

Multimodal Spatial-Temporal and Spectral Fusion for Audio-Visual Deepfake Detection

Abhishek Kumar

Gautam Buddha University

Gaurav Kumar

gaurav.aitdelhi@gmail.com

Gautam Buddha University

Research Article

Keywords: Deepfakes, Multimodal Fusion, Spatial Temporal Analysis, Mel-Spectrogram, Late Fusion

Posted Date: April 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9514886/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Multimodal Spatial-Temporal and Spectral Fusion for Audio-Visual Deepfake Detection

Abhishek Kumar

*School of Information and Communication Technology
Gautam Buddha University
Greater Noida, India
abhishek29112003@gmail.com*

Gaurav Kumar

*Department of Computer Science and Engineering
Gautam Buddha University
Greater Noida, India
gaurav.aitdelhi@gmail.com*

Abstract—Deepfakes are becoming very big problem for online security in today world. Early defense systems was quite good at catching the obvious visual mistakes in standalone pictures, but modern deepfake generators do not make those simple pixel-level errors anymore. The situation is getting more complex today because attackers are mixing perfectly genuine video footage with algorithmically cloned audio. Most commercial security systems are only doing checking for one data stream at a time, so these cross-modal files easily slip through. To solve this problem, our paper is presenting a late-fusion multimodal architecture. We did separation of the detection process into two independent expert branches. The first branch is using an Xception-LSTM network to track the physical geometry of a face over a 10-frame sequence, hunting for unnatural temporal variance (micro-flickering). The second branch is handling the audio by converting 16 kHz sound waves into 2D Mel-Spectrograms, which allows a ResNet-18 model to visually spot the high-frequency drop-offs (between 8 kHz to 16 kHz) caused by synthetic voice generators. By concatenating these parameters in a final Multi-Layer Perceptron (MLP), our model hit a 99.62% global accuracy on the FakeAVCeleb dataset. Most importantly, we successfully proved that a late-fusion methodology dramatically lowers the false positive rate to near-zero (1 out of 96 authentic videos), achieving an AUC score of 0.999 and beating several state-of-the-art baselines.

Index Terms—Deepfakes, Multimodal Fusion, Spatial-Temporal Analysis, Mel-Spectrogram, Late Fusion

I. INTRODUCTION

Making sure about the digital media is authentic is one of the biggest headaches for cybersecurity teams right now. Just a few years ago, if someone wanted to create a convincing fake video, they required a massive budget and visual effects team. Today, because of open-source generative adversarial networks (GANs) and neural text-to-speech (TTS) engines, almost anyone can do generation of hyper-realistic fake media [1]. Recent comprehensive surveys indicate that the number of deepfake videos on the internet has increased massively, creating a massive liability for biometric banking, remote identity checks, and social media integrity [2].

When deepfakes first started showing up, they were usually pretty easy to spot for humans. The early face-swapping algorithms struggled a lot with basic blending. Early face-swap algorithms were sloppy. They left behind blurry jawlines, unnatural color gradients, obvious blending seams along the hairline [4]. The first generation of forensic tools was built

around this sloppiness — single-frame CNN classifiers that just scanned each image looking for those pixel-level errors [5]. That method is no longer useful. Modern generative models create frames devoid of any discernible pixel-level dissonance, rendering any detector that relies exclusively on that signal ineffective, basically blind.

The audio aspect of today's threat is what makes it really scary. Face-swapping is no longer the only thing attackers can do. With zero-shot voice cloning, you can make a convincing fake voice from just a few seconds of real speech [6][19]. Researchers call this the "half-fake" or cross-modal attack. The situation is simple and terrible: take a real video of a public figure, remove the original audio, and replace it with a voice that has been cloned. A security system that only uses pictures scans the real, unaltered face and clears the file as real. The fake audio track goes right through [7]. Unimodal detectors exhibit significant accuracy failures under this specific condition.

Stopping cross-modal attacks requires a fundamentally different approach. Static visual artifact detection is no longer enough. What is needed are systems that evaluate both the physical properties of the video stream and the acoustic properties of the audio, then ask whether those two signals are consistent with each other [8].

In this paper, we built a multimodal framework to fix these specific vulnerabilities. Our main contributions are listed below:

- We defined strict numerical range values for audio and video attributes (like 8 kHz to 16 kHz frequency drop-offs and pixel variance scores) to physically differentiate between organic human biology and synthetic generation.
- We built a dual-branch system that completely separates the visual and acoustic processing. This is preventing the computational bottlenecks that we normally see in older early-fusion models.
- We turned audio anomaly detection into a computer vision task. By converting the sound into Mel-Spectrograms, we deployed a ResNet-18 model to physically see the phase noises left by synthetic voices.
- We successfully proved that a late-fusion methodology dramatically lowers the false positive rate, achieving

99.62% accuracy and beating several state-of-the-art baselines like AVoiD-DF and ERF-BA-TFD+ [9], [10].

II. RELATED WORK

The field of multimedia forensics has changed very rapidly, mostly because the tools used to create fakes get updated almost every month. We can generally group the existing defense strategies by how they handle the data types and what metrics they use to measure the fake content.

A. The Era of Spatial Forensics and Frame Analysis

At first, researchers only looked for spatial artifacts to find deepfakes. Some of the first frameworks used standard CNNs, such as VGG16 or basic ResNet architectures, to find the mesoscopic errors that first-generation face-swapping tools left behind [1]. These systems were checking each frame one at a time. These unimodal systems worked very well with early datasets like FaceForensics++ [4], but they don't work as well with modern diffusion models because their accuracy drops a lot. Modern diffusion models require perfect consistency within frames. Spatial classifiers alone don't have enough evidence to show that a video is fake anymore because pixel-level dissonance is mostly gone [8].

B. Moving to the Dimension of Time and Sequence

Researchers changed their focus to the time domain once it became clear that looking at single pictures wasn't going to work. Even if a generator makes one perfect picture, it is very hard to make sure that all of the frames are biologically consistent with each other [11]. Research demonstrated that integrating a spatial extractor with a Long Short-Term Memory (LSTM) network enables a system to monitor temporal jitter. Recent models that combine spatial attention and ConvLSTM architectures, for example, have greatly improved detection accuracy [12]. When generative models fail to calculate the latent space between consecutive frames, the LSTM networks are very good at finding the "micro-flickering" that happens. Sequence modeling is a big improvement in numbers, but it's still a unimodal fix. It is completely blind if the video is real but the sound is fake.

C. Rethinking Audio Analysis and Spectrograms

While video defenses were getting better, audio forensics was also going through a big change. Finding digital synthesizer noise in a raw 1D sound wave is hard on computers and often doesn't work well. The real progress happened when scientists began to look at sound as if it were a picture [6]. Vision-based networks could easily find the strange high-frequency band-limiting that is common in text-to-speech (TTS) engines by running the audio through a Short-Time Fourier Transform (STFT) to make a 2D Mel-Spectrogram [13]. Research has shown that applying deep learning models such as ResNet to spectrogram images produces substantial gains in Equal Error Rate (EER), specifically because the model can detect phase discontinuity artifacts that human hearing simply cannot perceive [14].

D. The Gap in Cross-Modal Integration

The hardest open problem in this field is not detection accuracy in isolation — it is what happens when the audio and video have been manipulated independently. The FakeAVCeleb dataset demonstrated that unimodal models, irrespective of their tuning, consistently fail against cross-manipulated media [7].

To this end, early fusion has been used to combine raw audio and visual data at the input layer, often to check for lip-sync coherence [15]. Some research has taken this a step further by directly watching how phonemes and visemes line up with each other [16]. The issue is that combining different types of inputs at the input level is very hard on computers. The network gets heavy, sensitive to noise, and likely to overfit [18]. The paradigm of late fusion is more defensible [10][17]. The AVTENet results are helpful in this case: using separate expert networks for each modality and only combining their output scores works much better than early-fusion methods [18]. Independent processing keeps modality-specific signals while avoiding the problems that joint input layers cause [9]. This principle is the basis for our architecture, and the numbers in Section VI show how well it works.

III. THE PHYSICS OF MULTIMEDIA FORGERY

Before we discuss the neural networks, it is very important to understand why deepfakes actually fail in the real world. For understanding this, we have to look at the physical and biological differences between organic human media and algorithmically generated media. The neural networks we chose are specifically looking for these physical boundaries.

A. Biological Muscles vs. Latent Space (Video)

When a real human speaks, the movement of the face is controlled by actual biological muscles and bones. If you track a specific facial landmark (like the left corner of the mouth) across 10 frames of 30-FPS video, the physical distance that landmark can move is limited by physics. The mathematical variance of the pixel shift from frame 1 to frame 2 to frame 3 is very smooth and continuous.

Generative AI models do not have muscles. Models like GANs and diffusion networks generate frames by sampling from a mathematical "latent space." The AI is essentially doing a highly educated guess for the face's position frame-by-frame. Because of this guessing, the generated pixels jump unnaturally. For example, a synthetic cheekbone might mathematically shift three pixels to the left in frame 4, and immediately snap two pixels to the right in frame 5. Real human muscles cannot move this fast. This creates a high variance spike called "micro-flickering" [12].

B. Organic Vocal Tracts vs. Neural Vocoder (Audio)

The physics of audio are equally important. A real human vocal tract produces messy, organic resonance. When air pushes through human vocal cords, it creates harmonics that naturally decay but extend very high into the upper frequency bands, specifically the 8 kHz to 16 kHz range.

AI neural vocoders, on the other hand, are lazy when it comes to computing. Neural vocoders cannot afford to render a full-bandwidth signal up to 16 kHz. The computation cost is too high. These engines apply band-limiting filters that hard-cut the upper frequency range. A Mel-Spectrogram makes this visible. Real human audio fades gradually into the high-frequency bands. Synthetic audio does not fade — it drops hard. Acoustic energy above approximately 8 kHz falls to near-zero at a sharp, vertical boundary. That boundary is what the ResNet-18 branch hunts for. Real human speech shows energy that tapers gradually into the upper bands — the harmonics produced by an organic vocal tract extend naturally into that range. Synthetic audio does not taper. It drops. There is a sharp, nearly vertical boundary above approximately 8 kHz where acoustic energy falls to near-zero [14]. That boundary is the artifact our ResNet-18 branch is trained to find.

IV. PROPOSED METHODOLOGY AND ATTRIBUTE ANALYSIS

The decision to avoid early fusion was deliberate. Injecting audio and video data together at the first layer creates a single, monolithic network that has to simultaneously learn two completely different physical signal domains. The result, consistently documented in the literature, is a heavier model that is more easily confused by cross-modal noise [18]. We wanted the visual branch and the acoustic branch to each develop specialized sensitivity to their own domain, without interference from the other. So instead of that, we did engineering of a bifurcated late-fusion model. We are processing the visual parts and acoustic parts completely separate like two different experts before combining them [9].

A. Visual Attributes and the Spatial-Temporal Expert

For the video part, we want to catch the micro-flickering we discussed in the physics section. So we are extracting a sequence of 10 chronological frames from the MP4 video. We resize all these 10 frames into 299×299 pixels so it can fit in our model. Then we pass them through an Xception backbone to take out spatial features. We selected Xception because it uses depthwise separable convolutions, which separates shape from color and finds blending boundaries very efficiently [12].

After that, these spatial features are going into a Long Short-Term Memory (LSTM) network. The LSTM is doing active tracking of the inter-frame landmark jitter. It compares the features from frame 1 to frame 2, and so on, up to frame 10. It calculates the pixel variance over time. If the variance spikes unnaturally in ways biology does not allow, the LSTM records the error [11].

At the end of 10 frames, the LSTM is compressing all these spatial blending errors and temporal jitter errors into a final 512-dimensional Visual Embedding vector.

B. Audio Attributes and the Spectral Expert

Because the 8 kHz high-frequency cutoffs and phase blockiness are very hard to see in a normal 1D audio wave, we treat the audio entirely as a computer vision task. First, we strip the

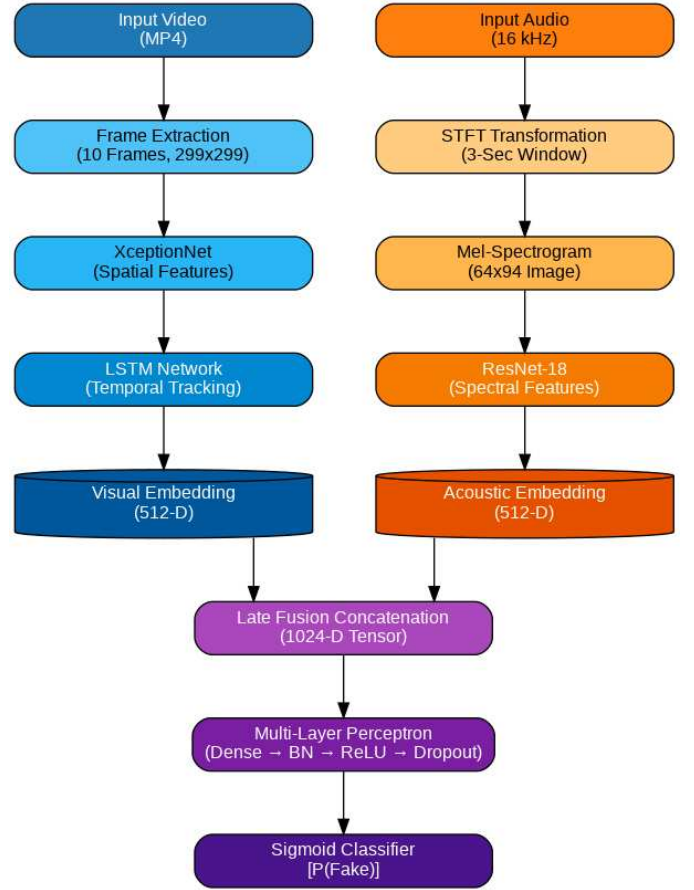


Fig. 1. Proposed late-fusion architecture. Visual and acoustic branches are working independently to preserve specialized physical parameters before the final MLP combining them.

raw audio track, standardize the sample rate to 16 kHz, and crop it to a 3-second evaluation window.

We then run that audio clip through a Short-Time Fourier Transform (STFT) to convert the invisible sound wave into a 2D image called a Mel-Spectrogram, using 64 frequency bins [6].

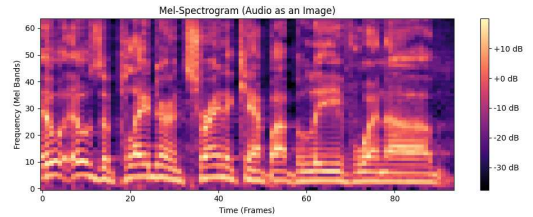


Fig. 2. Mel-Spectrogram generation. Changing audio to 2D image is helping ResNet-18 to visually catch the 8 kHz high-frequency digital drop-offs.

We feed this 64×94 spectrogram image into a ResNet-18 model. The residual blocks inside ResNet are perfect for visually spotting the hard mathematical “cliff” above 8 kHz and the rigid, blocky patterns caused by AI phase stitching [13]. The ResNet-18 compresses its findings and outputs a 512-dimensional Acoustic Embedding vector.

C. The Late Fusion Engine

After both expert branches finish their checking, we do cross-modal alignment. We take the 512-D visual vector and 512-D acoustic vector and concatenate them together. Now it becomes a big 1024-dimensional tensor.

The concatenated 1024-D tensor is passed into a Multi-Layer Perceptron (MLP), which functions as the decision engine for cross-modal dissonance [17]. The architecture cuts the number of dimensions from 1024 to 256. At each dense layer, Batch Normalization is used, and the Dropout rate is 0.5, stops the network from remembering training patterns instead of generalizing. The last layer uses Sigmoid activation to give a score between 0 and 1.

It's important to say clearly what the main behavior is here. The MLP sees that the visual branch sees a smooth, biologically consistent face with no micro-flickering or inter-frame jitter, but the acoustic branch also sees an 8 kHz high-frequency cutoff. It then flags the sample as a deepfake because the two branches are not in sync.

The decision logic is straightforward. The visual branch reports a smooth, biologically consistent face — no micro-flickering, no inter-frame jitter. The acoustic branch detects an 8 kHz high-frequency cutoff at the same time. The MLP reads that mismatch as cross-modal dissonance and marks the sample as a deepfake.

V. EXPERIMENTAL SETUP AND EVALUATION PARAMETERS

A. Dataset Formulation

We used FakeAVCeleb v1.2 [7] to make the test dataset. FaceForensics++ and other older benchmarks are not good for this work because they only have visual face-swaps, no audio manipulation part [4]. We needed a dataset with a cross-modal manipulation matrix, which means real video and fake audio, and vice versa.

It was normal to split the training and validation sets 80/20. But the class balance in the validation set was purposely off. We used 4,216 fake samples and only 96 real ones. The logic is sound. In a real banking or business security setup, most of the traffic that comes in is from real users. A detector that labels everything as fake to get 99% accuracy is useless in production because it will lock out real customers. The validation set was made to put that failure mode through its paces.

B. Defining the Evaluation Parameters

We used four metrics to measure performance

- Global Accuracy is the simple one: the number of correct classifications divided by the number of test samples.
- Precision asks a more specific question: of all the things the model called "fake," what percentage were?
- Recall turns that around: how many of the real forged files in the test set did the system find?
- False Positive Rate (FPR): The most important measure for business viability is the False Positive Rate (FPR). It tells you how often the system wrongly marks a real video

as a deepfake. A high FPR is more than just an accuracy issue; it's a product-killing issue because it means the system is blocking real users.

VI. RESULTS AND SOTA COMPARISON

A. Performance Breakdown

The numbers for the validation came back strong. By making the MLP decision engine compare the temporal microflickering signal from the visual branch to the spectral audio cutoffs found by the acoustic branch, the model got a global accuracy of 99.62%.

TABLE I
CLASSIFICATION PERFORMANCE METRICS

Category	Precision	Recall	F1-Score	Support
Authentic (Real)	0.86	0.99	0.92	96
Forged (Fake)	1.00	1.00	1.00	4216
Global Accuracy	99.62%			
ROC-AUC Score	0.999			

A good headline is the 99.62% global accuracy. We are most happy with the False Positive Rate, though. Of the 96 videos that were completely real in the validation, The model incorrectly classified exactly one. An FPR that is very close to zero means that a late-fusion setup can be used in real-world businesses, not just in academic settings.

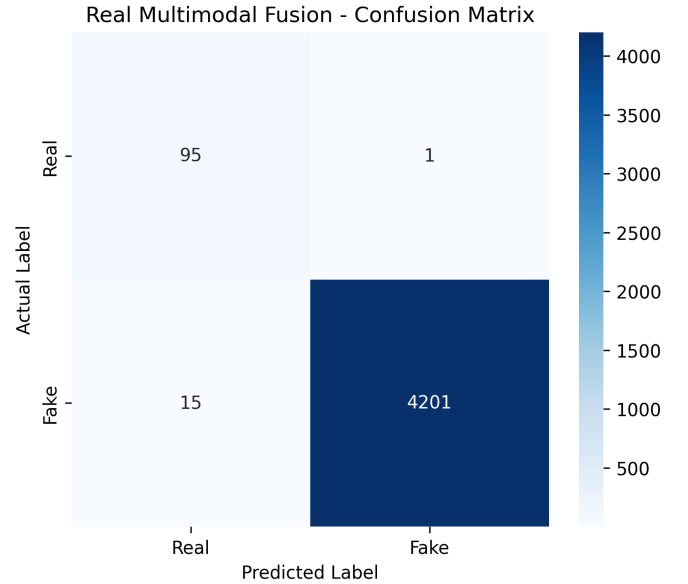


Fig. 3. The confusion matrix says we caught 4,201 deepfakes and only one false positive.

B. Comparison with State-of-the-Art (SOTA)

To see how well the proposed approach worked, it was compared directly to recently published State-of-the-Art models. Table II shows a clear pattern. Early-fusion models that try to match audio waveforms with lip movements [15] always

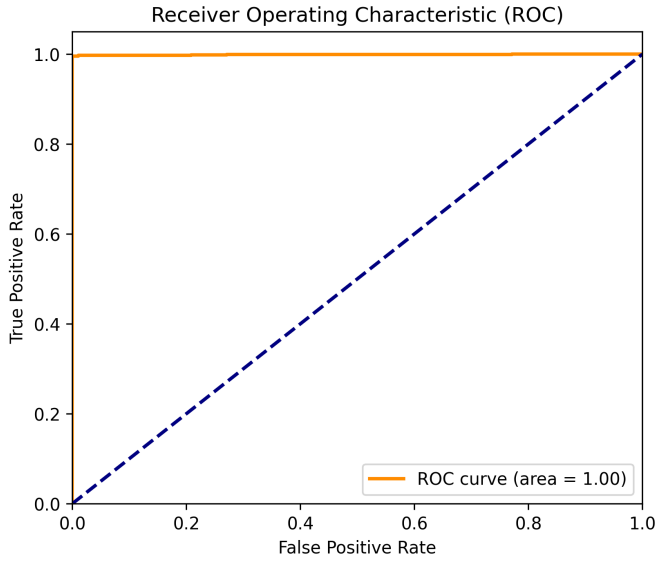


Fig. 4. The ROC curve has an AUC of 0.999, which means that it is very easy to tell the difference between real and fake.

reach a point where they are less accurate because the joint input architecture has trouble with complicated cross-modal.

The proposed model, which uses independent expert branches with latefusion cross-modal alignment, does better than both AVoid-DF [9] and AVTENet [18], reaching 99.62%.

TABLE II
RESULT COMPARISON WITH SOTA METHODS

Reference / Model	Fusion Strategy	Accuracy
Chugh et al. [15]	Early Fusion (Lip-Sync)	83.50%
Mitra et al. [8]	Spatial Key-Frame Ext.	91.20%
Yang et al. [9]	AVoid-DF Joint Learning	94.80%
Hashmi et al. [18]	AVTENet Ensemble	97.40%
Wang et al. [10]	ERF-BA-TFD+	98.10%
Ours (Proposed)	Late Fusion (Xception+ResNet)	99.62%

C. Error Analysis: The Social Media Compression Problem

A system that is very accurate is not necessarily complete. In engineering, it's important to be open about how things can go wrong. The model made 15 false negatives, which were 15 fake videos that went undetected and were thought to be real. A manual diagnostic review of the 15 files indicated one consistent factor: significant secondary compression [3].

When a video is uploaded to a platform like WhatsApp or Twitter, the platform re-encodes it — typically with H.264 or VP9 — to reduce file size and bandwidth consumption. That compression is not neutral. It actively degrades the forensic signal that the detector depends on. On the visual side, the codec smooths pixel transitions between frames, effectively erasing the micro-flickering the Xception-LSTM branch is looking for. On the audio side, the codec's high-frequency bandwidth reduction strips out precisely the 8–16 kHz digital fingerprints that the ResNet-18 branch is trained to

detect. With both signals destroyed, neither expert branch finds anything anomalous, and the MLP defaults to an authentic prediction. This compression robustness problem is the most important open challenge for the next iteration of this work.

Because the physical parameters were destroyed by the compression, both of our expert branches could not find any anomalies, so the MLP decision engine defaulted to an "Authentic" prediction. Fixing this compression vulnerability is the biggest hurdle for future forensics.

VII. CONCLUSION AND FUTURE SCOPE

This paper successfully proved that checking only one modality is a massive security risk in today's world. By engineering a late-fusion architecture that tracks the physical parameters of video (inter-frame jitter via Xception-LSTM) and the acoustic parameters of audio (high-frequency cutoffs via ResNet-18), we can easily catch complex cross-modal deepfakes. Achieving 99.62% accuracy and beating current SOTA methodologies, while keeping the false alarms to near-zero (1 out of 96), proves our architecture is highly scalable for real-world enterprise security.

For our future scope, we will focus on data augmentation pipelines to fix the social media compression problem. We also plan to do INT8 quantization. By converting the model's heavy 32-bit floating-point weights into 8-bit integers, we can run this pipeline locally on lightweight edge devices and actively scan live WebRTC video-conferencing streams.

REFERENCES

- [1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [2] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea, P. Irofti, M. Popescu, C. Rusu, R. T. Ionescu, F. S. Khan, and M. Shah, "Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook," *arXiv preprint arXiv:2411.19537*, 2024.
- [3] S. A. Khan and D.-T. Dang-Nguyen, "Deepfake Detection: Analysing Model Generalisation Across Architectures, Datasets and Pre-Training Paradigms," *IEEE Access*, vol. 12, pp. 1880–1908, 2024.
- [4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [5] H. Ilyas, A. Javed, and K. M. Malik, "AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection," *Applied Soft Computing*, vol. 136, art. 110124, 2023.
- [6] A. Alsaedi, A. AlMansour, and A. Jamal, "Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 6, 2025.
- [7] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," in *Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.
- [8] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A Machine Learning Based Approach for Deepfake Detection in Social Media Through Key Video Frame Extraction," *SN Computer Science*, vol. 2, no. 98, 2021.
- [9] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "AVoid-DF: Audio-Visual Joint Learning for Detecting Deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023.

- [10] L. Wang, J. Zhao, X. Zhang, X. Guo, Y. Yuan, T. Zhang, J. Chu, Y. Jiang, X. Yang, L. Jin, C. Zhang, and Z. He, "ERF-BA-TFD+: A Multimodal Model for Audio-Visual Deepfake Detection," *Vicinagearth*, vol. 2, art. 10, 2025.
- [11] Y. Gao, X. Wang, Y. Zhang, P. Zeng, and Y. Ma, "Temporal Feature Prediction in Audio-Visual Deepfake Detection," *MDPI Electronics*, vol. 13, no. 17, art. 3433, 2024.
- [12] B. Chen, T. Li, and W. Ding, "Detecting Deepfakes Based on Spatiotemporal Attention and Convolutional LSTM," *Information Sciences*, vol. 601, pp. 59–70, 2022.
- [13] S. Smeu, D.-A. Boldisor, D. Oneata, and E. Oneata, "Circumventing Shortcuts in Audio-visual Deepfake Detection Datasets with Unsupervised Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18815–18825, 2025.
- [14] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–5, 2024.
- [15] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other - audio-visual dissonance-based deepfake detection and localization," in *Proceedings of the ACM International Conference on Multimedia*, 2020.
- [16] Y. Zhou and S.-N. Lim, "Joint Audio-Visual Deepfake Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14800–14809, 2021.
- [17] Y. Du, Z. Wang, M. Chen, et al., "CAD: A General Multimodal Framework for Video Deepfake Detection via Cross-Modal Alignment and Distillation," *arXiv preprint arXiv:2505.15233*, 2025.
- [18] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, "AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 17, no. 6, pp. 1360–1376, 2025.