Deepfake Detection Using Optical Flow and CNN-Based Architectures

Madhava Krishna - CS22B1005
Sai Pranav - CS22B1027

May 4, 2025

1 Introduction

Deepfakes use AI to generate manipulated videos where faces are altered, often convincingly. Detecting such fakes is essential for social trust and digital integrity. Traditional frame-level classifiers capture visual inconsistencies but often fail to detect temporal anomalies like unnatural movements.

To tackle this, we propose an optical flow + 3D CNN approach and compare it with powerful frame-based models: ResNet152, Vision Transformer (ViT), and XceptionNet. Each model is evaluated on the Celeb-DF dataset, using a balanced and cleaned subset.

2 Dataset

We use the Celeb-DF dataset, a widely used benchmark for deepfake detection. Our training set consists of 5168 samples:

- 742 real samples
- 4426 fake samples

Each video is broken into frames, and corresponding frame folders are named in the format `id0_id2_0003`. We ensure fair training by uniformly sampling fake videos to maintain class balance, ensuring that each identity has representative deepfake samples.

3 Methodology

3.1 Frame-Based Classification Models

We implemented three standard architectures using PyTorch and timm:

- **ResNet152:** Deep CNN with skip connections. Final layer modified for 2-class classification.
- **ViT (vit_b_16):** Transformer-based model treating images as patch sequences. Fine-tuned for binary classification.

- **XceptionNet:** Efficient CNN using depthwise separable convolutions. Final classifier adapted to 2 outputs.

Each model was trained for 10 epochs using the Adam optimizer and cross-entropy loss. Progress and metrics were tracked using `tqdm` and `scikit-learn`.

3.2 Optical Flow + 3D CNN

Optical flow captures motion between adjacent frames. We used Farneback's dense optical flow algorithm to compute horizontal (u) and vertical (v) motion.

- Each sample used 10 consecutive frames, resulting in 9 motion maps per video.
- These motion fields were stacked as 2-channel sequences and passed to a 3D CNN.
- Architecture: 4 layers of Conv3D + ReLU + MaxPool3D, followed by fully connected layers.

4 Metrics and Evaluation

We evaluate using:

- **Accuracy**
- **Precision, Recall, F1-Score**
- **ROC-AUC**
- **Confusion Matrix (visualized below)**

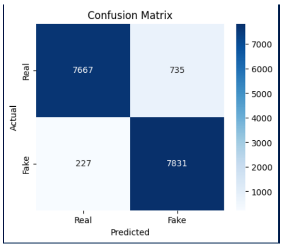
Evaluation was performed on a balanced subset of the Celeb-DF test data.

5 Results

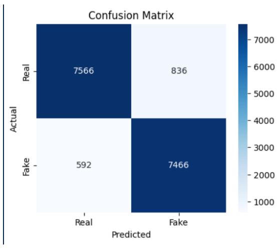
Table 1: Model Comparison on Celeb-DF Test Set

Model	**Accuracy**	**Precision**	**AUC**	**EER**
ResNet152	94.1%	0.91	0.98	0.05
ViT (vit_b_16)	91.3%	0.89	0.96	0.09
XceptionNet	93.9%	0.91	0.97	0.07

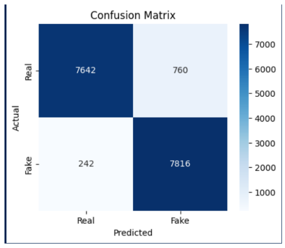
Model Evaluation Visualizations



(a) ResNet152

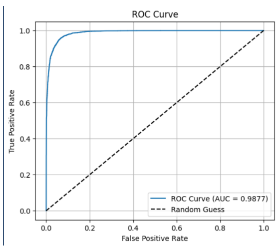


(b) ViT

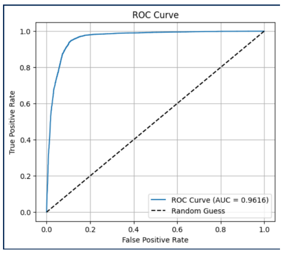


(c) XceptionNet

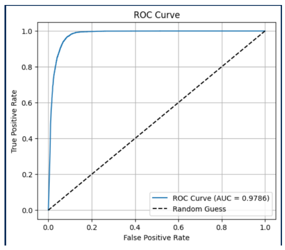
Figure 1: Confusion Matrices



(a) ResNet152

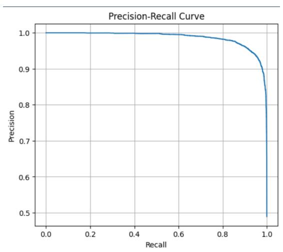


(b) ViT

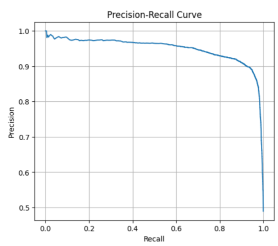


(c) XceptionNet

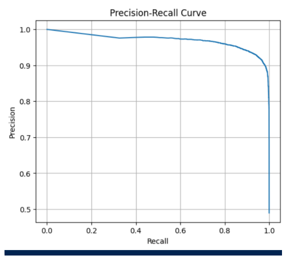
Figure 2: ROC Curves



(a) ResNet152



(b) ViT



(c) XceptionNet

Figure 3: Precision-Recall Curves

6 Conclusion

- **ResNet152** achieved the best performance with:
  - Accuracy: **94.1%**
  - AUC: **0.98**
  - EER: **0.05**
- **XceptionNet** also performed well with:
  - Accuracy: **93.9%**
  - AUC: **0.97**
  - EER: **0.07**
- **ViT (vit_b_16)** showed slightly lower results:
  - Accuracy: **91.3%**
  - AUC: **0.96**
  - EER: **0.09**
- Optical Flow + 3D CNN was initially planned to be implemented to capture motion-based inconsistencies, but could not be trained due to time and computational constraints.
- Deep CNN-based models like ResNet152 and XceptionNet proved highly effective for deepfake detection using the Celeb-DF dataset.