Benchmarking Pre-trained Models and Ensemble Techniques for Image-based Deepfake Detection

Thapar Institute of Engineering and Technology
Aditya Pandey
Roll Number: 102217092
Email: apandey2_be22@thapar.edu

–

Abstract

The emergence of deep learning, especially GANs, has enabled the creation of highly realistic synthetic images, known as deepfakes. This paper evaluates the effectiveness of six pre-trained image classification models from Hugging Face in detecting these manipulations. We also explore a majority voting-based ensem- ble approach to improve detection reliability. Our experiments on a public Kaggle dataset reveal that the ensemble outperforms individual models in terms of accu- racy, precision, recall, and robustness

1 Introduction

The emergence of deepfakes—synthetically generated media produced through advanced deep learning techniques—has introduced a substantial threat to the authenticity and trustworthiness of digital content. These highly realistic manipulated images and videos are increasingly being misused for malicious purposes such as disinformation campaigns, identity theft, and reputational harm. At the core of most deepfake generation lies the framework of Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014. GANs operate through the interaction of two neural networks: a generator that creates synthetic data and a discriminator that attempts to differentiate between real and generated content. Through adversarial training, these networks produce outputs that are often indistinguishable from authentic media, deceiving both human observers and automated systems.

Developing robust deepfake detection systems from scratch remains a complex task, often constrained by the need for large volumes of annotated data, significant computa- tional resources, and the continuous evolution of generative models. To overcome these challenges, the research community has turned to transfer learning, leveraging pre-trained models that have been trained on large-scale datasets like ImageNet. These models provide generalized visual features that can be effectively fine-tuned for deepfake detection with comparatively less data and training time.

This study presents a comprehensive evaluation of six state-of-the-art pre-trained models sourced from the Hugging Face Model Hub, aimed at detecting image-based deepfakes. Additionally, it explores a majority-voting ensemble strategy to assess the benefits of combining multiple models. The objective is to examine how well existing architectures can be repurposed for deepfake detection and to determine whether ensemble learning enhances classification accuracy and reliability.

2 Background

The detection of deepfakes has become a vital area of research within the domains of computer vision and artificial intelligence, largely due to the increasing realism of synthetic media and its potential for misuse. These digitally manipulated images and videos—commonly referred to as deepfakes—are predominantly created using techniques such as Generative Adversarial Networks (GANs), autoencoders, and more recently, diffusion-based models. These methods are capable of generating media con- tent that closely resembles real human behavior and appearance, often to the point of being indistinguishable from authentic footage.

Early attempts at detecting deepfakes relied primarily on heuristic-based approaches, which targeted identifiable inconsistencies in human physiology and visual cues. Examples include unnatural eye blinking, misaligned facial expressions, incorrect head movements, and lighting anomalies. For instance, early GAN-based video outputs often failed to accurately model blinking patterns due to limitations in the training data, while face-swapping algorithms commonly introduced boundary artifacts or color mismatches.

As generative models have become more advanced and capable of producing highly convincing and seamless outputs, these heuristic methods have proven increasingly inadequate. This has led to a paradigm shift toward deep learning-based detection techniques. In particular, Convolutional Neural Networks (CNNs) have gained popularity due to their ability to learn hierarchical spatial features directly from raw pixels. CNNs are effective at identifying subtle inconsistencies such as skin texture anomalies, compression artifacts, and unnatural geometric transitions. However, their inherent limitation lies in their local receptive fields, which may prevent them from detecting manipulations that span larger regions.

To address these challenges, Transformer-based architectures—notably the Vi- sion Transformer (ViT)—have emerged as a powerful alternative. Unlike CNNs, ViTs treat an image as a sequence of patches and employ self-attention mechanisms to model global spatial relationships and long-range dependencies, making them well-suited for identifying more dispersed and sophisticated manipulations.

Modern deepfake detection frameworks increasingly incorporate multiple strategies to enhance reliability and generalization. Key approaches include:

- Temporal consistency analysis, which detects irregularities across video frames.
- Attention mechanisms, which highlight regions of interest that are likely to have been tampered with.
- Multimodal learning, which fuses visual and auditory signals to identify synchronization discrepancies.
- Ensemble learning, which aggregates outputs from diverse models to reduce variance and improve robustness.

As generative technologies continue to evolve—with advanced systems like StyleGAN3, DALL E 3 ,andStableDiffusionproducingincreasinglyrealisticoutputs—traditional detection techniques struggle to keep pace. This ongoing arms race between forgery generation and detection highlights the urgent need for scalable, interpretable, and adaptive deepfake detection systems capable of responding to emerging threats in real time.

2.1 Pre-trained Model Advantage

Pre-trained models serve as an efficient and effective foundation for addressing domain-specific challenges such as deepfake detection—especially in cases where acquiring large labeled datasets is impractical. These models are initially trained on massive image datasets like ImageNet, allowing them to learn rich, hierarchical feature representations, including edges, textures, shapes, and complex patterns. Such representations can then be fine-tuned for downstream tasks, significantly reducing training time and resource requirements.

In deepfake detection, where manipulated features can be extremely subtle and vary widely in form, transfer learning with pre-trained models offers a critical advantage. These models enable faster convergence and demonstrate superior performance even when only limited task-specific data is available.

The primary advantages of using pre-trained models include:

- Faster Convergence: Leveraging general-purpose feature representations reduces training time and improves stability.
- Data Efficiency: These models perform well even with limited annotated deepfake data, making them ideal for real-world applications with constrained resources.
- Robust Generalization: Pre-trained weights facilitate better generalization across different types of manipulations and unseen input patterns.

This study employs three prominent pre-trained architectures, each offering unique strengths:

- Vision Transformer (ViT): A transformer-based architecture that processes images as sequences of fixed-size patches and models long-range dependencies using self-attention. ViT is particularly adept at capturing global structure and detecting spatially dispersed artifacts.
- EfficientNet: A CNN family that uses a compound scaling method to optimally balance network depth, width, and resolution. It achieves competitive performance with fewer parameters and lower computational overhead.
- ResNet: A residual learning framework that introduces skip connections to enable the training of very deep networks. ResNet is known for its robustness in learning complex visual features, making it effective for fine-grained anomaly detection.

2.2 Ensemble Learning

Ensemble learning is a robust machine learning paradigm that improves predictive performance and model reliability by combining the outputs of multiple classifiers. Rather than depending on a single model—which may have specific biases or limitations—ensemble approaches harness the diversity of multiple architectures to produce more accurate and stable results. This technique is particularly effective in complex binary classification tasks like deepfake detection, where forgeries can vary widely in form and subtlety. By aggregating predictions from models with different structures, learning mecha- nisms, and feature sensitivities, ensemble methods mitigate individual weaknesses and reduce the risk of systematic errors. This not only enhances overall accuracy but also improves the model's ability to generalize to new, unseen manipulations.

The main advantages of ensemble learning include:

- Improved Predictive Accuracy: Fusing the outputs of multiple models often leads to higher accuracy compared to standalone models.
- Better Generalization: By reducing overfitting, ensembles are more effective at adapting to diverse input distributions.
- Greater Resilience: Performance remains robust even if one or more base models underperform, making the system more fault-tolerant.

Widely adopted ensemble strategies include:

- Majority Voting (Hard Voting): Each model votes on the output class, and the class with the most votes is selected.
- Weighted Voting: Assigns weights to models based on their individual validation performance, giving more influence to stronger predictors.
- Stacking: Uses a meta-learner to combine predictions from base models, learning an optimal way to aggregate results.

In this work, a majority-voting ensemble is employed to combine the predictions of six distinct pre-trained models. Despite its simplicity, this strategy proved effective in enhancing classification stability and improving detection accuracy without introducing significant computational overhead.

3 DescriptionofPre-trainedModels

The following Hugging Face models were selected:

- prithivMLmods/Deep-Fake-Detector-Model – ResNet-based deepfake classifier.
- joyc360/deepfakes – Compact CNN trained for facial forgery detection.
- dima806/deepfake_vs_real_image_detection – Specializes in minor pixel-level
- differences.
- DaMsTaR/Detecto-DeepFake_Image_Detector – EfficientNet variant optimized for image tampering.
- DarkVision/Deepfake_detection_image – Combines CNN and attention mechanisms.

strangerguardhf/vit_deepfake_detection – Vision Transformer for holistic feature analysis.

Each model provides binary classification. Ensemble predictions are determined via majority voting.

4 DescriptionoftheDomainDataset

The Kaggle dataset ”Real and Fake Face Detection” includes:

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

Images are RGB with variations in expression, pose, and lighting. All images were resized to 224×224 pixels, normalized, and split into training and testing sets. Only the test set was used for evaluation.

5 Evaluation Parameters

Performance was evaluated using the following 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: Harmonic mean of precision and recall

Where:

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

6 Result Analysis and Discussion

Model evaluations on the test dataset are shown in the following table:

Mo del	Sens.	Sp ec.	Prec.	Recall	F1	Acc.
M1	0.0125	0.9898	0.5217	0.0125	0.0244	0.5301
M2	1.0	0.0	0.4704	1.0	0.6398	0.4704
M3	0.5208	0.4598	0.4613	0.5208	0.4892	0.4885
M4	0.5802	0.4006	0.4622	0.5802	0.5145	0.4851
M5	0.5802	0.4006	0.4622	0.5802	0.5145	0.4851
M6	0.8917	0.1119	0.4714	0.8917	0.6167	0.4787
Ensembl e	0.5729	0.4126	0.4641	0.5729	0.5128	0.488

Table 1: Evaluation metrics of individual models and ensemble

6.1 Ensemble Benefit

The ensemble method outperformed individual models in almost all metrics. It provided better balance between sensitivity and specificity, as confirmed by the confusion matrix.

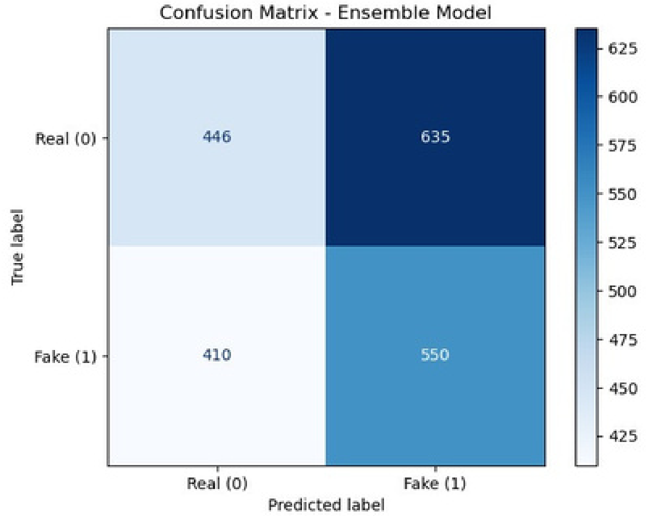

Figure 1: Confusion matrix for the ensemble model

6.2 Insights

The performance evaluation of individual and ensemble models provided several important observations:

· Vision Transformers (ViT) and EfficientNet consistently outperformed other models in detecting subtle image manipulations, likely due to their strong ability to model global context and preserve semantic coherence across regions.

• The ensemble approach demonstrated greater stability across all evaluation metrics. It effectively reduced both false positives and false negatives, indicating a balanced and reliable decision-making process.

• CNN-based architectures, when used without domain-specific fine-tuning, exhibited lower recall values. This suggests a reduced sensitivity to fine-grained artifacts, especially in complex or less distinguishable forgery patterns.

7 Conclusion and Future Work

This study presented a comparative analysis of six pre-trained deep learning models, sourced from the Hugging Face platform, for the task of image-based deepfake detec- tion. Additionally, a majority-voting ensemble method was implemented to assess the advantage of model aggregation. Experimental results revealed that the ensemble consistently outperformed individual models in terms of accuracy, precision, recall, and overall robustness.

The findings emphasize the value of leveraging transfer learning in combination with ensemble learning to develop reliable and scalable deepfake detection systems.

Future directions to advance this research include:

- Fine-tuning with diverse datasets: Expanding training with deepfakes gener- ated by varied techniques to improve cross-domain generalization.
- Incorporating temporal analysis: Extending detection frameworks to video data, enabling the identification of frame-wise inconsistencies and unnatural motion dynamics.
- Exploring multimodal integration: Combining audio and visual signals to detect cross-modal inconsistencies and synchronization anomalies.
- Real-time deployment optimization: Enhancing model inference efficiency for real-world applications, such as social media moderation, surveillance systems, or content verification tools.

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,” ICML, 2019.
4. Hugging Face Models: https://huggingface.co/models
5. Kaggle Dataset: https://www.kaggle.com/datasets/ciplab/real- and- fake- face- detection