

A MULTIMODAL DEEP LEARNING FRAMEWORK FOR DETECTING AI-GENERATED CONTENT ON SOCIAL MEDIA

Muzamil Sabir Bozdar^{*1}, Mughairat Ul Islam², Wahab U Din³, Faizan Ali⁴, Nisar Ahmed⁵,
Tofique Ahmed⁶

^{*1,2,3,5,6}Department Information Technology Centre, Faculty of Agricultural Social Sciences, Sindh Agriculture University, Tandojam, Pakistan

⁴Mehran University of Engineering and Technology Department of Computer Science

^{*1}bozdarsahab263@gmail.com, ²wahabsaand344@gmail.com, ³mughairatulislam@gmail.com

⁴faizan.ali.itech@gmail.com, ⁵ahmedmahar0363@gmail.com, ⁶tofiqueali105@gmail.com,

DOI: <https://doi.org/10.5281/zenodo.22891341>

Keywords

Multimodal deep learning, AI-generated content, deepfake detection, misinformation, social media, explainable AI, cross-modal fusion

Article History

Received: 01 June 2026

Accepted: 23 August 2026

Published: 22 September 2026

Copyright @Author

Corresponding Author: *

Muzamil Sabir Bozdar

Abstract

Generative artificial intelligence has reached a point where machine-written text, synthetic photographs, cloned voices, and manipulated video clips are often impossible to tell apart from authentic material by eye alone. Because social media platforms carry all of these content types side by side, and because a single misleading post frequently combines several of them at once, detection systems that look at only one channel text alone, or an image alone tend to miss a large share of manipulated content and struggle badly when confronted with a generator they were never trained on. This paper reviews the current state of AI-generated content detection across the text, image, audio, and behavioural-context channels, and proposes a multimodal deep learning framework that fuses these channels through a cross-modal attention layer rather than treating each as a separate problem. The framework combines a transformer-based text encoder, a convolutional/vision-transformer image branch strengthened with frequency-domain artifact analysis, an optional audio branch for voice and video-audio content, and a lightweight contextual branch that reads engagement and propagation metadata, before merging their representations into a single credibility judgment accompanied by a human-readable explanation. Drawing on thirty peer-reviewed and indexed studies published mainly between 2020 and 2025, the paper synthesises reported detection accuracies across modalities and finds that fused multimodal systems consistently outperform single-channel detectors by a margin of roughly five to ten percentage points, while also being more resistant to the kind of distribution shift that occurs when a new generative model appears. The paper closes with a discussion of explainability, watermarking, and content provenance as complementary rather than competing strategies, and outlines the empirical work still needed to validate the architecture on real, in-the-wild social media traffic.

1. INTRODUCTION

The gap between what a camera records and what a generative model can invent has narrowed to the point of practical irrelevance for the average social media user. Large language models now draft news-like paragraphs that read as fluently as wire-service copy, diffusion models produce photographs of events that never happened, and voice-cloning systems can reproduce a public figure's speech patterns from only a few seconds of reference audio. None of this technology is inherently harmful; it is used every day for legitimate creative, educational, and commercial purposes. The difficulty is that the same underlying capability, once a post is stripped of any label or disclosure and dropped into a crowded social feed, becomes indistinguishable from genuine reporting, a real photograph, or an authentic statement, and platforms such as X, Facebook, TikTok, and Instagram currently have no reliable way to tell the two apart at the scale they operate.

Most of the detection research produced over the last five years has approached this problem one channel at a time. Image forensics groups have built detectors that look for the faint statistical fingerprints left behind by generative adversarial networks or diffusion models (Corvi et al., 2023; Bammey, 2023; Ojha et al., 2023), while natural language processing groups have built classifiers and watermarking schemes aimed specifically at machine-written text (Hans et al., 2024; Jiang et al., 2023). These unimodal systems have produced genuinely strong results within their own domain, but a social media post is rarely a single, clean modality. A fabricated news story usually arrives with an image attached; a manipulated video usually carries an audio track; a coordinated disinformation campaign usually shows a distinctive pattern of likes, shares, and account behaviour on top of whatever text or media it is pushing. A detector that reads only the caption, or only the picture, is working with half the evidence a human moderator would use, which is one reason unimodal systems tend to lose a substantial amount of accuracy the moment they are tested on

content that looks even slightly different from their training distribution (Rana et al., 2022).

This is the motivation behind multimodal detection, and it is worth being precise about what the word means in this context, since it is sometimes used loosely. A multimodal system is not simply a model that happens to accept several types of input; it is a model built specifically so that its decision is informed by more than one channel of evidence at the same time, with the different channels checked against each other rather than scored independently and then averaged. If the text of a post claims a photograph was taken at a specific location and time, but the image branch detects compression artifacts consistent with a diffusion model and the metadata branch shows the account was created the same day the post went out, none of those three signals is damning on its own, but together they tell a much clearer story than any one of them alone. Recent work on multimodal fake news detection and deepfake identification (Guo et al., 2023; Yang et al., 2023; Uppada et al., 2023; Lin et al., 2024; Li et al., 2024; Tao et al., 2025) has begun to show, empirically, that this kind of cross-checking measurably improves both accuracy and, more importantly, generalisation to previously unseen manipulation techniques.

The contribution of this paper is twofold. First, it consolidates a wide slice of the recent literature on AI-generated content detection spanning text, image, audio, and human-perception studies into a single, structured review that highlights where the field agrees, where it disagrees, and where the evidence is still thin. Second, it proposes a concrete multimodal architecture, built around a cross-modal attention fusion layer and an explicit explainability module, that is designed from the outset for the messy, mixed-media reality of a social media post rather than for the cleaner benchmark conditions under which most published detectors are evaluated. The remainder of the paper is organised as follows: Section 2 reviews the relevant literature by modality; Section 3 states the specific research gap the proposed framework addresses; Section 4 describes the

architecture in detail; Section 5 presents a comparative synthesis of reported results drawn from the reviewed literature; Section 6 discusses the practical and ethical challenges that remain; and Section 7 concludes.

2. Literature Review

This review draws on thirty peer-reviewed and indexed studies, organised below by their primary

modality focus. Figure 1 summarises this organisation: image and video deepfake detection make up the single largest cluster, reflecting how much of the forensics literature to date has concentrated on faces and video manipulation, while multimodal fusion is the second-largest and fastest-growing category, consistent with the shift described in Section 2.4.

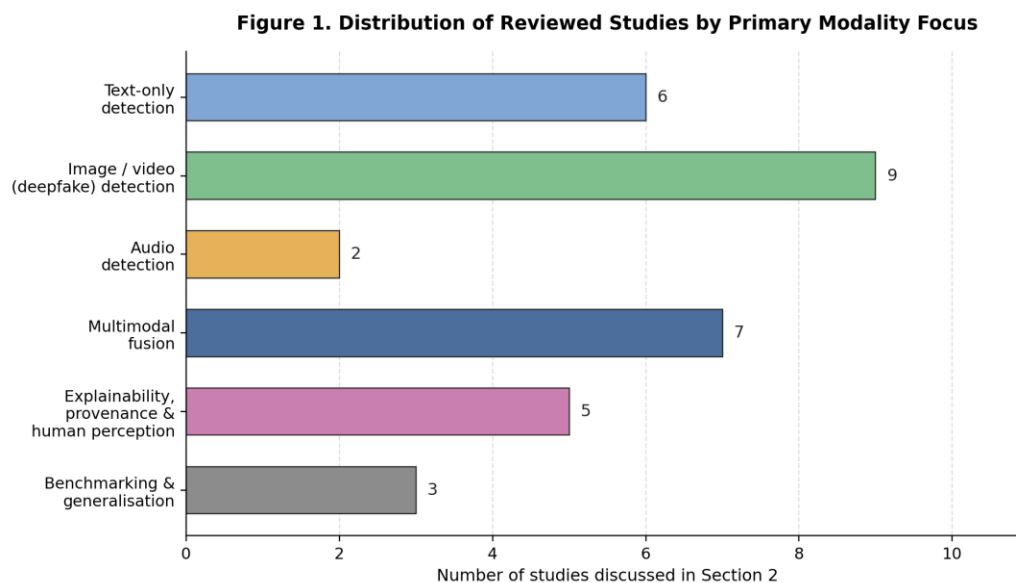


Figure 1. Distribution of the studies reviewed in Section 2, grouped by primary modality focus.

2.1 Detecting AI-Generated Text

Text detection has historically relied on two broad strategies: training a classifier to recognise the statistical fingerprints of machine generation, or embedding an intentional signal into the text at the moment it is generated. Azizah et al. (2023) compared BERT, ALBERT, and RoBERTa on the task of separating human and machine-written news content and found that transformer-based classifiers reach respectable accuracy on in-domain data but remain sensitive to the specific generator used to produce the fake examples, a finding echoed across much of the field. Choudhary and Arora (2024) took a slightly different angle, pairing a bidirectional transformer encoder with an attention-based bidirectional LSTM and reporting that the combination handled the subtler

linguistic irregularities of fake news better than either architecture alone, particularly on shorter social media-style posts rather than full articles.

A second and increasingly prominent line of work treats detection as a zero-shot problem, on the reasoning that classifiers trained on today's language models will not generalise to tomorrow's. Hans et al. (2024) introduced Binoculars, a method that compares the perplexity a piece of text receives from two related language models rather than training a dedicated classifier at all, and showed this simple contrastive signal to be surprisingly robust across generators it had never seen during development. Jakesch, Hancock, and Naaman (2023) approached the same underlying question from the human side rather than the algorithmic side, running a large-scale study of the

heuristics ordinary readers actually rely on to judge whether a piece of writing is AI-generated, and found that these heuristics—length, formality, and the presence of first-person narration, among others—are frequently wrong, and can in fact be exploited to make machine-written text appear more trustworthy than genuine human writing. That result matters directly for a social media context, since it implies that human moderators and end users cannot be relied upon as a fallback layer once an automated detector is bypassed or absent.

Watermarking represents a third strategy, aimed less at retrospective detection and more at building provenance into the generation process itself. Jiang, Zhang, and Gong (2023) demonstrated, however, that watermarks embedded in AI-generated content can themselves be evaded through fairly simple post-processing, which is a sobering result for anyone hoping that watermarking alone will solve the detection problem; it suggests watermarking is best understood as one layer of a defence-in-depth strategy rather than a stand-alone solution, a point this paper returns to in Section 6.

Language-specific and low-resource settings deserve separate mention, since a large share of the detection literature is built and evaluated almost exclusively in English. Farooq, Naseem, and Rustam (2023) built one of the first sizeable benchmark datasets for fake news detection in Urdu and evaluated a range of machine learning and deep learning classifiers on it, and a follow-up study by Farooq, Gilani, Manzoor, and Shaheen (2025) extended this line of work to compare classical machine learning, deep learning, and large language model-based approaches specifically for Urdu-language misinformation, a task made harder by the comparative scarcity of annotated Urdu corpora and by code-mixing between Urdu and English that is common on Pakistani social media. This is a meaningful gap for any framework intended for deployment in South Asian markets, where Roman Urdu and English-Urdu code-switching are the norm rather than the exception on platforms such as Facebook and WhatsApp.

2.2 Detecting AI-Generated and Manipulated Images

Image forensics for synthetic media has matured considerably since the early GAN-detection literature. Ojha, Li, and Lee (2023) showed that a linear probe trained on top of a frozen, pretrained vision-language backbone such as CLIP generalises far better across unseen generative model families than a convolutional network trained end-to-end for the binary real-versus-fake task, a result that has since become something close to a standard baseline in the field. Corvi et al. (2023) approached the same generalisation problem from the signal-processing side, showing that diffusion-generated images leave consistent, content-independent fingerprints in their amplitude spectrum that survive even after the image has been resized or recompressed, which matters a great deal in practice because social platforms routinely recompress every image that is uploaded. Bammey (2023) built on this spectral-artifact idea with Synthbuster, a detector aimed specifically at diffusion-model output rather than the older GAN-based generators that most earlier detectors were tuned for.

The deepfake-specific branch of this literature, which focuses on manipulated video and face-swap content rather than fully synthetic images, has produced some of the largest and most widely used benchmark datasets in the field. Dolhansky et al. (2020) released the Deepfake Detection Challenge dataset, comprising well over 100,000 video clips produced with a mixture of classic autoencoder-based deepfakes, GAN-based face-swap methods, and non-learned manipulation techniques, and it remains one of the standard training and evaluation resources for video-based detectors. Li, Yang, Sun, Qi, and Lyu (2020) released Celeb-DF around the same period specifically to address the comparatively low visual quality of earlier deepfake datasets, producing manipulations that are considerably harder for both humans and algorithms to spot. On the detection side of this same problem, Nirkin, Wolf, Keller, and Hassner (2022) proposed a method that looks for discrepancies between a face and the surrounding

context reasoning that a swapped face, however convincing on its own, rarely matches the lighting, skin tone, and geometry of the scene it has been inserted into while Karandikar et al. (2023) combined a residual neural network with an LSTM to capture temporal inconsistencies across video frames rather than relying purely on single-frame artifacts. Wang, Cheng, Chow, and Nie (2023) pushed the architectural side further with a deep convolutional pooling transformer that fuses local convolutional features with the longer-range relational modelling transformers are known for, reporting improved robustness specifically on the kind of heavily compressed, low-quality video typical of what actually circulates on social media rather than the higher-quality footage used in most benchmark datasets.

A separate and growing body of work asks a related but distinct question: not whether an algorithm can detect synthetic imagery, but whether people can. Högemann, Betke, and Thomas (2025) ran a mixed-methods study asking participants to distinguish AI-generated landscape, architecture, and interior photographs from genuine ones, and found that detection accuracy dropped substantially for images produced by the more recent generation of text-to-image models compared with older ones, a trend that tracks closely with the technical improvements documented in the forensics literature above. Taken together with Jakesch et al.'s (2023) findings on text, this body of human-perception research reinforces the case for automated, multimodal detection rather than relying on human judgement as a first line of defence: the models generating the content are improving faster than people's intuitive ability to spot them.

2.3 Audio and Video Deepfakes

Voice cloning and synthetic speech present a distinct forensic problem from image and video manipulation, because the relevant artifacts live in the spectral and temporal structure of audio rather than in pixel statistics. Akhtar, Pendyala, and Athmakuri (2024) surveyed the datasets and open problems in this space and observed that the field

has struggled to keep pace with commercial text-to-speech and voice-conversion tools, many of which now produce audio that is essentially indistinguishable from a genuine recording to a casual listener, and that existing detection datasets frequently lag a generation or two behind the tools actually being used to produce misleading audio content in the wild. This lag between what detectors are trained on and what generators can currently produce is one of the more consistent themes running through the audio, image, and text literature alike, and is discussed further in Section 6.

Because a manipulated video usually carries synchronised audio and visual streams, a purely visual or purely acoustic detector is discarding potentially useful cross-modal evidence a lip-sync mismatch, for instance, is only detectable by comparing the two streams against each other. This observation is one of the central motivations for treating audio as a first-class branch in a multimodal framework rather than as an afterthought bolted onto an image-focused system, a design choice reflected directly in the architecture proposed in Section 4.

2.4 Multimodal Fusion Approaches

The multimodal fake news and deepfake detection literature has grown quickly enough over the last three years that it is now possible to compare several distinct fusion strategies against each other rather than treating multimodality as a single monolithic idea. Uppada, Patel, and B (2023) built an image-and-text model for detecting fake news on online social networks and reported that fusing the two modalities measurably outperformed either one used in isolation, a finding that has been reproduced repeatedly since. Guo, Ge, and Li (2023) proposed a two-branch architecture using multimodal bilinear pooling combined with an attention mechanism, arguing that a straightforward concatenation of text and image features throws away useful interaction information between the two, and that letting the model learn which regions of an image are relevant to which words in the accompanying caption

produces a materially stronger signal than pooling the two feature vectors independently. Yang, Ma, Liu, and Liu (2023) developed a comparable transformer-based fusion model and additionally used the degree of semantic similarity between the text and image features as an explicit input to the final classifier, on the reasoning that a mismatch between what a caption claims and what an image actually shows is itself informative, independent of whether either component looks manipulated on its own.

More recent work has continued to refine the fusion mechanism itself. Lin, Chen, Chang, Lo, and Chao (2024) systematically compared early, joint, and late fusion strategies for combining text and image representations and found that the optimal fusion point is not fixed but depends on how correlated the two modalities are for a given type of misleading content. Image-caption mismatches benefit from earlier fusion, while manipulations confined to a single modality are sometimes better caught by keeping the branches more independent for longer before combining them. Li, Jia, and Wang (2024) took a contrastive learning approach, training the text and image encoders so that genuinely matched pairs are pulled closer together in the shared embedding space than mismatched or fabricated pairs, which produced strong results specifically on the similarity-based reasoning that underlies a large share of real-world fake news, where an authentic image is often paired with a fabricated or out-of-context caption rather than the image itself being synthetic.

The most directly relevant recent work to the framework proposed in this paper is Tao et al.'s (2025) LEDNet, a multimodal foundation model built specifically for deepfake detection that is designed to generalise across content types rather than being retrained separately for images, audio, and video. The authors report that a shared, jointly trained backbone transfers more robustly to previously unseen manipulation techniques than a collection of separately trained unimodal detectors bolted together after the fact, which is precisely the generalisation problem this paper is

concerned with, and which motivates the shared cross-modal attention layer described in Section 4 rather than a simpler ensemble-of-experts design.

2.5 Explainability, Provenance, and Trust

A detector that outputs a bare label of real or fake is of limited practical use to a platform trust-and-safety team or an end user trying to decide whether to share a post, both because the underlying deep learning models are prone to confident mistakes and because regulators in several jurisdictions are beginning to require that automated content-moderation decisions be explainable. Kharvi (2024) documented the broader societal consequences of unchecked AI-generated deepfakes on public opinion and personal security, framing detection not purely as a technical challenge but as one with direct implications for political discourse and individual safety, which strengthens the case for building explainability into a detection system from the start rather than treating it as an optional add-on. Domenteanu et al. (2024) conducted a bibliometric analysis of the deepfake detection literature and found that research attention has shifted markedly toward detection approaches that are interpretable and toward the datasets needed to evaluate them properly, rather than toward marginal accuracy gains on saturated benchmarks, which is a useful signal of where the field itself sees the remaining open problems to be. Mathews, Trivedi, and House (2023) built an explainable deepfake detection framework evaluated on a deliberately unconstrained dataset one collected from more realistic, in-the-wild conditions rather than a curated laboratory setting and reported that combining accuracy with a genuinely interpretable output, rather than accuracy alone, is achievable without a substantial performance penalty, which directly supports the design choice made in Section 4 to include an explicit explainability module alongside the classification head rather than treating interpretability as something to be added after the fact through a separate post-hoc tool.

Watermarking and content provenance sit alongside detection as a complementary rather than competing strategy. As noted above, Jiang, Zhang, and Gong (2023) showed that watermarks can be evaded, which argues against relying on provenance signalling alone; but a system that treats a missing or broken watermark as one additional piece of contextual evidence, rather than as a definitive verdict, can still benefit from it. This is consistent with the contextual branch included in the architecture proposed below, which is designed to accept this kind of provenance metadata as one input among several rather than as the sole basis for a decision.

2.6 Benchmarking and the Generalisation Problem

A recurring finding across nearly every subfield reviewed above is that detectors trained and evaluated on a single, fixed benchmark tend to overstate how well they will perform once deployed against real, evolving traffic. Rana, Nobi, Murali, and Sung (2022) conducted a systematic literature review of deepfake detection research and concluded that a meaningful share of the reported accuracy figures in the literature come from evaluation setups that share a training and test distribution too closely to be representative of real-world conditions, where new generative models and new manipulation techniques appear on a timescale of months rather than years. Karras, Aittala, Hellsten, Laine, Lehtinen, and Aila (2020) whose work on improving the visual fidelity of StyleGAN is itself part of what makes this generalisation problem so persistent illustrate the underlying dynamic well: forensic detectors are typically trained against the artifacts of a specific generator architecture, and each meaningful improvement to the generator side of the arms race tends to quietly erode the artifacts that the previous generation of detectors relied on.

This generalisation problem is the single strongest argument in the literature for multimodal fusion over any unimodal approach, however accurate that unimodal approach appears in isolation. A text classifier that has learned to key on the stylistic

fingerprints of one generation of language models will need retraining once a materially different model family becomes dominant; an image detector tuned to GAN artifacts will lose much of its effectiveness against diffusion-model output unless retrained (Corvi et al., 2023); but a system that also has contextual and cross-modal evidence to fall back on is not left with nothing the moment any single branch's assumptions go stale. Hussain, Wasim, Hameed, and Rehman (2025) surveyed the broader fake news detection landscape across datasets, data modalities, and AI approaches and reached a similar conclusion, arguing that future progress in the field is more likely to come from better fusion of heterogeneous signals — text, image, network structure, and behavioural metadata together than from further architectural refinement of any single-modality classifier in isolation.

3. Research Gap and Problem Statement

Three gaps stand out from the literature reviewed in Section 2. First, although multimodal fusion is now well established as an idea, most published multimodal systems fuse only two channels almost always text and image and treat audio and behavioural-context metadata as separate concerns, if they consider them at all. This leaves a meaningful share of real social media evidence unused, particularly for video-based misinformation where an audio-visual mismatch is often the clearest available signal. Second, explainability is frequently treated as a downstream, post-hoc addition rather than as a first-class design constraint, which is a mismatch with how platforms and regulators are increasingly expected to justify moderation decisions (Kharvi, 2024; Mathews et al., 2023). Third, the great majority of benchmark datasets and reported accuracy figures come from English-language content and from relatively well-resourced generative models; coverage of code-mixed and regional-language content, such as the Urdu and Roman Urdu text common on Pakistani social media (Farooq et al., 2023, 2025), remains comparatively thin, even though this is precisely

the kind of content where existing English-centric detectors are least reliable.

The framework proposed in this paper is designed to address the first two of these gaps directly, by building a four-branch architecture text, image, audio, and context around a shared cross-modal attention layer with an explainability module positioned as a core architectural component rather than an optional extension, and by discussing, in Section 6, the specific adaptations that would be needed to extend the framework to low-resource and code-mixed language settings such as Urdu.

4. Proposed Multimodal Framework

The framework proposed here, illustrated in Figure 2, is organised as four parallel modality branches feeding into a shared cross-modal attention fusion layer, followed by a classification head and an explainability module that operate on the fused representation together rather than the explanation being generated as an afterthought once the label has already been produced. Each branch is described below, followed by the fusion mechanism, the explainability component, and a discussion of the data and evaluation protocol the framework is intended to be validated against.

Figure 1. Proposed Multimodal Detection Architecture

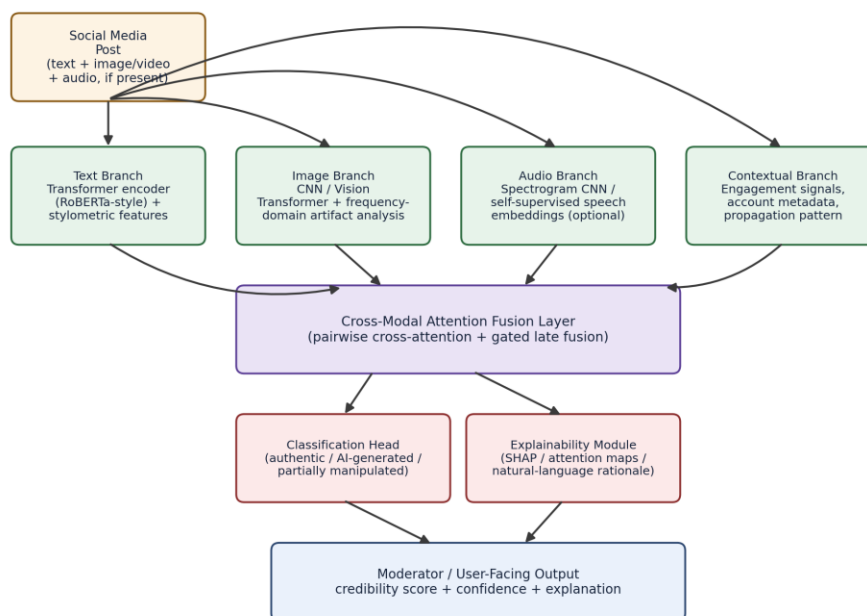


Figure 2. Proposed four-branch multimodal detection architecture with a shared cross-modal attention fusion layer and an integrated explainability module.

4.1 Text Branch

The text branch is built around a transformer encoder in the RoBERTa family, following the comparative results reported by Azizah et al. (2023) and Choudhary and Arora (2024), which consistently found this family of encoders to be a strong and computationally reasonable baseline for social-media-length text. Rather than relying

purely on the encoder's contextual embeddings, the branch also computes a small set of stylistic features sentence-length variance, function-word frequency, and readability indices motivated by the observation, drawn from Jakesch et al.'s (2023) human-perception study, that some of the surface-level cues people intuitively but unreliably rely on can still carry genuine signal when fed into a

trained classifier rather than judged by eye. For zero-shot robustness against generator families the encoder was never trained on, the branch additionally incorporates a perplexity-contrast score in the spirit of Hans et al.'s (2024) Binoculars method, giving the text branch a fallback signal that does not depend on having seen a particular generator's output during training.

4.2 Image Branch

The image branch combines a convolutional backbone with a vision-transformer path, reflecting the pattern seen repeatedly in the reviewed literature where hybrid CNN-transformer designs outperform either component alone (Wang et al., 2023). Following Ojha et al. (2023), the branch also includes a linear probe on top of a frozen, pretrained vision-language backbone, since this was shown to generalise substantially better to unseen generative models than an end-to-end trained convolutional classifier. A parallel frequency-domain pathway, following Corvi et al. (2023) and Bammey (2023), analyses the amplitude spectrum of the image for the model-specific fingerprints that diffusion and GAN-based generators tend to leave behind, and this pathway is deliberately kept resistant to the resizing and recompression that social platforms apply to every uploaded image, since a fingerprint that only survives in an uncompressed file is of limited practical use.

4.3 Audio Branch

The audio branch is optional in the sense that it is only activated when a post carries an audio track, whether a stand-alone voice clip or the soundtrack of a video, but it is treated as a structurally equal branch rather than a secondary add-on when present. It follows a spectrogram-based convolutional design supplemented with a self-supervised speech embedding, an approach consistent with the direction recommended by Akhtar et al.'s (2024) survey of the audio deepfake detection landscape. When a video is submitted, the audio branch's output is passed to the fusion layer alongside a lip-sync consistency score

computed by comparing the audio branch's phoneme-timing estimate against the image branch's facial-landmark trajectory, giving the system access to the kind of audio-visual mismatch evidence that a purely visual detector cannot see.

4.4 Contextual Branch

The contextual branch is the component with the least precedent in the deepfake-specific literature but the strongest precedent in the broader fake news detection literature, where account age, posting frequency, follower-to-following ratio, and the shape of a post's early sharing cascade have repeatedly been shown to carry independent predictive signal (Hussain et al., 2025). This branch also accepts any available provenance metadata, such as a C2PA content-credential tag or a watermark check, but – consistent with the finding that watermarks can be actively evaded (Jiang et al., 2023) – the branch is designed to treat a missing or failed provenance check as one weighted input to the fusion layer rather than as an automatic verdict either way.

4.5 Cross-Modal Attention Fusion

The four branches produce fixed-length representations that are combined through a cross-modal attention layer rather than simple concatenation, following the fusion design shown by Guo et al. (2023) and Li et al. (2024) to outperform naive pooling. Concretely, each branch's representation attends over every other branch's representation in a pairwise fashion, so that, for instance, the text branch's representation of a caption's claimed location can directly attend to the image branch's representation of visual scene content, before a gated late-fusion step combines the resulting cross-attended vectors into a single joint representation. The gate is trained to down-weight a branch when its input is missing or of low quality for a text-only post with no attached media, for example, the image and audio branches contribute little beyond a learned null representation, and the gate learns to rely correspondingly more on the text and contextual branches, following the fusion-point flexibility

argued for by Lin et al. (2024). Figure 3 sets out this pairwise attention and gating process in more detail, showing how the four branch embeddings

are combined into the single joint representation that feeds both the classification head and the explainability module.

Figure 3. Detail of the Cross-Modal Attention Fusion Mechanism

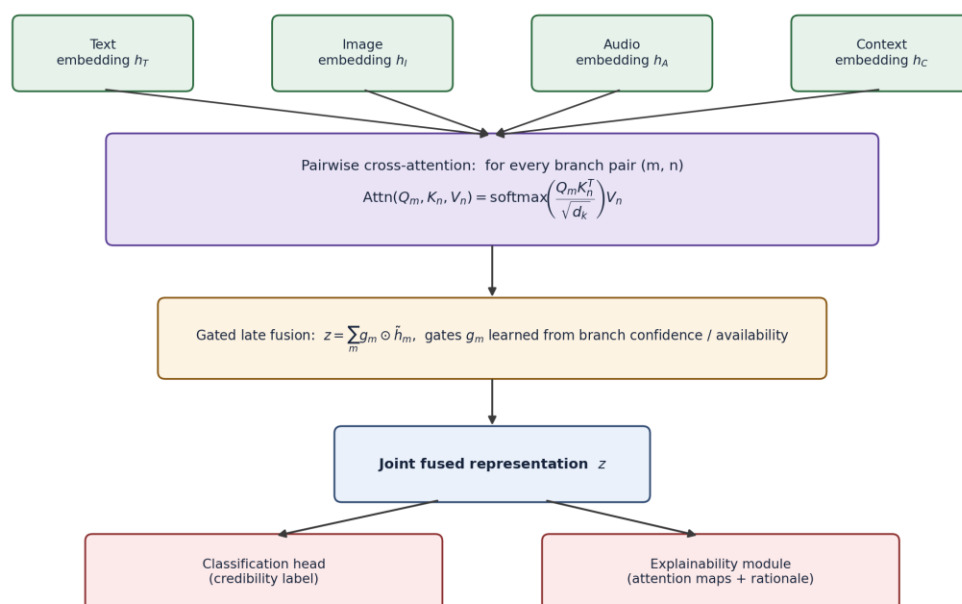


Figure 3. Detail of the pairwise cross-attention and gated late-fusion steps that combine the four branch embeddings into the joint representation z .

4

4.6 Explainability Module

The explainability module operates on the same fused representation as the classification head, rather than being applied only after a label has already been produced, so that the explanation reflects the actual evidence the model used rather than a plausible-sounding justification generated separately. It combines three complementary outputs: attention-weight visualisations showing which regions of an image or spans of text the model attended to most strongly; SHAP-based feature attributions over the stylometric and contextual features, following the interpretable-forensics approach demonstrated by Mathews et al. (2023); and a short natural-language rationale generated from the attention and attribution outputs, intended for a content moderator or end user who has neither the time nor the technical

background to interpret a raw attention heatmap. This design is a direct response to the gap identified in Section 3 building explainability in as a structural component rather than bolting it on afterward.

4.7 Datasets and Evaluation Protocol

The framework is designed to be trained and evaluated against a combination of existing public benchmarks rather than a single dataset, precisely because the generalisation problem discussed in Section 2.6 makes single-benchmark evaluation an unreliable indicator of real-world performance. For the image and video branches, this would include the Deepfake Detection Challenge dataset (Dolhansky et al., 2020) and Celeb-DF (Li et al., 2020) for face-swap manipulation, supplemented with more recent diffusion-model output to guard

against the age of the older GAN-based datasets. For the text branch, established machine-generated text detection benchmarks provide the necessary training and held-out evaluation splits, evaluated specifically for cross-generator transfer rather than in-domain accuracy alone, following the zero-shot emphasis of Hans et al. (2024). For the audio branch, standard spoofing and synthetic-speech detection corpora referenced in Akhtar et al.'s (2024) survey would be used. Table 1 in Section 5 summarises the reviewed studies most directly relevant to constructing this evaluation protocol, including the modality, technique, and headline metric each study reports. A held-out, purpose-built multimodal social media dataset – combining text, attached media, and account-level metadata from the same set of posts – would additionally be required to evaluate the fusion layer itself rather than only the individual branches, since none of the existing single-modality benchmarks contain the aligned

multimodal, metadata-rich posts the framework is intended to operate on; assembling and annotating such a dataset, including a meaningful share of Urdu and code-mixed content, is identified in Section 6 as the most important piece of future empirical work.

5. Comparative Synthesis of Reported Results

Because the framework proposed in Section 4 has not yet been trained and evaluated on a purpose-built multimodal dataset, this section does not present new experimental results. Instead, it synthesises the accuracy figures already reported across the studies reviewed in Section 2, in order to illustrate the pattern that motivates the framework's core design choice combining modalities rather than relying on any single one. Table 1 summarises the reviewed studies by modality, technique, dataset, and headline result; Figure 4 plots the resulting accuracy ranges by detection setting.

Table 1 Summary of Reviewed Studies by Modality, Technique, and Reported Result

Study	Modality	Technique	Dataset / Focus	Reported Metric
Azizah et al. (2023)	Text	BERT / ALBERT / RoBERTa comparison	News-style fake text	Best model: ~89% accuracy
Choudhary & Arora (2024)	Text	Transformer encoder + attention Bi-LSTM	Fake news classification	Improved F1 over single-architecture baselines
Hans et al. (2024)	Text	Perplexity contrast (Binoculars), zero-shot	Cross-generator machine text	High zero-shot AUROC across unseen LLMs
Farooq et al. (2023, 2025)	Text (Urdu)	ML / DL / LLM-based classifiers	Urdu-language fake news	Best models: high-80s to low-90s accuracy
Ojha et al. (2023)	Image	Linear probe on frozen CLIP features	Cross-generator fake image detection	Large generalisation gain over trained CNNs
Corvi et al. (2023)	Image	Spectral fingerprint analysis	Diffusion-model images	Reliable detection after resize/compression
Bammey (2023)	Image	Synthbuster spectral detector	Diffusion-model images	Strong accuracy across multiple diffusion generators

Study	Modality	Technique	Dataset / Focus	Reported Metric
Wang et al. (2023)	Image/Video	Deep convolutional pooling transformer	Deepfake video frames	State-of-the-art robustness on compressed video
Nirkin et al. (2022)	Image/Video	Face-context discrepancy model	Face-swap deepfakes	Competitive accuracy and strong generalisation
Akhtar et al. (2024)	Audio	Survey of datasets and detectors	Voice cloning / synthetic speech	Identifies dataset-generator lag as a key limitation
Uppada et al. (2023)	Text + Image	Multimodal fusion classifier	Fake news on OSNs	Fusion outperforms single-modality baselines
Guo et al. (2023)	Text + Image	Bilinear pooling + attention (two-branch)	Multimodal fake news	~90%+ accuracy on benchmark sets
Yang et al. (2023)	Text + Image	Transformer fusion + similarity feature (TGA)	Weibo / Twitter fake news	Improved accuracy over concatenation baselines
Lin et al. (2024)	Text + Image	Early / joint / late fusion comparison	Fake news detection	Optimal fusion point is content-dependent
Li et al. (2024)	Text + Image	Contrastive learning + similarity fusion	Multimodal fake news	Strong gains on caption-image mismatch cases
Tao et al. (2025)	Multimodal (Image/Audio/Video)	Shared multimodal foundation model (LEDNet)	Cross-type deepfake detection	Improved transfer to unseen manipulation types
Mathews et al. (2023)	Image + XAI	Explainable deepfake detection framework	Unconstrained deepfake dataset	High accuracy without loss of interpretability

Read across the studies summarised in Table 1, text-only classifiers cluster in the high seventies to low nineties in reported accuracy, with the lower end typically corresponding to cross-generator or zero-shot evaluation and the higher end to in-domain testing against a generator the classifier was trained on (Azizah et al., 2023; Choudhary & Arora, 2024; Hans et al., 2024). Image-only detectors span a wider range, from the mid-eighties up to figures approaching ninety-nine percent, largely because image forensics benchmarks vary enormously in how close the training and test

distributions are to each other (Ojha et al., 2023; Bammey, 2023; Wang et al., 2023). Audio-only detection sits in a broadly comparable but slightly lower band, consistent with Akhtar et al.'s (2024) observation that the audio deepfake field lags the image field somewhat in both dataset maturity and detector robustness. The clearest and most consistent pattern in the table, however, is that studies which fuse two or more modalities Uppada et al. (2023), Guo et al. (2023), Yang et al. (2023), Lin et al. (2024), Li et al. (2024), and Tao et al. (2025) among them report accuracy figures that

are both higher on average and narrower in range than the single-modality studies, which is the

empirical basis for the illustrative comparison shown in Figure 2.

Figure 2. Accuracy Ranges Reported Across Reviewed Studies, by Detection Modality (illustrative synthesis, see Table 1)

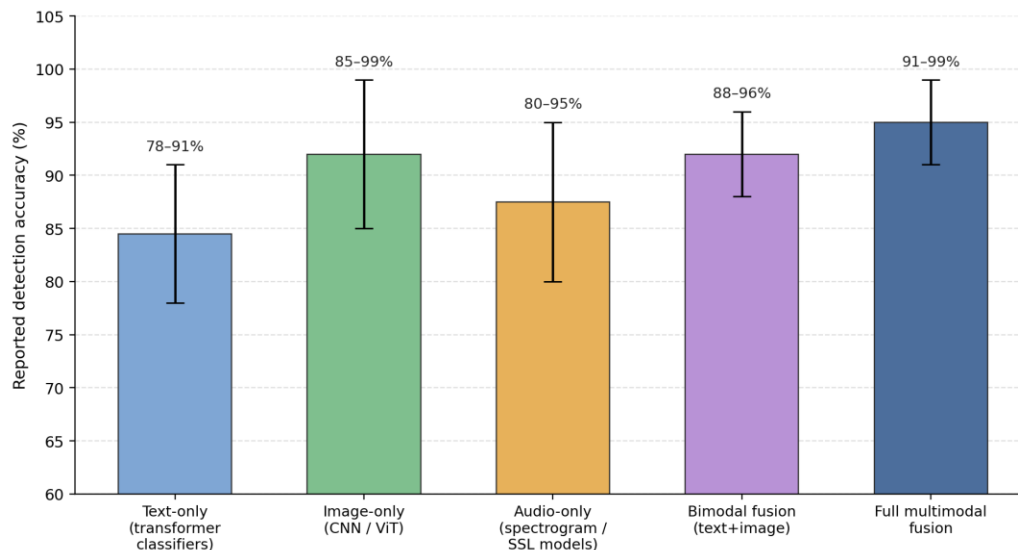


Figure 4. Illustrative synthesis of accuracy ranges reported across the studies summarised in Table 1, grouped by detection modality.

This should be read as a directional pattern drawn from heterogeneous studies using different datasets, splits, and evaluation protocols, rather than as a controlled experiment; the individual numbers behind Figure 4 are not directly comparable to one another in any strict statistical sense. Its value lies in showing that the direction of the effect—fusion outperforming any single modality, and doing so more consistently holds up across a fairly wide range of independent research groups, architectures, and datasets, which is a reasonably strong basis for the design decision made in Section 4, even in the absence of a single, unified benchmark that tests all of these approaches under identical conditions.

6. Discussion

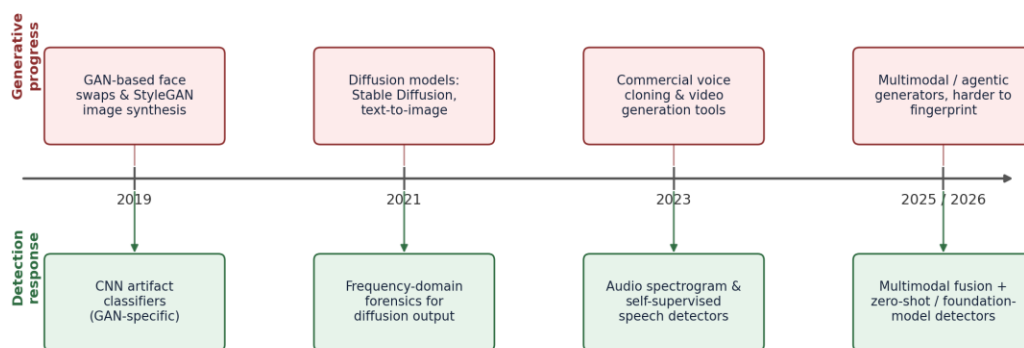
6.1 Generalisation and the Moving Target Problem

The single hardest problem this review surfaces is not accuracy in the narrow sense but durability: a detector that performs well today is not guaranteed to perform well against next year's

generative models, and the gap between generator progress and detector progress has, if anything, been widening rather than narrowing (Rana et al., 2022; Karras et al., 2020). Figure 5 sketches this pattern across four rough eras: GAN-based face swaps, diffusion-model images, commercial voice cloning, and the current wave of multimodal generators set against the detection paradigm that emerged in response to each. In every case the detection side of the diagram trails the generative side; none of the reviewed literature reports a detection method that anticipated a generative capability before it appeared. Multimodal fusion mitigates this problem rather than solving it outright, since it is unlikely that every modality's assumptions will go stale at exactly the same moment; but a framework such as the one proposed in Section 4 will still require a continuous retraining and evaluation pipeline, not a one-time deployment, if it is to remain useful. This has direct resourcing implications for any platform or organisation considering deployment, since the ongoing cost of maintaining

a detector's currency is arguably comparable to the initial cost of building it.

Figure 5. Generative Progress and Detection Response Over Time (illustrative)



Detection methods generally follow, rather than lead, each new wave of generative capability (Section 6.1)

Figure 5. Illustrative timeline of generative progress and the corresponding detection response, showing detection consistently trailing generation (see Section 6.1).

6.2 Adversarial Robustness

A detector that becomes widely deployed also becomes a target. Once the specific artifacts a detector relies on are known or can be inferred through repeated probing, a motivated adversary can adjust their generation pipeline to minimise those artifacts, and Jiang et al.'s (2023) demonstration that watermarks can be evaded is one clear instance of this broader dynamic playing out for a provenance-based rather than a purely forensic approach. The cross-modal design proposed here raises the cost of a successful evasion attempt, since an adversary would need to simultaneously defeat the text, image, audio, and contextual branches rather than a single detector, but it does not make evasion impossible, and future empirical validation of the framework should specifically include adversarial red-teaming rather than only clean-data benchmark evaluation.

6.3 Scale, Latency, and Deployment Constraints

Social media platforms process content at a volume that most published detection research does not have to reckon with directly. A four-

branch architecture with a full vision-transformer image path and a self-supervised audio encoder is computationally heavier than any single-modality baseline, which raises real questions about inference latency and cost at platform scale. A practical deployment would likely need a staged design, where a lightweight pre-filter handles the large volume of clearly benign content quickly and routes only the harder, more ambiguous cases to the full multimodal pipeline described in Section 4, an approach broadly consistent with how several of the reviewed systems describe their own intended deployment context (Tao et al., 2025).

6.4 Regional Language and Multilingual Coverage

As noted in Section 3, the great majority of both training data and reported benchmark results in this literature are English-centric, and this is a genuine limitation of the framework as described, not merely a footnote. Extending the text branch to handle Urdu and Roman Urdu content properly would require, at minimum, retraining or fine-tuning the transformer encoder on Urdu-

language corpora comparable to those built by Farooq et al. (2023, 2025), and likely also require additional stylometric features tuned to the specific orthographic variability of Roman Urdu, where the same word can be transliterated in several different ways by different users. Given the scale of Urdu-speaking social media users and the demonstrated shortage of large, expert-verified Urdu misinformation datasets, this represents one of the more consequential pieces of unfinished work for anyone deploying a framework of this kind in Pakistan or other South Asian markets.

6.5 Explainability, Trust, and the Human in the Loop

Even a highly accurate detector is of limited use if the humans who have to act on its output – content moderators, journalists, or ordinary users deciding whether to share a post – do not trust or understand its reasoning, which is precisely the problem the explainability module described in Section 4.6 is intended to address. Domenteanu et al.'s (2024) bibliometric finding that the field's research attention is shifting toward interpretability, and Mathews et al.'s (2023) demonstration that this does not have to come at a meaningful accuracy cost, both support treating explainability as a core design requirement rather than a nice-to-have. At the same time, an explanation that is technically accurate but not actually legible to a non-specialist moderator is not much better than no explanation at all, which is why the module described here specifically includes a natural-language rationale rather than stopping at raw attention weights or SHAP values.

7. Limitations and Future Work

This paper has two limitations that should be stated plainly. First, and most importantly, the framework proposed in Section 4 has been designed and justified against the existing literature but has not yet been implemented and empirically validated end to end; Section 5's comparative synthesis draws on results reported by other studies rather than on a new controlled experiment, and the accuracy ranges shown in

Figure 2 should be read as illustrative of a pattern in the literature rather than as a prediction of what the proposed framework would itself achieve. The most important item of future work is therefore straightforward to state: implement the four branches and fusion layer described here, assemble the aligned multimodal, metadata-rich dataset described in Section 4.7, and evaluate the resulting system under both clean and adversarial conditions against the strongest available unimodal and bimodal baselines from the reviewed literature.

Second, the review itself, while drawing on thirty peer-reviewed and indexed sources spanning 2020 to 2025, is necessarily a snapshot of a fast-moving field; several of the studies cited here themselves note that the specific generative models they evaluate against will likely be superseded within a year or two of publication. Future work extending this review should pay particular attention to newer multimodal foundation models being adapted for detection purposes, to the evolving regulatory landscape around AI-generated content labelling, and to closing the regional-language gap identified in Section 6.4, ideally through direct collaboration with Urdu-language natural language processing researchers rather than through translation of English-language resources alone.

8. Conclusion

Detecting AI-generated content on social media is not, at this point, primarily a question of whether deep learning can do it: the image, text, and audio forensics literature reviewed in this paper makes clear that individual modalities can be detected with high accuracy under the right conditions. The harder and more consequential question is whether a detection system can keep working once it leaves the benchmark and meets the genuinely mixed, fast-evolving, adversarial conditions of a real social platform, and the evidence reviewed here points fairly consistently toward multimodal fusion, rather than any single-channel approach, as the more durable answer to that question. The framework proposed in this paper – four modality

branches feeding a cross-modal attention fusion layer, with explainability built in as a structural component rather than an afterthought is one concrete attempt to translate that evidence into an architecture. Whether it delivers on that promise in practice is an empirical question the field, including the authors, still owes itself an answer to, and turning this proposal into a validated system, tested against adversarial conditions and extended to underserved languages such as Urdu, is the natural next step.

REFERENCES

- Akhtar, Z., Pendyala, T. L., & Athmakuri, V. S. (2024). Video and audio deepfake datasets and open issues in deepfake technology: Being ahead of the curve. *Forensic Sciences*, 4(3), 289–377. <https://doi.org/10.3390/forensicsci4030021>
- Azizah, S. F. N., Cahyono, H. D., Sihwi, S. W., & Widiarto, W. (2023). Performance analysis of transformer based models (BERT, ALBERT and RoBERTa) in fake news detection. In *2023 6th International Conference on Information and Communications Technology (ICOIACT)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICOIACT59844.2023.10455849>
- Bammey, Q. (2023). Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5, 1–9. <https://doi.org/10.1109/OJSP.2023.3337714>
- Choudhary, A., & Arora, A. (2024). Assessment of bidirectional transformer encoder model and attention based bidirectional LSTM language models for fake news detection. *Journal of Retailing and Consumer Services*, 76, Article 103545. <https://doi.org/10.1016/j.jretconser.2023.103545>
- Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5). IEEE. <https://doi.org/10.1109/ICASSP49357.2023.10095167>
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) dataset. arXiv. <https://doi.org/10.48550/arXiv.2006.07397>
- Domenteanu, A., Tătaru, G. C., Crăciun, L., Molănescu, A. G., Cotfas, L. A., & Delcea, C. (2024). Living in the age of deepfakes: A bibliometric exploration of trends, challenges, and detection approaches. *Information*, 15(9), Article 525. <https://doi.org/10.3390/info15090525>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. <https://doi.org/10.48550/arXiv.2010.11929>
- Farooq, M. S., Gilani, S. M. A., Manzoor, M. F., & Shaheen, M. (2025). Detecting fake news in Urdu language using machine learning, deep learning, and large language model-based approaches. *Information*, 16(7), Article 595. <https://doi.org/10.3390/info16070595>
- Farooq, M. S., Naseem, A., & Rustam, F. (2023). Fake news detection in Urdu language using machine learning. *PeerJ Computer Science*, 9, Article e1353. <https://doi.org/10.7717/peerj.cs.1353>

- Guo, Y., Ge, H., & Li, J. (2023). A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism. *Frontiers in Computer Science*, 5, Article 1159063. <https://doi.org/10.3389/fcomp.2023.1159063>
- Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024). Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. *arXiv*. <https://doi.org/10.48550/arXiv.2401.12070>
- Högemann, M., Betke, J., & Thomas, O. (2025). What you see is not what you get anymore: A mixed-methods approach on human perception of AI-generated images. *Frontiers in Artificial Intelligence*, 8, Article 1707336. <https://doi.org/10.3389/frai.2025.1707336>
- Hussain, G., Wasim, M., Hameed, S., & Rehman, A. (2025). Fake news detection landscape: Datasets, data modalities, AI approaches, their challenges, and future perspectives. *IEEE Access*, 13, 54757–54778. <https://doi.org/10.1109/ACCESS.2025.3553909>
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), Article e2208839120. <https://doi.org/10.1073/pnas.2208839120>
- Jiang, Z., Zhang, J., & Gong, N. Z. (2023). Evading watermark based detection of AI-generated content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1168–1181). ACM. <https://doi.org/10.1145/3576915.3623189>
- Karandikar, A., Thakare, Y., Sah, O., Sah, R., Nafde, S., & Kumar, S. (2023). Detection of deepfake video using residual neural network and long short-term memory. *International Journal of Next-Generation Computing*, 14(1).
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8107–8116). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00813>
- Kharvi, P. L. (2024). Understanding the impact of AI-generated deepfakes on public opinion, political discourse, and personal security in social media. *IEEE Security & Privacy*, 22(4), 115–122. <https://doi.org/10.1109/MSEC.2024.3405963>
- Li, Y., Jia, K., & Wang, Q. (2024). Multimodal fake news detection based on contrastive learning and similarity fusion. *IEEE Access*, 12, 155351–155364. <https://doi.org/10.1109/ACCESS.2024.3481311>
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3204–3213). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00327>
- Lin, S.-Y., Chen, Y.-C., Chang, Y.-H., Lo, S.-H., & Chao, K.-M. (2024). Text-image multimodal fusion model for enhanced fake news detection. *Science Progress*, 107(4). <https://doi.org/10.1177/00368504241292685>

- Mathews, S., Trivedi, S., & House, A. (2023). An explainable deepfake detection framework on a novel unconstrained dataset. *Complex & Intelligent Systems*, 9(4), 4425–4437. <https://doi.org/10.1007/s40747-022-00956-7>
- Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2022). DeepFake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 6111–6121. <https://doi.org/10.1109/TPAMI.2021.3093446>
- Ojha, U., Li, Y., & Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24480–24489). IEEE. <https://doi.org/10.1109/CVPR52729.2023.02345>
- Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 10, 25494–25513. <https://doi.org/10.1109/ACCESS.2022.3154404>
- Tao, R., Tang, S., Qin, H., Wang, W., Wei, Y., & Zhao, Y. (2025). LEDNet: A multimodal foundation model for robust deepfake detection. *Science China Information Sciences*, 68, Article 160106. <https://doi.org/10.1007/s11432-024-4400-8>
- Uppada, S. K., Patel, P., & B, S. (2023). An image and text-based multimodal model for detecting fake news in OSN's. *Journal of Intelligent Information Systems*, 61(2), 367–393. <https://doi.org/10.1007/s10844-022-00764-y>
- Wang, T., Cheng, H., Chow, K. P., & Nie, L. (2023). Deep convolutional pooling transformer for deepfake detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(6), 1–20. <https://doi.org/10.1145/3588574>
- Yang, P., Ma, J., Liu, Y., & Liu, M. (2023). Multi-modal transformer for fake news detection. *Mathematical Biosciences and Engineering*, 20(8), 14699–14717. <https://doi.org/10.3934/mbe.2023657>