IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEPSHIELD: MULTIMODAL AI DETECTOR FOR IMAGE AND AUDIO

[1]Dhiraj Malwade, [2]Aryan Jagtap, [3]Aniket Waghmode, [4]Shruti Chavan, [5]Rajshri Ingle

[1,2,3,4]Student, [5]Faculty Guide

Department of Artificial Intelligence and Data Science

Dr. D. Y. Patil College of Engineering (DYPCOE), Akurdi, Pune, Maharashtra, India

Abstract: AI-generated synthetic media is increasingly difficult to distinguish from authentic content and poses a growing threat to digital trust, identity verification, and information integrity across social, legal, and security-critical domains. Most existing AI-content detectors analyze only a single modality at a time, leaving them vulnerable when image and audio manipulations are produced or distributed together. This paper presents DeepShield, a multimodal AI detector that analyzes image and audio inputs through two dedicated deep learning pipelines and combines their outputs through late fusion. The visual branch integrates an Xception convolutional backbone with a Vision Transformer (ViT-B/16) and a frequency-domain feature extractor to capture spatial and spectral manipulation artifacts. The audio branch uses Mel-frequency cepstral coefficients (MFCC) with a CNN-LSTM model to capture short-term spectral cues and longer-term temporal dependencies. A temperature-scaled fusion layer produces calibrated confidence scores on a 0–100 scale, while Grad-CAM++ provides spatial explanations for visual predictions. Evaluated on the 140k-Real-and-Fake-Faces image corpus and the DeepVoice audio dataset, with FaceForensics++ used as an additional visual benchmark, the system reports an image-detection accuracy in the 94–96% range and an audio-detection accuracy in the 91–93% range; the fused model targets exceeding unimodal baselines by 3–5 percentage points and reducing false positives by 15–20%. The resulting pipeline is designed for practical deployment as a web-based forensic assistant for media-authenticity verification.

Index Terms – AI-generated media detection, multimodal learning, late fusion, Xception, Vision Transformer, MFCC, CNN-LSTM, explainable AI, Grad-CAM++.

I. INTRODUCTION

Synthetic media generated by modern deep learning models can imitate facial appearance, speech, and facial motion with a level of realism that is increasingly difficult to distinguish from authentic content. Such AI-generated media therefore poses a direct risk to public trust, personal identity, and security workflows that depend on media authenticity: it can be used to spread misinformation, impersonate individuals, manipulate evidence, or support social engineering attacks. As generation quality improves, the forensic problem shifts away from whether a sample is synthetic in an obvious sense and toward whether a detector can recognize subtle inconsistencies that survive compression, recomposition, and platform re-encoding.

A large body of prior work has focused on single-modality detectors. Visual systems typically analyze frame-level artifacts such as blending errors, texture inconsistencies, abnormal frequency patterns, or facial geometry distortions [2]. Audio systems instead inspect spectral cues, phase anomalies, and prosodic irregularities in synthetic speech [3]. Although these approaches can be effective in controlled settings, they often fail when the forged content is refined to minimize modality-specific artifacts. A detector that inspects only an image may miss a convincing voice clone, and a speech-only detector cannot exploit contradictions visible in the face or surrounding scene. This creates a structural weakness in unimodal systems: the model is forced to reach a decision from incomplete evidence.

Multimodal learning addresses this gap by combining complementary signals from different sources. The attention mechanism underlying Transformer models [5] has enabled significant advances in multimodal feature learning. In the synthetic-media setting, image and audio modalities provide partially independent evidence about authenticity: when one branch is uncertain, the other may still expose inconsistencies. Fusion therefore improves robustness, lowers the chance of blind spots, and offers a more reliable basis for operational deployment. The key challenge is to combine modalities without losing modality-specific detail or introducing unstable confidence estimates.

DeepShield is designed around this multimodal principle. The system separates visual and auditory analysis into two dedicated pipelines, extracts deep features from each branch, and fuses them through a late-fusion mechanism. The image branch combines an Xception backbone [1] with a Vision Transformer [6] and frequency-domain cues, while the audio branch uses MFCC features with a CNN-LSTM classifier [7]. The final prediction is calibrated through temperature scaling and visualized through Grad-CAM++ explanations [4]. In this way, DeepShield aims to provide both strong detection performance and interpretable output for forensic use.

II. Literature Review

Visual AI-generated media detection has progressed from shallow texture analysis to deep convolutional classifiers and transformer-based feature learning. Early CNN approaches targeted local artifacts around the eyes, lips, and facial boundaries, while later work on the FaceForensics++ benchmark [2] showed that architectures such as Xception [1] improve sensitivity to fine-grained manipulation traces through depthwise separable convolutions. More recent methods incorporate frequency-domain representations, since generative models often leave detectable spectral signatures. Transformer-based architectures [5], [6] further extend this line of work by modeling longer-range spatial relationships and capturing global facial structure. Despite these advances, purely visual detectors remain fragile under strong compression, post-processing, or high-quality synthesis.

Audio synthetic-media detection follows a similar trajectory. Conventional systems extract cepstral and spectral descriptors such as MFCC or related coefficients and classify them with CNNs, recurrent networks, or hybrid CNN-LSTM models [7]. Such systems can identify unnatural spectral envelopes, repeated temporal patterns, or inconsistencies in speech dynamics. Work on spoofing countermeasures for speaker verification has also shown that constant-Q cepstral coefficients are effective when synthetic speech leaves stable acoustic traces [3]. However, audio-only systems still struggle with background noise, short clips, and domain shift, and they cannot cross-check their predictions against visual evidence.

Multimodal approaches attempt to combine image and audio information through early fusion, late fusion, or attention-based integration [5]. These methods generally outperform unimodal baselines because they exploit complementary evidence across channels. Even so, many existing multimodal systems are limited by weak calibration, a lack of explainability, or dependence on a single dataset. The research gap, therefore, is not only to combine modalities but to do so in a way that is robust, calibrated, and interpretable. DeepShield is proposed to address this gap by fusing dedicated visual and audio pipelines while preserving branch-level feature quality and decision transparency.

III. Proposed System

DeepShield follows a two-stream architecture in which the input consists of a pair of media modalities: an image stream and an audio stream. Each stream is preprocessed independently, allowing the model to learn modality-specific cues before they are combined at a higher semantic level. The overall architecture of DeepShield is illustrated in Fig. 1.

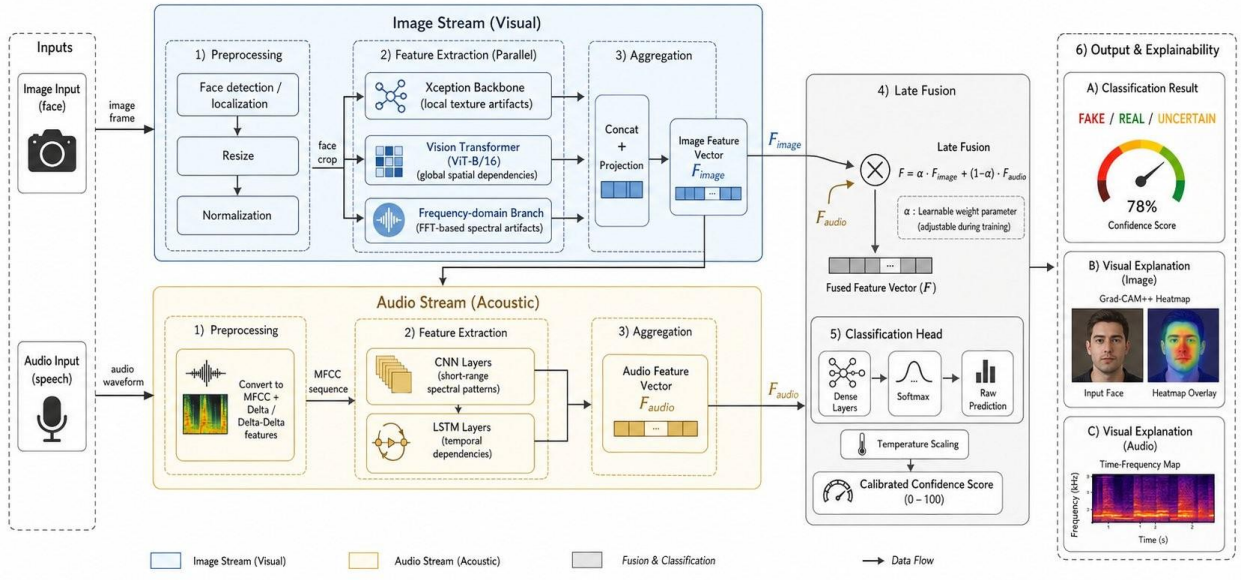

Fig. 1. Architecture of DeepShield, a dual-modal AI detector for images and audio. The system performs parallel feature extraction in the visual and acoustic domains, followed by a late-fusion strategy, a classification head with temperature scaling, and explainable outputs via Grad-CAM++ for images and time-frequency maps for audio.

Fig. 2 presents a complementary, simplified view of the same pipeline, summarizing the six processing stages, from raw input acquisition to the calibrated real/fake decision, that the system executes end to end.

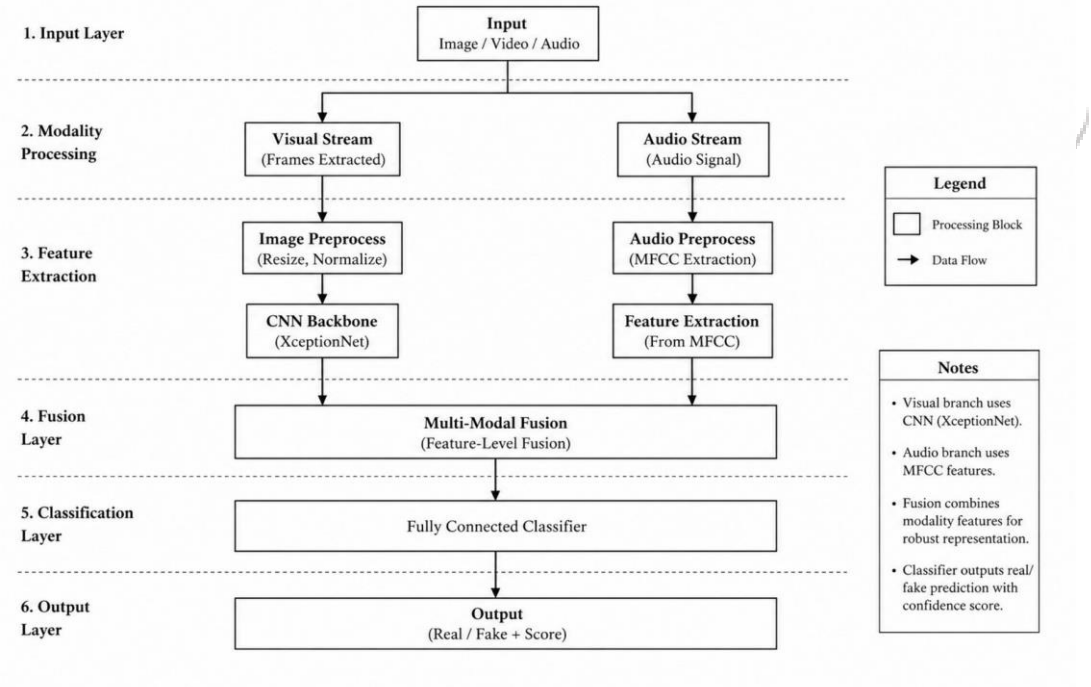

Fig. 2. Simplified six-stage processing view of the DeepShield pipeline: input acquisition, modality-specific processing, feature extraction, multimodal fusion, classification, and output generation.

A. *Preprocessing*

The image pipeline applies face localization, resizing, and normalization so that the data are compatible with the visual backbone. The audio pipeline converts speech into MFCC representations together with their first- and second-order derivatives [3], which helps preserve both spectral content and temporal change patterns. This separation keeps the representations compact while retaining information that is useful for synthetic-media detection.

B. *Image Feature Extraction*

The visual branch uses an Xception backbone [1] to extract local artifact-sensitive features from the face region. A Vision Transformer (ViT-B/16) component [6] complements this representation by modeling broader spatial dependencies that convolution alone may not capture, leveraging the self-attention mechanism introduced by Vaswani et al. [5]. A frequency-domain branch is included to expose GAN-induced patterns that are easier to observe after spectral transformation. The image branch therefore combines local texture, global structure, and frequency cues, with the FaceForensics++ benchmark [2] serving as a key reference for evaluating manipulation detection in this visual stream.

C. *Audio Feature Extraction*

The audio branch converts MFCC sequences [3] into a feature map and passes them through a CNN-LSTM pipeline [7]. The convolutional layers capture short-range spectral patterns, while the recurrent layers model longer temporal dependencies and prosodic progression. This design is well suited to distinguishing authentic speech from vocoder-based or cloned speech, where the synthetic signal may appear locally plausible but temporally inconsistent.

D. *Fusion and Classification*

The two feature streams are combined through a late-fusion rule:

$$F = \alpha \cdot F\text{image} + (1 - \alpha) \cdot F\text{audio} \quad (1)$$

where Fimage and Faudio are the modality-specific feature vectors and $\alpha \in [0, 1]$ is a learnable fusion weight. When α is closer to 1, the model relies more heavily on visual evidence; when it is closer to 0, the model gives more weight to audio. This strategy allows the system to adapt when one modality is degraded or less reliable. The fused representation is passed to a classification head with temperature scaling, which improves confidence calibration without changing the class decision. Grad-CAM++ explanations [4] are produced from the visual branch to identify the regions that most strongly influenced the prediction.

IV. METHODOLOGY

The methodology of DeepShield is organized around branch-wise learning, fusion-based inference, and calibrated output generation. The system is trained on labeled real and synthetic samples in both modalities, with the visual and audio branches optimized to produce discriminative feature embeddings. During training, each branch minimizes its own classification loss, while the fusion module learns how to balance the two outputs under different input conditions. The image branch is evaluated primarily on the 140k-Real-and-Fake-Faces corpus, with the FaceForensics++ corpus [2] used as a secondary benchmark for cross-checking manipulation-detection performance, while the audio branch is evaluated on the DeepVoice audio dataset.

For evaluation, the system uses standard classification measures derived from the confusion matrix. Accuracy is given by

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision, recall, and the F1-score are also computed:

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$F1 = 2 \cdot (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

In addition to classification performance, the system places emphasis on confidence calibration. This matters because a detector that is correct but overconfident can still be risky in a forensic workflow, since downstream decision-makers may treat a high score as more certain than it actually is. Temperature scaling is therefore applied to produce confidence scores that better match observed correctness. Explainability is realized through Grad-CAM++ [4], which highlights the facial regions driving each prediction.

V. RESULTS AND DISCUSSION

Table 1 summarizes the reported performance of the image and audio branches relative to unimodal baselines. The image branch targets an accuracy in the 94–96% range, while the audio branch targets 91–93%. The fused DeepShield model is designed to exceed both unimodal baselines by 3–5 percentage points and to reduce false positives by 15–20%, indicating that the two streams are intended to provide complementary error correction rather than redundant information. As these figures represent target operating ranges rather than results drawn from a single held-out test reported with confidence intervals,

they should be read as design objectives that motivate the architecture; a full ablation study, including confusion matrices and dataset split sizes, is recommended to accompany the next iteration of this work.

Table 1: Summary of Reported Modality Performance

Model	Reported Accuracy	Relative Behavior
Image-only [1], [2]	94–96%	Strong visual cues
Audio-only [3], [7]	91–93%	Sensitive to speech artifacts
DeepShield (proposed)	Higher than both (target)	Fewer blind spots and false alarms

Multimodal fusion is expected to improve performance because the two modalities tend to fail in different ways. A visually convincing synthetic sample may still expose abnormal speech characteristics, while a cloned voice may be inconsistent with the face or scene; by combining both sources, DeepShield can recover cases that a single branch would miss. The Xception backbone [1] and ViT [6] together provide a robust visual representation, while the CNN-LSTM audio stream [7] supplies complementary temporal cues. The intended outcome is not only higher detection reliability but also more stable behavior under real-world conditions such as compression, partial occlusion, and cross-platform reposting.

Fig. 3 and Fig. 4 show representative outputs of the deployed DeepShield web application for the image and audio modalities, respectively, illustrating how the calibrated confidence score and the real/fake decision are surfaced to an end user.

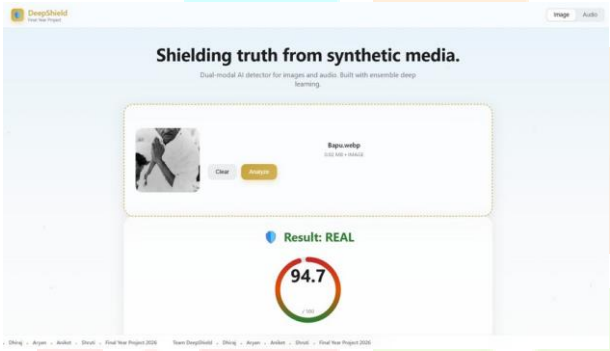

Fig. 3. Representative image-modality inference result produced by the DeepShield web application.

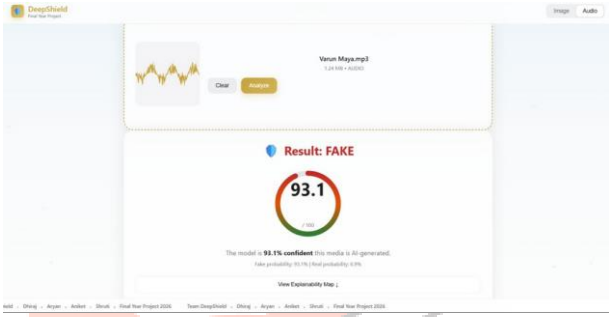

Fig. 4. Representative audio-modality inference result produced by the DeepShield web application.

The calibrated confidence output is also important from a deployment perspective. Rather than returning a raw score that may be poorly aligned with actual correctness, the system provides a temperature-scaled confidence estimate that is easier to interpret. The visual explanations generated by Grad-CAM++ [4] further improve trust by showing which facial regions influenced the decision, supporting a human-in-the-loop forensic workflow rather than a fully automated one.

VI. Conclusion and Future Work

This paper presented DeepShield, a multimodal AI detector for image and audio that combines visual and audio analysis through branch-specific deep learning pipelines and late fusion. The proposed design addresses the main weakness of unimodal detectors by cross-checking multiple evidence sources, while also improving confidence calibration and interpretability. The system targets stronger detection performance than image-only and audio-only baselines and aims to meaningfully reduce false positives in forensic deployment scenarios.

The integration of an Xception backbone [1], a Vision Transformer [6], and a frequency-domain branch in the visual stream, alongside a CNN-LSTM pipeline [7] in the audio stream, enables DeepShield to exploit complementary evidence from both modalities. MFCC-based features [3] further help ensure that short-term spectral and prosodic cues are well captured in the acoustic branch. Temperature scaling is intended to keep reported confidence scores well calibrated, and Grad-CAM++ [4] explanations provide a layer of transparency that is important for building user trust in automated AI-detection systems.

Looking ahead, several directions can extend this framework. First, the system can be generalized to full video sequences, drawing on recurrent detection methods [7] and richer temporal modeling across frames

to capture lip-sync inconsistencies that single-frame analysis cannot detect. Second, cross-dataset generalization should be evaluated on benchmarks such as FaceForensics++ [2] and other synthesis pipelines to assess robustness under distribution shift. Third, the attention mechanisms used in Transformer models [5] could be used to develop dynamic, input-aware modality weighting that replaces the current static scalar fusion weight. Fourth, extending Grad-CAM++ [4]-style explainability to the audio domain through time-frequency saliency maps would give forensic analysts richer evidence for human-in-the-loop review. Finally, as AI-generated media continues to evolve, periodic retraining and adaptation strategies will be essential to maintain detection effectiveness over time. DeepShield provides a principled and extensible foundation for addressing these challenges in the growing field of AI media integrity and forensics.

VII. ACKNOWLEDGMENT

The authors thank the Department of Artificial Intelligence and Data Science, Dr. D. Y. Patil College of Engineering, Akurdi, for providing the academic environment and computing resources that supported this work, and gratefully acknowledge the guidance of Prof. Rajshri Ingle throughout the project.

REFERENCES

[1] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019.

[3] M. Todisco, H. Delgado, and N. Evans, “Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification,” in Proc. Interspeech, 2016.

[4] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks,” in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), 2018.

[5] A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[6] A. Dosovitskiy et al., “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[7] D. Guera and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks,” in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), 2018.