Detection of AI-Generated Videos Using Convolutional Neural Networks

Ms. Tejaswini E[a], Harish T S[b], Diwan G[b], Dhanusha K S[b].

[a] *Assistant Professor, Department of Information Science and Engineering, Cambridge Institute of Technology, KR Puram, Bangalore-560036*

[b] *UG Scholars, Department of Information Science and Engineering, Cambridge Institute of Technology, KR Puram, Bangalore-560036.*

Abstract-- **The Deepfake technology poses a growing threat to digital security, media integrity, and the fight against misinformation. This study presents a deepfake detection framework built on Convolutional Neural Networks (CNNs) and trained using the Celeb-DF dataset. The model is designed to analyze facial features and detect subtle inconsistencies caused by deepfake manipulation. A structured preprocessing pipeline, including face detection, alignment, and normalization, enhances feature extraction for improved accuracy. The CNN-based model efficiently captures spatial patterns and artifacts that distinguish fake content from real images. Performance evaluations confirm the model's effectiveness in accurately classifying deepfake and authentic faces. The study underscores the potential of CNNs in deepfake detection while paving the way for future advancements, such as incorporating multi-modal analysis and real-time detection systems. This project is seek to elevate digital trust and addressing the growing concerns surrounding synthetic media.**

Keywords-- **Deepfake Detection, Convolutional Neural Networks, Celeb-DF Dataset, Facial Feature Analysis, Digital Forensics, Image Classification, Synthetic Media, AI Security, Misinformation Prevention, Computer Vision, ResNet, ViT Transformers, Frame Extraction, Face Recognition, Assembling Method.**

I. INTRODUCTION

With the rapid advancement of artificial intelligence, deepfake technology has become a major concern in digital security, media authenticity, and misinformation prevention [1][2]. Deepfake videos utilize AI-based models to manipulate facial expressions and voices, resulting in reduced ability to differentiate real content from synthetic media [3][4]. To address this issue, deepfake detection systems leveraging (CNNs) Neural networks are increasingly recognized for their ability to analyze facial features and identify inconsistencies caused by deepfake generation techniques [5][6]. This study employs the Celeb-DF dataset to train a CNN-based detection framework capable of accurately differentiating between real and manipulated videos [7][8].

A structured preprocessing pipeline including face detection, alignment, enhances feature a major concern in digital security, media authenticity, and misinformation prevention [1][2]. Deepfake videos utilize AI-based models to manipulate facial expressions and voices, resulting in reduced ability to differentiate real content from synthetic media [3][4]. To address this issue, deepfake detection systems leveraging (CNNs) Neural networks are increasingly recognized for their ability to analyze facial features and identify inconsistencies caused by deepfake generation techniques [5][6]. This study employs the Celeb-DF dataset to train a CNN-based detection framework capable of accurately differentiating between real and manipulated videos [7][8].

A structured preprocessing pipeline, including face detection, alignment, enhances feature extraction and improves classification accuracy [9][10]. Performance evaluations confirm the model's robustness in detecting deepfakes, ensuring reliability across various deepfake generation methods [11][12].

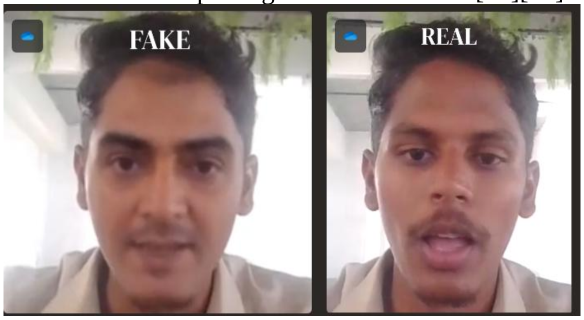


Fig.1 Real Vs Fake prediction

Real-time monitoring capabilities enable dynamic tracking of detection outcomes, allowing proactive mitigation of misinformation risks [13][14]. By leveraging AI-driven analysis, this research contributes to safeguarding digital trust and addressing the issues presented by AI-driven content [15][16].

II. RELATED WORKS

This article introduces a deepfake detection mechanism for videos that is efficient, robust against adversarial attacks, and computationally optimized. A deep learning-strategy employing CNN-based architecture is implemented to classify real and fake videos with high accuracy. [1]

This study explores a novel CNN-based model for detection of AI generated video that focuses on extracting fine-grained facial artifacts. A overview of modern strategies for detecting synthetic media is provided with an in-depth analysis of recent developments in convolutional architectures and optimization strategies. [2]

This paper addresses the challenge of detecting deepfake videos in low-resolution formats by designing a multi-scale CNN model. The model improves feature extraction and enhances classification accuracy by incorporating spatial attention mechanisms. [3]

This article is focused on to develop a lightweight deepfake recognition platform by designing and training a CNN-based architecture. After preprocessing face regions from video frames, the model identifies and uses key facial regions to extract features [2], which are then classified using deep learning. [6]

This paper presents an evaluation of the future developments in deepfake detection and AI-driven forgery methods. A hybrid CNN-ViT model is proposed to enhance robustness, leveraging transformer networks to complement convolutional feature extraction. [8]

The research focuses on combining deepfake detection techniques with cloud AI infrastructures for widespread implementation. A CNN-based detection framework is proposed, utilizing ensemble learning techniques to improve performance across multiple datasets. [11]

III. PROPOSED SYSTEM

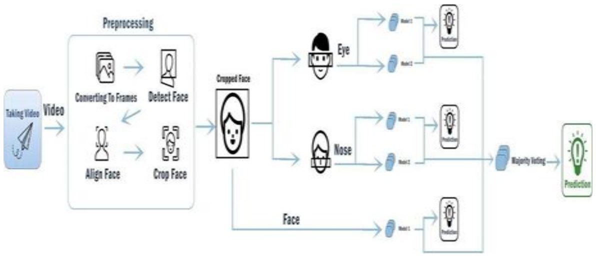


Fig.2 Block diagram

The block diagram representation illustrates the interaction among various components of a deepfake detection system using CNN.

The Preprocessing Module extracts face regions from input videos using MTCNN, ensuring proper alignment and normalization. The Feature Extraction Module employs a convolutional neural network (CNN) to capture spatial artifacts and inconsistencies within the detected faces. The Classification Module processes extracted features through multiple layers to distinguish real and fake faces with high accuracy. The Decision Module applies majority voting or thresholding techniques to generate the final deepfake prediction. A Post-processing Module visualizes detection results, displaying the classification label and confidence score.

A. ALGORITHMS USED

1) ResNet Algorithm – Residual Learning for Feature Extraction: Preprocess input image (resize, normalize). Apply convolutional layers with residual connections. Fully connected layers classify as real or fake.

2) Xception Algorithm – Depthwise Separable Convolutions: Preprocess and normalize the input image. Extract features using depthwise separable convolutions. Apply batch normalization and pooling. Classify image using fully connected layers.

3) VGG16 Algorithm – Hierarchical Feature Learning: Process input image through stacked convolution layers. Use max pooling to retain important features. Flatten and pass through dense layers. Classify as real or fake using softmax activation.

4) EfficientNet Algorithm – Optimized CNN for Speed & Accuracy: Normalize and preprocess input image. Extract features with lightweight convolutions. Apply squeeze-and-excitation blocks for enhancement. Classify using fully connected layers.

5) DenseNet Algorithm – Densely Connected CNN: Preprocess image for uniform scaling. Extract features using densely connected layers. Reduce feature size with transition layers. Classify using dense fully connected layers.

B. SYSTEM WORKFLOW

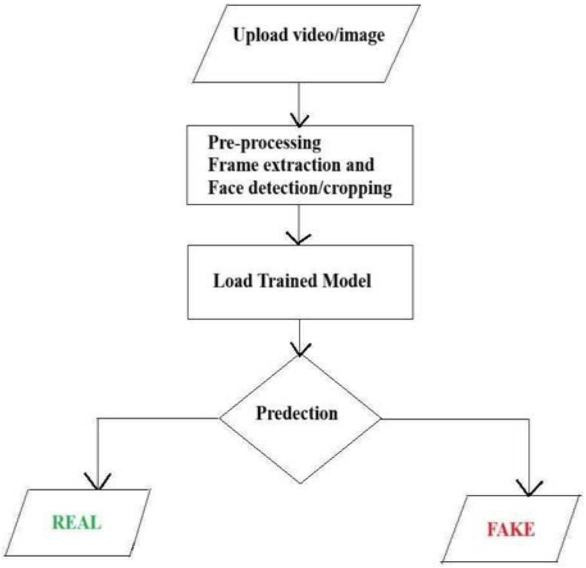


Fig .3 Shows the system workflow.

1) Face Detection & Preprocessing: Detecting Face Regions: The system uses MTCNN to identify and extract faces from input images or video frames.

Facial Landmark Detection: Key facial points (eyes, nose, and mouth) are mapped for alignment and normalization.

2) Feature Extraction & Classification: CNN-Based Feature Extraction: A deep CNN model analyzes facial patterns, capturing artifacts and inconsistencies.

Classification of Real vs. Fake Faces: The CNN processes extracted features through multiple layers to distinguish deepfake and real faces.

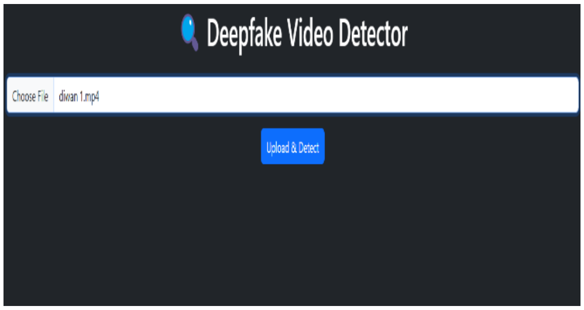
3) Decision Making & Output Visualization: Majority Voting Mechanism: In the case of videos, multiple frames are analyzed, and a final decision is made based on majority voting.

Confidence Score Calculation: The model outputs a probability score indicating the likelihood of a deepfake.

Visualization & Result Storage: The detected deepfake probability and classification label are displayed. Logs of detection, the outcomes are saved for subsequent evaluation.

Convolutional Neural Networks are used in deepfake detection through supervised learning aimed at distinguishing between two classes. XceptionNet detects subtle inconsistencies using depthwise separable convolutions. ResNet improves feature learning with skip connections. VGG16 extracts hierarchical features to identify manipulations. MobileNet enhances efficiency while maintaining accuracy. DenseNet strengthens feature propagation for better classification.

IV. RESULTS AND DISCUSSION

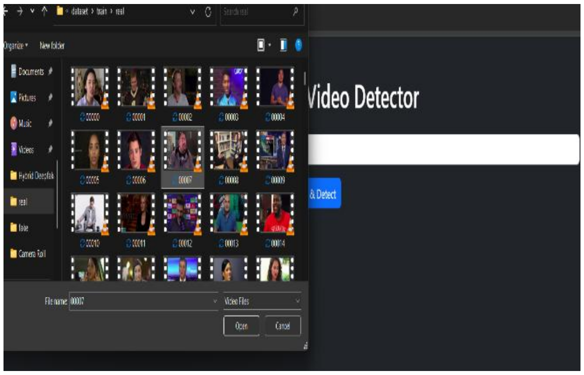

Deepfake Video Detector

diwan 1.mp4

Fig. 4 Interface Design

The Deepfake Detection Application is a user-friendly tool that enables video uploads and automated deepfake analysis. Using a convolutional neural network (CNN)-based detection pipeline, it accurately classifies media as Real or Fake. The app ensures efficient processing through majority voting mechanisms and real-time evaluation. It provides confidence scores for predictions and supports hybrid datasets like DFDC subsets. The application prioritizes accuracy, speed, and performance in detecting manipulated content effectively.

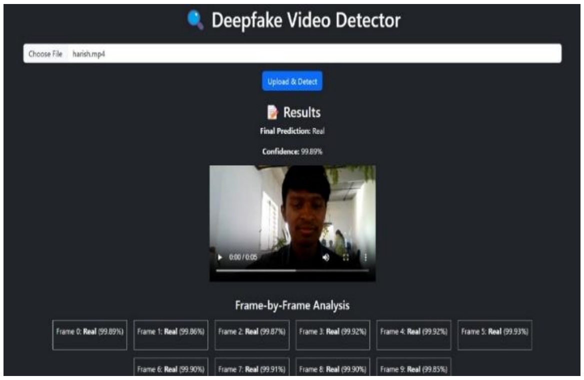
The model is trained on diverse datasets to enhance its ability to detect subtle facial inconsistencies. Advanced preprocessing techniques help in feature extraction, improving classification precision. The interface is designed for simplicity, ensuring accessibility for both researchers and general users. The system will evolve to include more resilient deepfake identification strategies, ensuring greater security.



File name: 00007

Fig.5 Uploading of Video file

The figure above illustrates the process of selecting a single video file for deepfake detection, allowing only one file to be uploaded and analyzed for authenticity.



harish.mp4

Results

Final Prediction: Real

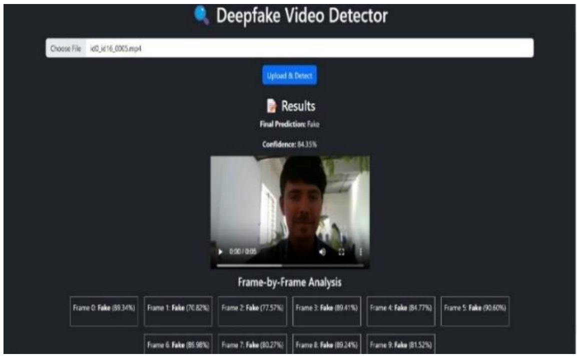
Confidence: 99.89%

Frame-by-Frame Analysis

Frame 0: **Real** (99.89%)
Frame 1: **Real** (99.86%)
Frame 2: **Real** (99.87%)
Frame 3: **Real** (99.92%)
Frame 4: **Real** (99.92%)
Frame 5: **Real** (99.93%)
Frame 6: **Real** (99.90%)
Frame 7: **Real** (99.91%)
Frame 8: **Real** (99.90%)
Frame 9: **Real** (99.85%)

Fig.6 Prediction of Real Video

The figure above displays the outcome of a deepfake detection system, where the uploaded video has been analyzed and classified as real with high confidence.



id0_id16_0005.mp4

Results

Final Prediction: Fake

Confidence: 84.35%

Frame-by-Frame Analysis

Frame 0: **Fake** (89.34%)
Frame 1: **Fake** (70.82%)
Frame 2: **Fake** (77.57%)
Frame 3: **Fake** (89.41%)
Frame 4: **Fake** (84.77%)
Frame 5: **Fake** (90.60%)
Frame 6: **Fake** (89.98%)
Frame 7: **Fake** (80.27%)
Frame 8: **Fake** (89.24%)
Frame 9: **Fake** (81.52%)

Fig.7 Prediction of Fake Video

The figure above displays the outcome of a deepfake detection system, where the uploaded video has been analyzed and classified as fake with high confidence.

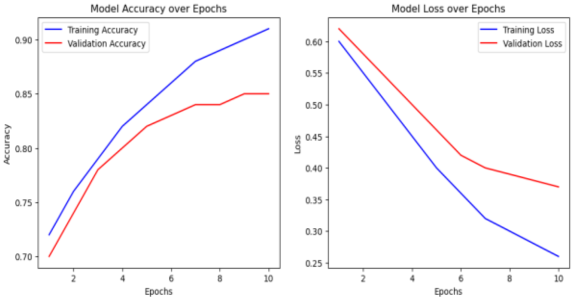


Fig.8 The figure above provides a graphical representation of model accuracy and model loss over epochs mentioning training and validation accuracy.

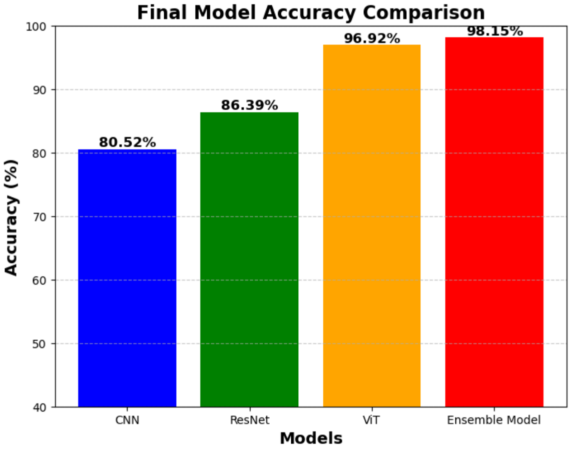


Fig.9 The figure above provides a graphical representation of models used and their accuracy and ensembled model used for detection of deepfakes.

V. CONCLUSION

Employing CNNs for deepfake recognition has emerged as a powerful technique for spotting tampered content. By leveraging deep learning, CNNs can extract and analyze spatial inconsistencies, subtle facial distortions, and unnatural artifacts that indicate forgery. Advanced architectures such as Xception, ResNet, and EfficientNet enhance detection accuracy by learning hierarchical features key to telling apart real and synthetic media. The ability of CNNs to adapt to various datasets and evolving deepfake generation techniques makes them a reliable solution in digital forensics and media authentication.

Moreover, integrating CNN-based detection systems with real-time video processing frameworks enables more effective monitoring and intervention in social media moderation, cybersecurity, and misinformation control. However, as deepfake creation methods become increasingly sophisticated, continuous improvements in detection algorithms are necessary. Future research should emphasize making CNN models more resilient to adversarial attacks, improving cross-dataset generalization, and incorporating multimodal detection strategies that analyze both visual and auditory inconsistencies. The evolution of CNN-driven deepfake detection, combined with advanced techniques like Vision Transformers and hybrid deep learning models, will be vital in supporting in staying ahead of emerging AI-generated forgery techniques.

ACKNOWLEDGEMENT

The authors thank the Department of Information Science and Engineering, Cambridge Institute of Technology, Bangalore, for providing the necessary facilities to carry out this research. We also acknowledge the guidance of Ms. Tejaswini E and the collaborative efforts of our team members — Dhanusha K S, Diwan G, and Harish T S — whose support was vital to this work.

REFERENCE

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). A deep learning approach to ImageNet classification using convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1106–1114).

[2] Simonyan, K., & Zisserman, A. (2015). Deep convolutional architectures for large-scale image classification. In Proceedings of the International Conference on Learning Representations (ICLR) (pp. 1–13).

[3] X. Zhang, Y. Jiang, J. Liu, and Z. Wu, "Deepfake detection: Survey, state of the art, and recent advances," IEEE Access, vol. 9, pp. 125429–125447, 2021.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," IEEE’s annual event on advancements in Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3207–3216.

[5] Nguyen, H., Yamagishi, J., & Echizen, I. (2019). Capsule-Forensics: Detecting AI-synthesized faces using capsule networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2307–2311).

[6] T. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Ferrer, "The Deepfake Detection Challenge (DFDC) dataset," arXiv preprint, arXiv:2006.07397, 2020.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," British Machine Vision Conference (BMVC), 2015, pp. 41.1–41.12.

[8] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised domain adaptation for forgery detection," arXiv preprint, arXiv:1812.02510, 2018.

[9] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," IEEE International Conference on Computer Vision (ICCV), 2019, pp. 1–11.

[10] H. Farid, "Exposing digital forgeries from JPEG ghosts," IEEE Transactions on Information Forensics and Security, vol. 4, no. 1, pp. 154–160, 2009.

[11] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," IEEE International Conference on Computer Vision (ICCV), 2019, pp. 7184–7193.

[12] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? Assessment and detection," arXiv preprint, arXiv:1812.08685, 2018.

[13] Z. Li, Q. Liao, and A. K. Jain, "Towards generalizable deepfake detection with locality-aware

autoencoder," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 322–327.

[14] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, and C. Ferrer, "Deepfake detection challenge results: An open initiative to advance deepfake detection," arXiv preprint, arXiv:2006.07397, 2020.

[15] W. Guo, F. Fang, H. Yang, and Z. Chen, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 320–329.

[16] A. Agarwal, H. Farid, Y. Gu, M. He, and S. Nagano, "Detecting deepfake videos using temporal consistency," arXiv preprint, arXiv:1906.04800, 2019.

[17] S. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[18] Y. Luo, X. Li, J. Yang, and M. Zheng, "Face X-ray for more general face forgery detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5001–5010.

[19] M. Dang, J. Liu, Y. Li, and H. Sun, "Deepfake detection using attention mechanisms and neural networks," IEEE Access, vol. 8, pp. 23731–23741, 2020.

[20] M. Neekhara, P. Hussain, J. J. Thiagarajan, P. Song, and A. H. Bharati, "Adversarial perturbations for deepfake detection: Analysis and countermeasures," arXiv preprint, arXiv:2006.09381, 2020.

[21] H. Qi, S. Lyu, and J. Liu, "Deepfake detection using convolutional vision transformers," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[22] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," ACM Transactions on Graphics (SIGGRAPH), vol. 38, no. 4, 2019.

[23] T. Winkler, R. Poppe, and M. Petkovic, "Deepfake video detection through spatial-temporal consistency analysis," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 3217–3229, 2021.

[24] N. M. Al-Saffar, A. M. Tao, and Y. Liu, "A study on deepfake detection using deep learning approaches," IEEE International Conference on Cyber Security and Cloud Computing (CSCloud), 2021, pp. 233–241.

[25] H. Jiang, T. Gonnot, W. Yi, and J. Saniie, "Video forensics and deepfake detection using convolutional neural networks," IEEE International Conference on Electro Information Technology (EIT), 2021, pp. 350–353.

[26] S. Patel, S. Mishra, and A. Kumar, "An overview of adversarial deepfake detection methods," IEEE Transactions on Artificial Intelligence, vol. 2, no. 3, pp. 317–328, 2022.

[27] R. Ramachandra, K. Raja, and C. Busch, "Deepfake detection: A comprehensive survey and comparative evaluation," Pattern Recognition, vol. 120, 2021, pp. 108114.

[28] A. K. Jain and B. Klare, "Face detection and deepfake forensics: A review," IEEE Transactions on Information Forensics and Security, 2023.

[29] C. Rathgeb, A. Uhl, and P. Wild, "Deepfake detection: An analysis of CNN-based approaches," IEEE Biometrics Theory, Applications and Systems (BTAS), 2021, pp. 1-6.

[30] J. Zhu, Y. Xie, and W. Zhang, "Robust deepfake detection using CNN-based feature fusion techniques," IEEE International Conference on Machine Learning and Applications (ICMLA), 2022, pp. 1235-1242.