Deepfake Shield: AI-Based Deepfake Detection System

Dr. Sumitha C
Professor, Head dept. of CS&E
G. Madegowda Institute of Technology.
Bharathinagara, Mandya,
India.

Chinmay L
Student, dept. of AI&ML
G. Madegowda Institute of Technology.
Bharathinagara, Mandya,
India.

Mahalakshmi M
Student, dept. of AI&ML
G. Madegowda Institute of Technology.
Bharathinagara, Mandya,
India.

Rakshitha M
Student, dept. of AI&ML
G. Madegowda Institute of Technology.
Bharathinagara, Mandya,
India.

***Abstract* - Deep fake images are AI-generated or digitally manipulated visuals that alter human identity with high realism, resulting in risks such as misinformation, identity theft, fraud, and privacy violation. Manual verification of image authenticity is time-consuming and unreliable. This paper presents Deep fake Shield, an AI-based deep fake image detection system designed to classify images as real or fake using deep learning and image processing techniques. Input images undergo preprocessing for normalization and noise handling, followed by feature extraction using a convolutional neural network (CNN) model. The proposed system effectively identifies visual artifacts, pixel inconsistencies, and unnatural facial characteristics to detect deep fake images. The system aims to support cyber security, social media verification, and safe digital communication.**

Keywords - Deepfake Detection, Image Processing, Machine Learning, CNN, Fake Image Identification, Computer Vision.

I. INTRODUCTION

Deep fake technology has rapidly evolved with advancements in generative artificial intelligence, enabling the creation of highly realistic fake images. These synthetic images can convincingly manipulate facial identity, expressions, and appearance, making it difficult for humans to distinguish between genuine and AI-generated visuals. Such manipulation creates serious challenges including misinformation, online fraud, reputational damage, and misuse of personal identity. Therefore, a reliable and automated detection mechanism is essential.

This paper proposes Deep fake Shield, an intelligent AI-based detection system focused exclusively on deep fake image identification. The system analyzes uploaded images, extracts discriminative visual features, and classifies them as Real or Fake. The objective is to enhance digital trust, prevent misuse of identity, and support secure online interactions.

Deep fakes are synthetic media generated using Artificial Intelligence (AI) and deep learning algorithms that manipulate or fabricate images, videos, or audio to make them appear realistic. These technologies are capable of swapping faces, modifying expressions, and generating entirely fake visuals that closely resemble real individuals. Due to their high realism, deep fakes are extremely difficult to detect through manual observation. The concept of deep fakes originated from the combination of "deep learning" and "fake." Advanced neural network models such as Convolutional Neural Networks (CNNs), Auto encoders, and Generative Adversarial Networks (GANs) are commonly used to generate such content. GANs consist of two competing neural networks — a generator and a discriminator — where the generator creates fake content and the discriminator attempts to distinguish it from real data. Over time, this adversarial training results in highly realistic fake media. Initially, deep fake technology was developed for creative and beneficial applications such as film production, animation, digital art, and virtual reality. In the entertainment industry, deep fake techniques are used for visual effects, face replacement, and dubbing. Educational institutions and research organizations also use synthetic media for simulation and training purposes. However, despite these positive applications, the misuse of deep fake technology has grown rapidly and poses serious ethical, social, and security concerns.

II. BACKGROUND AND RELATED WORK

A. *Literature Survey*

Deep fake detection has become an active research area due to the rapid increase in AI-generated media. Early detection approaches depended on handcrafted features such as texture variations, edge distortions, lighting inconsistencies, and facial geometry irregularities followed by traditional classifiers. However, these methods often lacked robustness and failed when faced with high-quality deep fakes.

Recent studies highlight the effectiveness of deep learning models, especially Convolutional Neural Networks, for deep fake identification. CNNs automatically learn spatial and structural patterns from images, capturing subtle inconsistencies that are difficult for humans to perceive. Researchers have also explored frequency-domain analysis, transformer-based models, and transfer learning approaches to improve detection accuracy and generalization capability when datasets are limited. These advancements demonstrate that combining deep learning with effective pre processing techniques leads to more reliable deep fake detection systems.

Other works focus on blending-boundary identification, where inconsistencies occur around facial regions such as eyes, mouth, and hair edges due to poor synthesis. Some studies utilize colour channel anomalies and compression artifacts to identify manipulations. Researchers have further explored attention-based deep learning networks to highlight suspicious regions in an image. With the growth of diffusion models and improved generative architectures, detection research continues to evolve to counter increasingly realistic deep fakes.

B. Existing Deepfake Detection Approaches

Earlier approaches for deep fake identification relied on visual observation and manual inspection, which were inefficient and prone to human error. With technological advancements, AI-based systems have become the preferred solution. CNN-based classifiers, hybrid deep learning architectures, and frequency-aware models are widely used for detecting facial manipulation patterns. Some studies combine image processing with deep learning to enhance robustness, whereas others rely on transfer learning to overcome dataset limitations.

However, challenges remain due to the continuous improvement in deep fake generation techniques, limited dataset diversity, variability in lighting and pose, and difficulty in handling extremely realistic synthesized images. Computational cost, need for large annotated datasets, and difficulty in interpreting decisions are additional constraints. These challenges motivate the development of systems like Deep fake Shield that focus on reliable detection, systematic architecture, and practical usability.

III. METHODOLOGY

The methodology of the proposed deep fake image detection system integrates structured data handling, preprocessing techniques, CNN-based feature extraction, and classification to ensure accurate and efficient detection.

Step 1: Data Collection

Deep fake and real facial images are collected from publicly available benchmark datasets. The dataset contains labeled samples categorized as Real and Fake. Damaged, low-quality, or improperly labeled samples are removed to improve data reliability. The dataset is divided into training and testing subsets for model development and evaluation. Data augmentation may be applied to handle imbalance and enhance generalization.

Step 2: Image Preprocessing

To improve detection accuracy, preprocessing is applied to standardize and enhance image quality. Images are resized to a fixed resolution, normalized to maintain consistent pixel intensity, and subjected to noise reduction techniques. Basic enhancement operations help highlight relevant features and reduce the effect of artifacts unrelated to manipulation. Facial region extraction may be optionally applied to focus analysis on the most relevant regions.

Step 3: Feature Extraction

A Convolutional Neural Network (CNN) is used for feature extraction. CNN layers automatically learn essential facial structures, textural characteristics, and spatial irregularities. Pooling layers reduce dimensionality while preserving meaningful information, enabling efficient extraction of high-level manipulated features and natural face characteristics. Depending on design, pre-trained architectures can be leveraged to improve accuracy and reduce training effort..

Step 4: Classification

Extracted features are passed to fully connected neural network layers for binary classification. The classifier outputs probability scores indicating confidence levels for Real or Fake categories. These probabilistic outputs improve interpretability and decision reliability. Threshold-based decision logic is applied to finalize classification.

Step 5: Performance Evaluation

The trained model is evaluated using unseen test images. Performance is measured using accuracy, precision, recall, and F1-score to assess detection reliability and robustness. Confusion matrix analysis helps identify misclassification trends, while computational time evaluation ensures practical usability.

IV. PROPOSED WORK

The proposed work focuses on designing an intelligent and user-friendly deep fake image detection system that ensures reliable classification and practical usability.

.

A. Image Acquisition

Images are obtained from standard deep fake datasets containing clearly labeled Real and Fake categories. The dataset ensures sufficient diversity in lighting, pose, expression, and image quality for effective model training and evaluation. The dataset is curated to eliminate duplicates and distorted samples to maintain quality.

B. Image Enhancement and Preprocessing

To ensure uniform input quality, preprocessing is applied. Images are resized to standard dimensions, normalized, and filtered to reduce noise. Enhancement techniques improve clarity and emphasize critical patterns required for classification. These steps ensure that the CNN receives clean and consistent input, improving learning capability.

C. Deep Learning-Based Detection.

The enhanced images are processed using a CNN model that learns hierarchical feature representations. Lower layers capture basic image edges and textures, whereas deeper layers capture complex structural irregularities associated with deep fake manipulation. Transfer learning techniques may be applied to improve learning efficiency. The model architecture is designed to balance accuracy and computational feasibility to support deployment.

D. Classification and Prediction

Based on extracted features, the classifier predicts whether the image is Real or Fake. The system displays the classification label along with prediction confidence, helping users understand the reliability of the result. The design ensures minimal user interaction and a straightforward decision output.

E. System Architecture Description

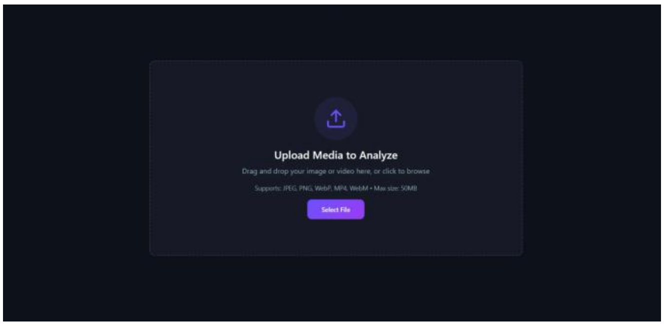
The system architecture consists of four primary modules: user interface module for image upload, preprocessing module for enhancement, CNN-based feature extraction and classification module, and result display module. Communication between modules is streamlined to ensure smooth execution and minimal delay.

V. RESULTS AND PERFORMANCE ANALYSIS

The system is evaluated based on its functional execution and detection capability. Testing confirms that the proposed model successfully processes deep fake and real images and produces meaningful classification outputs.

A. Image Upload and Analysis Process

The system accepts user-uploaded images and passes them through preprocessing and CNN-based detection processes. This confirms the system's ability to handle real-time user input and demonstrates user-friendly functionality.



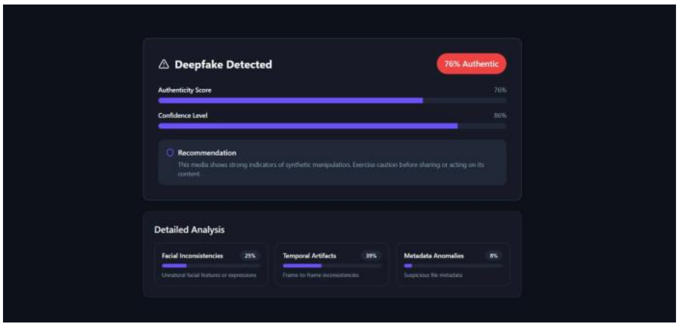
Drag and drop your image or video here, or click to browse

Supports: JPEG, PNG, WebP, MP4, WebM • Max size: 50MB

Fig. 1. Image upload interface used for submitting images for deep learning–based analysis

B. Fake Image Detection Result

The model identifies manipulated features in fake images and produces classification output with confidence scores, proving its effectiveness in detecting deepfake images. Visual inconsistencies such as blending errors and unnatural textures are successfully captured by the CNN layers.



76% Authentic

Authenticity Score 76%

Confidence Level 86%

This media shows strong indicators of synthetic manipulation. Exercise caution before sharing or acting on its content.

Facial Inconsistencies 25% — Unnatural facial features or expressions

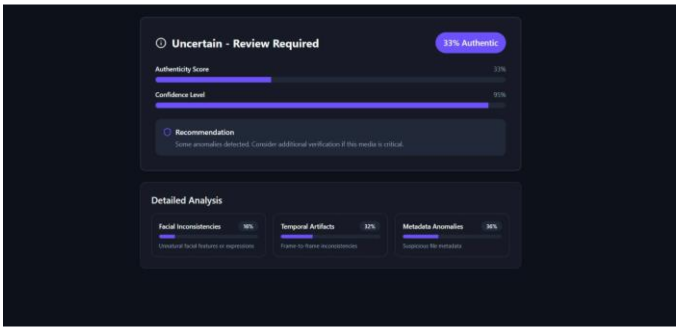
Temporal Artifacts 39% — Frame-to-frame inconsistencies

Metadata Anomalies 8% — Suspicious file metadata

Fig. 2. AI/Fake Image detection result generated by the proposed system with associated confidence score.

C. Real Image Detection Result

The system accurately recognizes genuine images and classifies them as Real, ensuring balanced and reliable classification. This reduces false alarms and improves trust in detection results.



33% Authentic

Authenticity Score 33%

Confidence Level 95%

Some anomalies detected. Consider additional verification if this media is critical.

Facial Inconsistencies 16% — Unnatural facial features or expressions

Temporal Artifacts 32% — Frame-to-frame inconsistencies

Metadata Anomalies 36% — Suspicious file metadata

Fig. 3. Real Image detection result generated by the proposed system with associated confidence score.

D. Performance Discussion

The results demonstrate that integrating preprocessing and CNN-based deep learning enables efficient and accurate deep fake detection.

The system shows strong capability in distinguishing manipulated images from real ones and can serve as a supportive digital authentication tool. With appropriate dataset expansion and tuning, performance can be further improved.

VI. CONCLUSION.

The rapid growth of deep fake technology has created serious challenges to digital authenticity and public trust. Manual identification of fake images is becoming increasingly difficult due to the high realism of manipulated content. This paper presented Deep fake Shield, an AI-based deep fake image detection system that integrates pre-processing techniques with CNN-based deep learning to classify images as Real or Fake in a reliable and systematic manner.

Experimental evaluation indicates that the system can successfully distinguish manipulated images from genuine ones while maintaining user-friendly operation. By enabling automated detection, Deep fake Shield supports cyber-security, social media verification, and digital forensics. Overall, the proposed system demonstrates that deep learning-based approaches are effective in addressing deep fake threats and can significantly contribute to safer digital communication environments.

REFERENCES

[1] Basanta Kumar Panigrahi, Siba Prasad Mishra, and Chinmay Kumar Samal, “Deepfake Detection Using Deep Learning: A Review,” *Journal of Artificial Intelligence Research and Applications,* 2025.

[2] R. Sunil, “Exploring Autonomous Methods for Deepfake Image Detection,” *ScienceDirect – Materials Today: Proceedings*, 2025.

[3] B. Carter, “Deepfake Image Detection via Facial Feature Extraction,” *arXiv Preprint,* arXiv:2507.18815, 2025.

[4] Y. Li, M. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Generated Fake Face Images,” *IEEE Transactions on Information Forensics and Security,* 2024.

[5] A. Rossler et al., “FaceForensics++: A Dataset for Deepfake Image Forensics,” *IEEE International Conference on Computer Vision (ICCV),* 2019.