




Multimodal System for Deepfake Detection



Team Members:

-  V Ashwin – 22BDS0002
-  M Thirunarayanan – 22BDS0342
-  Anand Vignesh – 22BDS0364

Introduction

- Deepfake is a media file like an image, video or audio that is altered using an artificial intelligence model
- This is done to depict another person by face swapping and mimicking expressions
- Alter a person's speech to manipulate the video and its context
- The goal is to make a manipulated media look and sound convincingly real
- Deepfakes can be used for entertainment but they are increasingly being used for malicious activities

Domain

- Artificial Intelligence
- Deep Learning – Computer Vision
- Cybersecurity
- Digital Forensics

Relevance to Industry and Society

Relevance to Society:

- Prevents the spread of misinformation and manipulation of public opinion
- Protects individuals from harassment, political smear campaigns, and identity fraud

Relevance to Industry:

- Social media platforms can automatically flag and remove harmful content
- News agencies can verify authenticity before publication
- Legal and law enforcement can use it for evidence validation
- Video streaming and content creation platforms can protect brand trust

Comparative Analysis of Models

Model Type	Strengths	Limitations	Datasets Used	ML Techniques Applied
XceptionNet	Fast inference (23ms/video) High accuracy on studio data	Fails on real-world compression	FaceForensics++ (HQ, LQ, raw)	CNN (deep convolutional)
Convolutional Vision Transformer	Good at capturing subtle artifacts	Needs very large datasets to generalize	DFDC, UADFV, FaceForensics++	Vision transformer + convolutional networks
EfficientNet-LSTM	Good Temporal consistency checks Can be mobile-optimized	Struggles with brief clips (<3sec) High RAM usage	Celeb-DF	EfficientNet (CNN backbone) + LSTM
Spatio Temporal Graph Networks	Captures spatial and temporal inconsistencies in video frames	Graph based processing requires high computing	Celeb-DF, DFDC, WildDeepfake	Spatial-Spectral-Temporal Graph Neural Network
Cross-Attentive Spatio-Temporal (CAST)	Integration of spatial and temporal cues using cross-attention	Model complexity may stop deployment	FaceForensics++, DeepfakeDetection, Celeb-DF (v2)	CNN + Transformer with cross-attention fusion

DEEPPFAKE



Datasets

Dataset Name	Fake Face Sequences	Real Face Sequences	Real Video Source	Link
DFDC	~100,000	~20,000	Volunteer Actors	Kaggle
Celeb-DF v2	5,639	590	YouTube	Kaggle
WildDeepfake	3,509	3,805	Internet	Hugging Face
Deepfake-Eval-2024	Video: 964 Image: 767 Audio: 710	Video: 1,072 Image: 1,208 Audio: 1,110	Social Media	Hugging Face

Member 1 Objectives – Anand Vignesh

1. Implement XceptionNet and testing model robustness

- Limitation: XceptionNet is trained on FaceForensics++ but fails on other recent datasets and real world uses
- Improvement: New benchmarks and wild-data to show better robustness for XceptionNet

2. Integrate confidence scoring and explainability mechanism system

- Limitation: Confidence scoring mechanism – Current detectors give binary results without reliability measures, making them less reliable in critical decisions
- Improvement: Show suspected manipulated region and give confidence score. General users can visually see the reason of model's detection

Member 2 Objectives – V Ashwin

1. Attention-based CNN Model

- Limitation: Traditional CNNs excel at local feature extraction but struggle at capturing global relationships
- Improvement: Attention mechanisms in CNNs helps the network focus on important regions, capturing both fine-grained local details and broader global context

2. Optimizing the detection model for real time inference

- Limitation: Recent deepfake detectors have high accuracy but large model size and slow inference make it impossible to use in real time deployment
- Improvement: Apply model compression techniques (pruning, quantization) to optimize models and get good accuracy

Member 3 Objectives – M Thirunarayanan

1. **Classify Generation method used in creating deepfake images**

- Limitation: Cannot know source of deepfake images
- Improvement: Usage of machine learning models to classify the deepfake generation model

2. **Website Development for Commercial usage**

- Limitation: Existing models don't focus on public usability
- Improvement: Smooth and Easy user interface. Can experiment the model

3. **Data Collection and Preprocessing. Check Generalizability**

- Data based on real world setting is needed for models like XceptionNet to generalize well. This can improve robustness of the model



Insights and Improvements

- Gap Identified: Current high-accuracy models (CAST, SSTGNN) are not deployment-friendly for real-time web services due to heavy compute requirements
- Use multiple datasets (DFDC, Celeb-DF, WildDeepfake, Deepfake-Eval-2024) to avoid overfitting to a single domain
- Usage of recent datasets and benchmarks to show robustness of model
- User-Centric Design: Confidence scores on detection, explanation visualizations - heatmaps showing manipulated areas



THANK YOU!