

Reality Shield: AI Generated Content Detector

Sharvari Desai¹, Swapnali Munde², Shital Patil³, Samruddhi Bhoite⁴, Vaishnavi Chougale⁵

Final Year Students

Prof. A. A. Mahaldar⁶

Assistant Professor

Department of Computer Science and Engineering

Ashokrao Mane Group of Institutions, Vathar, Maharashtra, India

Abstract - The rapid advancement of Artificial Intelligence (AI) has enabled the creation of highly realistic synthetic media, including images, videos, and audio, through technologies such as Generative Adversarial Networks (GANs), diffusion models, and voice cloning systems. While these innovations offer significant benefits in entertainment, education, and communication, they also introduce serious challenges such as deepfakes, misinformation, identity theft, cyber fraud, and digital manipulation. The increasing accessibility of AI-generated content has made it difficult for users to distinguish authentic media from manipulated content, creating a critical need for reliable verification systems.

This paper presents **Reality Shield**, an intelligent multimodal AI-generated content detection platform designed to identify manipulated images, videos, and audio files. The proposed system integrates advanced machine learning and deep learning algorithms, including Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forest, Logistic Regression, Bayesian classifiers, and Vision Transformer (ViT), to analyze multimedia content and classify it as real or fake. The system performs media preprocessing, feature extraction, classification, and confidence-score generation to provide accurate and explainable detection results.

Reality Shield is developed using a React-based frontend and REST API-driven backend architecture, ensuring a scalable, user-friendly, and real-time detection environment. Experimental evaluation demonstrates that the proposed framework achieves high detection accuracy and effectively identifies visual, temporal, and acoustic inconsistencies commonly present in AI-generated media. By supporting image, video, and audio verification within a single platform, Reality Shield contributes to enhancing digital trust, combating misinformation, strengthening cybersecurity, and promoting responsible use of artificial intelligence in modern digital ecosystems.

Keywords: Artificial Intelligence, Deepfake Detection, Multimedia Verification, CNN, Vision Transformer, Machine Learning, Cybersecurity, Fake Content Detection

1. Introduction

Artificial Intelligence has revolutionized digital content creation by enabling the generation of highly realistic images, videos, and audio. Technologies such as Generative Adversarial Networks (GANs), deep learning models, and transformer-based architectures are capable of producing synthetic media that closely resembles real-world content. Although these technologies provide benefits in entertainment, virtual reality, education, and communication, they also introduce serious challenges such as deepfakes, misinformation, identity theft, and cyber fraud.

In recent years, AI-generated content has become increasingly accessible due to powerful generative models and publicly available tools. As a result, manipulated media can now be created with minimal technical expertise, making fake content more difficult to identify. This has created serious concerns regarding digital trust, online safety, and information authenticity across various industries. To address these issues, the proposed system “Reality Shield: AI Generated Content Detector” is developed as an intelligent platform for detecting AI-generated or manipulated multimedia content. The system analyzes images, videos, and audio using advanced machine learning and deep learning algorithms including CNN, SVM, Random Forest, Logistic Regression, Bayesian models, and Vision Transformer (ViT).

The application is designed with a React and JavaScript frontend, while the backend is implemented using REST APIs integrated with machine learning pipelines. The project also incorporates a CI/CD pipeline for continuous integration, deployment, and automated testing to ensure scalability and reliability.

Reality Shield aims to provide a secure and efficient verification platform that helps users identify manipulated media quickly and accurately. The system not only focuses on detection accuracy but also emphasizes usability, scalability, and real-time performance. By combining modern AI models with responsive web technologies, Reality Shield contributes toward building safer and more trustworthy digital environments.

2. Literature Survey

This study utilized CNN, RNN, and GAN-based approaches to identify manipulated facial images and videos. The research demonstrated that combining spatial and temporal analysis significantly improves detection accuracy [2022]. Researchers proposed a scalable architecture incorporating CNNs, RNNs, GANs, and Autoencoders for automated deepfake detection. The study highlighted the importance of multi-model integration for improved detection reliability [2024]. The paper examined deepfake generation technologies and discussed detection techniques based on GANs and sequential neural networks. It emphasized ethical concerns and security risks associated with synthetic media [2022]. The research combined CNN and LSTM architectures for detecting manipulated videos. Results indicated that hybrid models outperform individual classifiers in recognizing deepfake patterns [2022].

3. System Architecture

The Reality Shield architecture processes images, videos, and audio through preprocessing, feature extraction, and classification stages to detect AI-generated content. It uses machine learning models and REST APIs to provide real-time real/fake prediction results.

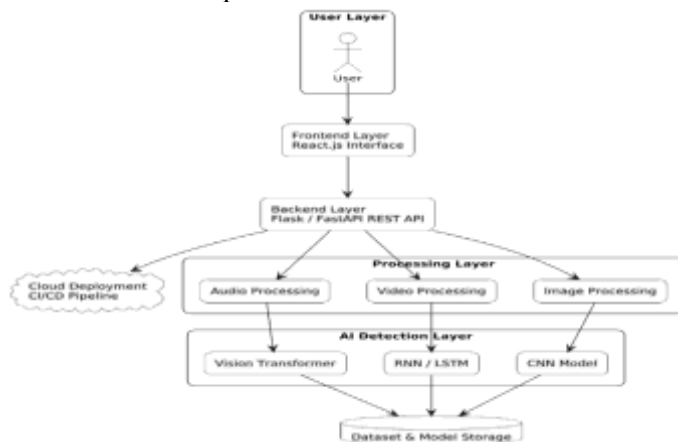


Fig. 3.1

3.1 Data Preparation

The collected multimedia datasets containing real and AI-generated images, videos, and audio samples are preprocessed before training the deepfake detection models. Data preparation improves the quality and consistency of the input data, which directly affects model performance and accuracy. Images and video frames are resized into fixed dimensions such as 224×224 pixels to ensure uniformity and reduce computational complexity. Normalization is applied by scaling pixel values between 0 and 1 to improve convergence during training. Data augmentation techniques including rotation, flipping, zooming, cropping, brightness adjustment, and noise injection are used to increase

dataset diversity and improve generalization capability. For video analysis, frames are extracted from uploaded videos to identify facial inconsistencies and temporal distortions. Audio preprocessing includes noise reduction, frequency normalization, pitch extraction, and spectrogram generation for identifying AI-generated voices.

3.2 Model Design

The proposed Reality Shield system uses machine learning and deep learning algorithms such as CNN, SVM, Random Forest, Logistic Regression, Bayesian classifiers, and Vision Transformer (ViT) models for multimedia deepfake detection. CNN models extract important visual features including textures, edges, GAN-generated artifacts, and pixel inconsistencies from images and video frames. Convolutional layers identify hidden manipulation patterns, while activation layers introduce non-linearity for learning complex features. Pooling layers reduce feature dimensions and computational complexity, and fully connected layers perform final classification of multimedia content as real or fake.

3.3 Model Training

The models are trained using labeled datasets containing authentic and manipulated multimedia samples. During training, multimedia inputs pass through neural networks, and the model updates its internal weights using backpropagation and optimization algorithms such as Gradient Descent and Adam Optimizer. Multiple training epochs help the system learn patterns related to facial distortions, GAN-generated artifacts, temporal inconsistencies, and cloned voice irregularities. The trained system generates prediction results along with confidence scores and validation metrics.

3.4 Model Architecture

The proposed Reality Shield architecture is designed as a multimodal deepfake detection framework capable of analyzing images, videos, and audio files. The architecture consists of input, preprocessing, feature extraction, classification, REST API communication, and output layers. Multimedia files uploaded through the React frontend undergo preprocessing operations such as resizing, normalization, frame extraction, spectrogram generation, and noise filtering. Extracted features are analyzed using CNN, Vision Transformer, and audio processing models, followed by classification using SVM, Random Forest, Logistic Regression, Bayesian classifiers, and deep learning models. REST APIs provide communication between frontend and backend modules, while the output layer displays prediction results, confidence scores, heatmaps, accuracy, and F1-score metrics.

4. Methodology

The proposed Reality Shield methodology focuses on detecting AI-generated multimedia content through stages such as data collection, preprocessing, feature extraction, model training, and evaluation. The methodology ensures accurate classification of manipulated images, videos, and audio files while improving multimedia verification efficiency.

4.1 Data Collection and Preprocessing

A diverse multimedia dataset containing real and AI-generated images, videos, and audio samples is collected from Kaggle datasets, deepfake benchmark datasets, public multimedia repositories, and research-based datasets. Each multimedia sample is labeled as real or fake based on authenticity, which helps the models learn classification patterns effectively.

During preprocessing, all image samples and video frames are resized into fixed dimensions such as 224×224 or 256×256 pixels to maintain consistency and reduce computational complexity. Normalization is applied by scaling pixel values between 0 and 1 to improve numerical stability and model convergence. Data augmentation techniques including flipping, rotation, zooming, cropping, and brightness adjustment are used to increase dataset diversity and improve model generalization capability.

For audio analysis, audio files are converted into spectrograms and frequency maps, enabling deep learning models to analyze voice characteristics and identify cloned or AI-generated audio patterns effectively.

4.2 Training the Deep Learning Model

The proposed system uses machine learning and deep learning models including CNN, Vision Transformer, SVM, Random Forest, Logistic Regression, and Bayesian classifiers for multimedia deepfake detection. During training, multimedia samples pass through the neural networks, and model parameters are updated using optimization algorithms such as Adam, SGD, and RMSprop.

Loss functions such as Binary Cross-Entropy and Categorical Cross-Entropy are used to measure prediction errors and improve model performance. Appropriate batch sizes and epoch values are selected to optimize training efficiency and detection accuracy. Through multiple training iterations, the models learn hidden manipulation patterns such as facial distortions, GAN-generated artifacts, temporal inconsistencies, and cloned voice irregularities.

4.3 Model Evaluation

The trained models are evaluated using separate validation datasets to ensure proper generalization on unseen multimedia samples. Performance evaluation is carried out using metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

The accuracy metric measures the percentage of correctly classified samples, while precision evaluates how many predicted fake samples are actually fake. Recall measures the ability of the system to correctly identify manipulated multimedia content. F1-score provides a balanced evaluation between precision and recall, especially for imbalanced datasets. Confusion matrix analysis helps visualize correct and incorrect classifications across multimedia categories.

The accuracy evaluation formula is:

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

The F1-score formula is:

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

Cross-validation techniques are also applied by dividing the dataset into multiple subsets and evaluating the model iteratively. This process improves model reliability, reduces overfitting, and enhances overall detection performance.

5. Result and Discussions

The proposed Reality Shield system successfully detects manipulated images, videos, and audio using machine learning and deep learning techniques.

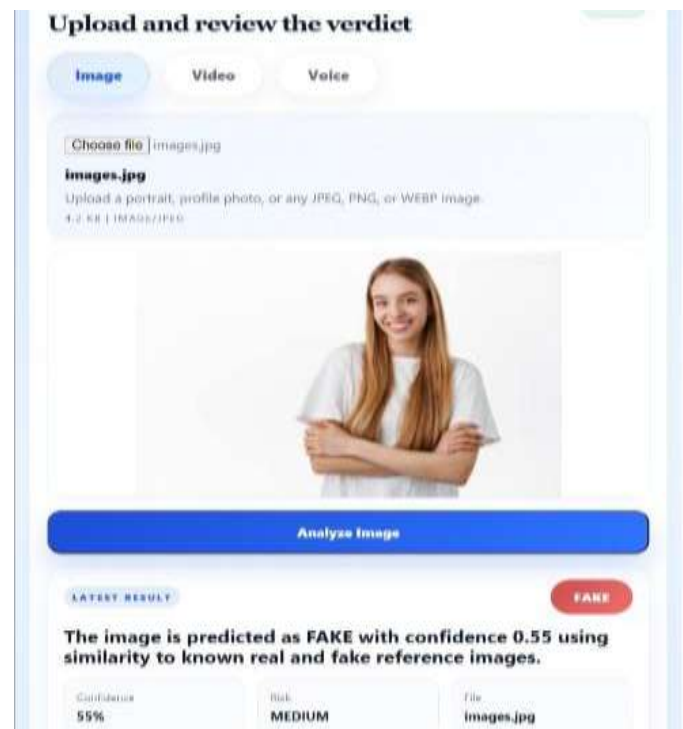


Fig. 5.1 Image Detection

The uploaded image was predicted as **FAKE** with a confidence score of 55%. The system identified visual inconsistencies and manipulation patterns associated with AI-generated images.

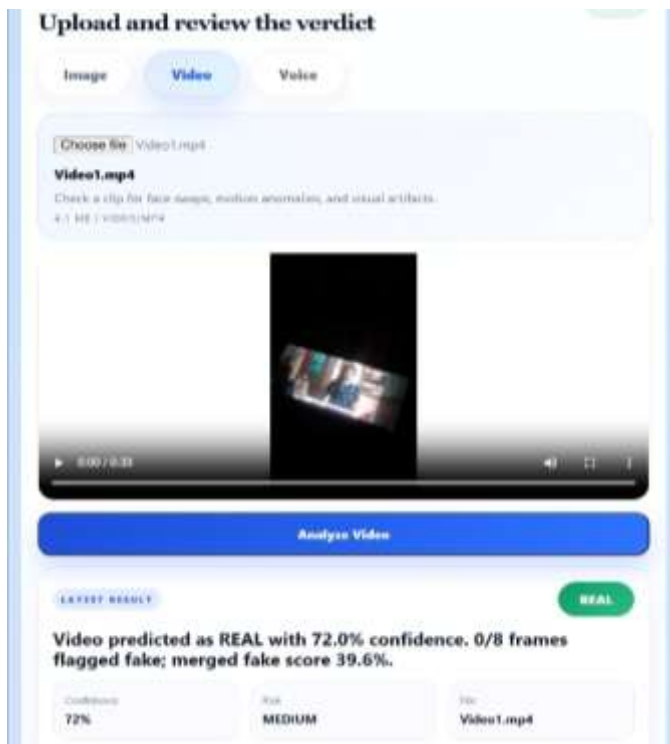


Fig 5.2 Video Detection

The uploaded video was predicted as **REAL** with a confidence score of 72%. The system analyzed video frames and detected no major facial manipulation, motion abnormalities, or deepfake artifacts.

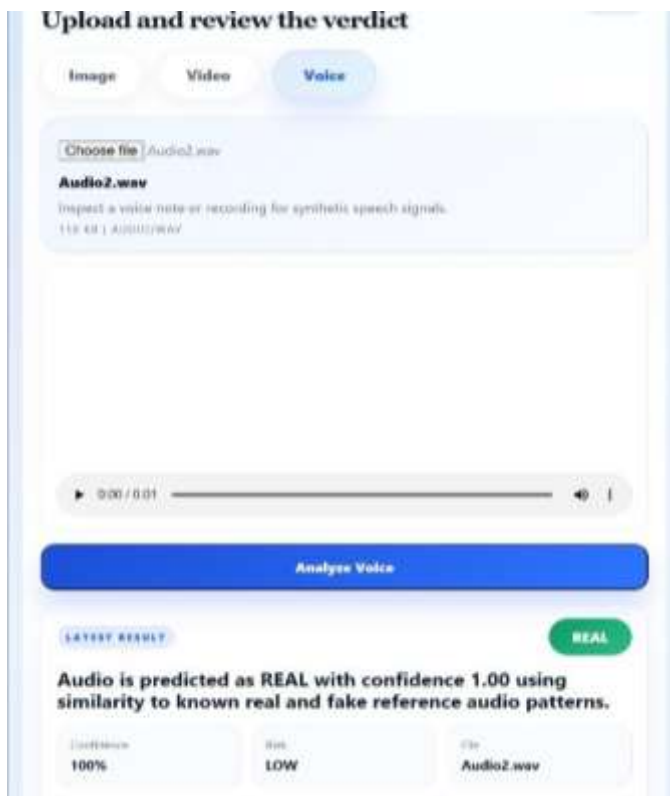


Fig. 5.3 Voice Detection

The uploaded audio sample was predicted as **REAL** with a confidence score of 100%. The system detected

authentic voice characteristics without significant signs of audio manipulation or cloning.

6. Conclusion

The proposed system, Reality Shield: AI Generated Content Detector, provides an effective solution for detecting AI-generated images, videos, and audio using machine learning and deep learning techniques. The system integrates algorithms such as CNN, SVM, Random Forest, Logistic Regression, Bayesian classifiers, and Vision Transformer to improve detection accuracy and reliability.

The project successfully performs multimedia preprocessing, feature extraction, classification, and real-time prediction using a React-based frontend and REST API backend. Experimental results show that the system can accurately identify manipulated multimedia content and provide confidence-based predictions. Reality Shield contributes toward reducing misinformation, improving cybersecurity, and strengthening digital trust in modern digital environments.

References

1. Rahman, A., Islam, M. M., Moon, M. J., Tasnim, T., Siddique, N., Shahiduzzaman, M., & Ahmed, S. (2022). A qualitative survey on deep learning based deep fake video creation and detection method. *Australian Journal of Engineering & Innovation Technology*, 4(1), 13-26.
2. Kingra, S., Aggarwal, N., & Kaur, N. (2023). Emergence of deepfakes and video tampering detection approaches: A survey. *Multimedia Tools and Applications*, 82(7), 10165-10209.
3. Geels, F. W. (2022). Causality and explanation in socio-technical transitions research: Mobilising epistemological insights from the wider social sciences. *Research Policy*, 51(6), 104537.
4. Chen, Y., Yang, X. H., Wei, Z., Heidari, A. A., Zheng, N., Li, Z., ... & Guan, Q. (2022). Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine*, 144, 105382.
5. Huang, Y., Lv, S., Tseng, K. K., Tseng, P. J., Xie, X., & Lin, R. F. Y. (2023). Recent advances in artificial intelligence for video production system. *Enterprise Information Systems*, 17(11), 2246188.
6. Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dute, I., Khan, S., & Parkinson, S. (2023). A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access*.
7. Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910-932. <https://doi.org/10.1109/JSTSP.2020.3000128>