International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Deep Learning Framework for Real-Time Detection of Deepfake Images, Audio and Videos

Srijib Samanta

Assistant Professor, Head of Department (HOD) Department: Bachelor of Computer Application (BCA)
Tamralipta Institute of Management & Technology

ABSTRACT:
Deepfakes have progressed from visibly defective face swaps to high-quality synthetic images, cloned speech and audio–visual manipulations that can survive social-media compression. A detector intended for real-time use must therefore do more than classify isolated frames: it must capture spatial traces, temporal inconsistency, acoustic artifacts and cross-modal synchronisation while operating within strict latency and memory limits. This paper develops a unified deep learning framework for detecting deepfake images, audio and videos. The study follows a structured research review and design-science methodology; it proposes an architecture and evaluation protocol rather than claiming fabricated experimental results. The framework contains a lightweight visual branch, a self-supervised acoustic branch, a temporal video branch and an audio–visual consistency branch. Reliability-aware gated fusion combines available modalities without forcing a missing or corrupted stream to contribute equally. A streaming inference layer uses face tracking, voice-activity detection, adaptive frame sampling, sliding windows, calibration and early exit to support real-time operation. The proposed assessment emphasises cross-dataset and cross-generator generalisation, area under the receiver operating characteristic curve, equal error rate, calibration, attack-wise recall, latency, throughput and memory. FaceForensics++, Celeb-DF v2, DFDC, ASVspoof 2021 and FakeAVCeleb are identified as complementary data sources, with identity-disjoint splitting and compression-aware augmentation required to prevent leakage. The analysis argues that multimodal fusion improves evidential coverage but does not automatically guarantee robustness: dataset bias, unseen synthesis methods, adversarial post-processing and demographic imbalance remain central threats. The paper concludes that operationally credible deepfake detection requires open-set evaluation, uncertainty-aware decisions, interpretable evidence and secure human review rather than a single opaque probability score.

Keywords: deepfake detection; multimodal learning; audio forensics; video forensics; real-time inference; cross-modal fusion; deep learning; digital authenticity.

1. INTRODUCTION:

Synthetic media can support film production, accessibility and creative communication, but deceptive use threatens identity, financial security, journalism and evidentiary trust. Image generators can synthesise plausible faces, voice cloning can imitate a speaker from limited audio, and talking-head systems can coordinate facial motion with generated speech. Detection consequently faces a moving-target problem: models trained on one generator often learn incidental traces that disappear under a new generator, codec or post-processing pipeline.

The research problem is to design a detector that is accurate, generalisable and fast enough for streaming use. Real-time performance is defined here as sustained processing at or above the incoming media rate under a declared hardware and latency budget—not merely a short average inference time measured without decoding, face tracking or audio preprocessing.

2. REVIEW OF RELATED LITERATURE:

2.1 Visual Deepfake Detection

Early visual detectors used convolutional neural networks to learn blending, colour and frequency artifacts. FaceForensics++ established a large supervised benchmark containing several facial manipulation families and compression levels (Rössler et al., 2019). Celeb-DF later increased realism and reduced obvious visual defects, exposing the weakness of detectors that depend on simple artifacts (Li et al., 2020). Contemporary systems combine local facial regions, frequency features, patch relations and transformer attention, but benchmark results remain sensitive to preprocessing and split construction.

2.2 Temporal Video Evidence

A video contains evidence unavailable in a single image: inconsistent identity features across frames, unstable boundaries, implausible motion and temporal frequency patterns. Three-dimensional CNNs, recurrent networks and temporal transformers model these dependencies. Dense processing of every frame, however, increases latency. Efficient deployment uses tracking, frame sampling and short temporal windows, while aggregating clip scores into an auditable video-level result.

2.3 Audio Deepfake Detection

Synthetic and voice-converted speech can reveal phase, prosody, excitation and vocoder artifacts. Spectrogram CNNs and raw-waveform networks remain useful, while self-supervised speech encoders provide representations learned from diverse natural speech. ASVspoof 2021 introduced realistic codec and channel variability for spoofed and deepfake speech (Liu et al., 2023). Recent work shows that multilingual pretrained representations may improve robustness to pitch, accent and language variation (Phukan et al., 2024), yet cross-domain evaluation continues to expose substantial degradation (Li et al., 2024).

2.4 Multimodal and Synchronisation-Based Detection

Multimodal systems examine whether mouth motion, phonetic content, speaker identity, emotion and timing agree. FakeAVCeleb supplies fine-grained real/fake audio and video combinations, enabling a detector to learn that either stream—or both—may be manipulated (Khalid et al., 2021). Audio–visual representation learning and cross-attention can detect inconsistency that a unimodal model misses. Nevertheless, fusion can overfit dataset-specific synchronisation errors, and modern generators can produce well-synchronised fakes. The detector should therefore combine intrinsic unimodal traces with cross-modal evidence.

2.5 Benchmarking and Generalisation

DeepfakeBench standardises data processing, models and metrics across multiple visual datasets, demonstrating the importance of uniform evaluation (Yan et al., 2023). High within-dataset accuracy can coexist with weak cross-dataset performance because identities, compression, generator artifacts and collection conditions leak into the label. Operational assessment must hold out generator families, identities and domains, then test recompression, resizing, noise, reverberation and partial manipulation.

Research stream	Strength	Unresolved limitation

Spatial visual	Fine local and frequency artifacts.	Generator- and codec-specific shortcuts.
Temporal video	Motion and frame-to-frame consistency.	Compute cost and short-window blindness.
Audio	Vocoder, phase, prosody and speaker traces.	Language, codec and channel shift.
Audio–visual	Lip–speech and identity coherence.	Synchronized fakes and missing modalities.
Benchmarking	Standardised comparison.	Real-world open-set drift remains.

3. OBJECTIVES AND RESEARCH METHODOLOGY:

3.1 Objectives

- To formulate a unified deep learning architecture for image, audio and video deepfake detection.
- To combine modality-specific forensic traces with audio–visual consistency evidence.
- To design a streaming inference process that satisfies stated latency, throughput and memory constraints.
- To establish a leakage-resistant, cross-dataset and cross-generator evaluation protocol.
- To incorporate calibration, explainability, abstention and human review into the operational decision process.

3.2 Research Questions

- How can spatial, acoustic, temporal and cross-modal evidence be fused without allowing a weak modality to dominate?
- Which design choices permit real-time processing while retaining cross-domain detection performance?
- How should generalisation, calibration, fairness and latency be measured together?
- How can detector evidence be communicated responsibly to investigators and platform moderators?

3.3 Research Design

The paper employs a structured narrative review and design-science approach. Peer-reviewed dataset papers, benchmark studies and recent multimodal/audio research inform the threat model and architecture. The design is evaluated conceptually against four requirements: evidential coverage, robustness to distribution shift, computational feasibility and decision accountability. No accuracy, latency or ablation result is claimed without a completed implementation.

3.4 Threat Model

The system considers fully synthetic images, face swaps, facial reenactment, lip-sync editing, voice conversion, text-to-speech cloning and mixed audio–video attacks. It also considers partial manipulation, recompression, resizing, background noise, reverberation and missing audio. The primary adversary seeks to make deceptive media appear authentic; adversarial examples specifically optimised against the detector are treated as an advanced threat requiring separate red-team assessment.

3.5 Datasets and Split Protocol:

Dataset	Modality and role	Required precaution
FaceForensics++	Images/video; multiple facial manipulations and compression.	Identity-disjoint split; test several compression levels.
Celeb-DF v2	High-quality celebrity face videos.	External evaluation; avoid identity overlap.
DFDC	Diverse large-scale manipulated videos.	Control actors and acquisition source.
ASVspoof 2021 DF	Synthetic/converted speech under codec variation.	Report EER by codec and attack.
FakeAVCeleb	Fine-grained real/fake audio–video combinations.	Evaluate every modality combination.
In-the-wild samples	Operational domain and post-processing.	Document provenance and consent.

3.6 Data Governance:

Dataset licences, subject consent, identity protection and permitted use must be documented. Splits will be performed by identity and source video before frame extraction. Near-duplicate detection will prevent the same underlying content from entering training and testing. Demographic attributes, when legitimately available, will be used only for performance auditing and not identity inference.

4. MULTIMODAL FRAMEWORK:

The proposed framework processes media through modality-specific encoders and a reliability-aware fusion layer. It accepts a still image, an audio clip or an audio–video stream and produces calibrated modality scores, temporal localisation and a final decision with uncertainty.

Stage	Operation	Output
1. Ingestion	Decode media; verify container; extract timestamps and metadata.	Aligned raw streams and integrity log.
2. Localisati on	Detect/track faces; identify speech; segment active windows.	Face tracks and speech segments.
3. Visual encoding	Spatial RGB/frequency encoder plus local artifact head.	Frame/patch embeddings and heat maps.
4. Audio encoding	Waveform or log-mel branch with self-supervised representation.	Segment embeddings and acoustic saliency.
5. Temporal encoding	Causal temporal transformer over sampled frame/audio tokens.	Clip-level dynamics.
6. Consistenc y	Audio–lip alignment, speaker–face and semantic coherence.	Cross-modal discrepancy vector.
7. Gated fusion	Weight modalities by quality, availability and predictive uncertainty.	Fused representation and per-stream contributions.
8.	Calibration, threshold/abstention and temporal	Probability, uncertainty,

Decision	aggregation.	evidence and review status.

4.1 Mathematical Formulation:

For visual, audio, temporal and consistency embeddings z_v, z_a, z_t and zc, a quality network estimates reliability weights α_m. Missing modalities receive a mask of zero; available weights are normalised:

$$\alpha_m = \frac{exp(g_m(z_m, q_m))}{\Sigma_{m \in A} exp(g_m(z_m, q_m))}$$

; $z = \Sigma_m \in A\ \alpha_m\ z_m\ q_m$

contains observable quality such as face size, blur, signal-to-noise ratio and synchronisation confidence, and A is the set of available modalities. The fused classifier estimates

$p(y = fake \mid z)$.

Temperature scaling on a held-out validation set produces a calibrated probability. An abstention rule sends high-uncertainty or out-of-distribution samples to human review.

4.2 Reliability-Aware Fusion:

Simple concatenation assumes that every modality is equally trustworthy. Gated fusion instead learns to reduce the influence of silent, noisy, occluded or corrupted streams. Modality dropout during training prevents the model from becoming dependent on audio–visual pairing. Auxiliary losses preserve separate image, audio and video competence so that cross-modal consistency complements rather than replaces intrinsic forensic evidence.

4.3 Temporal Localisation:

A video-level label does not reveal where manipulation occurs. The temporal head therefore returns window-level scores and highlights suspicious intervals. Multiple-instance learning can aggregate a small set of manipulated windows into a video decision. Localisation is evaluated through temporal intersection-over-union when ground truth exists, and otherwise reported as exploratory evidence rather than verified explanation.

5. MODALITY-SPECIFIC PROCESSING:

5.1 Image and Frame Branch

The visual path uses a lightweight convolutional or compact transformer backbone. Detected faces are aligned without excessive smoothing, because aggressive preprocessing can erase forensic traces. Two complementary views are used: RGB appearance and a frequency-residual representation. A local head samples eyes, mouth, boundary and skin regions; a global head models whole-face coherence. Training includes real-world transformations such as JPEG compression, resizing, blur, colour shifts and screen recapture.

5.2 Audio Branch

Audio is resampled consistently and divided into overlapping speech-active windows. A compact self-supervised speech encoder or spectrogram network captures local spectral and longer-range prosodic evidence. Channel augmentation includes codecs, microphone impulse responses, noise, reverberation and bandwidth restriction. The system reports EER and attack-wise recall because accuracy can conceal class imbalance and operational threshold behaviour.

5.3 Video Temporal Branch

A face tracker maintains identity across frames and avoids repeated detection. Adaptive sampling selects a base frame rate and adds frames around rapid motion or score changes. Per-frame embeddings enter a causal temporal transformer or temporal convolution, enabling streaming without future frames. A

bounded cache prevents memory from increasing with video duration.

5.4 Audio–Visual Consistency Branch

Mouth-region features are aligned with acoustic units over short time windows. Cross-attention estimates whether visible articulation and audio timing agree. Additional heads may compare speaker and face identity embeddings only as consistency signals; they must not expose identity labels. An important safeguard is to train with authentic dubbed, noisy and naturally asynchronous material so the model does not label every mismatch as malicious.

5.5 Feature and Decision Losses:

The total training objective can combine fused classification, modality-specific classification, temporal localisation, contrastive consistency and calibration regularisation:

$$L = Lfused + \lambda^1 \Sigma_m Lm + \lambda^2 Ltemporal + \lambda^3 Lconsistency + \lambda^4 Lcalibration$$

Focal or class-balanced loss may be used when attack categories are uneven. Supervised contrastive learning groups authentic content across codecs while separating diverse manipulations. Generator labels may support auxiliary training but must never appear across both train and test when evaluating unseen-generator generalisation.

5.6 Preventing Shortcut Learning:

- Split identities and source videos before extracting clips or frames.
- Balance acquisition source, codec and resolution across real and fake classes.
- Randomise backgrounds and non-face regions so labels cannot be inferred from context.
- Audit saliency for watermarks, borders, subtitles and dataset-specific artefacts.
- Use generator-held-out and dataset-held-out tests as principal evidence.
- Create counterfactual pairs in which only one modality is manipulated.

Input condition	Active evidence	Expected behaviour
Still image	Spatial/local/frequency	Calibrated image score.
Silent video	Spatial + temporal	Clip and video score; no audio penalty.
Audio only	Acoustic/prosodic	Audio score and EER-oriented threshold.
Audio–video	All branches	Fused score plus consistency evidence.
Low quality/missing stream	Available branches + quality gate	Reduced confidence or abstention.

6. TRAINING AND EVALUATION PROTOCOL:

6.1 Training Strategy

The encoders are first trained or adapted on their modality-specific corpora, after which the fusion and temporal heads are trained on multimodal samples. Final end-to-end fine-tuning uses a low learning rate and balanced batches across real/fake status, manipulation family and quality level. Modality dropout, mixup within the same label and compression augmentation improve resilience. Validation data determine early stopping, temperature scaling and operating thresholds; the test set remains untouched.

6.2 Evaluation Scenarios

Scenario	Purpose	Minimum report
Within-dataset	Verify learning under familiar distribution.	AUC, F1, EER/audio, calibration.
Cross-dataset	Test acquisition and population shift.	AUC and attack-wise recall.
Unseen generator	Test open-set manipulation.	Macro AUC/recall by held-out generator.
Post-processing	Test codec, resize, noise and blur.	Performance-versus-severity curves.
Missing/corrupt modality	Test gated fusion and fallback.	Failure rate, calibration, abstention.
Streaming deployment	Test true operational feasibility.	End-to-end latency, FPS, memory, energy.

6.3 Metrics

- Area under the ROC curve (AUC) for ranking independent of a fixed threshold.
- Precision, recall, F1 and confusion matrices at a declared operating threshold.
- Equal error rate (EER) and min-tDCF where relevant to audio anti-spoofing.
- Expected calibration error and Brier score for probability reliability.
- False-positive rate at a high true-positive operating target for screening applications.
- Frame/clip throughput, p50/p95 latency, peak memory and model size.
- Temporal intersection-over-union and localisation average precision when labels exist.
- Subgroup and attack-wise performance with confidence intervals.

6.4 Statistical Analysis

Confidence intervals will be obtained by bootstrapping at the media-file level rather than treating correlated frames as independent samples. Paired comparisons between models will use the same test media and appropriate tests for AUC or error differences, with multiplicity control for several primary comparisons. Performance will be reported across at least three training seeds. Calibration and threshold selection will be conducted only on validation data.

6.5 Ablation Study

Ablation experiments will remove frequency features, temporal modelling, consistency learning, quality gating, modality dropout and calibration one at a time. Each ablation will report quality and latency so that a component is retained only when its gain justifies its computational cost. A lightweight unimodal baseline and a simple score-average baseline provide essential comparisons.

6.6 Success Criteria

The framework will be considered promising only if it improves cross-dataset or unseen-generator performance over efficient baselines, remains calibrated under common post-processing and sustains the incoming stream rate on declared hardware. No universal accuracy target is predetermined because datasets and operating costs differ; application thresholds will be based on acceptable false-positive and false-negative risks.

7. REAL-TIME IMPLEMENTATION:

7.1 Streaming Pipeline

A practical system separates ingestion, preprocessing, inference and reporting into asynchronous stages. Hardware decoding supplies timestamped frames and audio. Face detection runs periodically; tracking propagates boxes between detections. Voice-activity detection prevents silent audio from consuming inference. Short windows enter batched modality encoders, while a causal temporal cache maintains context. Results are emitted incrementally and revised as additional evidence arrives.

7.2 Latency Budget

Component	Illustrative budget	Optimisation
Decode and resample	5–10 ms/window	Hardware decode; zero-copy buffers.
Face detection/tracking	8–15 ms	Detect sparsely; track between detections.
Visual encoder	8–18 ms	Compact backbone; mixed precision; batching.
Audio encoder	5–12 ms	Voice-active windows; cached features.
Temporal/consistency	3–8 ms	Bounded causal attention.
Fusion/calibration	<3 ms	Small gated head.
Total target	Below media window duration	Pipeline overlap; p95 monitoring.

The values above are engineering targets, not measured results. The implemented study must report actual p50 and p95 end-to-end latency—including decoding and preprocessing—on named CPU/GPU hardware. Throughput must be tested under concurrent streams, because a single-file benchmark does not establish production capacity.

7.3 Efficiency Techniques

- Mixed-precision inference and hardware-supported quantisation after calibration.
- Knowledge distillation from large modality encoders to compact students.
- Dynamic batching constrained by maximum waiting time.
- Early exit when several independent high-confidence cues agree.
- Adaptive frame rate based on motion and uncertainty.
- Feature caching for stable face tracks and repeated prefixes.
- Model compilation and operator fusion on the target runtime.

7.4 Decision and Human Review

The output should not state that a person created or endorsed a deepfake. It should report that the media contains evidence statistically associated with manipulation, together with confidence, modalities examined, quality warnings and suspicious intervals. A three-way policy—authentic-like, suspicious and inconclusive—reduces pressure to force uncertain media into a binary label. High-consequence decisions require trained human review and corroborating provenance evidence.

7.5 Monitoring

Production monitoring will track score distribution, abstention rate, latency, modality availability and confirmed error cases. Drift alarms should trigger when new codecs, languages or generators alter embeddings or calibration. Confirmed samples may enter a controlled retraining queue only after provenance, consent and label quality checks. Model and threshold changes require versioning and regression tests.

7.6 Deployment Security

Detector endpoints can themselves be attacked through excessive requests, malformed media or probing designed to infer decision boundaries. Controls include file-type validation, sandboxed decoding, request limits, signed model artifacts, audit logs and separation between public-facing scores and internal forensic detail. The system should be red-teamed against adaptive perturbations without publishing instructions that facilitate evasion.

8. EXPECTED CONTRIBUTION, RISKS AND ETHICS:

8.1 Expected Technical Contributions

- A single framework that accepts image-only, audio-only, silent-video and audio–video inputs.
- Reliability-aware gated fusion that handles missing, noisy and manipulated modalities.
- A causal streaming design with explicit end-to-end latency and resource metrics.
- A leakage-resistant protocol centred on identity-, dataset- and generator-held-out evaluation.
- Calibrated abstention, interpretable temporal evidence and human-review integration.

8.2 Anticipated Findings

The multimodal model is expected to be strongest when independent forensic traces and cross-modal inconsistency are simultaneously available. The visual branch may dominate face swaps; the audio branch may dominate cloned speech; and the consistency branch may improve lip-sync detection. Gains are expected to narrow on fully synthesised, well-synchronised audio–video content. Quality gating should improve stability when one stream is absent or heavily degraded. These are testable expectations, not claimed results.

8.3 Generalisation and Failure Modes

Failure mode	Cause	Mitigation
Unseen generator	Training on generator-specific traces.	Diverse generators, held-out testing, continual audit.
Codec/post-processing	Forensic traces suppressed or altered.	Compression/channel augmentation and severity curves.
False positive on authentic edit	Benign dubbing, enhancement or resampling.	Benign-edit negatives and inconclusive class.
Demographic disparity	Unequal subjects, language or capture quality.	Stratified audit, reweighting, targeted data collection.
Missing modality	Silent clip, damaged audio or occluded face.	Masks, quality gates and calibrated fallback.
Adaptive attack	Perturbations optimise against detector.	Ensembles, provenance, red-team tests, restricted detail.

8.4 Explainability

Heat maps and attention weights are not proof of causation. The framework will present them as model evidence and verify whether highlighted regions remain stable under small benign transformations. Audio explanations may indicate time–frequency regions; video explanations may identify suspicious intervals. Explanations should be accompanied by modality quality, uncertainty and known limitations.

8.5 Fairness and Privacy

False accusations can harm individuals and groups. Performance must be audited across gender presentation, skin tone, age, language, accent and recording quality where reliable annotations and

lawful use permit. Face and voice embeddings are sensitive biometric data and must be minimised, encrypted and retained only for the authorised purpose. Public release of samples requires consent or a defensible research basis and removal of unnecessary identity information.

8.6 Responsible Use

The detector is a decision-support instrument, not an automated judge. It should not be used as the sole basis for criminal, employment, financial or platform sanctions. Provenance standards, source verification and investigative context should complement statistical detection. Researchers must disclose error rates and avoid claims that any detector can certify authenticity permanently as generation techniques evolve.

9. CONCLUSION AND FUTURE SCOPE:

9.1 Conclusion

Real-time deepfake detection is a systems problem as much as a classification problem. Images, speech and videos contain different forensic evidence, and a useful detector must integrate these sources without assuming that every modality is present or trustworthy. The proposed framework therefore combines spatial and frequency analysis, acoustic self-supervised representations, causal temporal modelling and audio–visual consistency through reliability-aware gated fusion.

The research design places generalisation at the centre. Identity-disjoint splitting, source deduplication, unseen-generator tests and external datasets are necessary because excellent within-dataset scores can be produced by shortcuts that have little forensic value. Compression, noise, resizing, dubbing, missing streams and partial manipulation must be treated as normal operating conditions rather than optional stress tests. Calibration and abstention are equally important: a well-calibrated inconclusive decision can be safer than a confident but unreliable label.

Real-time credibility also requires complete latency accounting. A model is not real time merely because its neural forward pass is fast. Decode, face detection, tracking, resampling, batching, fusion and reporting must together sustain the input rate on declared hardware. Adaptive sampling, causal caches, quantisation, distillation and early exit can reduce cost, but each optimisation must be tested for cross-domain accuracy and subgroup effects.

The principal contribution of this paper is thus a complete, testable blueprint rather than invented performance. It joins multimodal evidence, streaming efficiency, leakage-resistant evaluation and responsible decision support within one university-format research framework. Once implemented, its value should be judged through cross-dataset robustness, calibrated error, p95 latency and transparent failure analysis—not a single accuracy figure.

9.2 Future Research

- Open-world detection that recognises manipulation from generator families absent during training.
- Continual learning that incorporates new attacks without catastrophic forgetting or uncontrolled label noise.
- Multilingual and cross-cultural speech/face evaluation, especially for under-represented Indian languages.
- Detection of partial and localised edits in long-form audio and video.
- Joint use of content forensics, cryptographic provenance and capture-device signatures.
- Robust fusion when audio and video are separately genuine but deceptively recombined.

- Privacy-preserving federated learning across platforms and forensic laboratories.
- Energy-aware detection for mobile, edge and high-volume moderation systems.
- Formal evaluation of explanation stability and the effect of explanations on human decisions.

REFERENCES:

1. Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020). Not made for each other: Audio-visual dissonance-based deepfake detection and localization. Proceedings of the 28th ACM International Conference on Multimedia, 439–447. https://doi.org/10.1145/3394171.3413700
2. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge dataset. arXiv. https://arxiv.org/abs/2006.07397
3. Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5039–5049.
4. Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. NeurIPS Datasets and Benchmarks, 1.
5. Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. IEEE International Workshop on Information Forensics and Security, 1–7. https://doi.org/10.1109/WIFS.2018.8630787
6. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for DeepFake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3207–3216.
7. Li, Y., Zhang, M., Ren, M., Qiao, X., Ma, M., Wei, D., & Yang, H. (2024). Cross-domain audio deepfake detection: Dataset and analysis. Proceedings of EMNLP 2024, 4977–4983.
8. Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., & Lee, K. A. (2023). ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2507–2522. https://doi.org/10.1109/TASLP.2023.3285283
9. Phukan, O. C., Kashyap, G. S., Buduru, A. B., & Sharma, R. (2024). Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake. Findings of NAACL 2024, 2496–2506.
10. Raza, M. A., Malik, K. M., Sharif, M., & Bashir, S. (2023). Multimodaltrace: Deepfake detection using audiovisual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 993–1000.
11. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1–11.
12. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–
13. 148. https://doi.org/10.1016/j.inffus.2020.06.014
14. 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.2020.3002101
15. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., & Wu, B. (2023). DeepfakeBench: A comprehensive benchmark of deepfake detection. Advances in Neural Information Processing Systems, 36, 4534–

4565.
16. Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-stream neural networks for tampered face detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1831–1839.