CYBERHACK 2025

PROBLEM STATEMENT:

Fake Narrative: Internet is used for spreading fake narrative by spreading fake news and deep fake videos (using AI). Suggest a technical solution (or algorithm) for flagging deep fake videos circulating on internet and also a technical solution for highlighting fake news.

INTRODUCTION:-

The Internet has made communication and information sharing easier, but it has also led to the spread of fake news and deepfake videos. Fake news uses false information to mislead people, while deepfakes use AI to create realistic but fake videos. These can harm public trust and spread misinformation. To tackle this, technologies like computer vision, NLP, and machine learning can help detect and prevent fake content. This document explores solutions to fight misinformation, guiding hackathon participants in building effective tools.

The solutions presented here not only aim to identify and mitigate fake narratives but also emphasize the importance of ethical considerations and user trust. Hackathon participants are encouraged to use these insights to build scalable, transparent, and user-friendly tools that restore confidence in digital media and foster a more informed society.

PROBLEM STATEMENT:

How can we develop technical solutions to:

1. Detect and flag deepfake videos circulating on the internet?
2. Identify and highlight fake news to curb the spread of misinformation?

💡INNOVATIVE SOLUTION:

Colour-coded indicator system: Our innovative idea is to develop a colour coded indicator system

AI-powered video authenticity verification systems already exist, but our idea of a colour-coded indicator system for quick assessment enhances usability and accessibility.

Existing solutions, like Google's Deepfake Detection AI, Microsoft's Video Authenticator, and Reality Defender, focus on deepfake detection but do not widely implement a real-time, color-coded display for users. Our approach makes verification more intuitive and user-friendly, especially for social media and news platforms.

We will integrate it with browser extensions, mobile apps, or video-sharing platforms to provide instant verification before users interact with or share a video.

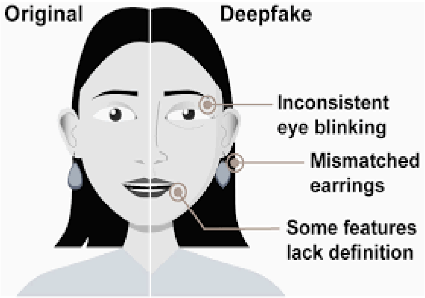
TECHNICAL SOLUTION FOR FLAGGING DEEPFAKE VIDEOS

Deepfake videos are created using AI models like Generative Adversarial Networks (GANs) to manipulate or synthesize realistic visual content. Detecting them requires a comprehensive approach that combines visual, audio, and metadata analysis.

AI-Based Deepfake Detection Algorithm

1. **Feature Analysis**

- Detect facial inconsistencies (blinking, eye alignment,distortions.
- Analyze lighting and shadow mismatches.



2. **Temporal Analysis**

- Identify flickering, jitter, and lip-sync issues.

3. **Audio-Visual Sync**

- Ensure lip movements match speech timing and tone.

4. **Machine Learning Pipeline**

- Train models on datasets like FaceForensics++.
- Use CNNs for images, RNNs/LSTMs for motion, and ViTs for deep patterns.

5. **Metadata Verification**

- Check timestamps and editing traces in video metadata.

Output:

- Assign a **truth probability score** to each video.
- Flag videos with scores below a predefined threshold for further review.

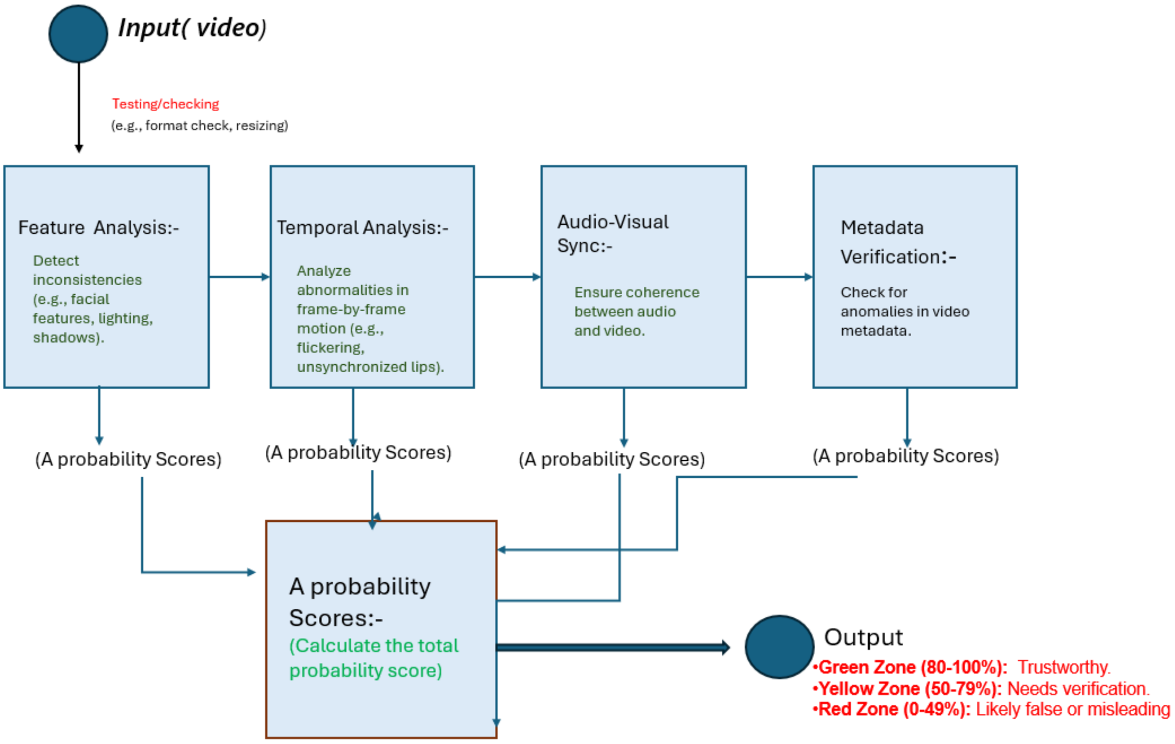

Input(video)
Testing/checking
(e.g., format check, resizing)
Feature Analysis:-
Detect inconsistencies (e.g., facial features, lighting, shadows).
Temporal Analysis:-
Analyze abnormalities in frame-by-frame motion (e.g., flickering, unsynchronized lips).
Audio-Visual Sync:-
Ensure coherence between audio and video.
Metadata Verification:-
Check for anomalies in video metadata.
(A probability Scores)
(A probability Scores)
(A probability Scores)
(A probability Scores)
A probability Scores:-
(Calculate the total probability score)
Output
•Green Zone (80-100%): Trustworthy.
•Yellow Zone (50-79%): Needs verification.
•Red Zone (0-49%): Likely false or misleading

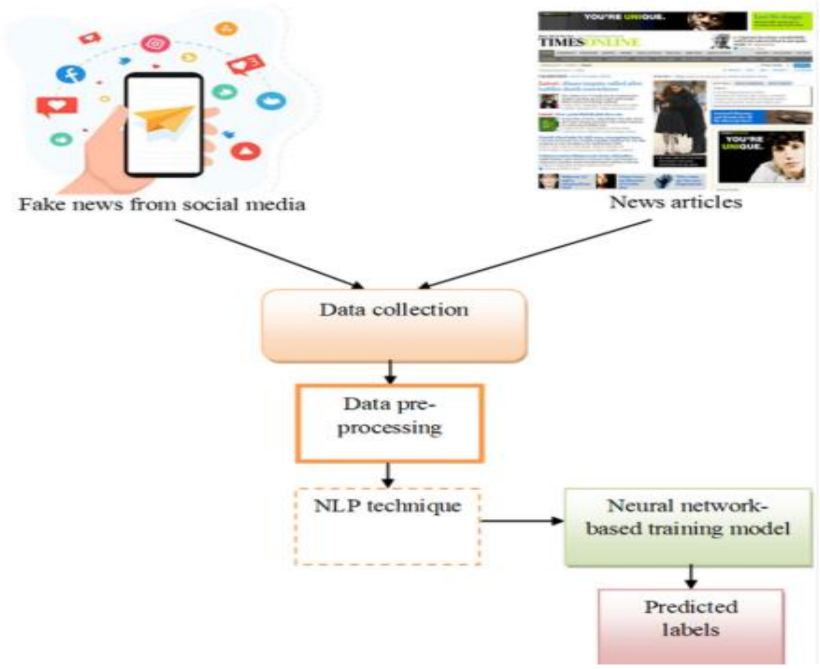
Figure 1: **AI-Based Deepfake Detection for Video**

TECHNICAL SOLUTION FOR HIGHLIGHTING FAKE NEWS

Fake news is often crafted to manipulate public perception using misleading or fabricated information. Detecting it involves analyzing the content and validating it against verified sources.

- **NLP-Based Fake News Detection Algorithm**

1. **Content Analysis (NLP)**
   - Detect sentiment polarity, writing style, and suspicious keywords.
2. **Fact-Checking via Knowledge Graphs**
   - Verify claims using databases, APIs (e.g., Google Fact Check), and fine-tuned LLMs.
3. **Source Reliability**
   - Assess domain reputation, history, and URL structure.
4. **Cross-Referencing**
   - **Compare content with trusted sources for consistency.**
5. **Network Analysis**
   - Detect bot activity and misinformation clusters on social platforms.
   - Assign a **truth probability score** to each article.
   - Highlight flagged sections and provide users with links to corroborating or refuting evidence.



https://media.springernature.com/lw685/springer-static/image/art%3A10.1007%2Fs13278-022-00995-5/MediaObjects/13278_2022_995_Fig4_HTML.png

Output:

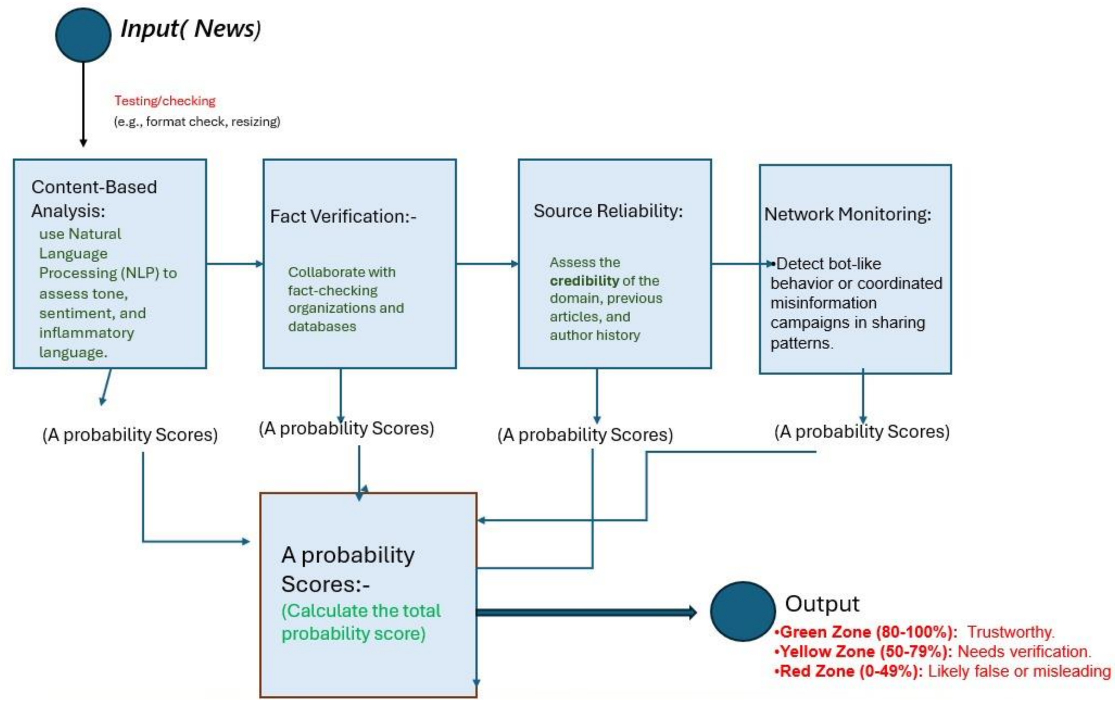

Input(News)
Testing/checking
(e.g., format check, resizing)
Content-Based Analysis:
use Natural Language Processing (NLP) to assess tone, sentiment, and inflammatory language.
Fact Verification:-
Collaborate with fact-checking organizations and databases
Source Reliability:
Assess the credibility of the domain, previous articles, and author history
Network Monitoring:
•Detect bot-like behavior or coordinated misinformation campaigns in sharing patterns.
(A probability Scores)
(A probability Scores)
(A probability Scores)
(A probability Scores)
A probability Scores:-
(Calculate the total probability score)
Output
•Green Zone (80-100%): Trustworthy.
•Yellow Zone (50-79%): Needs verification.
•Red Zone (0-49%): Likely false or misleading

Figure 2: **AI-Based Deepfake Detection for News**

Truth Probability Score Assignment:

1. **Scoring System**
   a. Use a weighted scoring model based on analysis across four domains:
      i. **Artifacts Detection (30%)**
      ii. **Temporal Analysis (25%)**
      iii. **Audio-Visual Sync (25%)**
      iv. **Metadata Verification (20%)**

Example :-

2. **Calculation**
   Consider a video evaluated on the following criteria:
   - Artifacts Detected: **Moderate inconsistencies** in facial regions and lighting → Score: 60/100.
   - Temporal Analysis: **Strong alignment** in most frames but occasional jitter → Score: 80/100.
   - Audio-Visual Sync: **Minimal misalignment** in lip movement → Score: 90/100.
   - Metadata Verification: **Clear markers of editing** detected → Score: 50/100.

Truth Probability Score Calculation:

Truth Probability Score =(0.3×60)+(0.25×80)+(0.25×90)+(0.2×50)=72.5%

- 🟢 **(Green Zone) Score 80-100**: Trustworthy
- 🟡 **(Yellow Zone) Score 50-79**: Needs verification
- 🔴 **(Red Zone) Score 0-49**: Likely False and Misleading

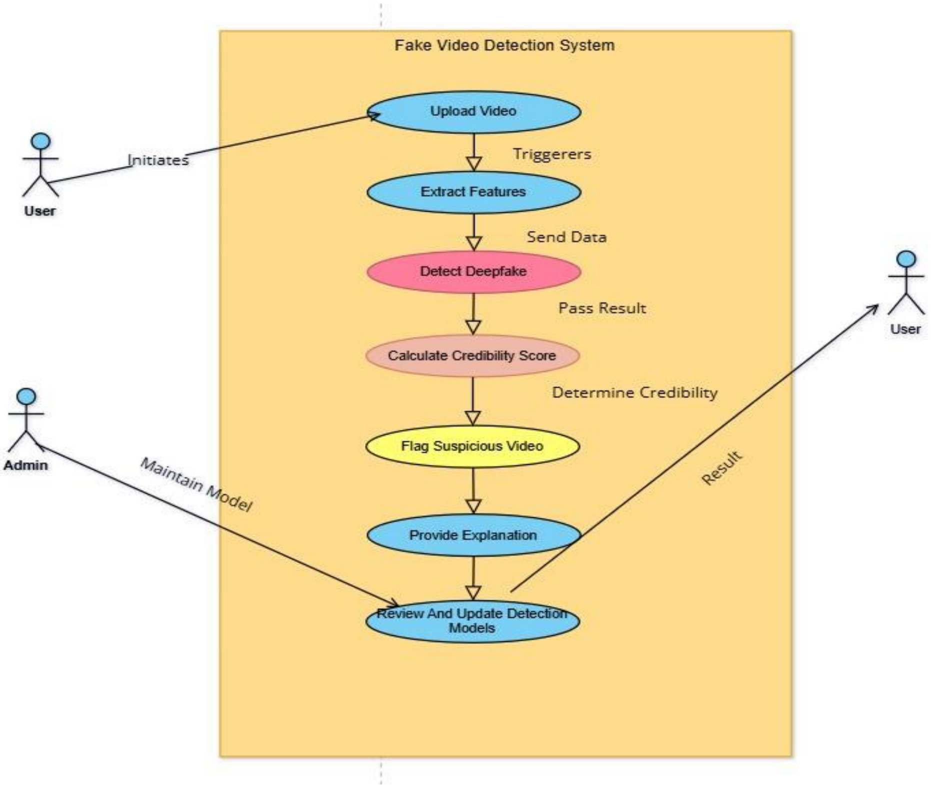


Figure 3: Use-Case Diagram for User Implementation

SOFTWARE REQUIREMENT:

1. **Programming Language:** Python 3.8+
2. **Deep Learning Frameworks:**
   a. TensorFlow 2.x or PyTorch
   b. OpenCV for image processing
   c. Dlib for facial landmark detection
3. **Machine Learning Libraries:**
   a. Scikit-learn for model evaluation
   b. NumPy, Pandas for data handling

4. **Computer Vision Tools:**
    a. MediaPipe for facial tracking
    b. Face-Recognition library for feature analysis
5. **NLP for Metadata Verification:**
    a. NLTK or SpaCy for text analysis
    b. Hugging Face Transformers for NLP-based fake detection
6. **Fact-Checking APIs (Optional):**
    a. Google Fact Check API
    b. Snopes API

3.Model Training & Datasets

1. **Datasets for Training:**
    a. FaceForensics++ – Real and fake video dataset
    b. Celeb-DF – Deepfake dataset for GAN-generated faces
    c. Deepfake Detection Challenge (DFDC) – Large-scale dataset
2. **Pretrained Models:**
    a. XceptionNet for deepfake detection
    b. EfficientNet or ResNet for image analysis
    c. Vision Transformers (ViT) for pattern recognition

4.Cloud & Deployment Services

1. Cloud Services (Optional for Large-Scale Deployment):
    a. AWS EC2 (GPU Instances)
    b. Google Cloud AI Platform
    c. Azure Machine Learning
2. **API Deployment:**
    a. FastAPI or Flask for serving the detection model
    b. TensorFlow Serving for scalable model deployment
3. **Database:**
    a. PostgreSQL or MongoDB for storing detection logs

Implementation Considerations

1. **Accuracy Improvement**
    a. Use diverse datasets to train detection models and minimize bias.
    b. Regularly update algorithms to adapt to evolving deepfake and misinformation techniques.

2. **Integration**
    a. Deploy solutions on:
        i. Social media platforms to detect and flag suspicious content.
        ii. Search engines and video-sharing platforms for real-time monitoring.
        iii. Browser plugins or mobile apps for individual users.
3. **User Education**
    a. Provide explanations for flagged content to build trust in the detection system.
    b. Encourage manual reporting of suspected deepfakes or fake news for review.
4. **Ethical Concerns**
    a. Ensure algorithms respect user privacy and avoid overreach.
    b. Maintain transparency about detection methodologies to prevent misuse or bias.

CONCLUSION:-

By leveraging advanced AI technologies like computer vision, NLP, and knowledge graphs, participants can create robust solutions to combat the spread of fake narratives on the internet. These tools, combined with user education and ethical practices, have the potential to restore trust in digital media and promote informed decision-making.

TEAM NAME: ALPHA

TEAM LEADER: Akansha A Pawar,

Semester:6th, dept: CSE , Email-id:akansha.pawar.cse@ghrce.raisoni.net

TEAM MEMBER 1: Ankit Baghel,

Semester:6th, dept: CSE, Email-id: ankitbaghel9975@gmail.com

TEAM MEMBER 2: Utkarsh Gupta,

Semester :6th, dept: CSE , Email-id:utkarshgupta2023@gmail.com