International Journal of Technology and Applied Science (IJTAS)
E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

Explainable Multimodal Deepfake and Media Forensics Platform

Mrs. Vedashree M R[1], Harshita M[2], Taniesh S Kumar[3], Yashas Rao Jadhav[4]

[1,2,3,4]Department of Artificial Intelligence and Machine Learning
SJB Institute of Technology, Bangalore, India

Abstract:
Synthetic media can preserve realistic visual and acoustic characteristics while altering the identity, speech, or context represented by a file. This paper describes a mediaforensics platform that routes images, videos, and audio recordings through modality-specific detection and forensic pipelines. The workflow combines learned detectors with compression inspection, frequency analysis, metadata extraction, cryptographic hashing, and explanation maps. Rather than treating a detector score as conclusive evidence, the platform records the score together with the observations that produced it and exports the case information as a structured report. A FastAPI gateway, PostgreSQL storage, Redis messaging, and Celery workers separate user-facing requests from computationally intensive analysis. The prototype demonstrates how detection, evidence preservation, and result interpretation can be assembled in one investigationoriented workflow. Reported model comparisons are treated as preliminary until the datasets, partitions, checkpoints, and repeated-run statistics are fully documented.

Index Terms: Deepfake detection, digital media forensics, explainable artificial intelligence, multimodal analysis, media authentication

I. INTRODUCTION

Generative models now produce convincing faces, voices, and audiovisual sequences. These capabilities support legitimate creative applications, but they also make it easier to fabricate identity and context in digital evidence. Public benchmarks such as FaceForensics++, Celeb-DF, ASVspoof 2021, and DeepfakeBench illustrate both the growth of the field and the difficulty of comparing detectors under different data and evaluation protocols [6], [7], [10], [11].

A single detector rarely captures every useful trace. Visual artifacts may be weak in compressed frames, while timing, speech, metadata, or file-integrity information may provide separate leads. Multimodal research therefore examines relationships across media streams instead of relying only on one representation [1], [2], [4]. At the same time, a probability score by itself does not explain which evidence influenced a decision. Explanation methods such as Grad-CAM can expose influential image regions, although such maps must be interpreted as model evidence rather than proof of manipulation
[3], [8].

This work develops an investigation-oriented prototype that coordinates image, video, and audio analysis with evidence preservation and report generation. The main contributions are: 1) a routed workflow for

modality-specific processing; 2) integration of detector output with metadata, hash, compression, and frequency observations; 3) explanation views for supported models; and 4) an asynchronous backend that records case results and generates a structured forensic report.

II. MOTIVATION

Media authentication is not equivalent to binary classification. An investigator must be able to relate a decision to the submitted file, preserve identifiers for that file, and inspect supporting observations. This requirement becomes important when a manipulation is visible in one stream but less apparent in another. For example, a video may contain plausible frames while exhibiting inconsistent speech, timing, or metadata.

The prototype is therefore organized around evidence aggregation rather than a single universal detector. Each modality is analysed with tools appropriate to its signal characteristics, and the resulting observations are retained separately. This design avoids presenting Error Level Analysis (ELA), frequency components, metadata, or a neural-network score as independently conclusive. Instead, the report records multiple indicators that an investigator can assess together.

Interpretability provides a second motivation. A heatmap can show where a visual classifier concentrated its response, while a spectrogram or pitch trace can support examination of an audio sample. Such outputs do not establish authenticity on their own, but they make the computational result easier to inspect and challenge [3], [8].

III. RELATED WORK

FaceForensics++ and Celeb-DF provide widely used visual corpora for studying manipulated facial imagery and crossdataset generalization [6], [7]. DeepfakeBench later introduced a common framework, data pipeline, and evaluation protocol for multiple visual detectors, addressing inconsistencies that can otherwise make reported comparisons misleading [11]. M2TR and prototype-based methods such as PUDD investigate multimodal or representation-level strategies intended to improve robustness beyond a single visual cue [1], [4].

Audio spoofing research addresses text-to-speech, voice conversion, replay, and other synthetic-speech conditions. AASIST models spectro-temporal relationships with graph attention [9]. The ASVspoof 2021 challenge includes a dedicated speech-deepfake task and publishes evaluation protocols for bona fide and spoofed audio [10]. WavLM representations.

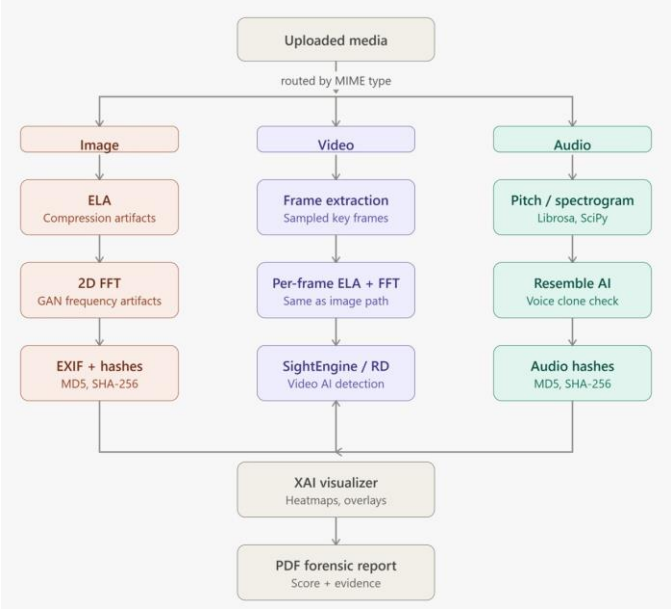


Fig. 1. Modality-specific analysis and reporting workflow.

have also been used in recent audio-visual deepfake systems, including an FPN-Transformer approach that models crossmodal temporal features [2].

Explainability research attempts to expose evidence used by a detector. Grad-CAM produces class-sensitive localization maps from convolutional features [8]. DF-P2E extends this direction by combining classification, Grad-CAM visualisation, and language-based explanation for non-expert users [3]. These studies motivate the explanation layer in the present prototype, but explanation maps remain dependent on the selected model and should not be interpreted as ground-truth localization.

The reviewed systems usually emphasize a detector, benchmark, or explanation method. The present work instead focuses on operational integration: routing heterogeneous media, preserving metadata and hashes, executing analysis asynchronously, and combining model and forensic outputs in a case report.

IV. PROPOSED SYSTEM

A. Overall Architecture

The application identifies the uploaded file by validated media type and directs it to an image, video, or audio workflow. Each workflow produces detector output and supporting observations. The explanation component receives compatible model outputs, while the reporting component stores the case record and assembles the final PDF. Fig. 1 summarizes this flow.

B. Image Processing

The image path computes ELA and a two-dimensional Fast Fourier Transform (FFT) representation. ELA is used as a compression-difference view, not as an automatic forgery verdict, because recompression, editing history, and file format can alter the observed error pattern. The FFT view is retained as another diagnostic representation that may reveal unusual periodic or spectral structure. Available EXIF fields are extracted, and cryptographic hashes are generated to associate later findings with the submitted file. These observations are stored separately from the learned detector score.

C. Video Processing

The video path samples frames for visual examination and applies the image diagnostics to the selected frames. Framelevel detector responses are retained with their timestamps before aggregation. This preserves the distinction between a local anomaly and the video-level result. Where supported, audio and visual timing can be analysed jointly because crossmodal temporal relationships may reveal manipulations that are weak in either stream alone [1], [2].

D. Audio Processing

The audio path produces spectrogram and pitch-related views and computes a file hash. A detector then estimates whether the signal resembles bona fide or synthetic speech. The system records preprocessing parameters and the detector version with the result. This is necessary because resampling, compression, segment length, and dataset mismatch can materially affect audio-spoofing performance [9], [10].

E. Explainable AI and Evidence Presentation

For compatible visual models, Grad-CAM maps are generated from the selected target layer [8]. The interface presents the map beside the original input and detector score. The report labels the map as model-attribution evidence and avoids describing highlighted regions as confirmed forgeries. Audio evidence is presented through signal views and detector output rather than a visual Grad-CAM analogy.

F. Forensic Report Generation

The reporting module records the file hash, case metadata, detector identity, model score, processing parameters, visual or acoustic evidence, and any warnings raised by the pipeline. This structure supports review and repeatability better than a stand-alone real/fake label.

V. SYSTEM IMPLEMENTATION

A. Backend Architecture

An investigator submits media through the user interface. A FastAPI gateway validates the request, records case metadata in PostgreSQL, and places an analysis task on a Redis-backed Celery queue. Fig. 2 shows the backend components and data flow.

B. Asynchronous Processing

Celery workers consume queued tasks and execute the required detector and forensic modules outside the requestresponse cycle. The API can therefore acknowledge ingestion without keeping the client connection open for the full analysis duration. Task state and failures are recorded so that an incomplete run is not presented as a completed examination.

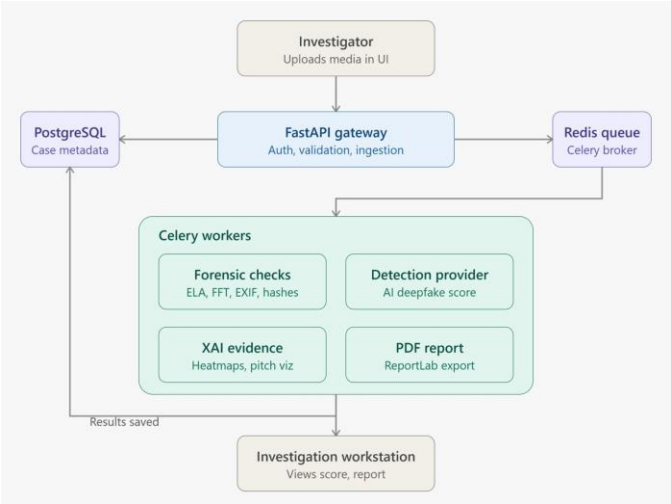


Fig. 2. Backend components for ingestion, asynchronous analysis, storage, and report delivery.

C. Forensic and Detection Processing

A worker selects operations according to the validated media type. Image and video jobs can include ELA, FFT views, EXIF extraction, hashing, frame sampling, and model inference. Audio jobs can include resampling, spectrogram generation, pitch-related analysis, hashing, and model inference. The output schema keeps observations, warnings, and scores in separate fields to prevent a diagnostic artifact from being silently converted into a definitive conclusion.

D. Database and Report Generation

PostgreSQL stores case identifiers, file hashes, model versions, task state, scores, and report locations. ReportLab assembles the completed record into a PDF after all mandatory tasks have succeeded. If a component fails, the report identifies the missing stage instead of omitting it without notice.

VI. EXPERIMENTAL SETUP

A. Dataset and Partitioning

The experiments were carried out using datasets and media samples prepared for the respective image and audio detection tasks. The data was organized according to the requirements of each modality and divided into suitable training and evaluation sets. Publicly available deepfake datasets such as FaceForensics++ and Celeb-DF have been widely used for developing and evaluating visual deepfake detection systems [6], [7]. These datasets provide manipulated and genuine media samples that support the evaluation of deepfake detection models under different manipulation conditions.

For audio-based detection, spoofed and genuine speech samples were prepared according to the requirements of the evaluated audio models. Audio anti-spoofing research such as AASIST demonstrates the importance of analyzing temporal and spectral characteristics for distinguishing genuine speech from spoofed speech [9]. The prepared image and audio data were subsequently processed using the preprocessing requirements of their respective models.

B. Preprocessing and Model Configuration

For image-based deepfake detection, five models were evaluated: ResNet50, EfficientNet-B4, ConvNeXt-Small, SwinBase, and ViT-Base. For audio-based detection, AASIST, WavLM, and P3 were evaluated. Each model was provided with the input format and preprocessing required by its respective architecture.

The evaluated models were selected to represent different deep learning architectures used for visual and audio deepfake detection. AASIST, in particular, is designed for audio antispoofing by modeling artifacts across temporal and spectral regions [9]. The trained models were then integrated into the corresponding processing pipelines of the proposed platform, where their outputs were used to obtain the deepfake prediction and confidence score.

C. Evaluation Metrics

For a binary test set with true positives *TP*, true negatives *TN*, false positives *FP*, and false negatives *FN*, the reported metrics are computed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$
$$\text{Precision} = \frac{TP}{TP + FP},$$
$$\text{Recall} = \frac{TP}{TP + FN},$$
$$F_1 = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}.$$

The area under the receiver operating characteristic curve (AUC) summarizes ranking performance across thresholds.

D. System Environment

The platform was implemented using Python-based backend and media-processing components. FastAPI was used to provide the backend API, while PostgreSQL was used for storing case information and analysis results. Redis and Celery were used to manage asynchronous processing tasks. The system also uses the required libraries for image, video, and audio analysis, along with ReportLab for automated PDF forensic report generation.

VII. RESULTS AND DISCUSSION

A. Image Detection Results

The results are presented as a comparison of the recorded model evaluations. Differences in datasets, preprocessing, and evaluation procedures should be considered when interpreting the values.

Within the recorded values, ConvNeXt-Small has the largest accuracy and AUC, whereas ViT-Base has the smallest values. This statement is limited to the entries in Table I; it does not establish statistical superiority or cross-dataset generalization. Those claims require repeated runs, uncertainty estimates, and a disclosed evaluation protocol.

TABLE I- PERFORMANCE COMPARISON OF IMAGE DETECTION MODELS

Model	Accuracy (%)	AUC (%)
ResNet50	94.67	98.07

EfficientNet-B4	93.73	98.34
ConvNeXt-Small	95.15	99.30
Swin-Base	92.93	95.59
ViT-Base	83.28	89.36

TABLE II- PERFORMANCE COMPARISON OF AUDIO DETECTION MODELS

Model	Accuracy (%)	AUC (%)	F_1 (%)
AASIST	84	91	78
WavLM-based model	82	88	86
NII's P3 (XLSR-large) + MLP	86	92	81

B. Audio Detection Results

AASIST, WavLM, and NII P3 (XLSR-large) + MLP were evaluated using accuracy, AUC, and F_1-score. As shown in Table II, P3 achieved the highest accuracy and AUC of 86% and 92%, respectively, while WavLM achieved the highest F_1score of 86%.

The recorded table assigns the largest accuracy and AUC to P3 and the largest F_1 value to the WavLM-based model. No broader effectiveness claim is made because the dataset, threshold, and run-to-run variation have not yet been documented.

C. Discussion and Limitations

The prototype demonstrates functional integration of detectors, forensic views, asynchronous processing, and report generation. The current evidence does not yet demonstrate that multimodal fusion improves accuracy over unimodal baselines, because the manuscript contains no ablation study or paired experiment. ELA and spectral views are diagnostic aids and may be affected by benign processing. Grad-CAM describes model attention, not causal proof of manipulation. Generalization to unseen generators, codecs, languages, capture devices, and social-media recompression remains to be evaluated.

VIII. FUTURE WORK

Future evaluation should use fixed, disclosed partitions and include cross-dataset testing. The system should compare unimodal models with explicit fusion baselines, report latency and resource use, and measure failure rates for unsupported or corrupted inputs. Video analysis should preserve temporal localization instead of only aggregating frame scores. The reporting module should add signed audit logs, model-card information, and confidence intervals. External detection services should be isolated behind versioned interfaces so that provider changes do not silently alter an experiment.

IX. CONCLUSION

This paper described an integrated prototype for examining image, video, and audio files in a common media-forensics workflow. The implementation combines model inference with hashes, metadata, compression and frequency views, explanation maps, asynchronous task execution, persistent case records, and PDF reporting. The design deliberately separates a detector score from the supporting

observations used to interpret it. The current numerical tables remain preliminary until their datasets, partitions, checkpoints, and evaluation procedures are documented. Accordingly, the contribution established here is the platform architecture and evidencehandling workflow, not a claim of universally superior detection performance.

REFERENCES:

1. J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, S. N. Lim, and Y.G. Jiang, "M2TR: Multi-modal multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimedia Retrieval (ICMR)*, 2022, pp. 615–623, doi: 10.1145/3512527.3531415.
2. C. Zheng, R. Suo, Z. Ji, J. Deng, F. Yi, C. Lin, and C. Shen, "Multimodal deepfake detection and localization with FPN-Transformer," arXiv:2511.08031, 2025, doi: 10.48550/arXiv.2511.08031.
3. S. Tariq, S. S. Woo, P. Singh, I. Irmalasari, S. Gupta, and D. Gupta, "From prediction to explanation: Multimodal, explainable, and interactive deepfake detection framework for non-expert users," in *Proc. 33rd ACM Int. Conf. Multimedia (MM)*, 2025, doi: 10.1145/3746027.3755786.
4. L. Pellicer, Y. Li, and P. Angelov, "PUDD: Towards robust multi-modal prototype-based deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, doi: 10.1109/CVPRW63382.2024.00385.
5. V. Nadimpalli and A. Rattani, "Facial forgery-based deepfake detection using fine-grained features," arXiv:2310.07028, 2023.
6. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niess-¨ ner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1–11.
7. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3207–3216.
8. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
9. J. Jung, H. Heo, H. Tak, H. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and T. Kinnunen, "AASIST: Audio anti-spoofing using integrated spectrotemporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022.
10. J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 47–54, doi: 10.21437/ASVSPOOF.2021-8.
11. Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, doi: 10.52202/0752800201.