IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEPFAKE DETECTION USING FACIAL ARTIFACT ANALYSIS IN DEEP LEARNING

"A Deep Learning Approach for Detecting Manipulated Media Using Spatial Feature Analysis"

Azhaguselvan G[1], Balaji V[2], Dinesh Kumar P[3], Divan G[4], Vaishnavi M[5]

[1234]Final year Student, [5]Assistant Professor - Department of Information Technology,
The Kavery Engineering College, Mecheri, Salem – 636453, India

Abstract: Deepfake technology has advanced rapidly in recent years, making it increasingly difficult to distinguish between real and manipulated media. This paper presents an efficient and explainable deepfake detection system that leverages facial artifact analysis and deep learning techniques. The proposed system integrates spatial feature extraction using a pretrained Xception Convolutional Neural Network (CNN) with fine-grained facial landmark analysis using MediaPipe FaceMesh. Video inputs are processed through an optimized frame extraction strategy, and facial regions are isolated using MTCNN. The model is trained on FaceForensics++ and Celeb-DF datasets and evaluated on the DFDC dataset to assess generalization capability. Experimental results demonstrate strong performance with high accuracy and robustness across datasets. Additionally, explainability is incorporated through probability-based outputs and visualization techniques such as Grad-CAM. The proposed approach effectively identifies deepfake content and provides interpretable results, making it suitable for real-world forensic applications.

Index Terms - Deepfake Detection, Facial Artifacts, Xception CNN, FaceMesh, Explainable AI, Grad-CAM

I. INTRODUCTION

The rapid evolution of artificial intelligence and deep learning has led to the emergence of deepfake technology, which enables the creation of highly realistic synthetic media. While this technology has applications in entertainment and media production, it also poses significant threats, including misinformation, identity theft, and digital fraud.

Traditional detection techniques struggle to identify modern deepfakes due to their high visual quality and temporal consistency. Therefore, there is a need for robust and explainable detection systems capable of identifying subtle inconsistencies in facial features and dynamics.

This research proposes a deep learning-based framework that combines spatial feature extraction with facial artifact analysis. The system aims to provide accurate classification along with interpretable insights, making it suitable for real-world deployment.

II. LITERATURE SURVEY

Recent studies have explored various approaches for deepfake detection. Convolutional Neural Networks (CNNs) have been widely used for extracting spatial features from images. Models such as XceptionNet have demonstrated strong performance in detecting manipulated media.

Other research has focused on temporal inconsistencies, including eye blinking patterns and facial movements. Landmark-based approaches using facial keypoints have shown effectiveness in detecting unnatural expressions.

However, many existing methods lack generalization across datasets and fail to provide explainable outputs. This research addresses these limitations by combining CNN-based feature extraction with facial landmark analysis and explainability techniques.

III. RESEARCH METHODOLOGY

3.1 Data Acquisition

The datasets used in this study include FaceForensics++ and Celeb-DF for training and validation, while the DFDC dataset is used exclusively for testing. This ensures no data leakage and enables evaluation of real-world performance.

3.2 Data and Sources of Data

Video data is processed using OpenCV to extract frames at a rate of 2 frames per second (FPS). This optimized sampling reduces redundancy while preserving essential information. Each video contributes approximately 100–120 frames.

3.3 Face Detection

MTCNN is used to detect and align facial regions. The detected faces are cropped and resized to a uniform format suitable for model input.

3.4 Facial Feature Analysis

Media Pipe Face Mesh is employed to extract detailed facial landmarks. These landmarks capture micro-expressions such as eye blinking, lip movement, and subtle inconsistencies that are often present in deepfake content.

3.5 Model Architecture

A pretrained Xception CNN model is used for feature extraction. Transfer learning is applied to fine-tune the model for binary classification (real vs fake). The model processes cropped facial images and outputs a probability score.

3.6 Training Strategy

The model is trained using Binary Cross Entropy loss and optimized with AdamW. Regularization techniques such as dropout and weight decay are applied to prevent overfitting. Validation is performed during training to monitor performance.

3.7 Evaluation Metrics

1. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision

$$Precision = \frac{TP}{TP + FP}$$

3. Recall

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. Binary Cross Entropy Loss

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

3.8 Explainability

Explainability is incorporated using Grad-CAM visualization, which highlights manipulated regions in images. Probability scores are also provided to indicate confidence levels in predictions.

IV. Results And Discussion

4.1 Performance Evaluation

The proposed model achieved the following results on the validation dataset:

- Accuracy: 87.6%
- Precision: 89.9%
- Recall: 88.2%
- F1-Score: 87.5%
- ROC-AUC: 0.90

On the DFDC dataset, the model achieved an accuracy of 83.3%, demonstrating good generalization capability.

4.2 Discussion

The results of our research indicate that the proposed system performs effectively in detecting deepfake content. The slight drop in performance on the DFDC dataset highlights the challenges of real-world variability.

The integration of facial landmark analysis enhances detection accuracy by capturing micro-level inconsistencies. Additionally, explainability features improve trust and usability in forensic applications.

4.3 output

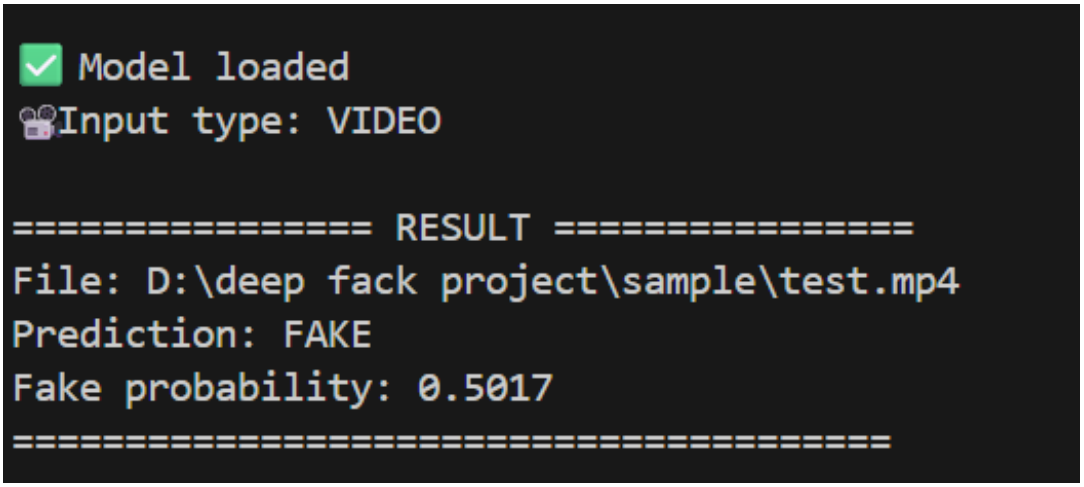

Fig:4.1 image detection output

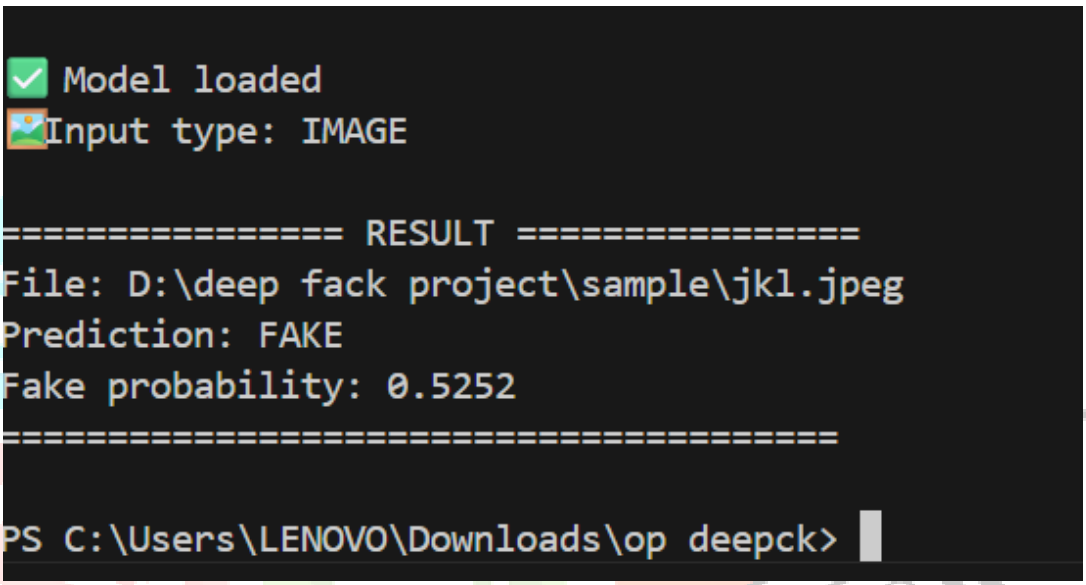

Fig: 4.2 video detection output

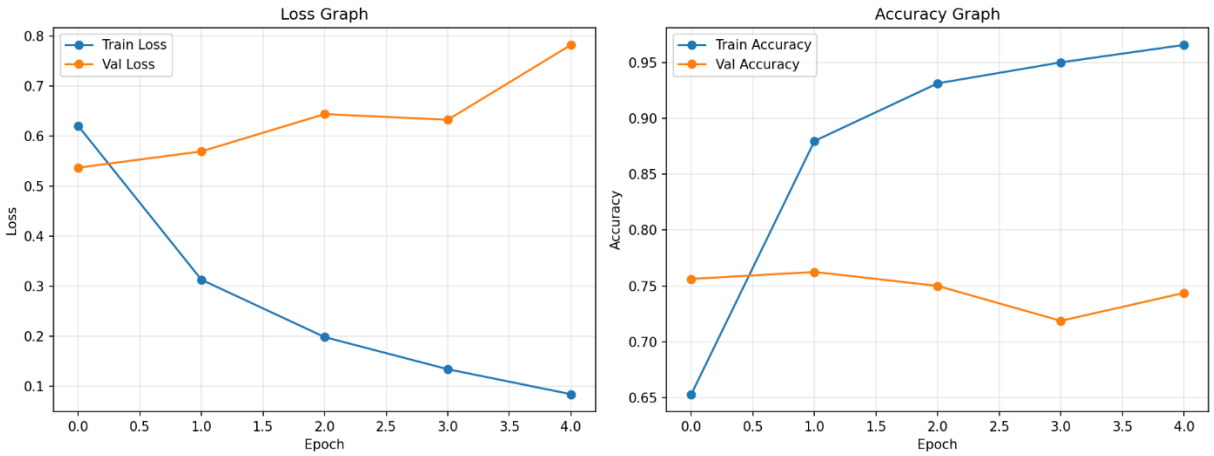

Fig: 4.3 training graph

V. Conclusion

From our observation this study presents a robust deepfake detection system that combines deep learning and facial artifact analysis. The model demonstrates strong performance across multiple datasets and provides interpretable results through explainability techniques.

Our proposed system is effective in identifying high-resolution deepfakes and can be applied in areas such as digital forensics, media verification, and cybersecurity.

VI. Future Enhancement

Future improvements may include:

- Integration of transformer-based architectures
- Enhanced temporal modeling using LSTM or video transformers
- Improved robustness against adversarial attacks

VII. Acknowledgment

Our sincere gratitude to **Mrs. Vaishnavi M., M.E., Assistant Professor, Department of Information Technology, The Kavery Engineering College**, for her valuable guidance and continuous support throughout this research work.

References

[1] Rossler et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” ICCV, 2019.
[2] Dolhansky et al., “The DeepFake Detection Challenge Dataset,” 2020.
[3] Chollet, F., “Xception: Deep Learning with Depthwise Separable Convolutions,” CVPR, 2017.
[4] Zhang et al., “Detecting Deepfake Videos with Temporal Inconsistencies,” 2020.
[5] Afchar et al., “MesoNet: A Compact Facial Video Forgery Detection Network,” 2018.
[6] Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks,” ICCV, 2017.
[7] Kingma & Ba, “Adam: A Method for Stochastic Optimization,” 2015.
[8] Sandler et al., “MobileNetV2: Inverted Residuals,” 2018.