International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deepfake Detection of AI Images and Audio Using Deep Learning

Prof. P. I. Kinikar, Shruti Jadhav, Srushti Gawali, Tanuja Kharat
Professor, Department of Electronics and Telecommunication Engineering Dr. Daulatrao Aher College of Engineering, Karad, Maharashtra, India Student Members: Shruti Jadhav, Srushti Gawali and Tanuja Kharat

t

ABSTRACT

The rapid development of generative artificial intelligence has changed the way digital images and audio can be produced, edited and shared. Modern generative systems can create realistic portraits, scenes and human-like speech, making it increasingly difficult to distinguish authentic recordings from synthetic or manipulated media through visual or auditory inspection alone. Such content is commonly described as deepfake media. While generative technologies support creative production, accessibility and entertainment, their misuse can contribute to misinformation, identity impersonation, financial fraud and harm to personal or institutional reputation. Therefore, dependable methods for analysing the authenticity of digital media are becoming increasingly important.

This review paper studies deepfake detection for AI-generated images and audio using deep learning techniques. It reviews Convolutional Neural Networks (CNNs) for learning spatial patterns in images, along with audio feature extraction and CNN/RNN-based models for analysing synthetic speech. The proposed framework accepts an image or audio file, applies preprocessing suitable for that media type, extracts useful patterns and classifies the input as likely real or fake. A result interface presents the predicted class with a confidence score. The review also considers datasets, model training, testing procedures and evaluation measures such as precision, recall, F1-score and ROC-AUC. Important limitations include the rapid evolution of generation methods, compression, noise, dataset bias and weak generalization to unfamiliar media. The intended system is a practical first-level screening tool; its output should support further verification rather than be treated as conclusive proof of authenticity.

Keywords: Deepfake detection, AI-generated images, synthetic audio, deep learning, CNN, RNN, spectrogram, image forensics, classification, confidence score.

1. INTRODUCTION

Digital images and audio recordings are used every day in news reporting, social media, online education, business communication and personal messaging. They are often accepted as evidence of events or as representations of what a person said or did. Generative artificial intelligence has made it possible to create convincing images and voices from limited inputs, and some tools can modify existing media with only small visible or audible changes. As these tools become easier to access, people may have difficulty deciding whether a file is genuine, edited or generated entirely by AI. Repeated sharing, resizing, filtering and compression can further hide the traces that might reveal manipulation.

Image deepfakes may contain unusual textures, inconsistent lighting, irregular edges, unnatural details or statistical patterns left by a generation process. Synthetic audio may show irregularities in spectral structure, phase information, rhythm, voice dynamics, pronunciation or background characteristics. However, these clues are not always visible or consistent. Genuine media can also contain blur, noise or unusual lighting, while advanced synthetic media may appear highly realistic. Manual inspection is therefore useful but cannot provide a consistent or scalable solution for every file.

Deep learning provides a data-driven approach by learning distinguishing patterns from labelled examples of authentic and manipulated media. A CNN can learn spatial features from image pixels, including local textures and more complex visual structures. For audio, a model can analyse raw waveform segments or time-frequency representations such as log-Mel spectrograms. CNNs can learn local acoustic patterns, while recurrent neural networks (RNNs), including LSTM or GRU variants, can model how those patterns change across successive frames. These approaches do not guarantee perfect detection, but they provide a systematic way to classify large amounts of media.

The aim of this review is to describe the main image and audio detection approaches and present a common framework that selects a suitable detector according to the uploaded file. The paper discusses relevant research directions, the proposed processing pipeline and practical limitations that must be considered before deployment. A reliable system should be evaluated on unseen examples, should communicate uncertainty clearly and should not make definitive claims based on a single model prediction.

2. LITERATURE SURVEY

2.1 Image deepfake detection. Research in visual forensics has explored CNN-based classifiers, frequency-domain analysis, attention mechanisms and transformer-based models. CNNs learn hierarchical representations, beginning with edges and textures and progressing to more complex visual features. These features can help identify blending artefacts, inconsistent details or other patterns associated with manipulated images. Frequency-based approaches examine statistical information that may be less apparent in ordinary visual inspection. Benchmark datasets such as FaceForensics++ and the DeepFake Detection Challenge (DFDC) have supported the development and comparison of image and face-manipulation detectors. Nevertheless, results can depend heavily on the dataset, the generation methods represented in training and the quality of the input file. A detector trained on a limited group of generators may perform poorly when it encounters a new synthesis model, an unfamiliar image style or media that has been heavily compressed.

2.2 Audio deepfake detection. Audio deepfake detection research investigates both waveform-based methods and feature-based approaches. A common method converts speech into a log-Mel spectrogram or another time-frequency representation, enabling a CNN to learn local acoustic patterns. Temporal models, including RNN and LSTM architectures, can capture changes across frames and provide information about the continuity of speech. Some approaches use learned representations from raw audio to reduce dependence on manually selected features. The ASVspoof challenges and WaveFake dataset are used in research on synthetic speech and spoofing detection. Evaluation remains difficult because audio varies with speaker identity, language, microphone, recording environment, codec and background noise. A model may learn irrelevant recording conditions instead of robust signs of synthesis if the training data is not sufficiently diverse.

2.3 Combined lessons from existing work. Findings from visual and speech deepfake research indicate that high performance on a single benchmark does not necessarily transfer to real-world media. The distinction between real and fake may become harder after resizing, recompression, filtering or re-recording. Dataset construction and evaluation procedures are consequently as important as model selection. Training and test partitions should be separated carefully to reduce leakage, and performance should be measured on examples that were not used for training or model tuning. Multiple measures, including precision, recall, F1-score and ROC-AUC, provide a more informative picture than accuracy alone. A practical detector should display a confidence estimate, allow an uncertain or inconclusive result and explain that its output is a prediction rather than a guarantee. These lessons motivate the proposed framework, which uses separate processing branches for images and audio and evaluates each branch according to the type of media it receives.

3. PROPOSED SYSTEM

The proposed system is a software-based screening framework designed to identify media that may have been generated or manipulated by AI. It combines an input interface, preprocessing stage, separate deep learning models for images and audio, a classification module and a result interface. The use of separate branches is important because images and audio contain different kinds of information and require different preprocessing methods. The image branch uses a CNN-based classifier to learn visual features, while the audio branch transforms sound into suitable features and applies a CNN, RNN or hybrid model. The output from either branch is converted into a predicted category and a confidence score.

The system is intended to follow a consistent workflow. First, the user uploads an image or audio file. The application checks the file type and whether the file can be processed. It then routes the input to the correct branch. Image preprocessing prepares pixel values and image dimensions for the visual model. Audio preprocessing standardizes the audio format and prepares waveform segments or spectrogram features for the audio model. The selected model extracts patterns learned during training and produces a classification result. The interface presents the result in a clear form and includes a caution that the score should not be interpreted as absolute proof of authenticity.

Model development uses labelled real and fake examples. Data is divided into training, validation and test sets so that the model can learn patterns, model settings can be selected, and final performance can be measured independently. The image model is trained to distinguish visual characteristics in authentic and synthetic images. The audio model is trained to distinguish acoustic patterns in genuine and generated speech.

Diverse sources and realistic variations should be included to reduce dependence on one dataset or recording condition. The system should be tested with unseen samples and media affected by compression or noise before any performance claims are made.

The final result interface should show the predicted class, confidence estimate and a short explanation of the system's limitations. If a file is corrupted, unsupported or outside the expected input conditions, the application should request another file rather than return an unreliable prediction. A high confidence score means the model strongly favours one of its learned classes; it does not prove that a file is genuine or fake. For sensitive decisions, results should be checked against trusted sources, provenance information and expert forensic analysis. In this way, the proposed system provides a common entry point for screening two media types while retaining modality-specific processing.

3.1 Block diagram

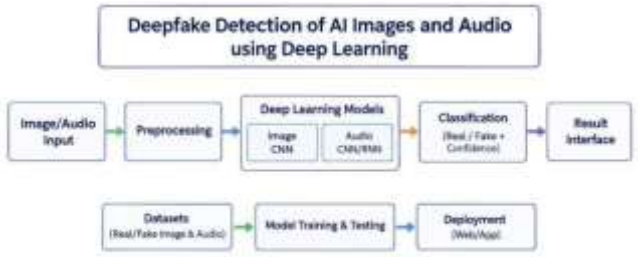


Figure 1. Block diagram of the proposed deepfake detection system.

3.2 Image preprocessing and CNN model

For image input, the application validates the file type, resizes the image to the model's required dimensions and normalizes pixel values. Depending on the implementation, face or region detection may be used, but the detector should not assume that every image contains a face. Augmentation such as mild resizing, compression and brightness variation can reduce overfitting to a narrow set of examples. The CNN learns hierarchical features: early layers detect edges and textures, while deeper layers combine them into more complex patterns. The final layer estimates the likelihood of the real and fake classes.

3.3 Audio preprocessing and CNN/RNN model

Audio input is converted to a consistent sample rate and channel format, then divided into manageable segments. A log-Mel spectrogram can represent energy across time and frequency, allowing a CNN to learn spectral artefacts. An RNN or other temporal model can capture changes across frames. During training, the dataset should include different speakers, recording conditions and synthesis methods. Silence trimming and loudness normalization may be applied carefully so that preprocessing does not remove useful clues.

3.4 Training and testing

The dataset is divided into training, validation and test partitions, avoiding leakage between partitions. The model learns from labelled real and fake samples; validation data supports model selection and threshold tuning, while the final test set estimates performance on unseen examples. Useful metrics include accuracy, precision, recall, F1-score, ROC-AUC and, for audio anti-spoofing, equal error rate (EER). Confusion matrices help identify false positives and false negatives. Cross-dataset evaluation is important because deployment data may differ from the training distribution.

4. SYSTEM WORKING AND RESULT INTERFACE

First, the user uploads an image or audio file through the interface. The application checks the file and selects the corresponding processing branch. Image files are normalized and passed to the image CNN. Audio files are transformed into waveform or spectrogram features and passed to the audio detector. The classifier produces class probabilities, and the interface displays the predicted category, confidence and a brief caution. If the file is unsupported, corrupted or outside the model's expected input conditions, the system should request a new input instead of presenting an unreliable result. The confidence score indicates how strongly the model favours a class under its learned decision process; it is not proof that a person is lying or that media is definitively fake. For important decisions, results should be checked using provenance information, trusted sources and expert forensic analysis.

5. APPLICATIONS

Potential applications include preliminary screening of suspicious social-media posts, support for digital-forensics teams, protection against voice impersonation attempts, media-literacy tools, newsroom verification workflows and research into synthetic-media detection. In educational settings, the system can demonstrate how deep learning learns patterns from labelled data. The detector can also be integrated into a web or mobile application so users can submit a file and receive an initial assessment.

6. CHALLENGES AND LIMITATIONS

Deepfake generation techniques change quickly, so a detector may become less effective against new generators. Compression, cropping, denoising, re-recording and background noise can hide or alter forensic traces. Datasets may overrepresent certain identities, languages, devices or manipulation methods, creating bias. A detector may also mistake unusual but genuine media for fake content. Model size and inference time can limit deployment on low-power devices. Finally, adversarial modification and the absence of a universally reliable ground-truth signal mean that no single detector can guarantee perfect decisions. These limitations can be reduced—but not eliminated—by using diverse datasets, testing on unseen generation methods, reporting more than one metric, calibrating confidence, documenting model limitations and periodically evaluating the deployed system. Privacy must also be considered: uploaded media should be handled securely, retained only as necessary and not reused for training without appropriate permission.

7. FUTURE SCOPE

Future work may explore transformer-based image encoders, self-supervised audio representations, multimodal models, explainable AI and efficient models for mobile devices. Provenance tools and cryptographic content credentials can complement learned detectors. The system can be improved with broader multilingual audio, more diverse image sources and a clear "inconclusive" category. Real-world testing across platforms and compression levels should be prioritized over relying solely on benchmark accuracy.

8. CONCLUSION

This review has examined deep learning approaches for detecting potentially AI-generated images and audio and outlined a framework that brings both tasks into one application. Image CNNs can learn spatial features such as texture, boundaries and visual inconsistencies, whereas audio models can analyse waveform or spectrogram features and use temporal networks to represent changes across speech. By applying separate preprocessing steps and modality-specific models, the proposed system can accept either type of media and present a real/fake prediction with a confidence estimate through a common result interface. This design makes the detection workflow easier to use while respecting the different characteristics of image and audio data.

The reliability of such a system depends not only on its model architecture but also on the quality and diversity of its datasets, the separation of training and testing data, and the realism of evaluation conditions. Compression, noise, unfamiliar generators and dataset bias can reduce performance, so evaluation should use independent samples and several metrics rather than a single accuracy figure. Confidence scores should be communicated carefully, and an inconclusive result should remain possible when the evidence is weak. Since synthetic-media techniques continue to evolve, ongoing testing and updating are necessary. Overall, deep learning can provide useful support for preliminary deepfake screening, but it should not replace human judgement, provenance checks or expert forensic analysis. Responsible data handling and transparent limitations are essential for practical use.

9. REFERENCES

[1] H.-H. Nguyen-Le et al., "Deepfake detection across image, video, and audio: a comprehensive survey with empirical evaluation of generalization and robustness," *Artificial Intelligence Review*, vol. 59, article 203, 2026.

[2] "A Survey on Speech Deepfake Detection," *ACM Computing Surveys*, vol. 57, no. 7, article 165, 2025, doi: 10.1145/3714458.

[3] J. Yi et al., "Audio Deepfake Detection: A Survey," arXiv:2308.14970, 2023.

[4] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[5] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.

[6] ASVspoof, "Automatic Speaker Verification Spoofing and Countermeasures Challenge," challenge datasets and evaluation resources.

[7] "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review," available through PubMed Central, 2024.