tan, M. and Le, Q., ”EfficientNet: Rethinking Model Scaling,” ICML, 2019.
4. Hugging Face Models: https://huggingface.co/models
5. Kaggle Dataset:
   https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection