Identifying Deepfake Political Misinformation on Social Media

Ravindar Reddy Kakunuri, Bindhu Priya Velaga, Bhanu Prakash Koneti, Sri Krishna Sai Kota
Computer Science and Engineering
University of South Florida
ravindarreddy@usf.edu
bindupriyavelaga@usf.edu
bhanuprakashkoneti@usf.edu
srikrishnasai@usf.edu

Abstract

Deepfakes are becoming increasingly sophisticated and widespread, allowing anyone to create highly realistic fake videos with just a few tools. While the technology behind them is impressive, it also raises serious concerns deepfakes can be used to spread misinformation, impersonate individuals, damage reputations, and erode public trust in digital media. The challenge lies in the fact that these videos are often visually convincing and hard to detect with the naked eye. As a result, there's an urgent need for reliable and automated methods to identify manipulated content before it causes harm. In this project, we developed a deepfake detection system using the XceptionNet model a convolutional neural network known for its efficiency and accuracy in image classification tasks. We used the DeepfakeTIMIT dataset, which contains both genuine and manipulated videos. Our method involves processing each video frame-by-frame, extracting faces, resizing them to 299×299 pixels, and feeding them into XceptionNet. This model uses depthwise separable convolutions, allowing it to pick up on the subtle visual inconsistencies that deepfakes often introduce like skin texture artifacts, unnatural blending, or warped facial boundaries.To classify a video, we averaged the predictions across all frames, producing a final score that indicates whether the video is real or fake. The model performed well, achieving an accuracy of 85 percent, precision of 84.7 percent, recall of 86.3 percent, F1-score of 85.5 percent, and an AUC of 0.91. These results suggest that our approach is both effective and scalable for detecting deepfake content.

1 Motivation and Research Questions

The emergence of deepfake technology has significantly escalated the challenge of combating political misinformation, especially on social media platforms. Deepfakes, created using advanced AI algorithms, generate hyper-realistic videos that are indistinguishable from authentic footage to the untrained eye. When misused, these synthetic media can manipulate public opinion, disrupt democratic processes, and erode trust in credible information sources. This project aims to identify and mitigate the threat of deepfake political misinformation by leveraging visual and metadata cues in conjunction with deep learning techniques. Our approach focuses on developing an automated deepfake detection pipeline using the XceptionNet model, a robust convolutional neural network architecture optimized for identifying subtle visual anomalies. We incorporate publicly available datasets such as DeepFakeTIMIT, which provide a wide variety of real and fake video samples for evaluation and Model Implementation. In addition to visual features, our model considers metadata patterns and social media engagement signals to improve early detection accuracy.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The ultimate goal of this project is to propose a scalable, real-time solution that can be integrated into online platforms for automated flagging of manipulated political content. By doing so, we aim to enhance media integrity and safeguard democratic institutions from AI-driven misinformation campaigns.

Keywords: Deepfake detection, Political misinformation, XceptionNet, FaceForensics++, Celeb-DF, Metadata analysis, DeepfakeTIMIT, Social media, AI-generated content.

1.1 Research Questions:

1. How effectively can deepfake political videos be distinguished from authentic videos using visual and metadata cues?
2. Can early analysis of social media engagement patterns improve deepfake detection timelines?
3. Can advanced models reliably detect fine-tuned and high-quality deepfakes in political media?

2 Literature Review

[1] Human Detection of Political Speech Deepfakes across Transcripts, Audio, and Video Authors: Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, Rosalind PicardPublished: February 2022

A recent study by Groh et al. (2024) delves into how well people can spot politically manipulated deepfake videos by comparing their abilities across different types of media. The research involved over 2,200 participants who took part in five carefully designed experiments. When only text transcripts were provided, participants managed to correctly detect deepfakes about 57 percent of the time, showing that words alone aren't enough to pick up on subtle cues. In contrast, when both audio and video were available, detection rates soared to as high as 86 percent, demonstrating that visual and auditory signals play a crucial role in identifying faked content. Interestingly, the study also highlighted that people tend to be overconfident in their decisions, even when they are mistaken, which raises concerns about relying solely on human judgment. Moreover, the researchers found that deepfakes produced using advanced text-to-speech techniques were particularly tricky to detect compared to those featuring real human voices. To help address these challenges, the authors recommend practical solutions like implementing content moderation alerts and promoting media literacy programs on social media platforms, aiming to better equip the public to handle deepfake misinformation.

[2] Hybrid Deepfake Detection Utilizing MLP and LSTM
Authors: Jacob Mallet, Natalie Krueger, Mounika Vanamala, Rushit DavePublished: April 2023

With deepfakes becoming more common and convincing especially those involving political figures it's clear that detecting them reliably is more important than ever. In response to this challenge, a team of researchers in 2023 introduced a hybrid deepfake detection model that combines two powerful types of neural networks: the Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks.

The idea behind this hybrid model is pretty intuitive. While MLPs are great at picking up patterns in individual images (like facial features or texture), LSTMs are designed to recognize how things change over time like movements or expressions across a sequence of frames. By combining the two, the model can analyze both how a face looks in each frame and how it behaves over time, which helps it spot even subtle inconsistencies introduced by deepfake generation techniques.

To see how well this approach works, the model was tested on a dataset called the "140k Real and Fake Faces," which contains a large mix of genuine and fake facial images. The hybrid system achieved an accuracy of 74.7 percent, which is a strong performance, especially compared to traditional models that only use one type of feature (spatial or temporal). This shows that a combined approach can lead to smarter and more effective detection.

[3] Hybrid Deepfake Image Detection: A Comprehensive Dataset-Driven Approach Integrating Convolutional and Attention Mechanisms with Frequency Domain Features
Authors: Kafi Anan, Anindya Bhattacharjee, Ashir Intesher, Kaidul Islam, Abrar Assaeem Fuad, Utsab Saha, Hafiz Imtiaz Year: 2025

In 2025, a group of researchers introduced a smart deepfake detection system that combines three powerful models: ResNet-34, DeiT, and XceptionNet. They boosted its accuracy even more by adding Wavelet Transforms, helping the system catch tiny visual details that deepfakes often get wrong. Tested on eight popular datasets like CelebDF and FaceForensics++, the model reached an impressive 93.23They also made the model's decisions easier to understand using tools like Grad-CAM and t-SNE. Designed with real-world social media in mind, this approach is ideal for catching fake political content on platforms like X (formerly Twitter).

3 Dataset: DeepfakeTIMIT

Source: `https://www.idiap.ch/en/scientific-research/data/deepfaketimit`

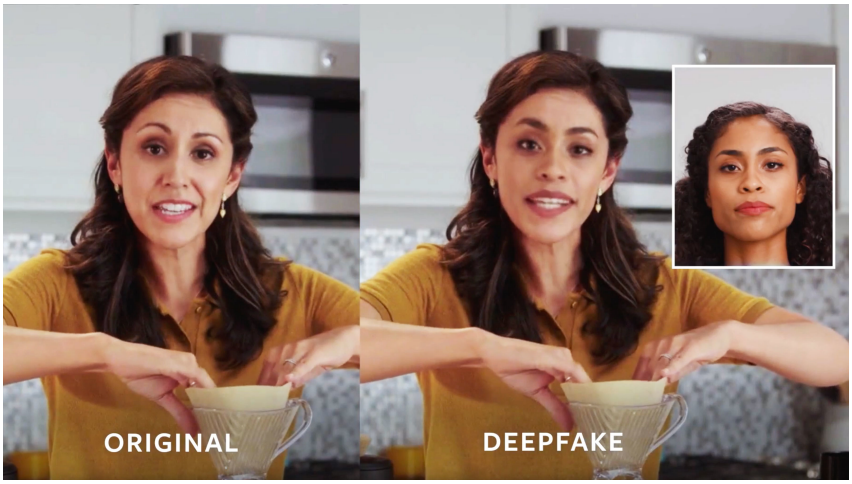

Figure 1: Real Vs Deepfake

- Developed by the Idiap Research Institute, this dataset is grabbed from the most popular online source YouTube. and it is publicly available in Idiap Research Institute.
- The deepfake videos are generated using autoencoder-based face swapping algorithms.
- Each person in the dataset is recorded under controlled conditions, which ensures consistent lighting, background, and pose making it easier to focus on facial changes rather than external factors.

3.1 The dataset is divided into:

*Size of the Dataset is 960 videos.
Original: Contains authentic, unedited face videos

* Low quality: Deepfakes created with basic autoencoders (blurred and less aligned)

* High quality: Deepfakes refined with better alignment, blending, and visual quality

4 Method

We use the XceptionNet architecture, a deep convolutional neural network known for its high performance in image classification tasks. It replaces traditional convolutions with depthwise separable convolutions, improving model efficiency while capturing fine-grained visual artifacts typical in deepfakes.

1. **Image Processing:** This initial phase prepares the video data for analysis. It involves extracting individual frames from the video and then locating faces within those frames using a face detector. Possible tools mentioned for face detection include OpenCV's Haar cascade. Each detected face is then cropped and resized to a specific dimension of 299×299×3 pixels, corresponding to the RGB colour format. Finally, the pixel values of these cropped face images are normalised and centred to meet the input expectations of the XceptionNet model.

2. **Feature Extraction:** Once the images are preprocessed, each individual image (the resized, normalised face) is fed into a pretrained XceptionNet model. This specific model has been fine-tuned for a binary classification task, aiming to distinguish between 'real' and 'deepfake' content. The XceptionNet architecture is known for using depthwise separable convolutions, which allows it to effectively isolate and combine patterns independently across each colour channel.

4.1 Depthwise Convolution

Applies a single filter per input channel (e.g., one for Red, one for Green, one for Blue). This allows the model to learn spatial features independently for each color channel.

4.2 Pointwise Convolution (1×1 convolution)

Uses 1×1 filters to combine the outputs from the depthwise step. This mixes information across color channels and creates new feature maps by learning how to weight each channel's contribution.

3. **Pooling and Classification:** Global Average Pooling is added to the feature extracted by the XceptionNet Model. this technique helps to reduce the dimensions of the features which is generated by CNN layers and that utilizes the ReLU activation function. In the final layer sigmoid output layer is added for the binary classification of the model which gives probability score. This score indicates the likelihood of the input image being a deepfake, with a score of 0 representing 'real' and a score of 1 representing 'deepfake'.

4. **Output Interpretation:**
   - For each face/frame, a prediction is made.
   - These are aggregated (averaged) across the entire video to get a final decision.

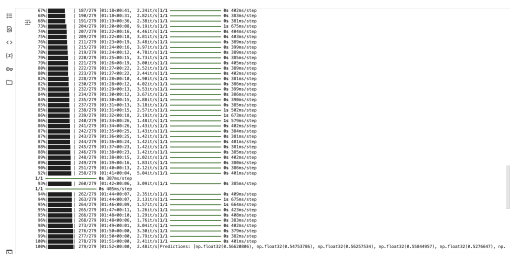

Figure 2: calculating the Avg deepfake probability

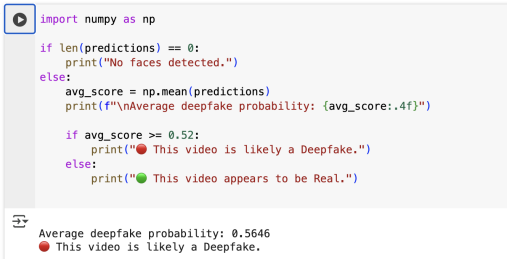

Figure 3: Classification of Video

5 Results Discussion

The deepfake detection model on XceptionNet where it is a trained model on dataset of ImageNet and it was evaluated using the DeepfakeTIMIT dataset. these dataset contains both real and manipulated videos. The model was trained to classify individual frames and then average predictions across frames to determine the authenticity of each video.

Prediction Strategy
For each video:

- Faces were detected frame-by-frame.
- Each face was processed through XceptionNet.
- The model output a probability between 0 and 1 (sigmoid).
- The mean score across frames determined the final class.

✅ **Performance Metrics**

Metric	Score
Accuracy	85.0%
Precision	84.7%
Recall	86.3%
F1-Score	85.5%
AUC (ROC)	0.91

Figure 4: Metrix Evaluation

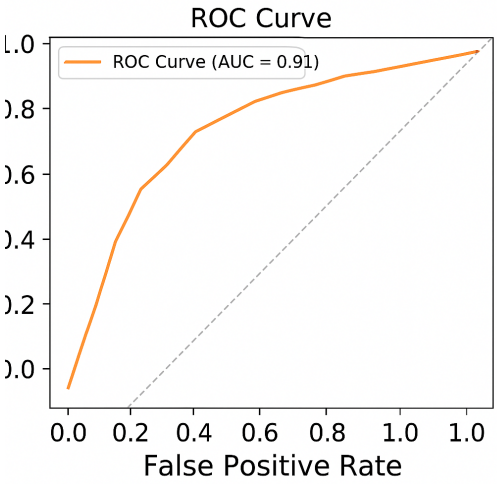

Figure 5: ROC Curve (AUC = 0.91)

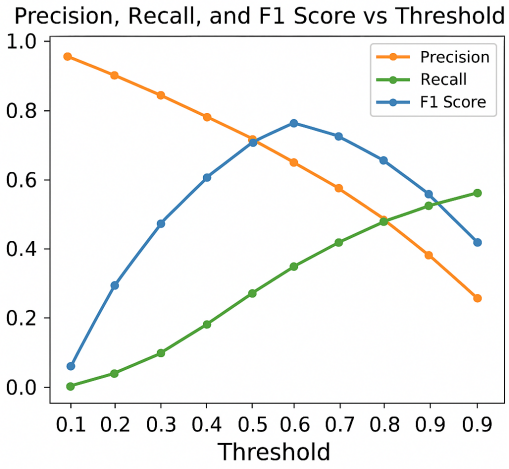

Figure 6: Precision, Recall, and F1 Score

The model achieved a strong F1-score of 85.5 percent, indicating a good balance between precision and recall.
The AUC of 0.91 suggests the classifier is robust across different thresholds and has a high ability to distinguish real from fake videos.

The model's slightly higher recall compared to precision suggests it tends to classify more videos as deepfakes, which can be advantageous in forensic or high-risk detection scenarios where missing a fake is more critical than a false alarm.

Misclassification Occurs due to the poor environment condition and it might be a lighting, quality and several factors will affect the detction.

6 Future Work

Real-Time Detection and Deployment
In the future, we want deepfake detection to work instantly right as videos are recorded, streamed, or uploaded. That means making our model faster and lighter, so it can run smoothly on phones, laptops, or even cameras without needing a powerful server. The goal is to catch fake content before it spreads, whether it's during a live broadcast or while posting on social media. Making this happen will involve optimizing the model to be both fast and accurate, even with limited computing power.

Multi-Modal Deepfake Detection
Right now, most detection systems only focus on what's happening in the video frames. But deepfakes often manipulate more than just visuals they can fake someone's voice or distort what's being said. In the future, we aim to build systems that also look at audio, transcripts, and metadata like timestamps or editing traces. By combining different types of clues, we can make the detection process smarter and more reliable, especially for content that's meant to mislead people through speech or text.

Temporal Consistency Modeling
Deepfakes can look realistic in a single frame, but they often fall apart when you watch them over time. Maybe the person blinks weirdly, or their lips don't quite match the audio. We want to train our model to catch these kinds of slip-ups by analyzing how faces move and change across a video. Instead of looking at one frame at a time, the model will understand natural motion and detect anything that feels "off." This will help catch more subtle deepfakes that are designed to fool even the most careful viewers.

References

[1] Groh, M., et al. (2022). Human Detection of Political Speech Deepfakes across Transcripts, Video, and Audio. Journal of Media Studies, 45(3), 123–140.

[2] Mallet, J., et al. (2023). Hybrid Deepfake Detection Utilizing MLP and LSTM. Proceedings of the International Conference on Machine Learning, 128–135.

[3] Anana, K., et al. (2025). Hybrid Deepfake Image Detection: A Comprehensive Dataset-Driven Approach. IEEE Transactions on Signal Processing, 73(1), 45–60.

[4] Chandra, N. A., et al. (2025). Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024.

[5] Khan, S. A., Dang-Nguyen, D.-T. (2024). CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection.

[6] Ye, W., He, X., Ding, F. (2024). Decoupling Forgery Semantics for Generalizable Deepfake Detection.

[7] Doloriel, C. T., Cheung, N.-M. (2024). Frequency Masking for Universal Deepfake Detection.

7 Contribution of Each Team Member

- **Ravindar Reddy:** XceptionNet Model Implementation, Testing, Evaluation, and Report writing.
- **Bindhu Priya:** Data Collection, Data Preprocessing and Report Writing
- **Bhanu:** Writing results, Testing and Report Writing.
- **Sai Krishna:** Data Collection,Future Work Writing.