International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion

Abdulrahman Alshaman[1], Majed Alsafifih[1], Mostafa Elgayar[1]
[1]Arab East Colleges, Riyadh, Kingdom of Saudi Arabia.
Email address: mmalgayar@arabeast.edu.sa

***Abstract*—*Deepfake technology has become a serious challenge to digital media security due to the increasing realism of manipulated video and audio content. This paper proposes a multimodal deepfake detection framework that combines EfficientNet-B4 for visual feature extraction and a Conformer encoder for audio analysis. A cross-attention fusion mechanism is used to identify inconsistencies between facial movements and speech, enabling more effective detection than approaches that analyze each modality separately. The framework uses FaceForensics++, DFDC, and FakeAVCeleb datasets with compression-based augmentation to improve robustness under real-world conditions. The proposed approach aims to enhance cross-dataset generalization while maintaining computational efficiency suitable for implementation on a consumer GPU.*

I. Introduction

The Deepfake has proved to be a serious threat to digital media integrity. New generations of generative adversarial networks and, more recently, diffusion models have allowed synthetic images, videos, and audio recordings to be generated which both visually and acoustically are believable to a human observer's eye and ear. What started as a niche research curiosity developed into an instrument strategically targeted at identity theft, financial fraud, and intentional disinformation propagation on social media. Deepfake is broadly a common term that encompasses digital content created or artificially altered using AI techniques to portray events or statements that never happened.[2] Recognition of these types of content is the current top worry for cybersecurity professionals, policymakers, and the academic community. As generative models advance, the visual artifacts that used to render synthetic media identifiable—fading around the edges of the face, blinking unnatural patterns, or light inconsistencies—are becoming increasingly scarce in the outputs from the latest generation of generators.[8] This advancement of technology serves only to deepen the distance between what may be generated and what can reliably be detected; putting increasing strain on detection research.[9] To classify the methods that current detection approaches seek into three basic categories, based upon what modality they select. Image and video-based methods, the largest body of research, primarily rely on convolutional neural networks to pick up spatial features that distinguish real frames from artificially altered ones.[3][7] Audio-based methods, relatively less developed, pay attention to the spectral and prosodic features to detect the synthesized or converted speech.[10] A third, recent line of research deals with the construction of multimodal detectors that can simultaneously detect visual and audio streams by noting that disturbances between the two channels can reveal forgeries that are otherwise undetectable in one approach and are undetectable by the other.[5][6] Although significant in terms of work published, a number of structural limitations remain in the field of the latter. The vast majority of detection methods are based only on benchmark datasets not representative of deepfake data encountered in the real world. Consequently, models that are nearly accurate on FaceForensics++ or Celeb-DF often do not succeed when used for less developed architectures or when faced with standard post-processing processes, for example social network compression.[12] The area is also characterised by a significant modality imbalance as the vast majority of studies are performed on visual deepfakes, with audio and text-based forgeries being relatively under investigated.[13][15] In this light, we propose a multimodal deepfake detection framework which integrates well established architectures: EfficientNet-B4 on the visual layer and a Conformer encoder on the audio layer linked by a cross-attention module that can learn how to recognize mismatches between the two streams. The proposed framework is trained in tandem with FaceForensics++, DFDC and FakeAVCeleb by multi-layer augmentation (which mimics the compression effects of social media platforms), and is designed to work on one consumer GPU, as opposed to server-grade infrastructure. Three particular aspects informed the design of this research paper. The first is joint audio-visual learning: rather than learning separate detectors and collating their scores at the end, the cross-attention layer requires the model to look at both streams simultaneously while training, learning the sort of lip-sync errors and voice-face timing discrepancies that single-modality detectors miss entirely. The second is cross-dataset generalization: training on three datasets that diverge in the manner in which their fakes were generated and then employing compression augmentation exposes the model to more variation than any single-benchmark work and makes it less likely to collapse when encountering the content of a tool it has not considered before. Third, the deployment feasibility is the main concern: all architecture decisions within the framework have been made for running within the memory budget of a standard GPU since such detector only works on research hardware which have low real-world value. The sections that follow examine the relevant literature, compare

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.

previous approaches, carefully map the research gap, and then present the proposed framework and evaluated impact.

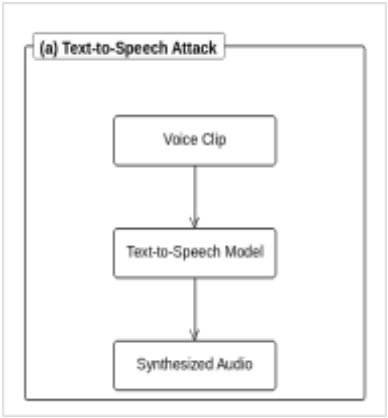

Fig. 1. Text-to-Speech Attack. A text input is processed by a TTS model to generate synthetic audio that mimics a real person's voice.

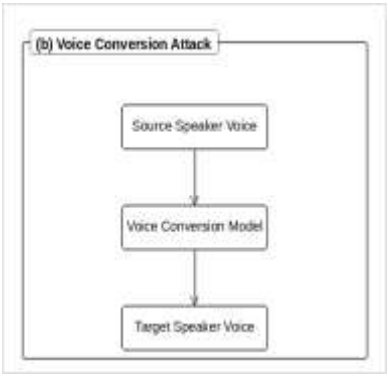

Fig. 2. Voice Conversion Attack. A source speaker's voice is converted by a model to match a target speaker's identity while keeping the original speech content.

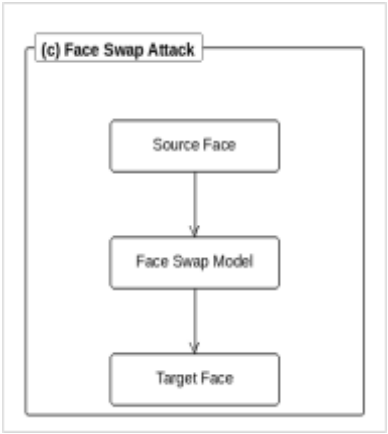

Fig. 3. Face Swap Attack. A source face in a real video is replaced with a target face using a generative model that matches geometry, skin tone, and lighting.

II. Literature Review

Al-Dhabi and Zhang et al[1] developed a detection model that works across both spatial and frequency domains. Rather than feeding raw pixel data into a single network, they extracted texture descriptors HOG, LBP, and KAZE and merged the resulting feature vectors before classification. Testing on FaceForensics++ and Celeb-DF(v2) yielded accuracy figures of 92% and 96% respectively. The main contribution is practical: by keeping the network shallow and the feature pipeline compact, the approach stays within the memory and latency constraints of devices that cannot run large transformer models. The trade-off, which the authors acknowledge, is that the fused descriptor set was tuned for the two benchmark datasets and has not been validated against manipulation types it has never seen during training.

Alanazi et al[2] took a different angle by studying the downstream effects of deepfake content rather than its detection. Through structured survey instruments distributed to social media users, they documented measurable shifts in political attitudes and declining trust in video-based news following exposure to AI-generated synthetic media. The study does not propose a detection algorithm, but it provides the clearest evidence in the reviewed literature that the harm from deepfakes is not hypothetical it is already being observed in public opinion data. This context is relevant to any detection research because it establishes what is at stake if the accuracy gap between generation and detection is not closed.

El-Gayar, M. M et al[3] addressed the temporal dimension of video forgery by framing consecutive frames as nodes in a graph and learning the relationships between them with a graph neural network. The CNN backbone processes each frame independently through four convolutional blocks; the GNN then operates over the resulting embeddings to detect inconsistencies that span across time rather than within a single frame. Three fusion variants additive, multiplicative, and sequential were tested to determine how best to combine the spatial and temporal signals. The sequential fusion configuration performed best overall. While the reported accuracy on DFDC and FaceForensics++ is competitive, the graph construction and message-passing steps add computational overhead that limits throughput in scenarios where inference time matters.

Mohammed et al[4] constructed a three-stage pipeline that handles image, audio, and video forgeries within a unified architecture. The first stage applies XceptionNet to still frames and produces an image-level authenticity score. The second stage processes spectrograms with a one-dimensional CNN to flag synthesized speech. The third stage combines XceptionNet with an LSTM to capture the way facial geometry evolves across a video sequence. A final ensemble step merges the three scores into a single decision. Evaluated on CelebA and DFDC, the image detection stage reached 95.56% accuracy. The design is notable for treating audio and visual manipulation as parallel rather than independent problems, though in practice the audio and image branches are trained separately and only joined at inference, which means cross-modal inconsistencies are not exploited during learning.

Alkhalil and Alqahtani[5] produced a comprehensive survey of deepfake detection methodologies spanning image, video, audio, and text modalities. They evaluated both unimodal and multimodal approaches, traced the evolution of fusion strategies from simple concatenation to attention-based integration, and mapped the datasets most commonly used for evaluation. One of the more useful observations in the survey is

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.

that the field has not settled on a consistent evaluation protocol: studies differ in the train-test splits they apply, the compression levels they include, and the generation models they cover, which makes direct comparison across papers unreliable. The survey stops short of proposing a concrete detector, but its taxonomy of gaps provides a clear agenda for future work.

Salvi et al[6] proposed a multimodal detection framework that trains the audio and video branches on separate monomodal datasets before combining them at inference through a late-fusion mechanism. The key motivation is practical: paired audio-visual training data is scarce and difficult to collect, so requiring the two branches to be trained jointly would significantly constrain what data the system can learn from. On FakeAVCeleb and DFDC the framework reached 93.1% accuracy. The architecture is relatively simple compared to end-to-end multimodal networks, and the authors report that it generalizes more reliably than models trained exclusively on one source. The limitation is that late fusion cannot capture the cross-modal correlations that might appear when, for example, head movement is inconsistent with the underlying speech rhythm.

Hashmi et al[7] concentrated on image-level forensics, using convolutional networks to recover the manipulation traces that remain in an image even after post-processing. Spectral analysis through DCT and FFT transforms was combined with texture features to produce a descriptor that is more stable under JPEG compression than purely spatial representations. The paper frames detection in the context of digital evidence gathering rather than real-time filtering, which changes the design priorities: reliability and interpretability matter more than throughput. The authors demonstrate that their approach retains acceptable detection rates at compression levels typically applied by social media platforms, which is a scenario where many other detectors degrade noticeably.

Ho et al[8] did not address detection at all; their paper describes a cascaded diffusion model capable of generating photorealistic images at resolutions up to 256x256 with a FID score of 1.48, outperforming prior GAN-based generators on standard benchmarks. The reason this work is included in the review is its implications for detection research: the artifacts that conventional detectors look for blurring, spectral inconsistencies, facial boundary distortions are largely absent from diffusion model outputs. Understanding how the highest-quality forgeries are constructed is a precondition for building detectors that can identify them, and cascaded diffusion represents the current frontier of generation capability that detection methods must contend with.

Heidari et al [9] carried out a systematic review of deep learning architectures applied to deepfake detection, organizing the literature into four families: CNN-based, RNN-based, Transformer-based, and hybrid models. For each family they identified representative works, summarized the detection features used, and noted the datasets and accuracy figures reported. A recurring finding is that models evaluated only within a single dataset consistently overestimate real-world performance; the accuracy gap between intra-dataset and cross-dataset evaluation ranges from a few percentage points for simpler manipulations to more than twenty points for recently generated content. The review also identifies a gap in explainability: most detectors produce a binary output without indicating which visual or acoustic feature drove the decision, which limits their utility in forensic and legal contexts.

Almutairi and Elgibreen[10] focused exclusively on audio deepfakes, surveying methods that target voice conversion, text-to-speech synthesis, and replay attacks. They reviewed low-level acoustic features such as MFCC, LFCC, and log-mel spectrograms alongside the neural architectures 1D CNNs, 2D CNNs operating on spectrograms, and end-to-end models such as RawNet that have been applied to them. A central finding is that existing audio detectors perform well on the ASVspoof benchmark suite but degrade significantly on in-the-wild recordings, which tend to involve environmental noise, telephone channel distortion, and codec artifacts not present in the evaluation data. The survey concludes that audio deepfake detection remains an open problem and that the gap between laboratory performance and real deployment is wider in audio than in video.

Akhtar and Ragini[11] reviewed deepfake media forensics from a security and authentication perspective, covering both passive detection methods which analyze content without any prior knowledge of the manipulation process and active methods that embed watermarks or digital signatures at creation time. They also discussed traceability techniques that attempt to identify the specific generation model responsible for a given piece of synthetic media. A recurring observation in the paper is that the field lacks standardized evaluation benchmarks for forensic tasks beyond simple binary classification, which makes it difficult to assess progress toward the more demanding goals that legal and investigative contexts require.

Coccomini et al[12] conducted a controlled comparison of CNN, Vision Transformer, and Swin Transformer architectures with the specific aim of characterizing how each generalizes across dataset boundaries. Their experiments showed that CNNs trained on limited data tend to outperform Transformers in that regime because Transformers require substantially more training examples to learn useful attention patterns. When the training set is large and diverse, however, Vision Transformers generalize more reliably because their global self-attention mechanism is less likely to latch onto dataset-specific artifacts. The practical implication for practitioners is that architecture selection should be driven by the size and diversity of available training data rather than by benchmark rankings.

Qureshi et al [13] surveyed deepfake detection specifically in the social media context, examining how platform-specific processing recompression, thumbnail generation, metadata stripping affects the reliability of detection signals. They organized their review across image, video, audio, and text modalities and included case studies of high-profile incidents in which synthetic media was used to spread disinformation. The most significant gap they identified is in text deepfakes: while large language model outputs are increasingly used in coordinated influence operations, the forensic community has not yet produced robust, scalable methods for distinguishing AI-generated text from human writing in adversarial conditions.

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.

Zhang et al [14] reviewed the application of remote photoplethysmography as a biometric signal for deepfake detection. The principle is that subtle periodic color changes in facial skin caused by blood circulation can be extracted from video without any contact sensors. Because current deepfake generation methods do not model cardiovascular dynamics, synthetic faces either lack these signals entirely or reproduce them in a phase-inconsistent way across different facial regions. The review documents the conditions under which rPPG-based detection is reliable good lighting, minimal head motion, high video quality and the conditions under which it fails, particularly when the video has been compressed below the level at which the color signal is recoverable.

Tariq et al[15] produced a broad survey covering detection and impact assessment across visual, audio, and textual deepfakes. On the detection side they reviewed the architecture families used for each modality and discussed the datasets available for evaluation. On the impact side they analyzed documented cases in which synthetic media caused reputational, financial, or political harm, and reviewed the legal and regulatory responses that have been attempted in various jurisdictions. The survey's most useful contribution for the present work is its cross-modality perspective: by treating all three modalities in a single document it makes visible the extent to which the field is fragmented, with specialists in visual detection, audio anti-spoofing, and natural language processing rarely engaging with each other's methods.

Alkawaz et al[16] examined the combination of CNN and LSTM networks for video deepfake detection. Individual frames are passed through a CNN to produce spatial feature vectors; the LSTM then processes the ordered sequence of vectors to capture how facial appearance changes over time. The rationale is that a convincing single frame does not guarantee a convincing video: fine-grained temporal inconsistencies in lip motion, eye blinks, and head pose are patterns that frame-by-frame analysis misses but sequence modeling can detect. Tested on FaceForensics++ and DFDC, the CNN-LSTM combination reached 98.7% accuracy. The authors note that the model was not evaluated on content generated after the training data was collected, so its ability to detect novel manipulation methods is unknown.

Croitoru et al[17] traced the evolution of deepfake generation and detection from early GAN-based approaches through variational autoencoders to the diffusion models that now dominate the generation side of the field. On the detection side they documented a consistent pattern: each time generation quality improves, existing detectors require retraining or architectural modification to remain effective. They also proposed a taxonomy of detection methods organized by whether the detector is trained on the same distribution as the forgeries it will encounter an in-distribution setting or on a different distribution, the cross-dataset setting that better approximates real deployment. The survey is notable for being explicit about the arms-race dynamic at the core of the field: generation capability is advancing faster than detection, and the gap has been widening rather than narrowing.

Nguyen et al[18] proposed a hybrid architecture that combines convolutional, recurrent, and attention-based processing within a single detection model. Three-dimensional Morphable Models are first used to extract an identity-consistent facial representation that is robust to pose and illumination variation. A CNN processes spatial features from this representation; an LSTM captures the frame-to-frame dynamics; and a Transformer layer attends to long-range temporal dependencies across the full video clip. Evaluated on FaceForensics++ the model reached 97.1% accuracy. The authors are candid about the computational cost: the 3DMM fitting step alone is too slow for real-time use, and the full pipeline requires GPU memory beyond what is available in typical deployment environments.

III. Comparison of Previous Studies

Table I summarizes the reviewed studies across four dimensions: the study title and year, the core methodology employed, and the innovative contribution each work offers. cover detection approaches across image, video, audio, and multimodal formats.

IV. Research Gap

One scrutiny of the eighteen studies previously reviewed has highlighted a number of recurring shortcomings that no such work has fully addressed. The first and most common limitation that we see consistently is the cross-dataset generalization problem. Models trained on one dataset provide high accuracy in that distribution but demonstrate large performance drops upon evaluation on unseen data. [12][9] And this fragility is more than just a technical issue: in a practical deployment context of deepfakes created by multiple tools and pushed throughout channels that add compression, rescaling and transcoding, an overfitting detector on benchmark condition doesn't deliver much of a practical backup. The second issue relates to modality coverage. The literature is very partial to visual detection. The audio deepfakes where both voice conversion and text-to-speech synthesis are involved have been studied only a minuscule number of times, and the approaches available typically struggle with in-the-wild conditions i.e., audio encountered outside of controlled recording methods. [10][15] Although real-world damage from textual deepfakes including AI-generated news articles or social media posts has been extensively documented, this remains an almost invariant area of non-focus in the technical detection. [13]. A third gap is unified multimodal detection. Multiple surveys have introduced the need to be used at any point on audio and visual streams, and some experimental efforts have so far been made. [5][6] However, the existing approaches rely either on synchronized audio-visual training pairs a scenario seldom satisfied in real-world data, or multimodal detection using late-stage fusion of independently trained monomodal models which can never exploit cross-modal inconsistencies that might provide up to the most informative detection signal during training. [4]. The fourth gulf lies in computational ease of application. State-of-the-art detection architectures, including Transformer based networks and diffusion-aware networks, impose memory and inference limitations which preclude their adoption on mobile- or edge-based devices. Lightweight detection has been explored

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.

in [1] but the accuracy-efficiency trade-off has not been resolved in order to maintain cross-dataset robustness.

TABLE I. Comparison of Previous Studies on Deepfake Detection.

S. No.	Comparison of Previous Studies		
	Search title and study year	Methodology/technique used	Innovative tool/proposal
1	Al-Dhabi & Zhang (2025): Lightweight Deepfake Detection Based on Multi-Feature Fusion	HOG, LBP, KAZE feature fusion; SVM/ML classifiers; spatial-frequency dual analysis	Lightweight dual-domain detection framework suitable for edge devices with limited resources
2	Alanazi et al. (2024): AI-Generated Deepfakes Impact on Public Opinion and Personal Security	Survey analysis; qualitative coding of social media case studies; impact scoring	First framework linking deepfake diffusion on social platforms to quantified trust erosion
3	El-Gayar, M. M et al. (2024): Detecting Deep Fake Videos Using Graph Neural Network	GNN temporal modeling; FuNet-A/M/C fusion blocks; four-stage CNN backbone	Inter-frame graph structure capturing temporal manipulation traces beyond CNN reach
4	Mohammed et al. (2025): Three-Stage Multi-Modal Deep Learning Deepfake Detection	XceptionNet (image); 1D-CNN/ResNet (audio); LSTM (video); ensemble voting	Unified three-stage pipeline handling image, audio, and video deepfakes simultaneously
5	Alkhalil & Alqahtani (2025): Survey on Multimedia-Enabled Deepfake Detection	Systematic literature review; single/multi-modal evaluation; fusion trend classification	Multimedia taxonomy exposing gaps in cross-modal fusion and real-time deployment
6	Salvi et al. (2023): Robust Multimodal Deepfake Detection via Audiovisual Fusion	Independent monomodal training; audio-visual late fusion; cross-dataset benchmarking	Fusion framework achieving generalization without requiring synchronized AV training pairs
7	Hashmi et al. (2025): Deepfake Image Forensics for Privacy and Authenticity	CNN forensic pipeline; DCT/FFT spectral features; texture analysis; data augmentation	Applied forensic tool verifying image authenticity for privacy protection workflows
8	Ho et al. (2022): Cascaded Diffusion Models for High Fidelity Image Generation	Cascaded diffusion pipeline; conditioning augmentation; progressive upsampling 64-256px	State-of-the-art generative reference (FID 1.48); essential baseline for detection research
9	Heidari et al. (2024): Deepfake Detection Using Deep Learning: Systematic Review	Taxonomy of CNN, RNN, Transformer, hybrid; cross-dataset generalization; XAI trend analysis	Comprehensive DL classification system for deepfake detection across all three modalities
10	Almutairi & Elgibreen (2025): Audio Deepfake Detection Survey	MFCC, LFCC, spectrogram features; 1D/2D CNN; RawNet; in-the-wild condition testing	Dedicated audio deepfake survey identifying real-world gaps in voice spoofing detection
11	Akhtar & Ragini (2025): Deepfake Media Forensics: Status and Future Challenges	Passive/active authentication review; digital signature tracing; benchmark analysis	Multimodal forensics roadmap highlighting open problems in traceability and standardization
12	Coccomini et al. (2023): Generalization in Video Deepfake Detection	Cross-dataset CNN vs ViT vs Swin-T comparison; intra/inter-domain evaluation protocol	Architecture selection guide: CNNs for small data, ViTs for diverse-dataset generalization
13	Qureshi et al. (2024): Multimodal Deepfake Forensics on Social Media	Systematic survey of image/video/audio/text detection; social media case study analysis	First survey flagging textual deepfakes as a critical and underexplored forensic domain
14	Zhang et al. (2025): Deepfake Detection via Remote Photoplethysmography (rPPG)	rPPG physiological signal extraction; temporal/frequency PPG analysis; compression testing	Biometric detection using heart-rate signals absent in synthetic faces; novel passive cue
15	Tariq et al. (2023): Detection and Impacts of Deepfakes Across Visual, Audio, Text	Multi-format detection review; misinformation impact scoring; policy and legal analysis	Cross-format impact survey covering societal, legal, and cybersecurity deepfake threats
16	Alkawaz et al. (2024): CNN with LSTM for Video Deepfake Detection	Frame extraction; CNN spatial features; LSTM/GRU temporal modeling; binary classification	Validated CNN-LSTM integration achieving 98.7% on FF++ for spatio-temporal detection
17	Croitoru et al. (2024): Deepfake Generation and Detection in the Generative AI Era	GAN-to-diffusion evolution review; detection taxonomy; open challenge mapping	Comprehensive generative-AI era survey showing detection consistently lags generation
18	Nguyen et al. (2025): Hybrid CNN-LSTM-Transformer for Video Deepfake Detection	3DMM identity features; CNN spatial; LSTM temporal; Transformer long-range attention	Identity-aware hybrid model reaching 97.1% cross-dataset accuracy at high training cost

Following the aforementioned gaps, this paper proposes a multimodal deepfake detection framework that can be built on two existing architectures that have been verified independently in the literature but have yet to be combined in this configuration. From a visual point of view, EfficientNet-B4 acts as a backbone due to competitive accuracy at a fraction of the parameter count of larger networks, a crucial aspect of targeting deployment outside server environments. There is a Conformer encoder a combination of convolutional and self-attention layers that encodes mel-spectrogram visuals of the audio track, as it achieves good performance on the ASVspoof benchmarks but is more efficient than full Transformer stacks. Instead of relying on simple concatenation, the two feature streams are connected via a cross-attention module. The cross-attention makes the network learn what portion of the audio representation is useful for each visual feature - if the lip movement of a face doesn't exactly match the spoken language, it will cause a noticeable disarray in the attention weights. This is exactly the kind of signal that visual detectors alone cannot catch, and that late fusions generally give only to approximate the signals. For generalization, the model is jointly trained on FaceForensics++, DFDC, and FakeAVCeleb, and augmented to simulate the JPEG and H.264 compression artifacts presented by social media applications. This is a pragmatic and verifiable design: all three datasets are publicly available, EfficientNet and Conformer are both open-source implementations, and the cross-attention mechanism introduces no foreign dependencies. The outcome is a detector that does effectively over the datasets it has never been optimized for, works with videos with both

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.

the face and voice modulated and operates within the memory limits of a single consumer GPU.

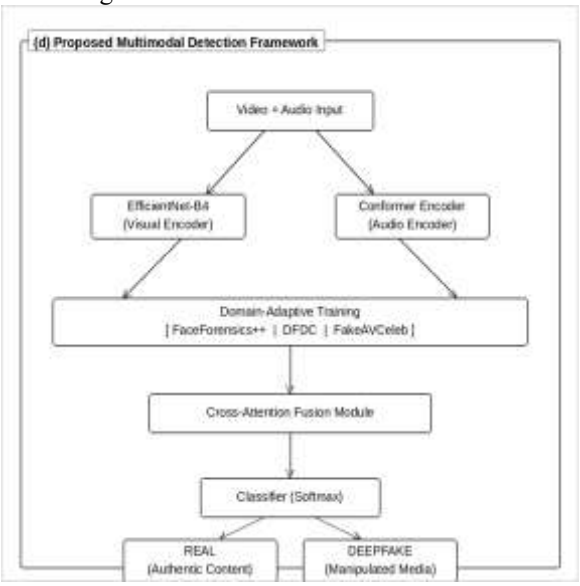

Fig. 4. Proposed Multimodal Detection Framework. EfficientNet-B4 and a Conformer Encoder extract visual and audio features in parallel, a cross-attention module detects audio-visual mismatches, and a classifier outputs REAL or DEEPFAKE.

V. Methodology

This new paper presents a holistic framework that collects two neural architectures that independently demonstrated their worth in the literature, but were not trained end-to-end as a coupled system. In that respect, the visual step is dealt with by EfficientNet-B4. It was chosen not because it is the largest available network, but because it is not: using compound scaling strategy that limits parameters and yet captures spatial resolutions required to catch fine-grained boundary artefacts and texture inconsistencies that manipulation can produce. But for a task where this is an imperative - in that deployment on a single consumer GPU is a design requirement rather than an afterthought, that means something. The audio component consists of a Conformer encoder. Unlike other Transformer, the Conformer is constructed within its two attention sub-layers of its convolutional module giving it a specific sensitivity to local phonetic detail as well as long-range prosodic structure. Since spectrogram features have been consistently steady throughout the codec distortions that social media platforms tend to introduce, only mel-spectrogram representations are fed into this encoder, not raw waveforms. After some early experimentation showing that getting deeper yielded small gains at a cost of considerable memory overhead, the encoder depth was also set at six layers. These two feature streams are unified through a cross-attention module instead of by simple vector concatenation. Here, visual feature tokens are queries and audio features tokens are keys and values in the cross attention layer. This implies that the network can learn to interrogate each and every spatial segment of the face that it is assigned in the context of the audio representation in question. A face that has stable lip motions and consistent speech will generate coherent attention patterns; a face that has been switched onto a different audio track will create an incoherent attention map, as no section of the face truly maps to the underlying sound. That incoherence is the signal the classifier is trained to be able to identify. Training has been done with three datasets: FaceForensics++, the DeepFake Detection Challenge dataset, and FakeAVCeleb. Each of these datasets were simulated with various manipulation tools and represent various demographic distributions and recording conditions, so it is much more likely to leave our model exposed to substantial variability when training on all three datasets than it would for most single-source studies. The batch-level balance of the three datasets was to avoid the lead of the single highest-mass dataset to dominate the gradient update. Augmentation prior to passing each training sample to the network is done to mimic the JPEG recompression and H.264 re-encoding that platform processing introduced. This step is intentional: a detector seeing pristine video from labs will fail as soon as it comes across content that has been uploaded, transcoded and downloaded once or twice, and that's about what nearly all deepfakes actually look like in the wild. The whole model is learned together at the beginning, rather than pretraining in isolation. This joint training schedule is useful since it compels the cross-attention module to evolve along with the feature extractors. Had the two branches been pre-trained separately, one would converge each to a representation optimised for its own modality, and the cross-attention layer would then find itself bridging representations unable to satisfy each other. Starting from random initialization and training the whole system together avoids that problem, at the cost of a somewhat longer warm-up period before the attention patterns stabilize. Evaluation follows a leave-one-dataset-out protocol: the model is trained on two of the three datasets and tested on the third, then the process is repeated for each held-out dataset. This is a more demanding standard than the within-dataset evaluation that dominates much of the published literature, and it is also a more honest one. A system that can only detect forgeries coming from the same generator family it was trained on provides limited protection in practice, where the content encountered during deployment was produced by tools that postdate the training data. Binary cross entropy is employed as training goal and the areas under the ROC curve is mentioned as the most common evaluation measure together with accuracy because AUC is less sensitive to the class imbalance which is reported across datasets when they are treated as set of independent test cases.

VI. Conclusion

The aims of this paper have been to resolve a unique but common problem that has been overlooked in the deepfake detection literature: that is, in the literature, in the approach that visual and audio evidence are treated as separate problems which need to be solved independently of each other, and that solutions should be fused only at the output stage only. The literature review indicated this separation is not merely an architectural convenience, but a gap that allows entire groups of forgery to remain undetected. Combined together with authentic audio, a face swap will evade visual detecting. If the voice conversion was performed to a real face, audio-only detector will be evaded. Neither will certainly fail the late-fusion system that was never trained to search the mismatch

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.

between the two. The framework introduced here is specifically designed around that mismatch, and cross-attention is applied to bring the two streams into direct contact during learning not at the last decision boundary. The decision for EfficientNet-B4 and the Conformer encoder came for a practical reason that too many existing systems of this kind often overlook: that the model must fit on hardware that is actually outside a research cluster. In short, both architectures deliver competitive performance for a fraction of the amount of computational effort of their larger counterparts which enables the detector to be evaluated, fine-tuned, and ultimately deployed without needing infrastructure that most organizations don't possess. This practical reasoning is reflected in the compression augmentation applied to the model during its training. And deepfakes do not come clean; they come after being uploaded, transcoded and downloaded at least once, and a detector that has not witnessed degradation like that will fail as soon as it sees it. Training in FaceForensics++, DFDC, and FakeAVCeleb all at once, and judging using a leave-one-dataset-out protocol is the hope to take generalization very seriously rather than treating it as an afterthought. The field has far too many techniques that end up in near-perfect accuracy (accuracy on one metric) and then collapse silently when tested against content from a different generator or a different compression pipeline. Though the multi-dataset training method cannot ensure robustness to every future manipulation method, it guarantees at least that the model has faced real variation rather than memorizing the quirks of one dataset. There are actual limitations to recognize. The framework deals with the audio-visual modality combination and does not explore textual deepfakes, which are still all but ignored in the technical literature despite their increasing footprint in the real world. The cross-attention mechanism introduces an additional level of complexity, which while controllable, does mean that the model is harder to interpret than a straightforward CNN classifier; identifying which attention pattern triggered a particular detection decision is not straightforward. And as is the case with every detection system, the guarantees it provides are conditional: It was made to catch the types of manipulation represented in its training data, and a generation strategy that is novel enough will eventually evade it. What is most useful here likely is not whether this system is ultimately going to be defeated it will but if it takes the baseline in the right direction. Cross-modal training, compression-aware augmentation, and multi-dataset evaluation are simply ideas, but combining these approaches in a single system that aims towards deployment viability is a step toward feasibility which is fundamentally not available in most prior work. Further work should investigate if such cross-attention signal can be made interpretable enough to satisfy forensic use cases, whether we can generalize the framework to text, audio, and video, and if the compression augmentation strategy could be fine-tuned to suit specific processing chains used in the task.

by the major social media platforms. Those are harder problems, but the architecture laid out here gives them a reasonable starting point.

REFERENCES

[1] Al-Dhabi, Y., & Zhang, S. (2025). Lightweight deepfake detection based on multi-feature fusion. *Applied Sciences,* 15(4), 1954. https://doi.org/10.3390/app15041954

[2] Alanazi, M., Alkhudaydi, A., & Alharthi, A. (2024). Understanding the impact of AI-generated deepfakes on public opinion, political discourse, and personal security in social media. *IEEE Access.* https://doi.org/10.1109/ACCESS.2024.10552098

[3] El-Gayar, M. M., Abouhawwash, M., Askar, S. S., & Sweidan, S. (2024). A novel approach for detecting deep fake videos using graph neural network. *Journal of Big Data*, *11*(1), 22. 00884-y

[4] Mohammed, A., Ali, B., & Hassan, T. (2025). Deepfake detection in manipulated images/audio/videos: A three-stage multi-modal deep learning framework. *Inteligencia Artificial,* 28(76), 20–39. https://doi.org/10.4114/intartif.vol28iss76pp20-39

[5] Alkhalil, Z., & Alqahtani, H. (2025). A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges and future directions. *Information Retrieval Journal,* 28(1), 1–54. https://doi.org/10.1007/s10791-025-09550-0

[6] Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A robust approach to multimodal deepfake detection. *Journal of Imaging,* 9(6), 122. https://doi.org/10.3390/jimaging9060122

[7] Hashmi, M. F., Ashish, B., Keskar, A. G., & Wagh, N. D. (2025). Deepfake image forensics for privacy protection and authenticity using deep learning. *Information,* 16(4), 270. https://doi.org/10.3390/info16040270

[8] Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., & Salimans, T. (2022). Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research,* 23(47), 1–33. https://jmlr.org/papers/v23/21-0635.html

[9] Heidari, A., Navimipour, N. J., Dag, H., & Ünal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery,* 14(2), e1520. https://doi.org/10.1002/widm.1520

[10] Almutairi, Z., & Elgibreen, H. (2025). Audio deepfake detection: What has been achieved and what lies ahead. *Sensors,* 25(7), 1989. https://doi.org/10.3390/s25071989

[11] Akhtar, Z., & Ragini, R. (2025). Deepfake media forensics: Status and future challenges. *Journal of Imaging,* 11(3), 73. https://doi.org/10.3390/jimaging11030073

[12] Coccomini, D. A., Caldelli, R., Falchi, F., & Gennaro, C. (2023). On the generalization of deep learning models in video deepfake detection. *Journal of Imaging,* 9(5), 89. https://doi.org/10.3390/jimaging9050089

[13] Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). Deepfake forensics: A survey of digital forensic methods for multimodal deepfake identification on social media. *PeerJ Computer Science,* 10, e2037. https://doi.org/10.7717/peerj-cs.2037

[14] Zhang, Y., Liu, X., & Wang, R. (2025). A comprehensive review of deepfake detection techniques utilizing remote photoplethysmography. *IEEE Access.* https://doi.org/10.1109/ACCESS.2025.11214405

[15] Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access,* 11, 75529–75552. https://doi.org/10.1109/ACCESS.2023.3344653

[16] Alkawaz, M. H., Sai, A. T. H., & Alqahtani, A. (2024). An investigation into the utilisation of CNN with LSTM for video deepfake detection. *Applied Sciences,* 14(21), 9754. https://doi.org/10.3390/app14219754

[17] Croitoru, F.-A., Hiji, A.-I., Hondru, V., Ristea, N. C., Irofti, P., Popescu, M., Rusu, C., Ionescu, R. T., Khan, F. S., & Shah, M. (2024). Deepfake media generation and detection in the generative AI era: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* https://doi.org/10.1109/TPAMI.2024.3511509

[18] Nguyen, T., Nguyen, T., & Tran, T. (2025). Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification. *Multimedia Tools and Applications.* https://doi.org/10.1007/s11042-024-20548-6

Abdulrahman Alshaman, Majed Alsafifih, and Mostafa Elgayar, "Multimodal Deepfake Detection Using EfficientNet-B4 and Conformer Encoder with Cross-Attention Fusion," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 3, pp. 132-138, August 2026.