ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1847

Design and Implementation of an AI/ML Framework for Identifying Face-Swapped Deepfake videos

Anagha Thorat[1]; Bhagyashree Kadam[2]; Pratiksha Rampure[3]; Shreya Patil[4]; Vijay Sonawane[5]

([4]Professor)

[1,2,3,4,5]JSPM'S Bhivrabai Sawant Institute of Technology & Research Wagholi

Publication Date: 2025/06/11

Abstract: Deep learning has revolutionized various complex tasks, including image interpretation, autonomous system control, and large-scale data analysis. However, its advancements have also facilitated the development of sophisticated tools capable of generating highly realistic yet fraudulent media content, known as deepfakes. These AI-generated images and videos can convincingly mimic real individuals, raising significant concerns regarding national security, democratic integrity, and personal privacy. Consequently, there is an urgent need for intelligent detection systems that can effectively identify and verify the authenticity of digital media. Such systems are crucial for distinguishing between genuine and manipulated content, ensuring the reliability of information, and preventing the dissemination of misleading visuals. This paper delves into the methodologies employed in creating prominent deepfakes and reviews the current literature on detection strategies. Furthermore, it discusses the inherent challenges posed by deepfake technologies and outlines prospective avenues for future research aimed at developing more robust and trustworthy detection mechanisms.

Keywords: *Deep Learning, CNN, Pre - Processing, Feature Extraction, Face Detection and Face Recognition.*

How to Site: Anagha Thorat; Bhagyashree Kadam; Pratiksha Rampure; Shreya Patil; Vijay Sonawane; (2025) Design and Implementation of an AI/ML Framework for Identifying Face-Swapped Deepfake videos. *International Journal of Innovative Science and Research Technology*, 10(5), 4314-4318. https://doi.org/10.38124/ijisrt/25may1847

I. INTROUCTION

With easy internet access and fast-growing technology, people and businesses can now communicate through social media. Lifelike digital content (text, video, and audio) may now be created using breakthroughs in generative artificial intelligence (AI). "Deepfakes" are fake content like images, videos, or sounds that look and sound real, made using advanced AI/ML. In recent years, big improvements in AI and machine learning have made it possible to create and edit photos and videos using new smart tools [1]. To disseminate fake information, provoke political unrest or violence, or even threaten and control individuals, extremely realistic bogus audio, video, or image content has been generated [2] [17]. The stunning, lifelike, and expertly edited movies are now known as "Deepfake." Recent methods for detecting deepfakes focus on looking at each frame of a video and checking for signs that it might be fake.

Recently, a free machine learning-based software application simplified the process of creating convincing face swaps in videos, resulting in "deepfake" videos that require minimal editing. Realistic bogus movies can be used to incite political unrest, blackmail, or stage terrorist acts [3] [18]. This article talks about a method that can automatically detect deepfake videos by paying attention to changes over time in the video [4].

II. RESULT

A. System Overview

➢ *Data Collection & Preprocessing*

- Gather datasets of real and deepfake videos (e.g., Face Forensics++, Deep-Fake Detection Challenge dataset).
- First, take out video frames, fix the faces so they face the same way, and prepare the data for the model. [5].

➢ *Feature Extraction*

- Use CNNs to extract facial features.
- Apply techniques like Optical Flow analysis to detect inconsistencies in facial movements [6] [19].
- Utilize frequency domain analysis to identify compression artifacts unique to deepfakes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1847

➢ *Model Selection & Training*

- Train machine learning models such as CNNs, LSTMs (for temporal analysis in videos), or Transformer-based architectures [7].
- Use ensemble learning by combining multiple models for improved accuracy.

➢ *Deepfake Classification & Detection*

- Apply binary classification (Real vs. Fake) or multi- class classification (Fake Type Detection: Face Swap, GAN-generated, etc.).
- Use anomaly detection models to flag unusual facial patterns.
- Implement adversarial training techniques to make models robust against evolving deepfake methods [8].

➢ *Performance Evaluation*

- Check how well the model works by using measures like Accuracy, Precision and Recall, F1-score, and AUC-ROC.
- Compare real-time inference efficiency for practical deployment.

➢ *Deployment & Integration*

- Create an online tool or API that can automatically detect deepfake videos.
- Integrate the system into social media platforms or forensic tools.

B. Flow Chart

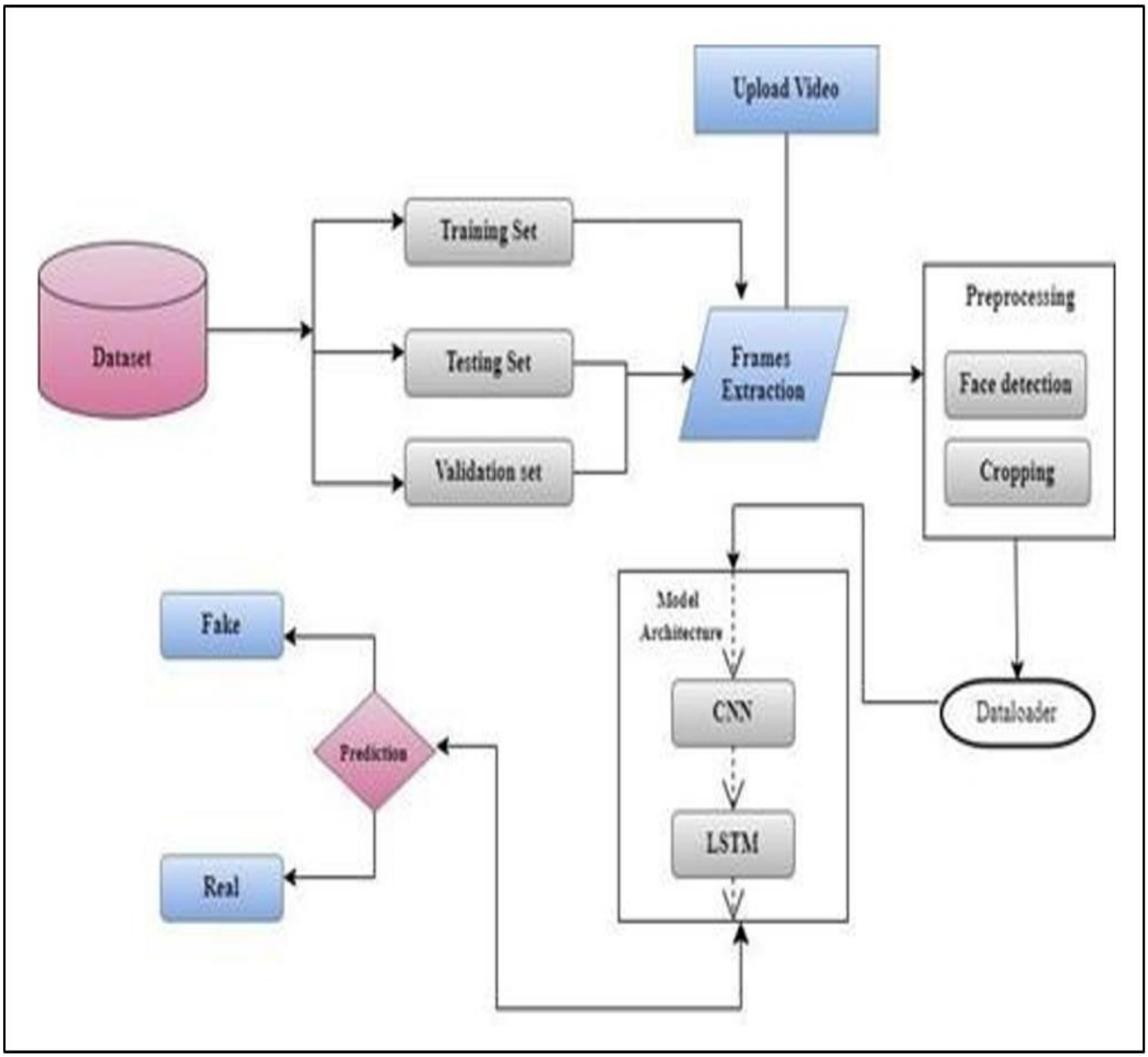

Fig 1 System Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1847

➢ *This System uses Multiple Steps to Detect Deepfakes, particularly face Swaps:*

- *Video Input:* The system begins by processing video content.
- *Frame Extraction:* The video is divided into frames for analysis.
- *Preprocessing:* Each frame undergoes preprocessing, such as adjusting brightness or filtering, to prepare it for the next steps.
- *Face Extraction:* The faces in the frames are extracted for further analysis.
- *Face Detection:* The system identifies and locates the faces in the extracted frames.
- *Feature Extraction Using CNN***:** CNN are used to extract key facial features, enabling the system to understand the structure of the face.
- *Deepfake Identification:* The system examines facial characteristics and matches them with established deepfake indicators to determine whether the video is genuine or altered.
- *Output Result:* The final output provides the detection result, identifying whether the video contains a deepfake.

C. System Architecture

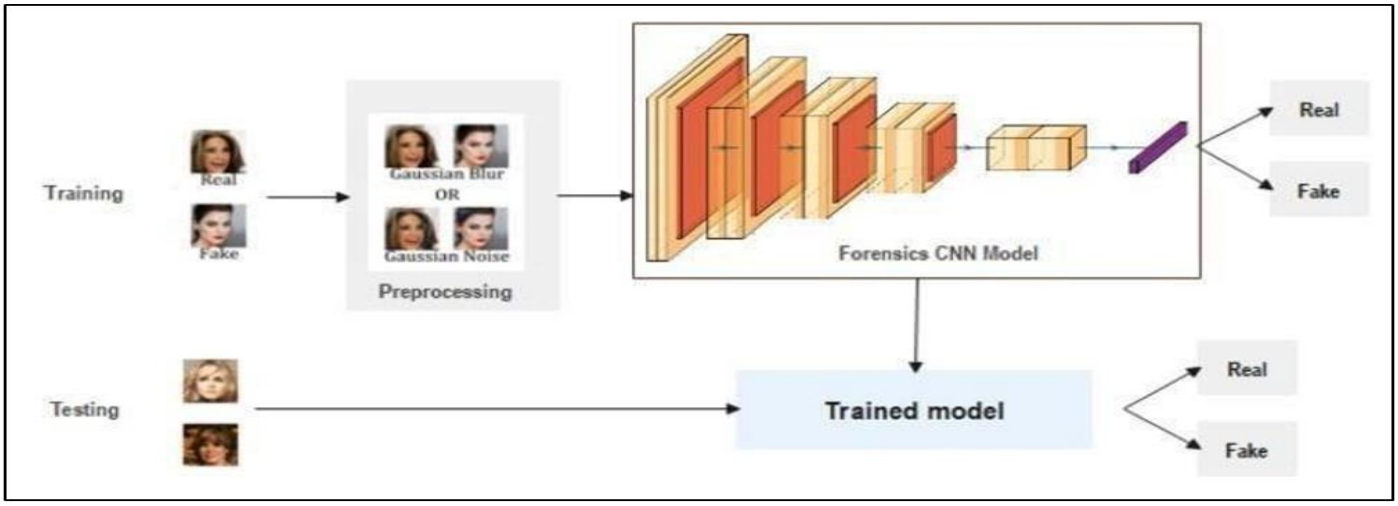

Fig 2 System Architecture

The proposed method's enhanced Convolutional Neural Network (CNN) architecture addresses problems in prior deepfake detection methods [9] [15].

➢ *Training Phase*

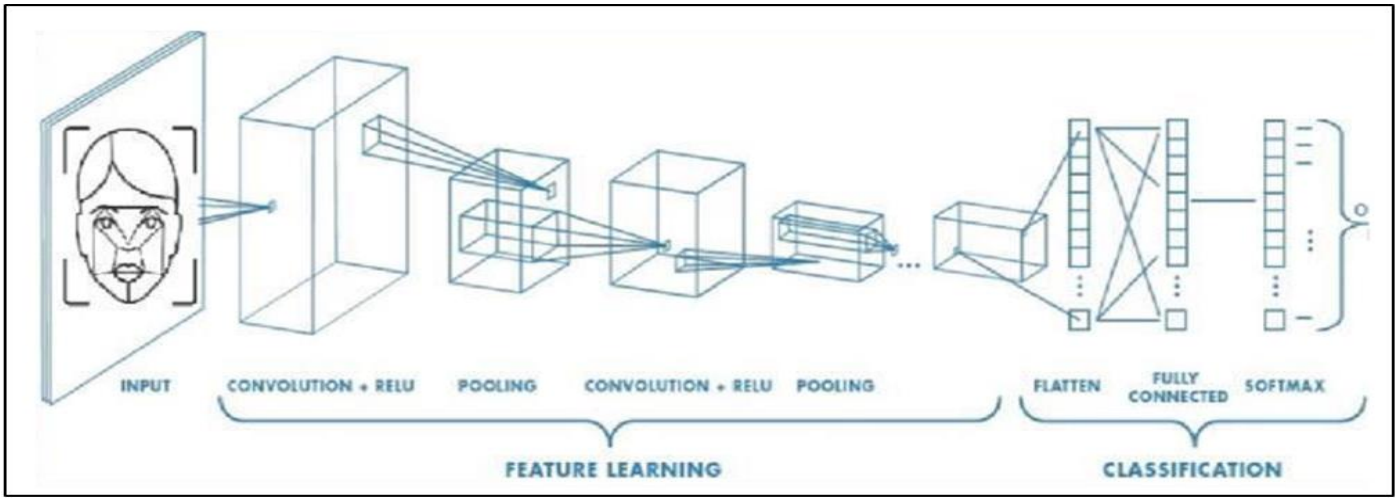

Fig 3 Training Phase

- *Input Data:* We teach the model using a dataset that has both real and edited (fake) images.
- *Preprocessing:* This step includes operations like face cropping, resizing, normalization, and possibly data augmentation to enhance the model's learning ability [10].
- *Feature Extraction via CNN*: A Forensics CNN Model (a deep learning-based Convolutional Neural Network) extracts important features from the preprocessed images.
- *Classification Layer:* The CNN learns patterns in real vs. fake images and classifies them accordingly.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1847

➢ *Testing Phase*

A fresh image, whether genuine or altered, is input into the trained model, which analyzes it using learned features andclassifies it as Real or Fake based on prior training. [11] [16].

III. ALGORITHM USED

A. Convolutional Neural Networks (CNNs) – Feature Extraction

CNNs analyze images and extract spatial features such as facial texture, lighting inconsistencies, and blending artifacts. Deep CNN architectures like Exception Net, Res Net, or Efficient Net can detect deepfake artifacts effectively. Used for analyzing individual frames in deepfake videos.

B. Long Short-Term Memory – Temporal Analysis

Since videos contain multiple frames, LSTMs help track temporal inconsistencies in facial expressions and movements [12]. Detects unnatural transitions or subtle deepfake glitches that CNNs alone might miss. Works well when combined with CNNs, where CNN extracts spatial features and LSTM processes the sequential data [13].

C. Generative Adversarial Networks (GANs) – Adversarial Training & Detection

GANs are used for both generating deepfakes and detecting them. Can be employed to train a model that learns to distinguish between real and fake faces [14]. Some approaches use auto encoders with GAN-based anomaly detection to flag manipulated content.

➢ *Integration Strategy for Deepfake Detection:*

- Step 1: Use CNNs to extract frame-level features.
- Step 2: Use CNN to get features from video frames, then send them to LSTM to find changes over time that might show a deepfake.
- Step 3: Use GAN-based adversarial training to improve robustness against evolving deepfake techniques.

IV. COMPARISON

Table 1 Comparison

Criteria	Existing Methods	Proposed CNN-based Method
Feature Extraction	Manual/Basic CNN	Multi-scale CNN Feature Extraction.
Accuracy	~80-88%	94.8%
Real-Time Detection	Limited	Improved via optimized architecture
Generalization Across Datasets	Weak (Overfitting)	Better Generalization (Trained on mixed data)
Temporal Artifact Detection	Rarely considered	Included in frame sequence analysis
Explainability	Black-box models	Intermediate visual pattern tracking possible
Computation Requirement	High	Moderate (using efficient edge computing)
Scalability	Low	Scalable and adaptive

V. CONCLUSION

Deepfakes have become more popular because there's so much content on social media, and tools to create deepfakes are now easier to get. Social media also makes it simple to share fake videos. One way to tell whether the video content is original or artificially altered is by using AI called a neural network. This AI uses something called CNN (Convolutional Neural Network) to look at the video and decide if it's real or fake, with a high level of confidence in the result.

FUTURE SCOPE

Deepfake detection systems that use machine learning need regular updates to keep up with new and advanced fake video techniques. With the help of AI, these systems can quickly spot deepfakes in live videos and stop them from spreading. We can add deepfake detection tools to social media platforms. These tools will check photos and videos before they are posted to see if they are real or fake. This can help improve safety, reduce the spread of false information, and build trust among users.

REFERENCES

[1]. Rana, Md Shohel, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. “Deepfake detection: A systematic literature review.” IEEE access 10 (2022): 25494-25513.

[2]. Lewis, John K., Imad Eddine Toubal, Helen Chen, Vishal Sandesera, Michael Lomnitz, Zigfried Hampel-Arias, Calyam Prasad, and Kannappan Palaniappan. “Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning.” In 2020 4. IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1- 9. IEEE, 2020.

[3]. Trinh, Loc, Michael Tsang, Sirisha Rambhatla, and Yan Liu. “Interpretable and trustworthy deepfake detection via dynamic prototypes.” In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1973-1983. 2021.

[4]. S. P and S. Sk, "DeepFake Creation and Detection:A Survey," 2021 Third International Conference on Inventive Research in Computing Applications.

[5]. S. R. B. R, P. Kumar Pareek, B. S and G. G, "Deepfake Video Detection System Using Deep Neural

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1847

[6]. Networks," 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2023, pp. 1-6, doi: 10.1109/ICICACS57338.2023.10099618.
[7]. M. S. Rana, B. Murali and A. H. Sung, "Deepfake Detection Using Machine Learning Algorithms," 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI), Niigata, Japan, 2021, pp. 458-463, doi: 10.1109/IIAI-AAI53430.2021.00079.
[8]. Swathi, P., and Saritha Sk. "Deepfake creation and detection: A survey." In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 584-588. IEEE, 2021.
[9]. Mahmud, Bahar Uddin, and Afsana Sharmin. "Deep insights of deepfake technology: A review." arXiv preprint
[10]. Busacca, Angela, and Melchiorre Alberto Monaca. "Deepfake: Creation, purpose, risks." In Innovations and Economic and Social Changes due to Artificial Intelligence: The State of the Art, pp. 55-68. Cham: Springer Nature Switzerland, 2023.
[11]. Chadha, Anupama, Vaibhav Kumar, Sonu Kashyap, and Mayank Gupta. "Deepfake: an overview."
[12]. In Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020, pp. 557-566. Springer Singapore, 2021.
[13]. Seow, Jia Wen, Mei Kuan Lim, Raphaël CW Phan, and Joseph K. Liu. "A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities." Neurocomputing 513 (2022): 351-371.
[14]. Westerlund, Mika. "The emergence of deepfake technology: A review." Technology innovation management review 9, no. 11 (2019).
[15]. Feeney, Matthew. "Deepfake laws risk creating more problems than they solve." Regulatory Transparency Project (2021).
[16]. Abu-Ein, Ashraf A., Obaida M. Al-Hazaimeh, Alaa
[17]. M. Dawood, and Andraws I. Swidan. "Analysis of the current state of deepfake techniques- creation and detection methods." *Indonesian Journal of Electrical Engineering and Computer Science* 28, no. 3 (2022): 1659-1667.
[18]. 16. Yadav, Digvijay, and Sakina Salmani. "Deepfake: A survey on facial forgery technique using generative adversarial network." In 2019 International conference on intelligent computing and control systems (ICCS), pp. 852-857. IEEE, 2019.
[19]. Shivale, N.M., Mahajan, R.A., Bhandari, G.M., Sonawane, V.D., Kulkarni, M.M., Patil, S.S., “Optimizing Blockchain Protocols with Algorithmic Game Theory”, Advances in Nonlinear Variational Inequalities, 2024, 27(4), pp. 231–246.
[20]. Sonawane, V.D., Mahajan, R.A., Patil, S.S., Bhandari, G.M., Shivale, N.M., Kulkarni, M.M., “Predicting Software Vulnerabilities with Advanced Computational Models”, Advances in Nonlinear Variational Inequalities, 2024, 27(4), pp. 196–212.
[21]. Kulkarni, M.M., Mahajan, R.A., Shivale, N.M., Patil, S.S., Bhandari, G.M., Sonawane, V.D., “Enhancing Social Network Analysis using Graph Neural Networks”, Advances in Nonlinear Variational Inequalities, 2024, 27(4), pp. 213–230.