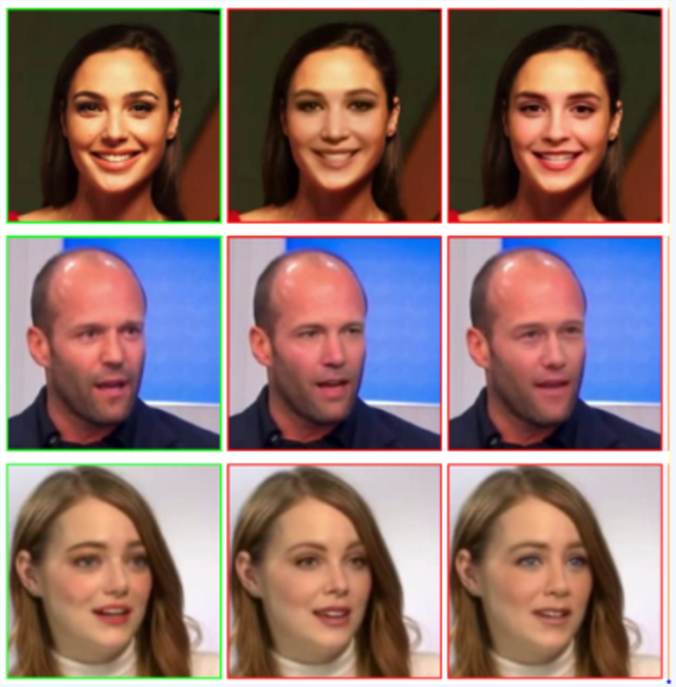


DEEPFAKE

Attacks Prevention

Akshada Malpure, Rutuja Rashinkar, Shreya Vaidya

Deepfakes: An Overview

1 **Definition**

Artificially generated media created using advanced neural networks.

2 **Functionality**

Replace or alter original images, audio, or video to fabricate new content.

3 **Applications**

Positive: Entertainment, creative media, and education

Negative: misinformation, and malicious activities

Privacy Concerns with Deepfakes

Identity Theft

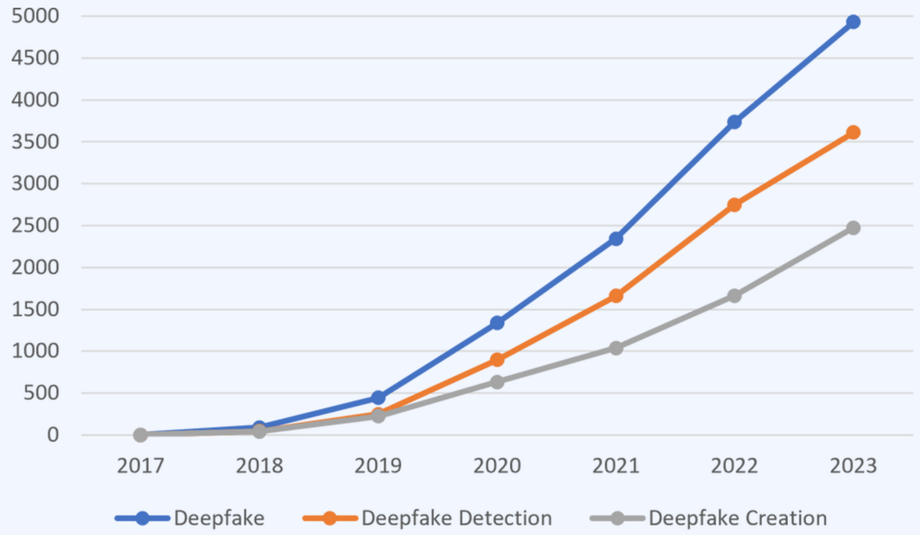
Impersonation through unauthorized use of voice or likeness.

Erosion of Trust

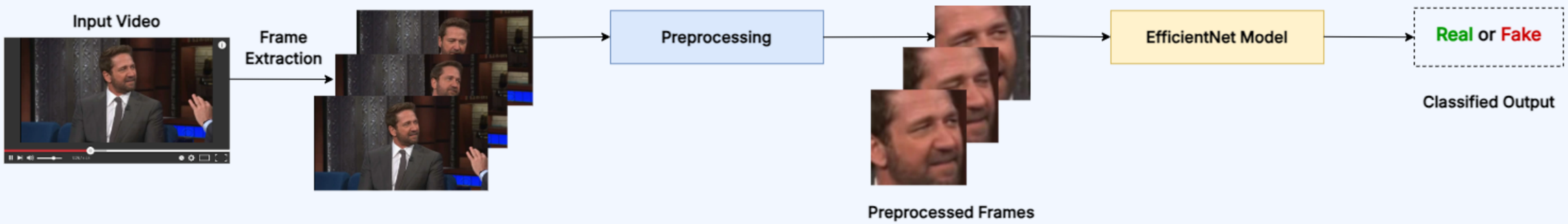
Difficulty discerning real from fabricated media undermines confidence.

Loss of Data Control

Individuals lose autonomy over their personal content.



Proposed Approach



Dataset

The DFDC Preview dataset consists of 5,000 videos.

Videos are sourced from YouTube and actors in controlled environments.

Designed specifically for training and evaluating deepfake detection models.

Preprocessing

Frames are resized to 224x224 pixels.

Data is normalized for consistency.

Augmentation (cropping, flipping and rotation) is applied.

Model

EfficientNet is used for feature extraction and classification.

Pretrained on ImageNet and fine-tuned on the DFDC dataset.

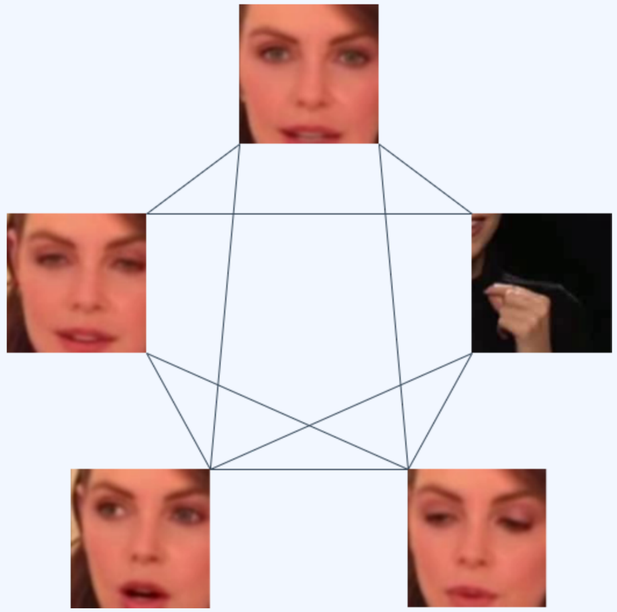
Outputs a binary classification (real or fake).

Testing Model on Adversarial Attacks

Testing models on adversarial attacks ensures their robustness, real-world applicability, and reliability in detecting manipulated content to prevent misuse and security risks.

Adversarial Attacks Implemented

1. Noise Injection
2. Pixel Modification
3. Compression & Blurring
4. Frame Manipulation



Findings and Results

	Original Data	Adversarial Data
Accuracy	0.7468	0.6122
Validation Loss	0.2444	0.5389

- The model is vulnerable to subtle perturbations from adversarial manipulations, especially in pixel areas critical for deepfake detection.
- Performance degradation underscores the need for adversarial training to improve robustness.

Limitations and Future Work

Adversarial Training

Current model lacks defenses against adversarial attacks like FGSM or PGD.

Dataset Scope

Limited to videos; multimodal deepfake detection (e.g., audio) needs to be explored.

Expand Dataset

Include larger, multimodal datasets like FaceForensics++ and Celeb-DF.

Questions?