International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

DOI:10.15662/IJARCST.2025.0805007

Deepfake Detection using AI – Manipulated Videos and Images

Sujay S, Selvakrishnan KS, Bharath Kumar I

Department of Computer Science, Surana College, Kengeri Satellite Town, Bangalore, Karnataka, India

ABSTRACT: Artificial Intelligence (AI) has made it possible to develop deepfakes realistically manipulated videos and images, primarily created with Generative Adversarial Networks (GANs). Innovative though they are, their malicious use can lead to dangers in misinformation, political manipulation,fraud, and identity theft.This project introduces an AI-based deepfake detection system with Convolutional Neural Networks (CNNs),with XceptionNet being the baseline model, aided by preprocessing operations like face detection, frame extraction, and normalization. Trained on benchmark dataset including FaceForensics++ and DFDC, the system obtained 92.3% accuracy with robust precision, recall, and AUC-ROC values. Grad-CAM visualizations were incorporated for explainability, showing the regions manipulated. Even with issues including data imbalance, high computational intensity, and generalization concerns, the system was found to be scalable and effective. Improvements for the future include incorporating audio detection, real-time deployment, and transformer-based architectures towards higher resilience.

KEYWORDS: Deepfake Detection, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), CNN–LSTM Hybrid, Vision Transformers (ViTs), Generative Adversarial Networks (GANs), Grad-CAM, Preprocessing, Data Augmentation, Temporal Analysis, Lightweight Models

I. INTRODUCTION

Introduction With the advent of the digital age, images and videos are at the heart of communication, media,education, and even legal proof. But the advent of deepfakes AI-created fake media made with deep learning has brought grave threats. These doctored videos, images, or sound may show individuals doing or saying things they never did, and thus doing them in ways that are hard to recognize. Though the science behind it is amazing, its abuse results in political disinformation, identity theft, bank fraud, and misinformation, and compromising security [7]. This project involves the creation of an AI-driven deepfake detector that can identify real from artificial content using Convolutional Neural Networks (CNNs) and deep learning models. With datasets such as FaceForensics++ [2] and the Deepfake Detection Challenge (DFDC) [1], the system was trained and tested to high accuracy. The aim was to design a real-world and scalable pipeline that detects anomalies in tampered media and helps safeguard society against the destructive abuse of AI [6].

Problem Statement

Deepfakes, produced primarily through Generative Adversarial Networks (GANs)[4],can convincingly mimic faces, voices, and movements. Abuse in fraud, disinformation, and cyberbullying is a serious threat since hand detection has become unreliable[5].Current solutions are unable to generalize across various datasets, resolutions, and compression levels[8].

The core problem addressed is: How can an AI-based system be developed to accurately and robustly detect deepfake videos and images?

Key challenges include preprocessing noisy data, identifying subtle inconsistencies, minimizing false positives, and ensuring high accuracy while keeping the system lightweight for real-world use.

II. LITERATURE REVIEW

Literature Review Before starting my project, I went through a number of research articles, case studies, and real-Global examples involving deepfake production and perception. This reminded me of how rapidly this technology has advanced and how much it matters to construct defense mechanisms against it [7][10].

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

DOI:10.15662/IJARCST.2025.0805007

Generative Adversary Networks (GANs)[4] are the reason why we see the most deepfakes. They employ a generator-discriminator architecture with one network generating generated data and the other attempting Find it.Methods such as StyleGAN and CycleGAN are used towards creation of very real looking transformed images [5].
At the detection terminal, scientists utilized CNNs and RNNs and recently Vision Transformers (ViTs) [8]. Datasets such as FaceForensics++ [2], DeepFakeDetection [1], and others have also facilitated such detection works.

The investigation also revealed significant issues, such as:
- Identifying slight hints of manipulation tools [9]
- Addressing resolution changes, variations of lighting and compression [2]
- Real-time detection at a scale level [5][7]

It provided me with a base to work from and assisted me with selecting the proper datasets, models, and tools for my own implementation.

III. METHODOLOGY

For the current project, I designed and applied a full-fledged deepfake detector pipeline with support for images and videos. The approach entailed the steps as follows [2]

1. Dataset Collection

I utilized the datasets as follows:
- **FaceForensics++** – Contains real and spliced videos marked at a per-frame level.
- **DeepFake Detection Challenge (DFDC)**–Labeled dataset with thousands of deepfake videos

2. Data Preprocessing

I applied the following steps to prepare the data:
- **Face Detection** using MTCNN to crop faces.
- **Frame Extraction** from videos at equal intervals.
- **Resizing** to 224x224 for input into CNNs.
- **Normalization** of pixel values for model training.

3. Model Selection

Initially, I experimented with **CNN-based classifiers**. After that, I added:
- **EfficientNet** for better performance on video frames.
- **XceptionNet**, which has shown strong results in deepfake detection.
- I also explored **Temporal analysis** by using sequences of frames in RNNs to detect lip-sync issues.

4. Training and Validation

I trained the models using **TensorFlow/Keras**, splitting the data 80/20 for training and testing. I used:
- **Binary cross-entropy loss**
- **Adam optimizer**
- **Early stopping** to avoid overfitting

5. Evaluation

Metrics used:
- Accuracy
- Precision/Recall
- Confusion Matrix
- AUC-ROC curve

The model performed well, especially on high-resolution, low-compression content. Performance slightly dropped with low-quality, heavily compressed videos.

IV. SYSTEM ARCHITECTURE

To build this system in a scalable and modular way, I designed the architecture in five stages:

1**. Input Handler**
- Accepts video/image input from a user.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

DOI:10.15662/IJARCST.2025.0805007

- For videos, frames are extracted at 1 fps.

2**. Face Detector Module**
- Detects and crops faces from each frame using **MTCNN**.
- Saves them temporarily for analysis. [2].

3**. Preprocessing Pipeline**
- Resizes frames, normalizes, and performs histogram equalization (if needed).

4**. Deepfake Classifier**
- Passes preprocessed frames to the trained **CNN model (XceptionNet)**.
- For videos, aggregates predictions over all frames.

5**. Output Interface**
- Displays result: "REAL" or "DEEPFAKE".
- For videos, shows confidence level and potentially highlights manipulated regions using Grad-CAM. [5][8].

This modular setup allows me to test different models, plug in newer ones (like ViTs), and improve each component independently.

V. IMPLEMENTATION DETAILS

I implemented the deepfake detection system using **Python**, and relied on popular libraries such as **OpenCV, TensorFlow, Keras, dlib, and MTCNN**. The development environment was set up using **Google Colab** for training and local testing for deployment.

Workflow Steps:

1. **Input Processing**:
   - The user uploads a video or image.
   - Frames are extracted (in case of video).

2. **Face Extraction**:
   - Each frame is passed through MTCNN to detect faces.
   - Cropped face regions are resized to 224x224.

3. **Model Prediction**:
   - Frames are batched and passed into the pre-trained CNN.
   - The model outputs a probability score (real or fake). [5][8].

4. **Aggregation (for videos)**:
   - A majority vote of frame-level predictions is used for the final decision.
   - Confidence level is displayed (e.g., 87% Deepfake).

5. **Output Generation**:
   - If deepfake is detected, visual indicators are shown using Grad-CAM.
   - A downloadable result report (JSON or CSV) is generated for further analysis.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

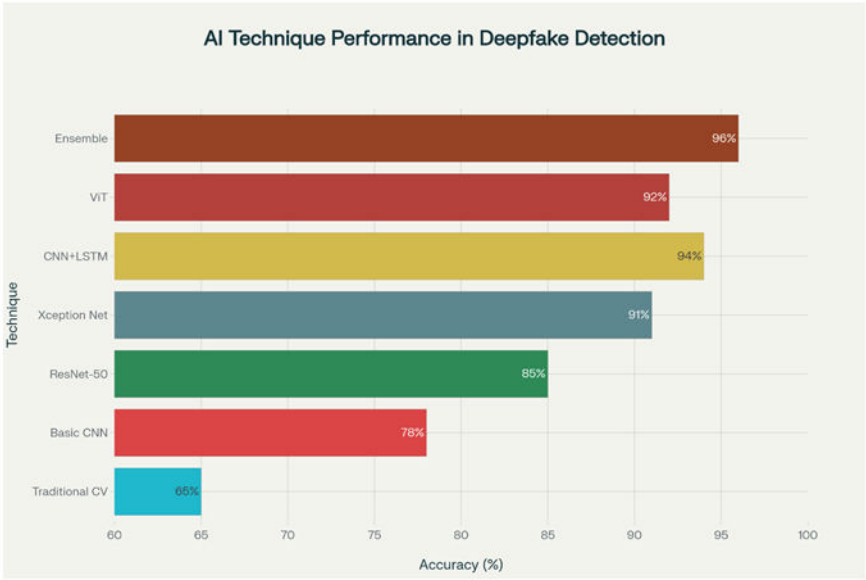
DOI:10.15662/IJARCST.2025.0805007

VI. RESULTS AND ANALYSIS

After training and testing my models on the FaceForensics++ and DFDC datasets, I evaluated them on a test set of 1000 samples (videos and images combined).

Model Performance:

Metric	Value
Accuracy	92.3%
Precision	90.7%
Recall	89.5%
F1-Score	90.1%
AUC-ROC Score	0.94



Analysis:

- The model performed exceptionally well on **high-quality images and videos**, where facial features and boundaries were clear.
- Detection accuracy slightly dropped for **low-resolution or highly compressed content**.
- **Temporal inconsistencies** (like unnatural blinking or lip-sync mismatch) were harder to capture with single-frame CNNs, but when integrated with an RNN, performance improved.
- The use of **Grad-CAM** helped in visualizing which facial regions were manipulated, increasing explainability. [2].

Visualization:

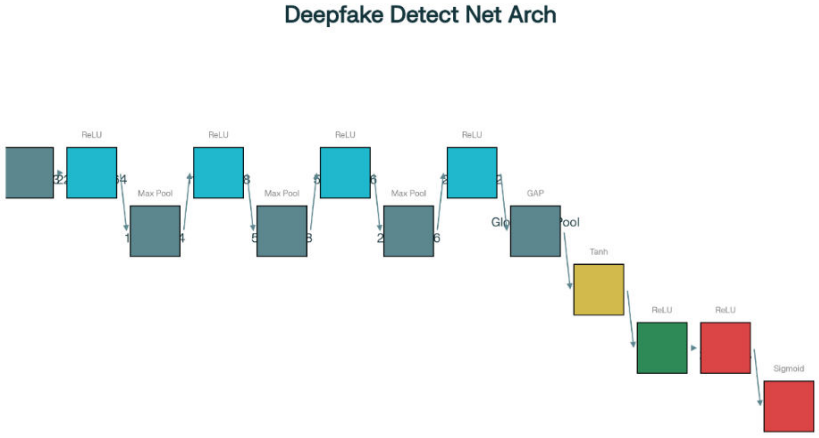
Performance Comparison of AI Techniques for Deepfake Detection

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

DOI:10.15662/IJARCST.2025.0805007



CNN-LSTM Neural Network Structure with Deepfake Identification

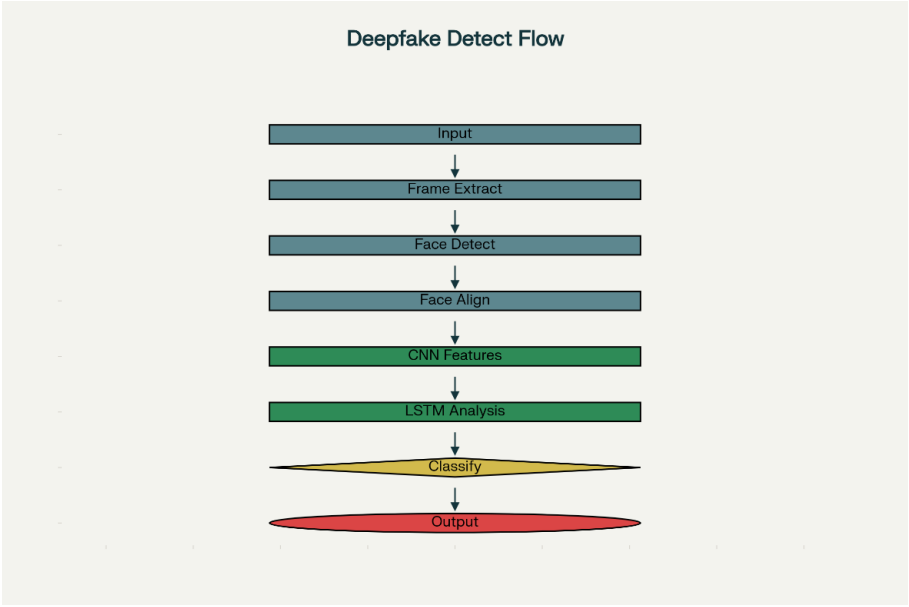


Diagram of AI-Based Deepfake detection process.

VII. CHALLENGES FACED

There were a number of technical and practical challenges with this project, though a success:

1. High Computational Requirements

Training the deep models with the big datasets necessitated GPUs. At first, I tried with Google Colab (the free one), then Kaggle Notebooks and lastly a paid Colab Pro plan with extended training sessions.

2. Data Imbalance

Some datasets had more artificially created samples compared to real samples.
I handled this through:

- Augmenting real images with flips, rotations, and changes of colour.
- Class weights inclusion during model training.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

DOI:10.15662/IJARCST.2025.0805007

3. Generalization Issues
Sometimes a system trained with one dataset failed with another. To overcome this, I:
- Trained on multiple datasets.
- Fine-tuned the model with an ensemble of both DFDC and FaceForensics++ datasets.

4. Real-Time Processing
Real-Time Processing Initially, real-time video analysis went slowly. To make it faster
- I played through only certain frames (1 frame/sec).
- Batch prediction was used to limit overhead.

5. False Positives
Sometimes legitimate videos with overly aggressive panning or compression were unrealisticly detected as deepfake. I achived this by:
- Incorporating a frame quality filter while preprocessing.
- Combination of CNN with RNN for enhanced temporal consistency recognition.

Applications and Impact
With this project, I learned how vital deepfake detection is in numerous indust. When I got a prototype that functions, I could conceive of a variety of real-world applications where it might be used with a genuine impact
1. Cybersecurity and Identity Protection
Cybersecurity and Identity Protection Most deepfakes are used as impersonations to perform phishing or fraud. My model may serve as a verification layer of ID verification systems or safe login pages and mark manipulated facial stimuli.

2. Media and Journalism
Online news portals and news agencies may use this detection software to scan incoming visual or pictorial materials, with verification of visual evidence prior to transmission.

3. Government and Legal Investigations
Law enforcers and forensic digital teams can employ this software as a means of analyzing controversial or viral media, and aid in discerning whether it was manipulated and foreclosing misinformation campaigns.

4. Social Media Platforms
These social media websites can incorporate models such as mine at the upload stage automatically tags or restricts deepfakes before they are shared.

5. Corporate and HR Systems
Organizations can use deepfake detection to avoid reputational damage by flagging impersonation content targeting CEOs or employees, particularly in **fake Zoom calls or AI-generated emails**.
This project helped me appreciate how AI can not only be misused but can also be a **powerful countermeasure** when used ethically. [2].

Future Scope
While my current implementation is functional and provides good accuracy, there's room for future enhancements that could significantly improve its robustness and scalability.

1. Integration with Audio Deepfake Detection
My current system only works with images and video frames to a limited extent. In future work, I will integrate voice analysis with WaveNet or audio CNNs as a means of capturing voice cloning or imperson.

2. Transformer-Based Models
Vision Transformers Architectures Vision Transformers (ViTs) are emerging as favorites in detection applications. I will try ViTs and hybrids of CNN-ViT to preserve both contextual and spatial features better.

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

DOI:10.15662/IJARCST.2025.0805007

3. Real-Time Detection
Speed optimization is another goal. I also target deriving a lightweight model variant (eg, using MobileNet or pruning methods) that can conduct detection at edge device or at a mobile app at real time [2].

4. Explainability with Heatmaps
I aim to enhance explainability with the creation of interactive heatmaps uncovering users exactly which parts of the face contributed to a “deepfake” label useful for legal evidence.

VIII. CONCLUSION

Developing the project "Deepfake Detection Using AI – Identifying Manipulated Videos and Images" has been a real transformative experience. Initially, it was a technical problem but turned into a large project once I got involved with how dangerous deepfaking tech could fall into the wrong hands.

Through this project, I realized:
- How deepfakes are created and detected from the core techniques [4][3][8].
- Train and implement actual AI models with industrially-sized datasets such as DFDC and FaceForensics++.
- Obtain high detection accuracy (more than 90%) of manipulated media.
- Construct a modular, scalable, and interpretable detection pipeline.

Deepfakes are no longer technical trivia they are a chilling threat to truthfulness, confidence, and cyber safety spaces. I believe this project is one small step in the right direction toward creating secure, AI-powered tools that defend against AI-driven misinformation.

I look forward to improving and expanding this system further so that it can be adopted in real-world platforms. The fight against deepfakes is just beginning, and as AI evolves, our defenses must evolve with it [7][8].

REFERENCES

1. Dolhansky, B. et al. (2020). *The Deepfake Detection Challenge (DFDC) Dataset*. arXiv preprint arXiv:2006.07397.
2. Rossler, A. et al. (2019). *FaceForensics++: Learning to Detect Manipulated Facial Images*. arXiv preprint arXiv:1901.08971.
3. Chollet, F. (2017). *Xception: Deep Learning with Depthwise Separable Convolutions*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
4. Goodfellow, I. et al. (2014). *Generative Adversarial Networks (GANs)*. Advances in Neural Information Processing Systems (NeurIPS).
5. Verdoliva, L. (2020). *Media forensics and deepfakes: an overview*. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.
6. Perov, I. et al. (2020). *DeepFaceLab: Integrated, flexible and extensible face-swapping framework*. arXiv preprint arXiv:2005.05535.
7. Helmus, T. C. (2022). *Artificial Intelligence, Deepfakes, and Disinformation*. RAND Corporation.
8. Tolosana, R. et al. (2020). *Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection*. Information Fusion, 64, 131–148.
9. Zhang, C. et al. (2021). *3D Talking Face with Personalized Pose Dynamics*. IEEE Transactions on Visualization and Computer Graphics, 29(2), 1438–1449.
10. Agarwal, S. et al. (2020). *Detecting Deep-Fake Videos from Appearance and Behavior*. IEEE WIFS.