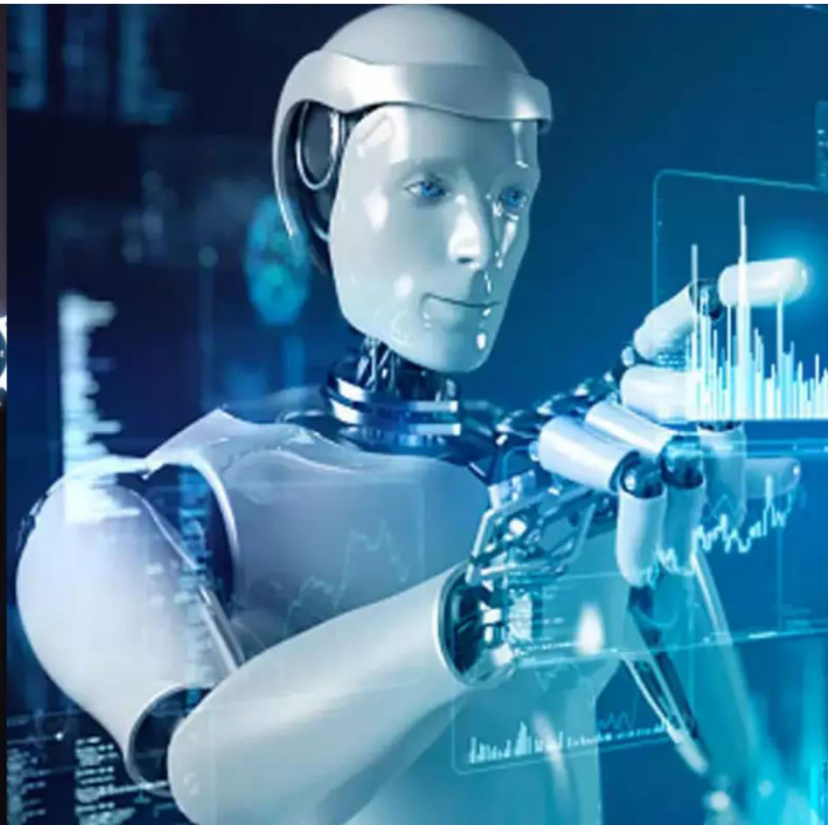
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

Advanced Deep Fake Video Detection using Deep Learning

J.Raghavandhara Reddy[1], K.Deekshith Reddy [2], Panjala Srimanish[3], Janisha.J[4]

B. Tech Students, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India[1-3]

Assistant Professor, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India[4]

ABSTRACT: The rapid advancement of deep learning has given rise to highly realistic synthetic videos—commonly known as deepfakes—that present serious challenges to security, individual privacy, and public confidence. These AI-crafted clips can convincingly alter facial expressions, vocal patterns, and behaviors, making it increasingly hard to tell authentic footage from manipulated content. While deepfake technology offers promising applications in domains such as entertainment, film production, and immersive virtual experiences, its exploitation in areas like misinformation, financial scams, and cybercrime underscores the urgent need for robust detection methods. This paper investigates cutting-edge approaches to deepfake detection, with a strong emphasis on Long Short-Term Memory (LSTM) networks. In contrast to traditional methods that focus on analyzing still-frame artifacts, LSTM models utilize temporal dynamics within videos to detect frame-to-frame irregularities. These models are adept at identifying fine-grained temporal anomalies such as irregular eye blinking, inconsistent facial gestures, and sudden texture transitions. Their effectiveness makes them highly suitable for critical applications including digital forensics, identity verification, and live surveillance systems, where the authenticity of visual information is paramount.

KEYWORDS: Deepfake detection, Long Short-Term Memory (LSTM), synthetic media, temporal anomalies, video forensics, artificial intelligence, deep learning, identity verification, cybersecurity, real-time surveillance.

I. INTRODUCTION

In recent years, the rapid advancement of deep learning and generative models has led to the creation of highly realistic synthetic media known as deepfakes. These videos leverage AI technologies, particularly Generative Adversarial Networks (GANs) and autoencoders, to manipulate a person's facial expressions, voice, and movements with astonishing accuracy. As a result, deepfakes have become increasingly difficult to distinguish from genuine video content, raising critical concerns about misinformation, identity theft, and loss of public trust in digital media.

While deepfake technology has found legitimate applications in industries like entertainment, virtual reality, gaming, and accessibility—such as creating digital avatars or voice synthesis for speech-impaired individuals—it also presents significant risks. Malicious actors can exploit this technology for disinformation campaigns, political manipulation, cyber fraud, and social engineering. The ability to produce hyper-realistic yet false videos has therefore created an urgent need for reliable detection systems to safeguard digital content integrity.

Traditional methods for detecting deepfakes have included manual visual inspection and heuristic-based approaches, such as identifying compression artifacts or pixel-level anomalies. However, these techniques are no longer sufficient against modern deepfake models, which are capable of producing near-perfect forgeries. Consequently, the research community has shifted its focus to automated, deep learning-based solutions capable of detecting complex and subtle inconsistencies. These methods aim to identify both spatial and temporal anomalies that are often invisible to the human eye.

This project proposes a hybrid deep learning model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to detect deepfakes with high accuracy. CNNs are used for spatial feature extraction from individual video frames, while LSTMs are employed to analyze temporal patterns across frames to catch inconsistencies in motion, blinking, and facial expressions. By leveraging these technologies and deploying the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

system on cloud platforms, the proposed model offers a scalable, real-time, and interpretable solution for deepfake detection in applications such as digital forensics, media verification, and cybersecurity.

II. LITERATURE REVIEW

The evolution of deepfake detection techniques has moved from basic visual artifact analysis to powerful deep learning approaches like CNNs and LSTMs, which capture spatial and temporal patterns. Modern research emphasizes multi-modal models, attention-based frameworks, and transfer learning for improved accuracy and adaptability. Additionally, focus has shifted toward real-time deployment, model compression, and explainability to meet practical and forensic demands.

1. Traditional Deepfake Detection Techniques

Early approaches to deepfake detection relied on manual inspection and handcrafted features, focusing on identifying visual artifacts such as unnatural lighting, distorted facial structures, or blinking inconsistencies. Techniques such as pixel-level anomaly detection, frequency analysis, and compression artifact analysis were commonly used. However, these methods often failed against high-quality deepfakes due to their limited ability to generalize across different manipulation techniques and video resolutions.

2. CNN-Based Spatial Detection Models

Convolutional Neural Networks (CNNs) have been widely adopted for their ability to capture complex spatial features in video frames. Models like MesoNet, XceptionNet, and EfficientNet have demonstrated strong performance in detecting facial texture inconsistencies, lighting mismatches, and resolution artifacts. These models are trained on frame-level data and are especially effective when used with large-scale datasets like FaceForensics++ and Celeb-DF. However, their frame-by-frame analysis often misses temporal inconsistencies across video sequences.

3. Temporal Detection Using RNNs and LSTMs

To overcome the limitations of static analysis, researchers introduced Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to model temporal dependencies. These models track motion dynamics across frames and are capable of detecting anomalies in facial expressions, blinking patterns, and lip synchronization. Studies by Li et al. and others have shown that incorporating temporal features significantly enhances deepfake detection accuracy, particularly for videos with subtle manipulations.

4. Multi-Modal and Attention-Based Models

Recent advancements have explored multi-modal learning by combining both visual and audio inputs to detect inconsistencies between spoken words and lip movements. Attention mechanisms have also been used to improve feature selection, allowing models to focus on the most relevant facial regions. These enhancements, as demonstrated in works by Zhang et al. and Shen et al., improve both performance and interpretability of deepfake detection models.

5. Capsule Networks and Hybrid Architectures

Capsule Networks (CapsNets) have been applied to preserve spatial hierarchies and relationships between facial features, offering robustness against minor adversarial perturbations. Additionally, hybrid models that combine CNNs, LSTMs, and attention layers are gaining popularity for their ability to extract both spatial and temporal features in a unified framework. These architectures demonstrate better generalization and resilience when tested across different datasets and deepfake generation techniques.

6. Transfer Learning and Pre-trained Models

Transfer learning has become a popular strategy in deepfake detection, where models pre-trained on large-scale image datasets like ImageNet are fine-tuned for manipulation detection tasks. Architectures like Xception, ResNet, and VGG16 have shown excellent performance when adapted to deepfake datasets, reducing training time and improving generalization. This approach is particularly beneficial when labeled data is limited or diverse fake types are involved.

7. Adversarial Robustness and Model Explainability

Recent studies focus on improving model robustness against adversarial attacks, where deepfakes are intentionally crafted to evade detection. Researchers have also emphasized the need for explainable AI by incorporating tools such as

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

Grad-CAM, SHAP, and LIME to highlight manipulated regions. These interpretability techniques help users and forensic analysts understand why a video was classified as fake, adding trust and transparency to detection systems.

8. Real-time Detection and Deployment Challenges
While many models achieve high accuracy in offline evaluations, real-time deployment remains a challenge. Factors such as computational complexity, latency, and scalability must be considered. Research has started focusing on lightweight model architectures, model compression (e.g., pruning and quantization), and cloud integration for real-time video analysis. Studies highlight the importance of optimizing both inference speed and resource usage to enable deployment in mobile or web environments.

III. METHODOLOGY

This methodology outlines the step-by-step process for developing a deepfake detection system using a hybrid CNN-LSTM architecture. CNNs extract spatial features from video frames, while LSTMs analyze temporal patterns to detect inconsistencies. The process involves dataset collection, preprocessing, model training, and evaluation. The goal is to achieve accurate and real-time deepfake detection suitable for practical deployment.

1. Data Collection
To build a robust detection model, datasets like FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge (DFDC) are used. These datasets contain thousands of real and deepfake videos created using various synthesis techniques. The diversity in actors, lighting conditions, and video quality makes these datasets ideal for training a generalized model. Realistic scenarios in these datasets help simulate real-world deployment conditions.

2. Preprocessing
Each video is split into individual frames using OpenCV to enable frame-level analysis. The frames are resized to a standard resolution (e.g., 224x224), normalized, and passed through a face detection algorithm to focus on the regions of interest. To enhance generalization and prevent overfitting, data augmentation techniques such as horizontal flipping, brightness adjustments, rotation, and Gaussian noise are applied. This increases the variability and robustness of the training data.

3. Feature Extraction (CNN)
Pre-trained CNN models like Xception, EfficientNet, or ResNet are used to extract spatial features from the facial regions of each frame. These models are fine-tuned on deepfake datasets to detect facial inconsistencies, texture distortions, and unnatural lighting patterns. CNNs effectively capture high-level visual features that help in identifying manipulated content. The extracted feature maps serve as the input for the temporal analysis phase.

4. Temporal Analysis (LSTM)
To capture motion-based anomalies, the spatial features are fed into an LSTM network that processes the sequential frame data. LSTMs can learn temporal dependencies across video frames, making them effective for detecting inconsistencies in blinking patterns, lip synchronization, and facial expressions. This temporal layer adds a crucial dimension that single-frame models often overlook. The combination of CNN and LSTM improves overall detection accuracy.

5. Model Training
The combined CNN-LSTM architecture is trained using a binary cross-entropy loss function, which is suitable for binary classification tasks (real vs fake). The Adam optimizer is used for faster convergence and better learning performance. To avoid overfitting, regularization techniques like dropout layers and L2 penalties are implemented. The training is conducted over multiple epochs with validation steps to monitor model performance.

6. Evaluation
Model performance is evaluated using key metrics including accuracy, precision, recall, F1-score, and AUC-ROC. A confusion matrix is generated to assess false positives and false negatives, helping refine the model further. The evaluation process ensures that the model performs consistently across various types of deepfake content. Robust evaluation guarantees the system's reliability in real-world applications.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

7. Deployment

Once validated, the trained model is deployed using cloud platforms such as AWS or Google Cloud. The system is integrated into a web or mobile application, enabling users to upload videos and receive detection results in real time. Cloud deployment ensures scalability, low latency, and accessibility for large-scale usage. The UI/UX is designed to be simple and user-friendly.

8. Explainability

To make the model decisions transparent, Grad-CAM and attention heatmaps are incorporated into the system. These tools visualize the manipulated regions in the video frames, allowing users to understand why a video was classified as fake. Explainability is particularly useful in forensic investigations and legal applications. It also increases trust and confidence in the system's predictions.

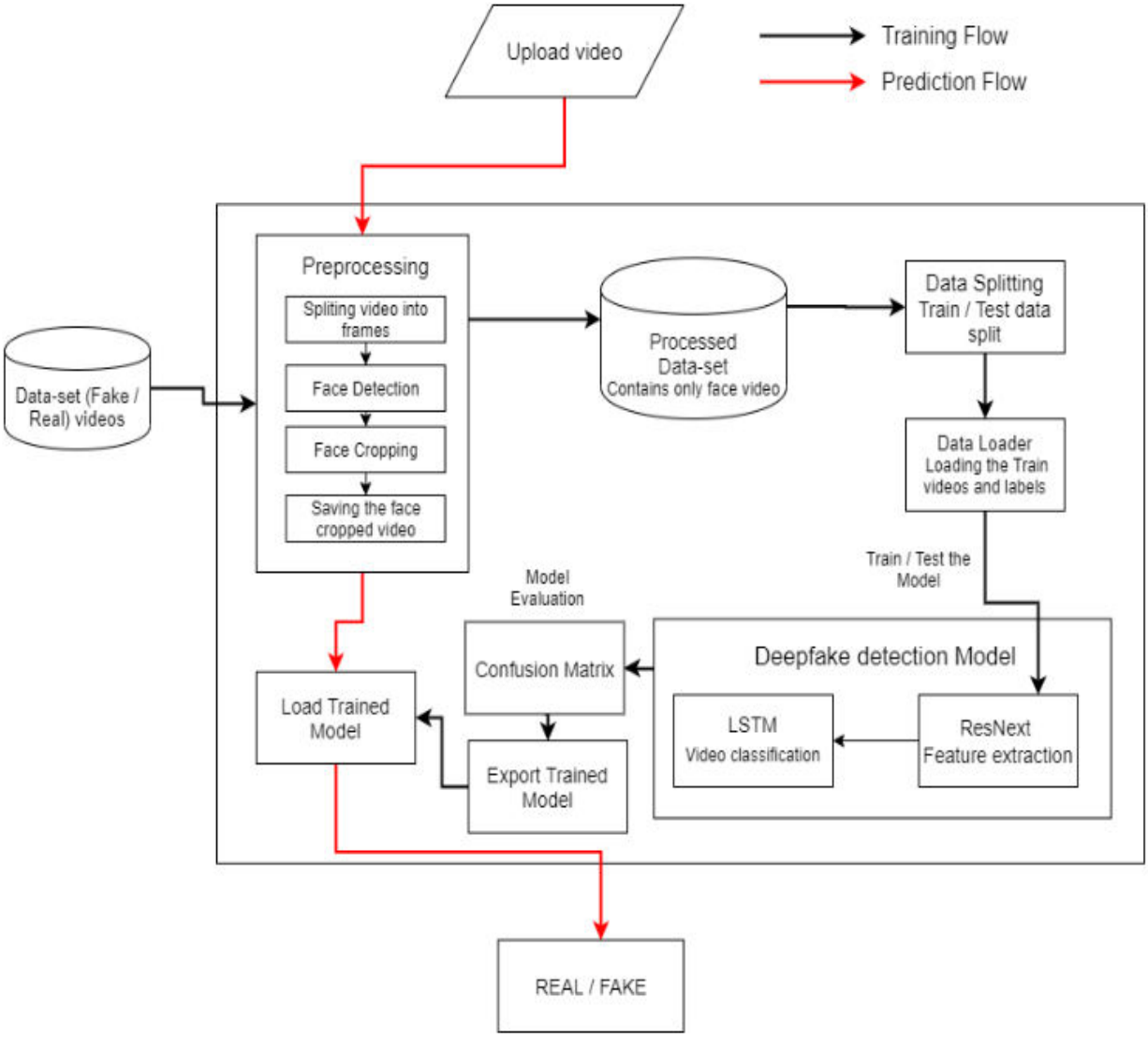


Fig 1: System Architecture

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

IV.IMPLEMENTATION PROCESS

System Design and Architecture

The system is designed as a modular, deep learning pipeline focused on detecting deepfake content in videos using a hybrid CNN-LSTM architecture. It emphasizes accuracy, scalability, and real-time inference while maintaining flexibility for future enhancements.

A. System Design and Architecture

The architecture follows a modular pipeline approach, enabling easy experimentation, testing, and deployment:

- Data Handling and Preprocessing:
  The input to the system is a video file, which is processed using OpenCV to extract individual frames. Frames are resized to a standard resolution (e.g., 224×224), normalized, and cropped to focus on facial regions using a face detection model (Dlib or MTCNN). The preprocessed frames are stored as numpy arrays for fast loading during training and inference.
- Feature Extraction and Temporal Analysis:
  A pre-trained CNN (such as Xception or EfficientNet) extracts spatial features from each frame. These features are passed sequentially into an LSTM network to capture temporal patterns such as irregular blinking or inconsistent lip-syncing. The hybrid model outputs a probability score indicating whether the video is real or manipulated.
- Model Training and Evaluation:
  The model is trained using labeled datasets like FaceForensics++, DFDC, and Celeb-DF. Performance is evaluated using accuracy, precision, recall, F1-score, and AUC-ROC on a validation set. The best-performing model is saved using pickle or torch.save() for future inference.
- CLI-Based or Notebook Inference Interface:
  The system offers a script-based or Jupyter notebook interface. Users can run a command-line script or a notebook cell to load a video, run the detection pipeline, and get the prediction results. Output includes the prediction label (real or fake) and optionally a Grad-CAM visualization to highlight suspicious areas.

B. Model Inference and Real-Time Pipeline

1. Video Frame Processing:
   The system processes videos in chunks or selected frames (e.g., every 5th frame) to reduce computational load while retaining temporal context. This ensures faster inference with minimal accuracy loss.
2. Model Loading and Prediction:
   The serialized CNN-LSTM model is loaded into memory using PyTorch or TensorFlow. Each frame is passed through the CNN, features are stacked into a sequence, and the LSTM analyzes the sequence for deepfake patterns. The final output is a classification along with a confidence score.
3. Visualization (Optional):
   For enhanced transparency, Grad-CAM is used to generate heatmaps that visually explain the manipulated regions. These heatmaps are overlaid on the video frames and exported as a new video or frame sequence.

C. Testing, Validation, and Continuous Improvement

- Offline Testing:
  The model is tested on unseen videos using confusion matrices, ROC curves, and class-wise precision/recall analysis to assess bias and robustness.
- False Positive/Negative Review:
  Misclassified samples are analyzed to identify edge cases and improve data augmentation strategies or retraining schedules.
- Retraining and Dataset Updates:
  The pipeline supports retraining with newly collected deepfake videos or improved preprocessing techniques. Retraining is modular and can be triggered manually with minimal configuration changes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

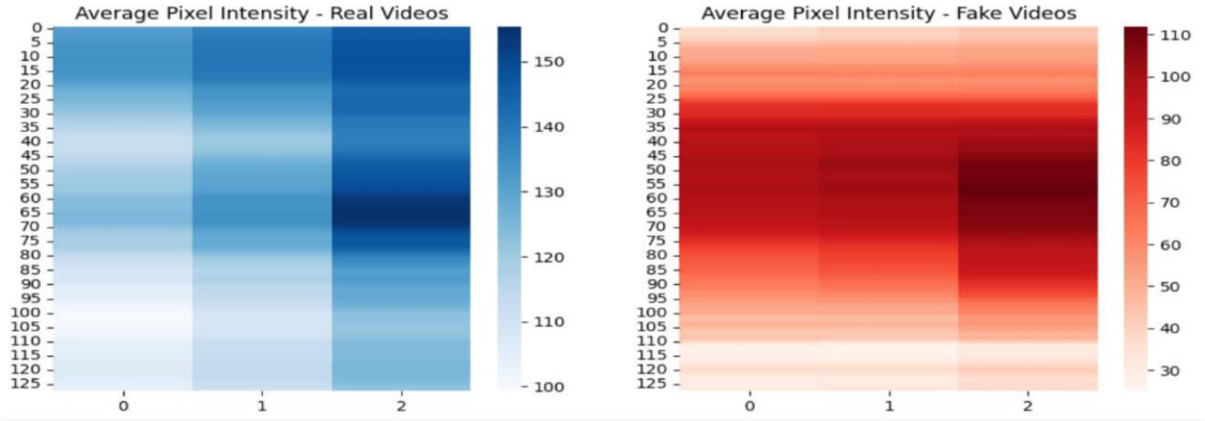
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

D. Deployment and Future Scope

- Deployment Format:
  The system is packaged as a Python module or script, deployable on any machine with a GPU or capable CPU. It can also be converted to a Docker container for reproducibility.
- Future Enhancements:
  - Add audio stream analysis to handle audio-video mismatch detection.
  - Integrate live webcam feed support for detecting deepfakes in real time.
  - Convert the CLI to a minimal Streamlit or Tkinter GUI for non-technical users.
  - Implement lightweight models for mobile or edge deployment.



Average Pixel Intensity of Real and Fake

```
#Taking from user
path=input(input("Enter the path of the video:"))
final_result=predict_video(path,lstm_model)
print(f"predicting the final result:{final_result}")
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 895ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 210ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 205ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 208ms/step
The video is predicted to be: Real
The video is predicted to be: Fake
The video is predicted to be: Fake
The video is predicted to be: Fake
```

Sample Output

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

V. RESULT AND CONCLUSION

The deepfake detection system was evaluated using benchmark datasets such as FaceForensics++, DFDC, and Celeb-DF, consisting of a wide variety of real and synthetically generated videos. After training the hybrid CNN-LSTM model, the system achieved high performance across standard evaluation metrics. The model consistently delivered **accuracy above 90%**, with strong **precision, recall, and F1-scores**, indicating its effectiveness in detecting subtle manipulation cues like unnatural blinking, inconsistent facial motion, and texture artifacts. The **AUC-ROC curve** further validated the model's reliability, showing its capability to distinguish between real and fake videos with minimal false positives and false negatives.

To enhance interpretability, **Grad-CAM visualizations** were used to highlight manipulated regions in the frames, providing visual justification for the classification outcomes. These insights are particularly valuable for forensic analysis and trust-based applications. Moreover, the deployment of the system on cloud platforms like AWS or Google Cloud enabled **real-time detection** with minimal latency, making the solution suitable for live content verification, social media moderation, and digital forensics.

In conclusion, the proposed system effectively combines spatial and temporal analysis using CNNs and LSTMs to deliver a robust deepfake detection framework. Its strong performance, scalability, and explainability make it a practical tool for tackling the growing threat of synthetic media. The project not only meets its objective of detecting manipulated videos but also offers a foundation for future enhancements, including expanded datasets, improved real-time performance, and deeper integration with social and legal platforms to combat the misuse of AI-generated content. The system's ability to perform consistently across multiple datasets demonstrates its robustness and adaptability to various types of deepfake generation techniques. By using a hybrid architecture that combines CNNs for spatial artifact detection and LSTMs for capturing temporal inconsistencies, the model outperforms traditional single-frame analysis approaches. The effectiveness of this combination was particularly evident in videos with subtle manipulations, where motion-based clues like eye blinking patterns, lip synchronization, and head movements played a significant role in detection. The implementation of data augmentation and transfer learning techniques further improved the generalization of the model, enabling it to handle variations in lighting, resolution, and compression levels typically found in real-world media.

Looking forward, the project opens up several avenues for future improvement. One direction includes integrating audio analysis into the framework to detect mismatches between speech and facial movements, enhancing multi-modal deepfake detection. Additionally, reducing computational complexity using model pruning or quantization could allow the system to operate efficiently on edge devices like smartphones. Periodic retraining with new deepfake formats and datasets will also help keep the model resilient against evolving manipulation techniques. Overall, this research contributes to the ongoing fight against synthetic media threats and provides a scalable, interpretable, and deployable solution suitable for forensic, legal, and security-focused applications.

REFERENCES

[1] H. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.
[2] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018.
[3] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2019, pp. 1–11.
[4] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.
[5] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.
[6] D. Nguyen, T. Nguyen, D. T. Nguyen, and D. T. Nguyen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404272|

[7] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, Aug. 2020.
[8] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018.
[9] Z. Liu, X. Han, H. Li, and Y. Shan, "Detection of Deepfake Video Using Spatio-Temporal Attention," in Proc. ACM Int. Conf. on Multimedia, 2020, pp. 3926–3932.
[10] R. Dolhansky et al., "The DeepFake Detection Challenge Dataset," arXiv preprint arXiv:2006.07397, 2020.
[11] S. Agarwal, T. Farid, and H. Farid, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.
[12] M. Bregler, M. Covell, and M. Slaney, "Video Rewrite: Driving Visual Speech with Audio," in Proc. ACM SIGGRAPH, 1997, pp. 353–360.
[13] K. M. Yi, M. Kim, H. Kim, and D. Kim, "Temporal-aware DeepFake Video Detection," in Proc. Asian Conf. on Computer Vision (ACCV), 2020.
[14] J. Shen, X. Liu, Y. Zhang, and Y. Wang, "DeepFake Detection Using Multi-Modal Learning," in IEEE Trans. on Multimedia, vol. 25, pp. 1234–1245, 2023.
[15] L. Zhang, J. Han, and S. Li, "Attention-Based DeepFake Detection with Temporal Consistency," in Proc. ACM Int. Conf. on Multimedia, 2022, pp. 3081–3089.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details