DeepFake Detection

IT499 - Biometric Security (Winter - 2025)

Darpan Lunagariya
Dhirubhai Ambani Institute of
Information and Communication Technology
Gandhinagar, Gujarat, India
202201462@dau.ac.in

Neel Patel
Dhirubhai Ambani Institute of
Information and Communication Technology
Gandhinagar, Gujarat, India
202201494@dau.ac.in

***Abstract*—The rapid advancement of deep learning has facilitated the creation of hyper-realistic deepfake content, posing significant threats to digital security and information authenticity. This project focuses on detecting deepfakes in both images and videos using a comprehensive approach that combines state-of-the-art convolutional neural networks and physiological signal analysis. For image-based detection, multiple pre-trained models—DenseNet, EfficientNet, InceptionV3, ResNet, and XceptionNet—were employed and evaluated for their classification accuracy and robustness. To enhance video-based detection, a photoplethysmography (PPG) technique was implemented to analyze subtle physiological cues such as heart rate signals, which are often inconsistent or absent in deepfake videos. The fusion of deep learning models with physiological signal-based analysis significantly improves detection accuracy and generalizability across datasets. This hybrid approach demonstrates promising results and offers a practical solution for real-world deepfake mitigation.**

I. Introduction

Deepfakes, a term derived from "deep learning" and "fake," refer to synthetic media where an individual's likeness is altered or replaced using advanced artificial intelligence techniques, particularly deep learning. These manipulated images and videos can appear highly realistic, posing significant challenges in distinguishing them from authentic content. The proliferation of deepfakes threatens digital security, information integrity, and personal privacy, with potential misuse in spreading misinformation, defamation, and influencing political events.

The urgent need to counter these threats has spurred research into robust detection methods. This project aims to develop a comprehensive approach for detecting deepfakes in both images and videos. For image-based detection, we leverage multiple pre-trained convolutional neural network (CNN) models, including DenseNet, EfficientNet, InceptionV3, ResNet, and XceptionNet, to identify manipulation artifacts. For video-based detection, we employ remote photoplethysmography (rPPG) to analyze physiological signals, such as heart rate, which are often inconsistent or absent in deepfake videos. By integrating deep learning with physiological signal analysis, our hybrid approach seeks to enhance detection accuracy and generalizability, offering a practical solution for real-world deepfake mitigation.

II. Literature Review

The rapid advancement of deepfake technology has prompted extensive research into detection methods, which can be categorized into image-based and video-based approaches.

A. Image-based Detection

Image-based deepfake detection focuses on identifying artifacts in static frames. Convolutional neural networks (CNNs) are widely used due to their effectiveness in image classification. Rössler et al. [1] introduced the FaceForensics++ dataset, evaluating CNN architectures for detecting facial manipulations. Afchar et al. [4] proposed MesoNet, a compact CNN model achieving high accuracy on the FaceForensics dataset. Transfer learning with pre-trained models has also shown promise; Tariq et al. [5] fine-tuned VGG16 and ResNet50, demonstrating improved performance on deepfake datasets.

B. Video-based Detection

Video-based methods often exploit temporal inconsistencies or physiological signals. Remote photoplethysmography (rPPG) has emerged as a powerful technique by detecting heart rate signals from facial videos. Fernandes et al. [3] developed DeepFakesON-Phys, using a convolutional attention network to analyze rPPG signals, achieving over 98% AUC on Celeb-DF and DFDC datasets. Ciftci et al. [6] introduced FakeCatcher, which detects synthetic videos via biological signal inconsistencies. Additionally, Li et al. [7] explored temporal artifacts, such as unnatural eye blinking, to identify deepfakes.

C. Datasets

Key datasets facilitate deepfake detection research. The FaceForensics++ dataset [1] includes 1000 videos manipulated with techniques like DeepFakes and FaceSwap. The DeepFake Detection Challenge (DFDC) dataset [2], with over 100,000 videos, features diverse facial modification algorithms, making it one of the largest public benchmarks.

This project builds on these advancements, using FaceForensics++ for image-based detection and DFDC for video-based detection, integrating CNNs and rPPG to address gaps in robustness and generalizability.

III. Proposed Approach

Our approach combines image-based and video-based detection methods, leveraging deep learning and physiological signal analysis.

A. Image-based Detection

We employ pre-trained CNN models—DenseNet, EfficientNet, InceptionV3, ResNet, and XceptionNet—fine-tuned on the FaceForensics++ dataset.

1) Data Preprocessing:

- **Dataset**: FaceForensics++ with DeepFakeDetection manipulation.
- **Frame Extraction**: Extract the 19th frame per second from videos.
- **Face Cropping**: Use MTCNN to detect and crop faces.
- **Data Split**: 70% training, 15% validation, 15% testing.

2) Model Training:

- **Models**: InceptionV3 and ResNet18, selected for performance and efficiency.
- **Optimizer**: Adam.
- **Scheduler**: ReduceLROnPlateau, adjusting learning rate based on validation loss.
- **Loss Function**: Binary Cross Entropy.

$$L = \frac{1}{N}\sum_{i=1}^{N} -y_i \, log(\hat{y}_i) - (1-y_i) \, log(1-\hat{y}_i)$$

Where, y_i is the ground truth and $\hat{y}_i$ is the probability predicted by the model.

B. Video-based Detection

We use rPPG to extract heart rate signals, detecting inconsistencies in deepfake videos.

1) Data Preprocessing:

- **Dataset**: DeepFake Detection Challenge (DFDC).
- **ROI Extraction**: Haar Cascade classifier for face detection.
- **Feature Extraction**: Mean RGB and HSV values, temporal differences, and Error Level Analysis (ELA)-inspired features.
- **Normalization**: Normalize features for consistency.

2) Model Architecture:

- **Model**: PPGNet, CNN for sequence classification.
- **Architecture**: Three convolutional blocks with batch normalization, LeakyReLU, max pooling, dropout, and residual connections; classification head with linear layers.
- **Optimizer**: AdamW with weight decay.
- **Scheduler**: CosineAnnealingLR.
- **Loss Function**: Focal Loss for class imbalance.

$$L = \frac{1}{N}\sum_{j=1}^{N} -\alpha_{t_j}(1-p_{t_j})^{\gamma}\log(p_{t_j})$$

with pt defined as,

$$pt = \sum_{j=1}^{N} yi \, . \, softmax(xi)$$

Here,
α =weighting factor = 0.25
γ =focusing parameter = 2

C. Fusion Approach

While evaluated separately, combining image-based frame analysis and video-based temporal analysis via techniques like majority voting could enhance detection, though this is left for future exploration.

IV. Experiments and Results

A. Image-based Detection

1) Experimental Setup:

- **Dataset**: FaceForensics++ (DeepFakeDetection).
- **Models**: InceptionV3 and ResNet18, pre-trained on ImageNet.
- **Training**: Batch size 32, learning rate 0.001, 20 epochs.
- **Metrics**: Accuracy, ROC curve, confusion matrix, Training History.

2) Results: InceptionV3 achieved an AUC of 0.98, and ResNet18 reached 0.97, with low false positives and negatives, indicating robust performance.

TABLE I
Image-based Detection Results

Model	Accuracy (%)	AUC	Precision	Recall
InceptionV3	98.08	0.997	0.9821	0.9815
ResNet18	91.71	0.97	0.918	0.926

Results for InceptionV3

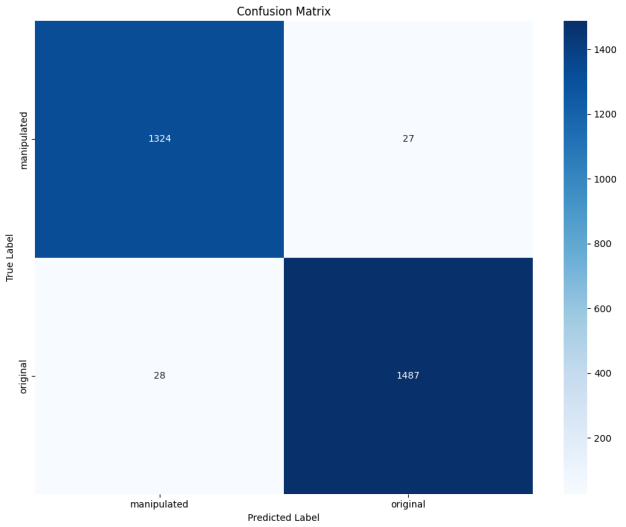

Fig. 1. Confusion Matrix for InceptionV3 Model

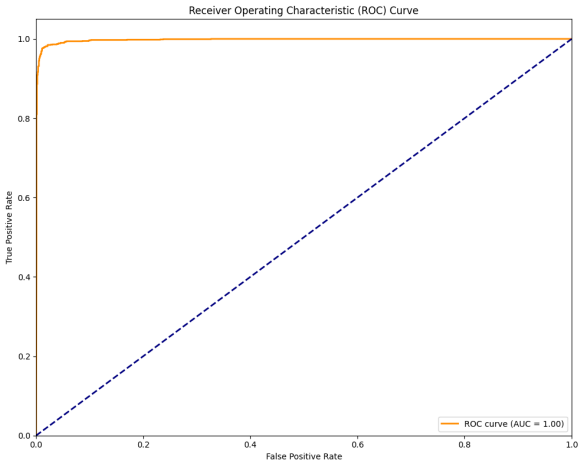

Fig. 2. ROC Curve for InceptionV3 Model

Results for ResNet18

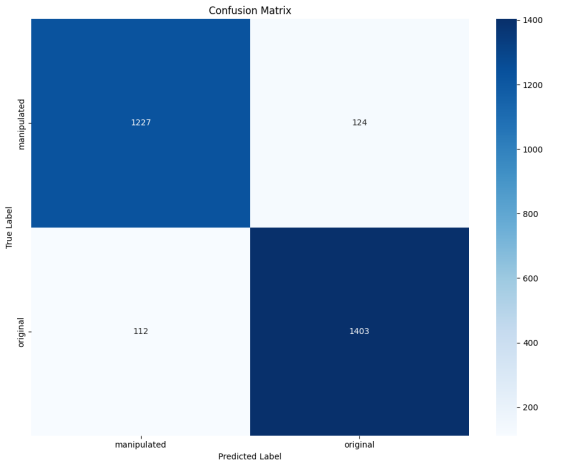

Fig. 3. Confusion Matrix for ResNet18 Model

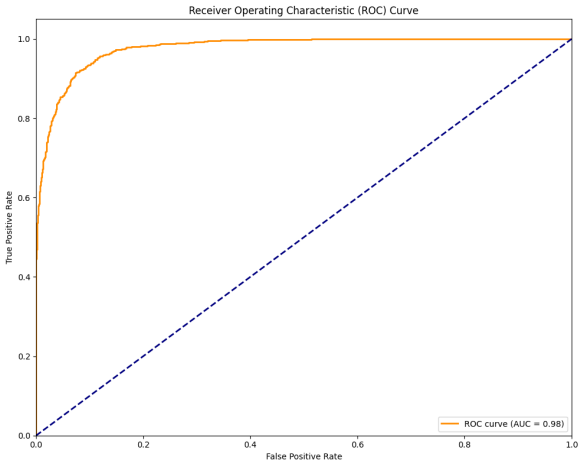

Fig. 4. ROC Curve for ResNet18 Model

B. Video-based Detection

Experimental Setup:

- **Dataset**: DFDC, 768 videos due to resource constraints.
- **Model**: PPGNet.
- **Training**: Batch size 64, learning rate 0.001, 15 epochs with early stopping.
- **Metrics**: Accuracy, precision, recall, F1-score, ROC-AUC.

Result

TABLE II
VIDEO-BASED DETECTION RESULTS

Model	Accuracy (%)	Precision	F1-score	AUC
PPGNet	55.45	0.5434	0.4827	0.577

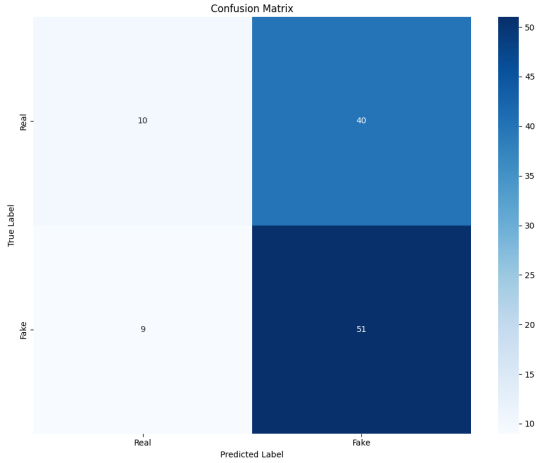

Fig. 5. Confusion Matrix for rPPG

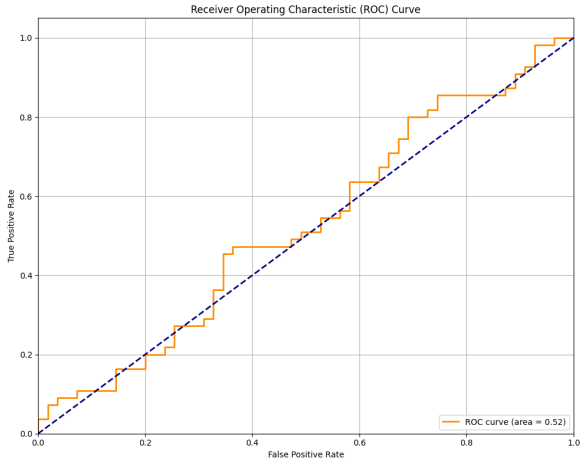

Fig. 6. ROC Curve for rPPG

V. Discussion

The PPG-based deepfake detection method showed lower performance, primarily due to several technical limitations. One major factor was class imbalance in the dataset—an unequal number of real and fake videos likely skewed the model toward the dominant class, even with the use of Focal Loss to mitigate this. Additionally, the model was trained on only 768 videos, which constrained its ability to learn robust patterns, especially given the complexity of extracting reliable PPG signals. Factors like video compression, lighting variability, and artifact noise further degraded signal quality, negatively impacting detection accuracy.

In comparison, image-based detection using pre-trained CNNs such as InceptionV3 and ResNet18 achieved significantly higher accuracies, that is, 98.08% and 91.71%, respectively. Leveraging transfer learning from large-scale image datasets, these models effectively captured manipulation artifacts in static frames, demonstrating strong performance when sufficient visual data is available.

VI. Future Work

Future work could explore strategies such as dataset augmentation, improved PPG signal extraction, and developing more robust models to address video detection challenges, enhancing overall performance.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

[2] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[3] S. Fernandes, S. Raj, A. Ortiz, and S. K. Jha, "DeepFakesON-Phys: DeepFakes detection based on heart rate estimation," *arXiv preprint arXiv:2010.00400*, 2020.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *2018 IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.

[5] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Detecting both machine and human created fake face images in the wild," in *Proc. 2nd Int. Workshop Multimedia Privacy Security*, 2018, pp. 81–87.

[6] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[7] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *2018 IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.

[8] Y. Li, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[9] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14185–14193.