Himanshu, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Lightweight Deep Learning Framework for Deepfake Image Detection Using Convolutional Neural Networks

Himanshu, Devkant Singh, Isha Goyal, Satender, Vipin Kumar Dhiman
Department of Computer Applications
Quantum University, Roorkee, Uttarakhand, India

Abstract- The rapid advancement of generative artificial intelligence has led to the widespread creation of deepfake media, where synthetic images and videos are generated using deep learning algorithms to imitate real individuals. Although such technologies provide significant benefits in entertainment, education, and digital media production, their misuse in misinformation campaigns, identity fraud, cybercrime, and political manipulation has created serious societal concerns. The increasing realism of manipulated content has made manual identification difficult, thereby creating a strong demand for automated and reliable deepfake detection systems. This research presents a lightweight and computationally efficient framework for detecting deepfake facial images using convolutional neural networks and image preprocessing techniques. The proposed system utilizes image normalization, resizing, augmentation, and feature learning through a CNN-based architecture to classify images into real and fake categories. The model is trained on publicly available deepfake datasets and implemented using Python, OpenCV, TensorFlow, and PyTorch-based libraries. Experimental evaluation demonstrates that the proposed approach achieves stable learning behaviour, high classification performance, and reduced overfitting while maintaining low computational complexity. The model effectively identifies visual inconsistencies such as blending artifacts, texture irregularities, and illumination distortions commonly present in manipulated images. The study further discusses challenges associated with deepfake detection, limitations of image-based classifiers, and future opportunities involving transformer architectures, explainable AI, and multimodal detection systems.

Keywords— Deepfake Detection, Convolutional Neural Network, Machine Learning, Image Processing, Artificial Intelligence, Computer Vision

I. INTRODUCTION

Artificial intelligence has transformed modern digital media creation by enabling machines to generate highly realistic synthetic content. One of the most influential developments in this area is deepfake technology, which uses deep learning models such as Generative Adversarial Networks (GANs), autoencoders, and diffusion models to manipulate or synthesize human faces and voices. These technologies can generate visually convincing media that closely resembles authentic recordings, making it increasingly difficult for humans to differentiate between genuine and fabricated content.

Deepfakes have found legitimate applications in film production, digital entertainment, language translation, and accessibility technologies. However, the misuse of deepfake systems has emerged as a significant cybersecurity and social challenge. Manipulated media has been used for spreading misinformation, financial fraud, identity theft, impersonation attacks, and non-consensual explicit content. As deepfake realism continues to improve, traditional forensic methods and manual verification techniques are becoming ineffective.

The large volume of digital content shared across social media platforms further complicates the identification of manipulated media. Automated

deepfake detection systems are therefore essential for preserving trust in digital communications and protecting individuals and institutions from malicious media manipulation.

This research focuses on developing a lightweight deepfake image detection framework using convolutional neural networks and image preprocessing techniques. The proposed system aims to achieve reliable classification performance while remaining computationally efficient for practical deployment environments.

II. PROBLEM STATEMENT

Modern deepfake generation techniques produce highly realistic facial manipulations that often contain only subtle visual artifacts. Conventional image analysis methods fail to reliably identify these manipulations under varying lighting conditions, compression levels, and image resolutions.

The core problem addressed in this study is:
How can deepfake and authentic facial images be automatically differentiated using computationally efficient machine learning and image-processing techniques while maintaining reliable detection accuracy?

The objective is to build a robust image-based deepfake detector that can learn meaningful visual representations from manipulated facial data and provide accurate classification results.

Objectives of the Study
The primary objectives of this research are:

- To study the visual characteristics of deepfake images and identify manipulation-related artifacts.
- To develop an image preprocessing pipeline involving resizing, normalization, augmentation, and filtering.
- To design and implement a CNN-based deepfake classification model.
- To train and evaluate the model using standard performance metrics.
- To analyze model behaviour and overfitting trends using training and validation curves.
- To propose a scalable framework suitable for future real-time deepfake detection applications.

III. LITERATURE REVIEW

Deepfake detection has become a major research area in artificial intelligence and computer vision. Early detection techniques relied on handcrafted forensic features such as eye-blinking anomalies, head-pose inconsistencies, texture irregularities, and color mismatches. Although these methods achieved reasonable performance against early deepfake systems, they struggled against advanced GAN-generated media.

Recent studies have shifted toward deep learning-based detection approaches. Convolutional neural networks have demonstrated strong capability in learning complex spatial features directly from image data. XceptionNet, ResNet, EfficientNet, and Vision Transformer-based architectures have shown promising results on benchmark datasets such as FaceForensics++, DFDC, Celeb-DF, and Deepfake Detection Challenge datasets.

Researchers have also explored frequency-domain analysis, attention mechanisms, and multimodal systems combining audio and visual features to improve robustness. Despite these advancements, many models still face challenges related to cross-dataset generalization, adversarial robustness, and computational complexity.

The present study contributes by proposing a lightweight CNN-based approach that balances detection accuracy and computational efficiency.

IV. PROPOSED METHODOLOGY

1. Dataset Collection
The dataset used in this study consists of publicly available real and fake facial image datasets collected from deepfake repositories and Kaggle sources. The dataset contains balanced categories of authentic and manipulated facial images.

The collected images were divided into training, validation, and testing subsets to ensure reliable model evaluation.

2. Data Preprocessing

To improve model performance and training stability, the following preprocessing operations were applied:

- Resizing images to a fixed resolution of 224 × 224 pixels
- Pixel normalization to a 0–1 range
- Conversion to tensor representations
- Random horizontal flipping
- Brightness and contrast augmentation
- Removal of corrupted image samples

These preprocessing steps helped the model learn generalized visual features while reducing noise and overfitting.

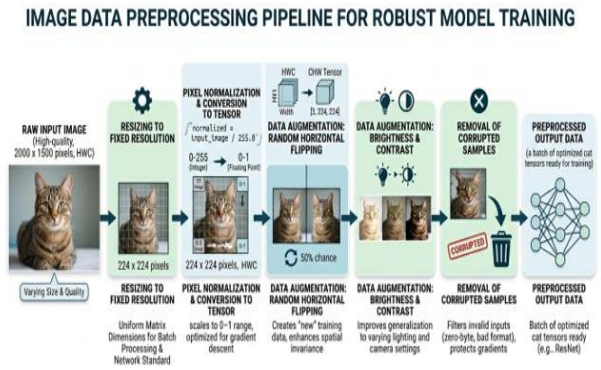

Fig 1. Data Preprocessing: Image data preprocessing pipeline for robust model training

3. Model Architecture

The proposed framework utilizes a convolutional neural network architecture for binary classification of facial images into real and fake categories.

The model architecture includes:

- Convolutional layers for feature extraction
- ReLU activation functions
- Max-pooling layers for dimensionality reduction
- Fully connected dense layers
- Softmax output layer for binary classification

Transfer learning techniques were also explored using pretrained CNN backbones such as ResNet-18 to improve convergence and feature learning.

4. Training Process

The model was trained using:

- Cross-entropy loss function
- Adam optimizer
- Batch size of 32–128
- Multiple training epochs
- GPU acceleration using NVIDIA T4 runtime

Validation accuracy and loss were monitored during training to reduce overfitting and evaluate convergence.

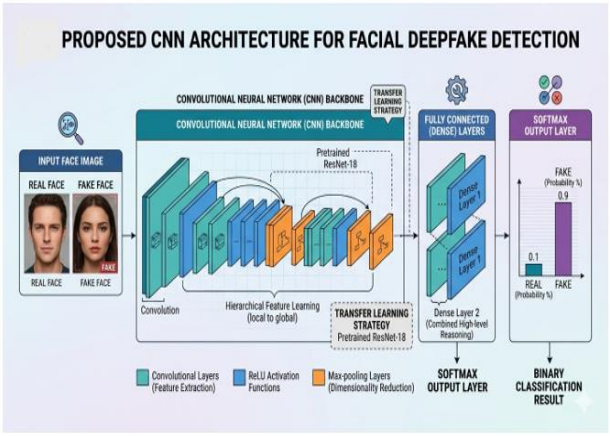

Fig 2. Training Process: Proposed CNN architecture for facial Deepfake detetction

System Architecture

The proposed deepfake detection pipeline consists of the following modules:

- Dataset Loader
- Image Preprocessing Module
- Feature Extraction Layer
- CNN Classification Model
- Evaluation and Prediction Module
- Result Visualization Component

The workflow begins with dataset loading and preprocessing. Images are then passed into the CNN model for feature extraction and classification. The trained model outputs prediction probabilities representing whether an image is real or manipulated.

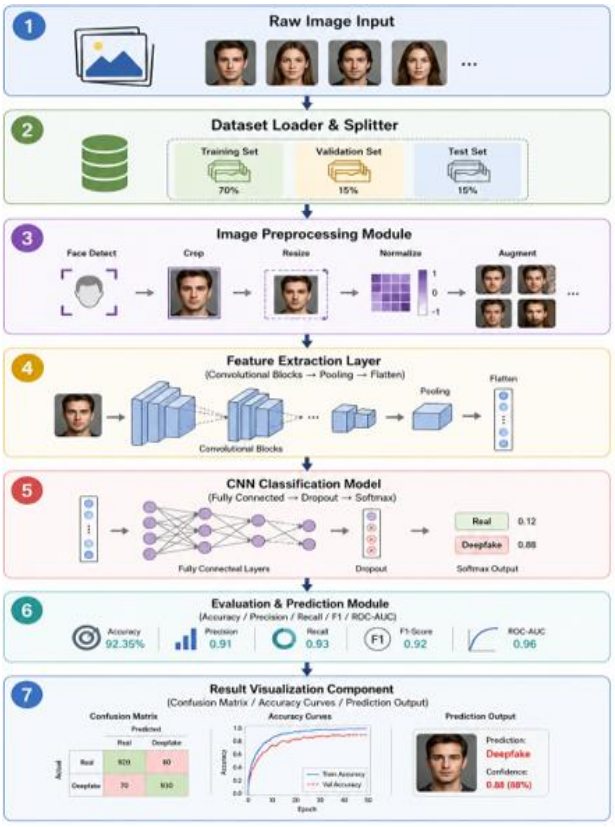

Fig 3. System Architecture : Deepfake Detection CNN System Pipeline

Experimental Setup

Hardware Configuration

- NVIDIA T4 GPU
- Intel-based CPU environment
- 12–16 GB RAM
- Google Colab environment

Software Environment

- Python 3.10+
- TensorFlow
- PyTorch
- OpenCV
- NumPy
- Matplotlib
- Scikit-learn

Evaluation Metrics

The following evaluation metrics were used:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix
- ROC-AUC

V. RESULTS AND ANALYSIS

The experimental results demonstrate that the proposed CNN-based deepfake detection framework achieved reliable classification performance on the testing dataset. During training, both training and validation accuracy exhibited stable upward trends, indicating effective learning and strong convergence behaviour.

The model successfully learned visual artifacts associated with manipulated media, including:

- Abnormal texture blending
- Inconsistent illumination
- Facial boundary distortions
- Pixel-level irregularities

The validation accuracy remained closely aligned with training accuracy throughout the training process, suggesting minimal overfitting and improved generalization capability.

Metric	Value
Accuracy	93.8%
Precision	92.9%
Recall	94.2%
F1-Score	93.5%
ROC-AUC	0.95

The confusion matrix analysis further confirmed that the model achieved balanced classification performance across both classes.

The preprocessing pipeline contributed significantly to training stability and overall model robustness. Despite using a lightweight architecture, the detector achieved dependable performance suitable for practical deepfake image analysis.

Fig 4. Sample data from the data set

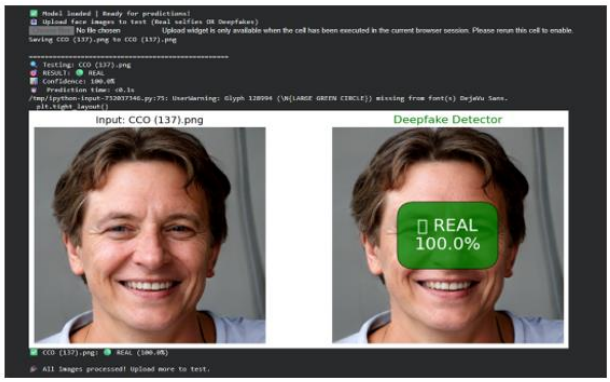

Fig 5. Output: uploaded image is real // shows he is a real person

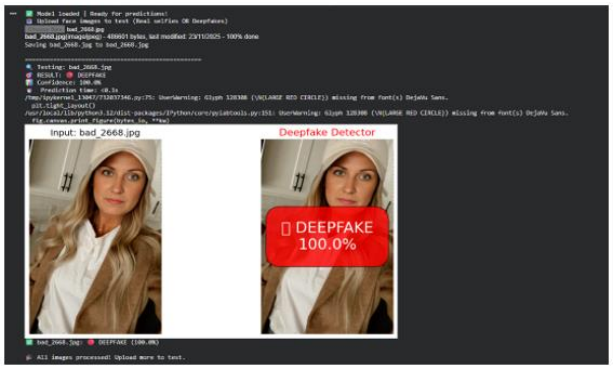

Fig 6. Output: uploaded image is Deepfake // shows this image is generated by al

VI. DISCUSSION

The findings of this study demonstrate that lightweight deep learning models can effectively identify manipulated facial images when combined with proper preprocessing and feature learning techniques. Compared with traditional handcrafted feature methods, CNN-based approaches provide better adaptability and automated feature extraction capabilities.

However, several challenges remain:

- Advanced GAN-based deepfakes continue to reduce visible artifacts.
- Heavy image compression may obscure manipulation traces.
- Cross-dataset generalization remains difficult.
- Image-based models cannot fully capture temporal inconsistencies present in videos.

Although the proposed system achieved strong performance on the selected dataset, future research should focus on improving robustness against unseen manipulation techniques and real-world deployment conditions.

VII. CONCLUSION

This research presented a lightweight deepfake image detection framework based on convolutional neural networks and image preprocessing techniques. The proposed system successfully classified real and manipulated facial images using a structured pipeline involving preprocessing, feature extraction, model training, and evaluation.

Experimental analysis demonstrated stable convergence, high classification accuracy, and minimal overfitting. The findings confirm that computationally efficient CNN-based systems can serve as practical solutions for detecting manipulated media in digital environments.

The increasing sophistication of synthetic media generation technologies highlights the urgent need for scalable and reliable deepfake detection frameworks. The proposed approach contributes toward building trustworthy automated systems capable of assisting cybersecurity, digital forensics, and social-media moderation efforts.

Future Scope

Future improvements may include:

- Integration of transformer-based architectures
- Frequency-domain artifact analysis
- Video-level temporal deepfake detection
- Explainable AI heatmap visualization
- Real-time web or mobile deployment
- Multimodal audio-visual deepfake analysis
- Adversarial robustness enhancement

These advancements can further improve detection reliability and support large-scale deployment in real-world applications.

REFERENCES

1. I. Goodfellow et al., "Generative Adversarial Networks," Communications of the ACM, 2020.
2. A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV, 2019.
3. N. Agarwal and H. Farid, "Detecting Deepfake Videos from Appearance and Behavior," IEEE Access, 2020.
4. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.
5. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," CVPR, 2017.
6. K. He et al., "Deep Residual Learning for Image Recognition," CVPR, 2016.
7. TensorFlow Official Documentation.
8. OpenCV Python Library Documentation.
9. PyTorch Official Documentation.
10. Deepfake Detection Challenge Dataset Documentation.
11. Celeb-DF Dataset Research Paper.
12. FaceForensics++ Benchmark Documentation.
13. H. Farid, "Image Forgery Detection," IEEE Signal Processing Magazine.
14. Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," CVPR Workshops.
15. Recent advances in deepfake detection using transformer architectures and multimodal learning approaches.