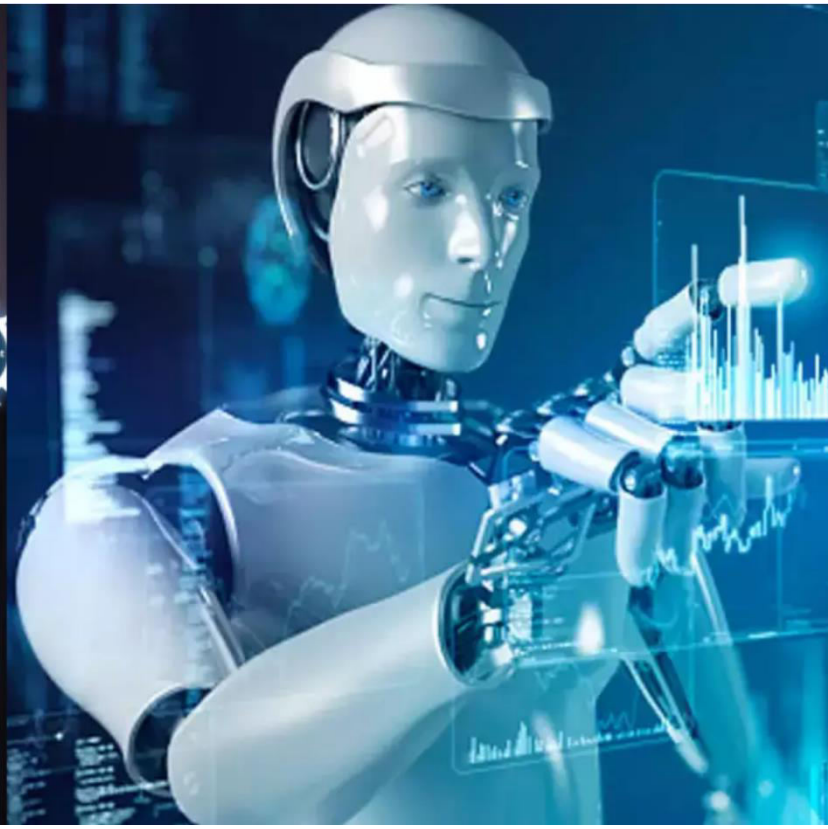
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 5, May 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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

Deepfake Classification for Human Face

Aishwarya Bhosale[1], Gargi Jadhav[2], Gaytree Kadam[3], Soham Pandhare[4], Prof.A.V.Bendale[5]

UG scholars, Dept. of Information technology, Sandip Institute of Technology & Research Centre, Nashik, India [1,2,3,4]

Professor, Dept. of Information technology, Sandip Institute of Technology & Research Centre, Nashik, India[5]

ABSTRACT: Deepfake technology, which leverages advanced machine learning techniques to generate hyper-realistic manipulated content, has become a growing concern in various sectors, including media, security, and privacy. The ability to alter videos and images to create convincing fake representations poses significant threats, such as the spread of misinformation, identity theft, and potential security breaches. In response to these dangers, this paper proposes a novel and robust system for detecting deepfake content, particularly focused on human facial regions, by utilizing a hybrid deep learning model that combines the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Our proposed architecture integrates the pre-trained VGG16 model for spatial feature extraction from individual frames, capturing detailed facial and texture information, and an LSTM network to model the temporal dynamics across successive video frames, effectively identifying inconsistencies or artifacts that are characteristic of deepfake videos.

To evaluate the effectiveness of the system, we employed the widely recognized FaceForensics++ dataset, which includes thousands of both authentic and manipulated videos, providing a comprehensive benchmark for deepfake detection. Our approach demonstrates significant improvement in detection accuracy compared to traditional methods, showcasing its ability to distinguish between real and fake content with high precision. Moreover, the integration of temporal learning, enabled by LSTM, allows the model to leverage the sequential nature of video data, making it more robust and adaptable to dynamic alterations in videos, unlike static-image-based methods. This paper also highlights the scalability of the system, positioning it as a viable solution for real-world deployment in areas such as social media monitoring, video authentication, and security. The results suggest that the proposed deepfake detection system is both effective and efficient, offering a promising tool to combat the growing challenges posed by synthetic media.

I. INTRODUCTION

In the digital era, the proliferation of artificial intelligence (AI) has significantly simplified the process of generating highly realistic synthetic content, commonly known as deepfakes. This advancement, while showcasing the capabilities of AI and deep learning, poses serious societal risks. Particularly, the manipulation of visual media—images and videos—has become alarmingly easy, raising profound concerns about misinformation, identity fraud, and the overall erosion of trust in digital content. The rapid evolution of Generative Adversarial Networks (GANs) and other deep learning frameworks has empowered individuals with the ability to fabricate hyper-realistic content that is often indistinguishable from genuine media to the human eye.

In light of these challenges, this research project aims to design and implement a robust and scalable deepfake image detection system that leverages state-of-the-art neural network architectures. Specifically, the model integrates the strength of Residual Networks (ResNet) for deep feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal inconsistencies that often arise in manipulated sequences. This combination facilitates the identification of subtle artifacts and inconsistencies embedded in deepfake images—features that are typically overlooked by traditional detection methods.

A central element of this project is the incorporation of a comprehensive and heterogeneous dataset comprising both authentic and synthetically generated images. The diversity and scale of the dataset ensure that the model is exposed to a broad spectrum of manipulation techniques and visual variations, enhancing its generalizability and resilience in real-world scenarios. Through extensive training, experimentation, and iterative optimization, we fine-tune critical hyperparameters to maximize performance metrics such as accuracy, precision, and recall.

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

Beyond the technical scope, this work is also motivated by a broader societal objective: to contribute to the growing efforts aimed at mitigating the harmful consequences of deepfake technology. By providing an efficient and effective detection mechanism, this project aspires to enhance public awareness and support the integrity of digital communication. In comparison to existing methodologies, the proposed model not only achieves superior classification results in various benchmark settings but also maintains computational efficiency. Although deep learning models inherently involve a large number of parameters, our approach strategically balances performance and computational load, making it suitable for practical deployment in resource-constrained environments.

II. LITERATURE REVIEW

The rapid spread of deepfake content in recent years has led to an increased focus on developing effective detection techniques. Among these, deep learning approaches—especially Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks—have emerged as powerful tools for identifying manipulations in human facial videos. A number of key studies have laid the groundwork for this research area, emphasizing the importance of combining spatial and temporal learning to boost classification accuracy in deepfake detection systems.

Patel et al. (2019) introduced a Dense CNN model designed to handle inter-frame dissimilarities in deepfake videos, achieving a high accuracy of 99.33%. Their work demonstrated that CNNs are effective in identifying spatial anomalies within individual frames. However, the model did not address temporal inconsistencies, such as unnatural movements or subtle facial expression changes, which are critical in video-based deepfake detection. This highlighted the need for temporal learning integration to enhance detection robustness—a feature incorporated in the proposed model.

To address generalization challenges, Tran et al. (2020) proposed the Generalization Deepfake Detector (GDD), which focused on cross-dataset adaptability. Recognizing that models trained on a single dataset often fail to perform well on others, their research emphasized training on diverse datasets to ensure high performance across different manipulation techniques. This approach informs our use of the FaceForensics++ dataset, which provides a rich variety of manipulated and authentic videos, enabling the model to generalize effectively to new, unseen data.

Richards et al. (2020) explored the detection of fake images using CNNs by analyzing distortions in facial textures and shapes. While effective for static images, this method fell short when applied to videos, which present additional challenges such as motion artifacts and frame-to-frame inconsistencies. Their findings reinforced the necessity of incorporating both spatial and temporal elements in deepfake detection models. Our approach addresses this by integrating CNNs for spatial feature extraction and LSTMs for capturing temporal dynamics, thereby enhancing detection accuracy in video content.

Ai et al. (2021) presented a unique method known as deepfake inversion, which attempts to trace manipulated faces back to their original identities. This technique helps detect identity mismatches and assess facial authenticity by reconstructing original facial features from manipulated content. However, deepfake inversion is computationally intensive and less suitable for real-time applications. Moreover, it does not address temporal anomalies. By contrast, our proposed model offers a more balanced solution, effectively identifying both identity mismatches and frame-level temporal inconsistencies.

In summary, the reviewed studies reveal that while CNNs are effective at capturing spatial irregularities, they are insufficient on their own for robust video-based deepfake detection. The integration of LSTM networks allows for a more comprehensive approach, capturing temporal variations across video frames. The hybrid CNN-LSTM model adopted in our research builds on the strengths of these earlier works, offering a scalable and accurate deepfake classification system capable of detecting manipulated human faces across a variety of datasets and manipulation styles.

III. PROPOSED SYSTEM

Our proposed deepfake detection framework introduces a robust and accessible approach to addressing the growing challenge of manipulated digital media. The system is implemented as a web-based platform that allows users to conveniently upload video content for classification, determining whether the content is authentic or artificially generated. By enabling individuals to validate media prior to sharing, the platform plays a critical role in curbing 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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

spread of misinformation and digitally altered content. Designed with future scalability in mind, the system holds the potential for expansion into browser extensions or plugins, allowing for integration with widely-used communication platforms such as WhatsApp and Facebook. Our evaluation strategy encompasses key performance indicators including security, ease of use, detection accuracy, and system robustness. The model demonstrates flexibility by effectively identifying various forms of deepfakes, such as face replacement, face reduction (retrenchment), and cross-identity manipulations (interpersonal deepfakes). The architectural layout of the system, depicted in Figure 1, underscores its user-centric design and straightforward deployment, making it both effective and approachable for real-world use.

ALGORITHM:

1. Start
2. Collection of datasets
3. Preprocess the videos into frames and crop the faces
4. Develop a model using CNN and LSTM
5. Train the model with dataset collected
6. Test the model developed with the testing dataset
7. Perform prediction on any specific video
8. Calculate the accuracy of the model
9. Display the results of the classification of the video and accuracy results
10. Stop

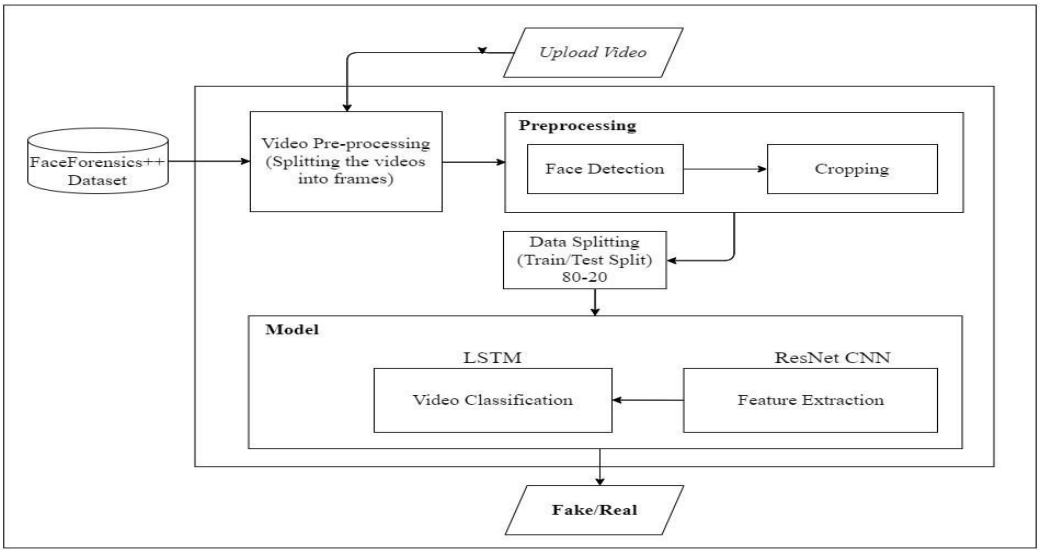


Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**-1:System architecture diagram for DeepFake Detection**

Our system architecture adopts a robust methodology to address the challenges of deepfake detection, utilizing a comprehensive framework that includes data preparation, model training, and prediction phases. Let's explore each component in detail:

A. Dataset

The dataset for this study was sourced from a variety of platforms, including YouTube, FaceForensics++, and the Deep Fake Detection Challenge dataset. It contains an equal distribution of genuine and altered deepfake videos to ensure a balanced dataset for analysis. To facilitate experimentation, the data was divided into 80% for training and 20% for testing.

B. Preprocessing

The preprocessing phase ensures uniformity and efficiency in data handling. First, video clips are segmented into individual frames to enable analysis at the frame level. Face detection algorithms are applied to extract and crop facial

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

regions in each frame. Frames without detectable faces are excluded to maintain consistency. To optimize computational efficiency, only the first 100 frames from each video are used for training. The processed frames are combined into new video formats for subsequent analysis.

C. Model Architecture

The model leverages a ResNet backbone integrated with an LSTM layer to ensure effective processing. A Data Loader module manages the processed videos, splitting them into training and testing batches. The model handles the frames in mini-batches during both training and testing phases.

D. ResNet CNN for Feature Extraction

The ResNet152 architecture serves as the foundation for feature extraction, employing its deep layers to capture complex features from images. Instead of redefining the classifier, the network is fine-tuned with additional layers and optimized learning rates to enhance performance. Extracted feature vectors, each with a dimensionality of 2048, are passed to the LSTM for further processing.

E. LSTM for Temporal Analysis

The LSTM layer processes the sequential feature vectors produced by ResNet. This enables the model to analyze time-based inconsistencies across frames and detect manipulations typical of deepfake content. A 2048-unit LSTM configuration with a dropout rate of 0.4 is employed for efficient sequence processing.

F. Prediction Phase

During the prediction phase, new videos undergo preprocessing to align with the input format of the trained model. Frames are extracted, facial regions are cropped, and the processed frames are fed directly into the trained model for classification. This streamlined workflow ensures precise detection of deepfake videos.

Our system architecture is carefully designed to leverage cutting-edge deep learning methodologies combined with thorough preprocessing techniques to ensure the accurate identification of deepfake videos. This approach plays a vital role in combating misinformation and preserving the authenticity of digital media. A graphical representation of our system workflow is presented in Figure 2.

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**-2:Flow Diagram of Deepfake Video Detection**

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

(a) (b)

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**-3:Example of real and fake frames. (a) is real frame, (b) is a fake frame**

Figure 3a showcases a frame extracted from the real video dataset provided by FaceForensics++. In contrast, Figure 3b illustrates a frame sourced from the Deepfakes category of the FaceForensics++ dataset. While at first glance, it may be challenging for a human observer to distinguish between genuine footage and manipulated deepfakes, subtle differences become apparent upon closer scrutiny. The real frame contains finer texture details, such as distinct double eyelids and delicate wrinkles around the eyes. On the other hand, the fake frame appears smoother with fewer intricate features, lacking these subtle details. These distinctions underscore the complexity of visually identifying deepfake content, emphasizing the necessity for sophisticated detection algorithms to curb the spread of falsified media and safeguard the reliability of digital information.

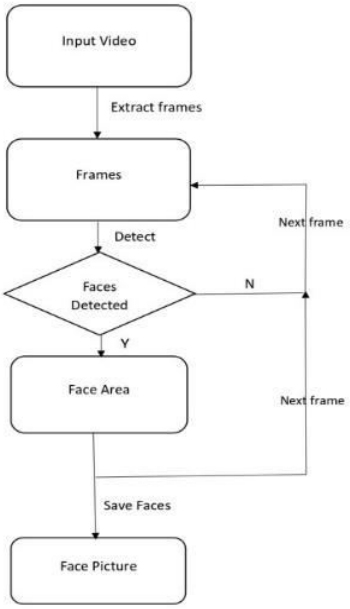


Figure 4. Preprocessing Flow

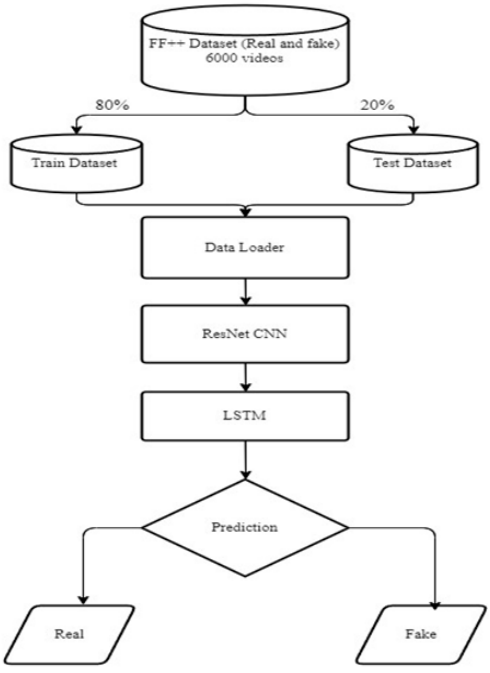


Figure 5. Training Flow

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

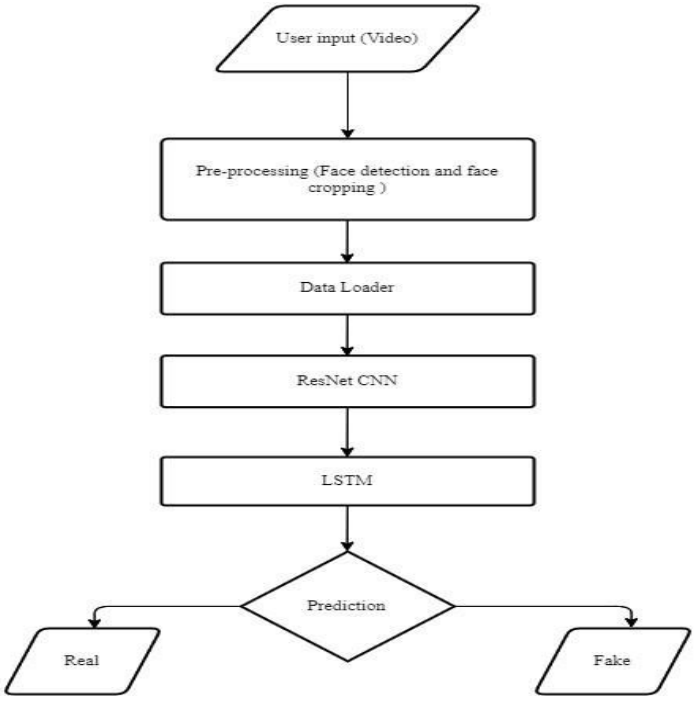


Figure 6. Prediction Flow

3.1 Dataset: FaceForensics++

The dataset utilized in this study is the **FaceForensics++** benchmark, which is widely adopted for evaluating deepfake detection systems. It comprises over **6,000 video samples**, equally split between real and manipulated content. The manipulated videos are generated using state-of-the-art face synthesis and manipulation techniques, including:

- **DeepFakes** – an autoencoder-based method that swaps facial identities.
- **FaceSwap** – a face replacement approach using 3D modeling.
- **Face2Face** – reenacts facial expressions by transferring them from source to target.
- **NeuralTextures** – leverages neural rendering to synthesize realistic faces.
- **FaceShifter** – a two-stage GAN-based face swapping method.

Each video in the dataset contains approximately **148 frames**, stored in **.mp4 format** with a **resolution of 112×112** pixels. For experimental consistency, we ensure that only frames with clearly detectable and centered faces are selected for training and evaluation.

Table1. FaceForensics++

Classes	Total No. of Videos	Train (80%)	Test (20%)	Total Frames	Image Size	Video Quality
DeepFakes	1000	800	200	148	112	C24
FaceShifter	1000	800	200			
FaceSwap	1000	800	200			
NeuralTextures	1000	800	200			
Face2Face	1000	800	200			
Original	1000	800	200			
Total	**6000**	**4800**	**1200**			

This dataset ensures balanced representation and diversity, making it ideal for training deep learning models capable of generalizing across multiple manipulation methods.

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

IV. RESULTS

Our deepfake detection system is designed to provide definitive conclusions on whether a video is real or manipulated, along with a confidence score to indicate the certainty of its predictions. This process is visually represented in Figure 7. When tested on the FaceForensics++ dataset, the system is expected to consistently and accurately distinguish between authentic and fake videos across diverse scenarios, showcasing its reliability. Furthermore, by offering confidence levels for its classifications, the system empowers users to make informed decisions regarding the credibility of video content. Overall, our solution aims to play a significant role in curbing the spread of fake videos and preserving the authenticity of digital media.

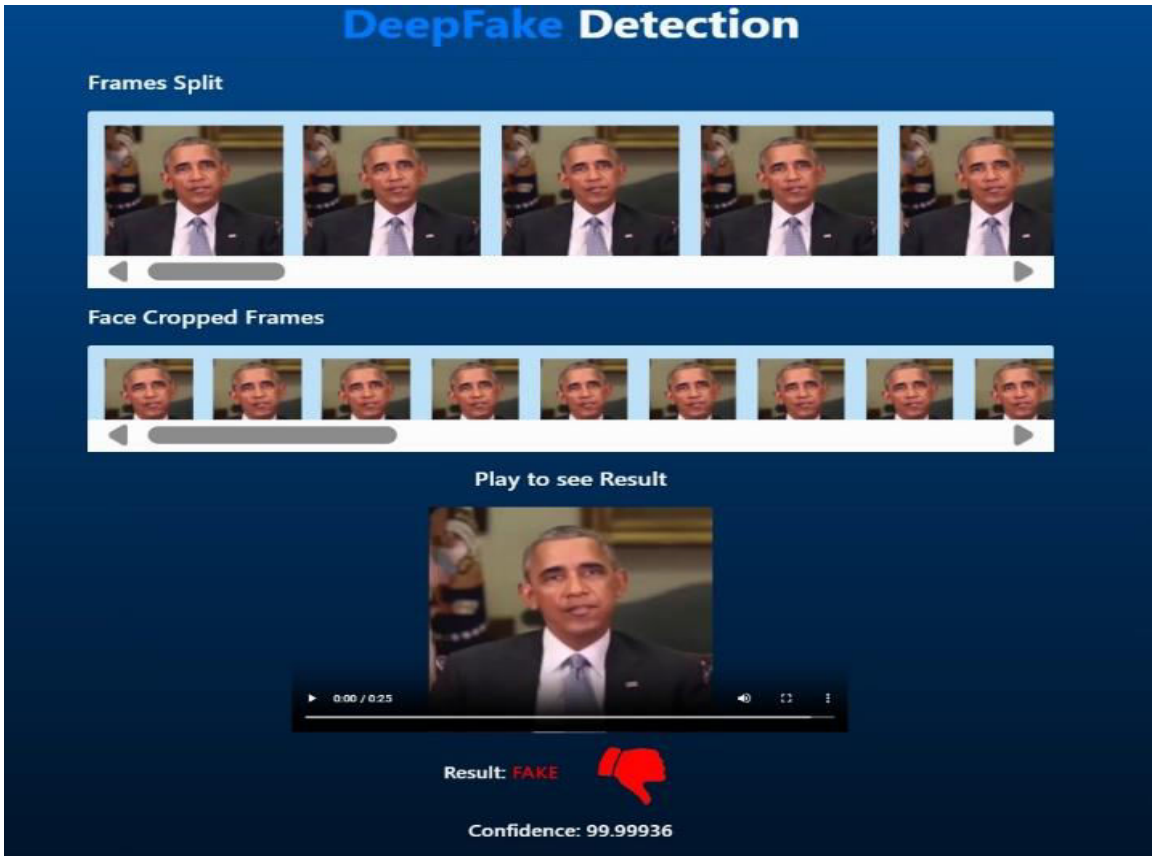


Figure 7. Expected Results

Deepfake Detection Process:

1. Frames Split (Top Section):
The video is divided into individual frames (still images). This allows the detection model to analyze each moment separately for inconsistencies typical of deepfakes.

2. Face Cropped Frames (Middle Section):
The system zooms in on the face in each frame. This helps focus on facial movements, expressions, eye blinks, and mouth synchronization, which are critical clues in identifying deepfakes.

3. Video Analysis (Bottom Section):
The cropped face frames are recompiled into a short video clip that the AI model evaluates for authenticity.
4. Result and Confidence Score:
Result: FAKE – The system has identified this video as a deepfake.

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

Confidence: 99.99936% – This is a very high confidence level, indicating the model is almost certain that the video is not real.

V. CONCLUSION

To summarize, our research was centred on building a reliable deepfake classification system by leveraging the FaceForensics++ dataset alongside advanced machine learning techniques. Using the ResNet model for feature extraction and incorporating LSTM layers for temporal analysis, we successfully developed a system capable of distinguishing between real and manipulated images. This hybrid architecture merges spatial and temporal insights, enabling our method to achieve high levels of accuracy in detecting deepfake content. Through rigorous experimentation and analysis, our study validated the effectiveness of this approach in identifying subtle inconsistencies within manipulated media. The synergy between ResNet's intricate feature extraction capabilities and LSTM's ability to analyze sequential data highlights the robustness of our proposed method. Overall, this work contributes significantly to the growing body of research aimed at addressing the challenges posed by deepfake content.

VI. FUTURE SCOPE

The results of our study emphasize the critical need for continued innovation in the field of digital forensics to counter the evolving threats posed by synthetic media. In the future, we aim to refine our methodology by integrating more sophisticated techniques to enhance the system's reliability and scalability. Expanding the dataset to include a wider variety of manipulation techniques and experimenting with advanced architectures may further improve the detection capabilities. Additionally, we envision applying this system to broader applications, such as real-time deepfake detection and large-scale content verification systems. By exploring these avenues, the proposed system can play a pivotal role in mitigating the impact of manipulated media, preserving the authenticity of digital content, and safeguarding public trust.

VII. LIMITATIONS

While our method demonstrated strong performance in detecting visual manipulations, it does not account for audio manipulations. As a result, the system is currently unable to identify audio-based deepfake content, which represents a growing concern in the domain of synthetic media. Audio deepfakes, often used to mimic voices or manipulate speech, require specialized detection mechanisms that analyze auditory features and temporal coherence in audio data. Future iterations of our system aim to incorporate audio analysis capabilities, enabling the detection of both visual and audio deepfakes. This advancement would allow for a more holistic approach to combating synthetic media, ultimately improving the system's effectiveness and applicability.

REFERENCES

1. Jadhav, A., Patange, A., Patel, J., Patil, H., & Mahajan, M. (2020). Neural Network-Based Detection Framework for Deepfake Videos. IJSRD.
2. Masood, M., Nawaz, M., Javed, A., Nazir, T., Mehmood, A., & Mahum, R. (2021). Utilizing Pre-Trained CNN Models for Deepfake Video Classification. IEEE Xplore.
3. Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., & Davidson, I. E. (2023). Enhanced Dense Convolutional Neural Network for Identifying Deepfake Images. IEEE.
4. Kharbat, F. F., Elamsy, T., Mahmoud, A., & Abdullah, R. (2020). Employing Image Feature Detection Techniques for Deepfake Video Identification. IEEE.
5. Guarnera, L., Battiato, S., & Giudice, O. (2020). Revealing CNN Traces in Images to Counter Deepfakes. IEEE Access.
6. Richards, A. M. J., Prakash, P., Varshini, K. E., Kasthuri, P., N., D., & Sasithradevi, A. (2023). Facial Deepfake Detection Using CNN-Based Architecture. ICoAC.
7. Vajpayee, H., Yadav, N., Jhingran, S., & Raj, A. (2023). A Comparative Transfer Learning Approach for Fake Face Image Identification. InC4.
8. Berjawi, A., Samrouth, K., & Deforges, O. (2023). Enhancing Deepfake Detection with Optimized Image Preprocessing Techniques. ACTEA.

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405099|

9. Tran, V. N., Le, S. H., Le, H. S., Kim, B. S., & Kwon, K. R. (2023). Generalizable Detection of Facial Forgeries in Unseen Domains. ICCE.
10. Afchar, D., Nozick, V., & Yamagishi, J. (2018). MesoNet: Lightweight Architecture for Detecting Manipulated Facial Videos. arXiv:1809.00888v1.
11. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Deepfake Detection through Multi-Attention Mechanisms. IEEE/CVF.
12. Carlini, N., & Farid, H. (2020). Bypassing Deepfake Image Detectors Using Black-Box and White-Box Attack Strategies. Proceedings of CVPR.

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