International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Explainable Deepfake Detection Using GRAD-CAM

Nrupen Kangutkar *(Kj's Trinity Polytechnic, Pune, India),
Email : nrupenkangutkar4@gmail.com
Saqib Halwai *(Kj's Trinity Polytechnic, Pune, India),
Email : halwaisaqib80@gmail.com
Kunal Ladke *(Kj's Trinity Polytechnic, Pune, India),
Email : ladkekunal20@gmail.com
Aliraza Shaikh*(Kj's Trinity Polytechnic, Pune, India),
Email : shaikh.aliraza.work@gmail.com

Abstract:
The system is developed using Python and integrates a Streamlit-based user interface for real-time interaction. The dataset consists of real and deepfake images and videos, which are preprocessed through face detection, frame extraction, and normalization techniques. A Convolutional Neural Network (CNN) model is trained on this data to classify media as real or fake with an observed accuracy of approximately 88–95% depending on dataset variations.
Performance evaluation shows that the average prediction time per input is approximately 0.8 to 1.5 seconds, ensuring near real-time response for users. The system also incorporates Grad-CAM (Gradient-weighted Class Activation Mapping), enabling visualization of important regions influencing the model's decision.
Experimental results indicate that the proposed system effectively detects deepfake content and provides visual explanations for predictions. The combination of efficient execution time, satisfactory accuracy, and explainability makes the system a practical solution for digital media verification. The system can be further extended with real-time detection and advanced deep learning models for improved performance.
Keywords — Deepfake Detection, Explainable AI, Grad-CAM, CNN, Image Processing, Computer Vision.

I. INTRODUCTION

In recent years, deepfake technology has become a major concern due to its ability to create highly realistic fake images and videos. This technology is widely used for misinformation, identity misuse, and digital fraud.
Existing detection systems mainly focus on classification accuracy but do not provide insights into how decisions are made, making them difficult to trust.
To address this issue, the Explainable Deepfake Detection system is proposed as an intelligent solution that detects manipulated media and provides visual explanations using Grad-CAM. The system integrates deep learning and explainable AI techniques to improve transparency.
The application is designed with a simple and user-friendly interface, allowing users to upload media files and receive instant predictions along with highlighted regions responsible for classification. This system enhances trust and contributes to secure digital environments.

II. SCOPE OF THE PROJECT

The scope of the Explainable Deepfake Detection system includes the development of a system capable of analyzing media data and identifying manipulated content.

1) Current Scope

- Classification of media into real and fake
- Use of deep learning models (CNN)
- Implementation of Grad-CAM for visualization
- Development of a user-friendly web interface
- Processing of images and video frames

2) Future Scope

- Integration with real-time video streams
- Deployment as a browser extension or mobile app
- Use of advanced models like transformers
- Integration with social media platforms
- Automated detection and reporting systems

The system can be extended further to support cybersecurity and digital forensics applications.

III. SYSTEM ARCHITECTURE

The system follows a modular architecture consisting of the following components:
1. User Interface Layer

- Developed using Streamlit

- Allows users to upload images/videos and view results

2. Data Processing Layer
   - Handles face detection and frame extraction
   - Converts raw media into structured input format
3. Prediction Module
   - Implements CNN-based deep learning model
   - Classifies media as real or fake
4. Explainability Module
   - Uses Grad-CAM technique
   - Highlights important regions influencing prediction
5. Dataset Layer
   - Contains real and deepfake datasets used for training and testing

Working Flow
User Input → Data Processing → Prediction Model → Grad-CAM → Output Display

Fig1: The system architecture consists of multiple modules. The user provides input through the Streamlit interface. The data processing module prepares the input data. The prediction module applies deep learning algorithms to classify the media. The Grad-CAM module highlights important regions, and the result is displayed to the user.

IV. METHODOLOGY

The system is developed using the following methodology:

1. Data Collection

Datasets containing real and deepfake images/videos are collected from available sources.

2. Data Pre-processing

The collected data is cleaned by:

- Removing corrupted or irrelevant data
- Extracting faces from images/videos
- Resizing and normalizing images

3. Feature Selection

Relevant features are automatically extracted by CNN, such as:

- Facial textures
- Pixel inconsistencies
- Visual artifacts

4. Model Development

Deep learning techniques are applied to detect deepfake content. The model is trained using labeled data and tested for accuracy.

5. System Implementation

- Backend is developed in Python
- Frontend is built using Streamlit
- Grad-CAM is integrated for visualization

6. Testing and Evaluation

The system is tested using sample inputs to evaluate performance, accuracy, and explainability.

V. CONCLUSION

The Explainable Deepfake Detection system provides an effective solution for identifying manipulated media using deep learning techniques. It enables users to detect fake content quickly and understand the reasoning behind predictions through visual explanations.

The system is simple, user-friendly, and capable of delivering accurate and interpretable results. Although it currently relies on pre-trained datasets, it can be further enhanced with real-time detection and advanced AI models.

Overall, the system has strong potential as a digital security solution and contributes significantly to the field of explainable artificial intelligence.

REFERENCES

[1] Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks,” ICCV, 2017.
[2] Ian Goodfellow et al., Deep Learning, MIT Press, 2016.
[3] FaceForensics++ Dataset, 2019.
[4] Python Software Foundation, Python Documentation, 2025.
[5] Streamlit Documentation, 2025.
[6] OpenCV Documentation, 2025.

