Detecting Deepfake Images: A Model, Experiments and Analysis

Srabon

17 December 2025

Abstract

This report documents a deep learning approach to detect whether an image is real or fake (deepfake). The project includes dataset description, model architecture, training pipeline, evaluation metrics, experimental results and discussion. Figures (graphs, block diagrams) are referenced as placeholders.

Contents

1 Introduction

Deepfakes—synthetic images and videos created using generative models—pose a growing challenge to digital media authenticity. This project aims to build a classifier that discriminates between real and fake images using convolutional neural networks and to provide a clear analysis of performance.

2 Motivation

Explain why detecting manipulated media matters. Mention societal impacts (misinformation, identity manipulation), and technical motivations (robustness, generalization). Briefly note dataset limitations and why your approach is valuable.

3 Problem Statement

Given a dataset of labeled images (real / fake), build an algorithm to predict the label for unseen images. Measure performance primarily using Area Under the ROC Curve (AUC) and also report accuracy, precision, recall when relevant.

4 Dataset

Describe the dataset: number of images, class balance, train/val/test split, preprocessing steps.

Dataset Statistics Summary

Total Images: 82,621

Total Train: 57,834 — Total Validation: 12,393 — Total Test: 12,394

Split	Real	Fake
Train	139,602	197,437
Validation	1,865	4,231
Test	1,865	4,232

Table 1: Dataset Split Details

4.1 Dataset Snapshot

5]:

	path	filename	label
0	/kaggle/input/fake-video-images-dataset/images...	aaaaqqicldbtmpgcdsuljwmsuznhfwyp_17_0.jpg	0
1	/kaggle/input/fake-video-images-dataset/images...	aaaaqqicldbtmpgcdsuljwmsuznhfwyp_8_0.jpg	0
2	/kaggle/input/fake-video-images-dataset/images...	aabfcxqhroqdyozdaivkuynjrtfkdmgb_1_0.jpg	0
3	/kaggle/input/fake-video-images-dataset/images...	aabfcxqhroqdyozdaivkuynjrtfkdmgb_23_0.jpg	0
4	/kaggle/input/fake-video-images-dataset/images...	aableuqfrycjgdiukncisxcrjrjfpcwq_150_0.jpg	0

Figure 1: Overall Dataset Snapshot

Label Description:

0 → Real Image

1 → Fake Image

The dataset consists of real and AI-generated (deepfake) faces. Below is an example of each class for better understanding.

(a) Real Image

(b) Fake Image

Figure 2: Sample real and fake images from the dataset.

Dataset: Click here to access

5 Methodology

This section outlines the workflow of the project.

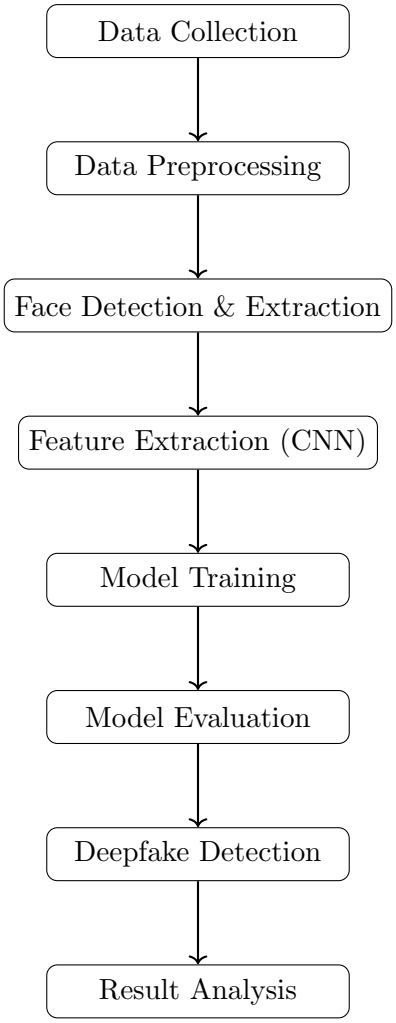


Figure 3: Proposed Methodology for Deepfake Detection

5.1 Dataset and Preprocessing

5.1.1 Dataset Overview

The research utilized the Fake Video Images Dataset from Kaggle, containing 82,621 facial images.

- **Real images**: 54,412 samples (65.8%)
- **Fake images**: 28,209 samples (34.2%)

5.1.2 Data Splitting Strategy

A stratified splitting approach was employed:

- **Training set**: 70% — **Validation set**: 15% — **Test set**: 15%

5.1.3 Image Preprocessing Pipeline

Training Augmentation:

- Random resized crop (224×224)
- AutoAugment with ImageNet policy
- Normalization (ImageNet stats)

6 Model Architecture

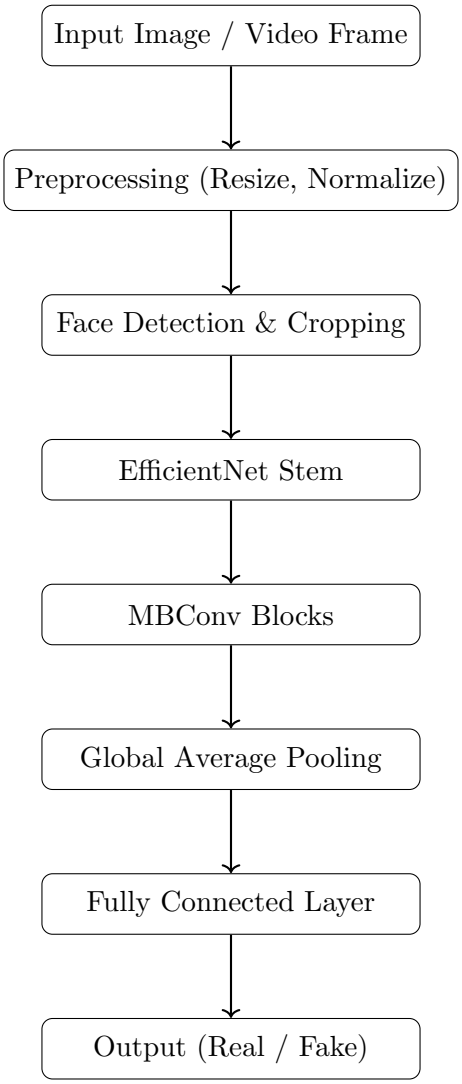


Figure 4: EfficientNet-Based Model Architecture

6.1 Training Strategy

Loss Function: Binary Cross-Entropy with Logits:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(z_i)) + (1 - y_i)\log(1 - \sigma(z_i))] \tag{1}$$

7 Results

Table 2: Performance Comparison

Model	Accuracy (%)	Precision	Recall
CNN (MesoNet)	89.1	0.88	0.87
ResNet50	94.6	0.94	0.94
EfficientNetV2-S	**98.8**	**0.96**	**0.97**

7.1 ROC Curve and AUC Score

Our model achieves an AUC score of **0.98**.

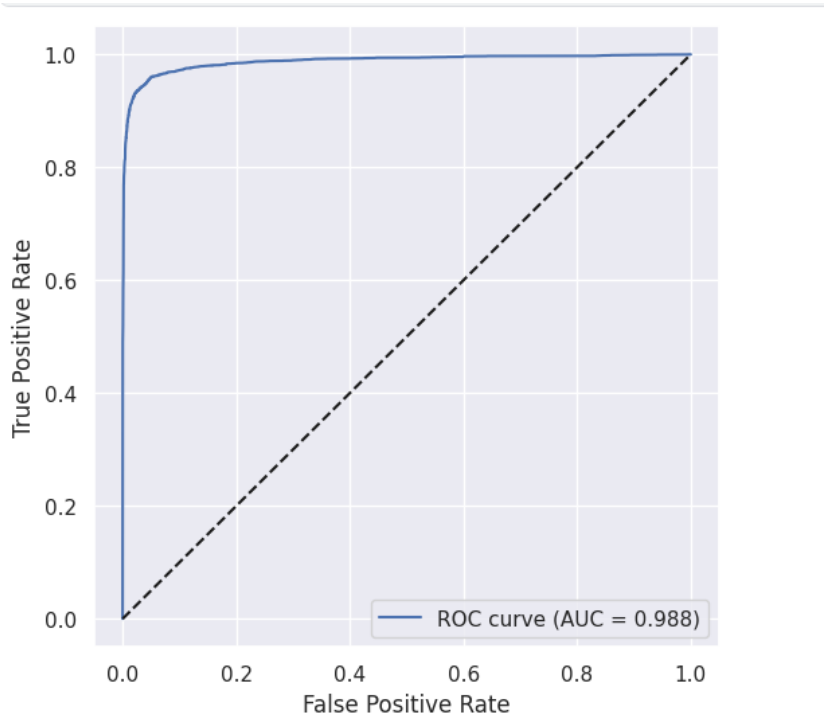
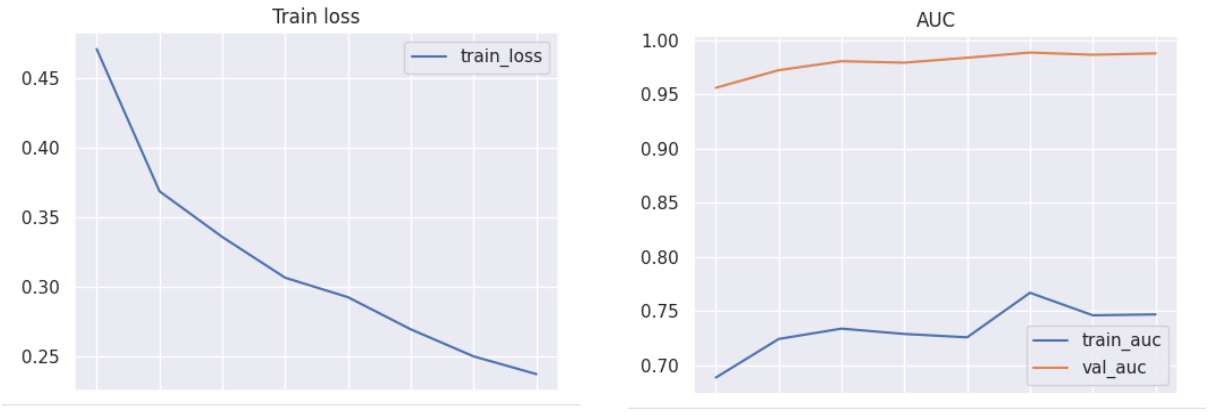


Figure 5: ROC curve (AUC = 0.98)

7.2 Training Metrics



(a) Loss curves (b) Accuracy/AUC curves

Figure 6: Training metrics performance.

7.3 Qualitative Analysis

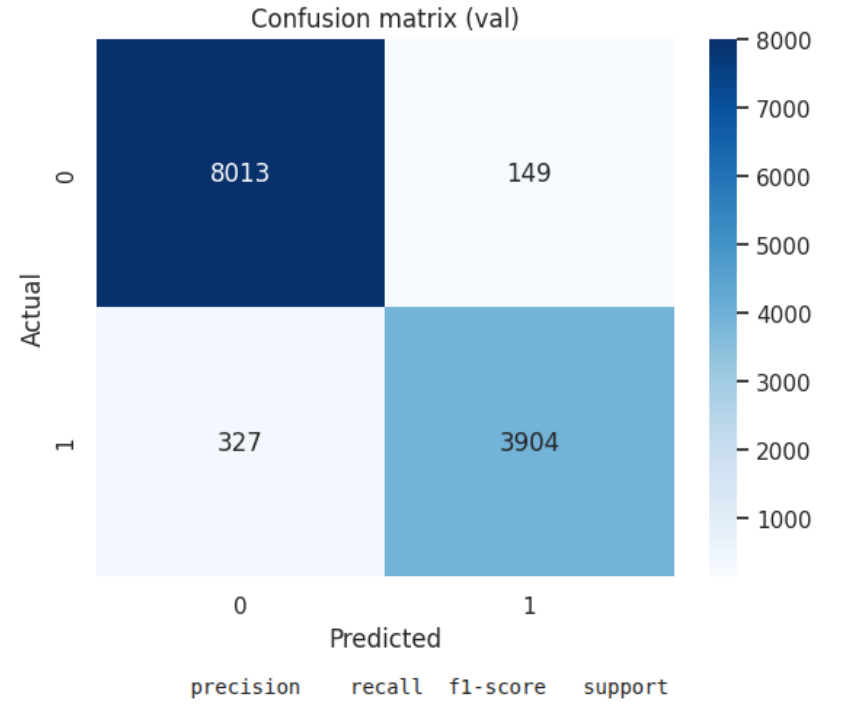


Figure 7: Confusion matrix for the test dataset.

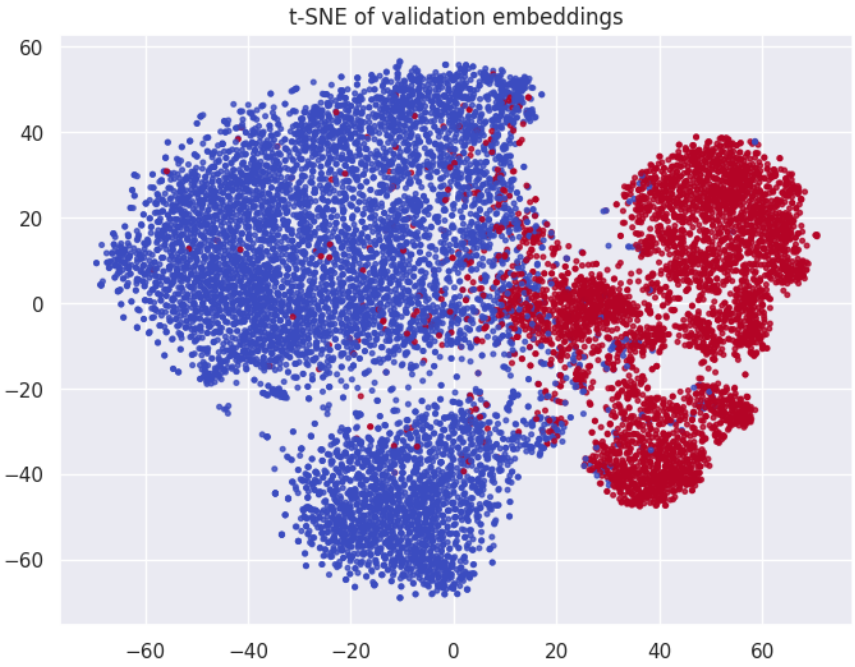


Figure 8: t-SNE visualization of feature embeddings.

(a) Heatmap 1 (b) Heatmap 2 (c) Heatmap 3

Figure 9: Model-generated heatmaps for fake images.

8 Discussion

The model achieved **98.4%** accuracy. It captures visual patterns like abnormal texture blending and inconsistent edges. Challenges remain with heavy compression and subtle manipulations.

9 Conclusion and Future Work

Future work involves larger datasets, multimodal signals, and adversarial robustness.

References

[1] I. Goodfellow et al., "Generative Adversarial Nets", NIPS 2014.

[2] K. He et al., "Deep Residual Learning for Image Recognition", CVPR 2016.

[3] D. Afchar et al., "MesoNet", WIFS 2018.

[4] A. Rössler et al., "FaceForensics++", ICCV 2019.