AUDIO-VISUAL DEEPFAKE DETECTION WITH CROSS MANIPULATION EVALUATION ON FAKEAVCELEB

[1]REHAM MOHAMED ABDULHAMIED, [2]MONA M. NASR, [3]FARID ALI MOUSA, [4]SARAH NAIEM

[1, 2, 4] Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

[3] Faculty of Computers and Artificial Intelligence, Department of Information Technology, Beni Suef University, Beni Suef, Egypt

ABSTRACT

Deepfake technologies have advanced rapidly in recent years, enabling highly realistic manipulations of both audio and video. While such technologies offer creative potential, they also present major risks in misinformation, fraud, and privacy violations. This paper explores the problem of detecting deepfake content using the FakeAVCeleb dataset, which provides both authentic and manipulated audio-video samples. We present experiments using state-of-the-art audio models (AASIST, RawNet2, ECAPA-TDNN), video models (Vision Transformers and SyncNet/Wav2Lip for lip-sync consistency), and multimodal fusion approaches. In particular, we evaluate robust detection under cross-manipulation scenarios, where models are tested on manipulation types unseen during training. Our results highlight the performance drop in cross-manipulation settings, emphasizing the importance of robust multimodal fusion. Fusion methods achieved improved generalization, indicating that combining complementary cues across modalities is key to resilient deepfake detection.

Keywords*: Deepfake Detection, FakeAVCeleb Dataset, Audio-Visual Forensics, Cross-Manipulation Evaluation, Video Manipulation, Audio Deepfake. Visual Deepfake Introduction.*

1. INTRODUCTION

Deepfake technology have changed the way digital media works. Attackers can make realistic false audio and video recordings using generative models like Generative Adversarial Networks (GANs) and advanced voice cloning. These fake media are bad for politics, journalism, privacy, and cybersecurity. For instance, deepfakes have been used to distribute false information about politics, make explicit content without permission, and even commit financial fraud by tricking people into giving them money over the phone.

Types of Deepfake Audio:

Deepfake audio generally refers to any audio in which important attributes have been manipulated via AI technologies while still retaining its perceived naturalness. Previous studies mainly involve five kinds of deepfake audio [1]: text-to-speech, voice conversion, emotion fake, scene fake, partially fake.

1. Text-to-speech (TTS) [2],

Figure 2 shows what is generally called voice synthesis. Its goal is to use machine learning-based models to create genuine and understandable speech from any text. Deep neural networks have made it possible for TTS models to make speech that sounds real and human-like [1]. The basic parts of TTS systems are text analysis and speech waveform production. The two main ways to make speech waveforms are concatenative [3], [4] and statistical parametric TTS [5]. The latter usually has an acoustic model and a vocoder. Recently, some end-to-end models have been suggested to make audio sound better, like Variational Inference with adversarial learning for end-to-end Text-toSpeech (VITS) [6] and FastDiff-TTS [7].

2. Voice Conversion

Voice conversion (VC) [2] is the process of digitally copying someone's voice. It tries to change the speaker's voice and prosody to sound like someone else's, while keeping the content of the speech the same. A VC system takes in a natural speech from the person speaking. There are three primary types of VC technologies: statistical parametric [8], [9], frequency warping [10], and unit-selection [11]. Statistical parametric models also include a vocoder that is similar to the one in statistical parametric TTS [12, 13]. End-to-end VC models have also been suggested in recent years to imitate the vocal characteristics of an individual [14].

3. Emotion Fake

Emotion fake [15] wants to change the audio such that the speech's emotion changes, but other things stay the same, like who is speaking and what they are saying. Changing the voice's emotions often changes the meaning of words. Figure 2 is an example of artificial emotion. Speaker B's original statement was made with a joyful tone. The fake utterance is the sound when the pleasant emotion has been transformed into a sorrowful emotion. There are two types of approaches for emotional VC called "emotion fake" [16]: ways that use parallel data and methods that do not use parallel data.

4. Scene Fake

Scene fake [17] is when speech augmentation tools change the acoustic scene of the original utterance while keeping the speaker's identity and speech content the same. An. The acoustic scene of the true speech is "Office." The artificial sound comes from an airport. If you change the scene of an original audio with another one, it will be hard to check the audio's authenticity and integrity, and the meaning of the original audio could even change.

5. Partially Fake

Partially fake [18] merely changes a few words of a sentence. The false statement is made by changing the real statements with real or synthetic audio samples. The individual who said the original and false footage is the same. The synthetic audio samples don't affect who the speaker is.

From a research standpoint, identifying deep-fakes constitutes a complex multimodal challenge. Initial research focused on visual artifacts, including facial blending borders [19], whereas contemporary investigations integrate temporal inconsistencies and physiological data. Researchers first focused on conventional replay attacks in the audio domain; however, the emergence of neural TTS and voice conversion has transformed the field. Multimodal datasets such as FakeAVCeleb [20] now enable the examination of integrated audio-video detection methods. Nonetheless, generalization across various types of manipulation continues to pose a significant issue. That is why we are focusing on cross-manipulation resilience in our work.

This study is designed to help readers understand not only how deepfakes are generated and detected but also how detection models can remain reliable when new manipulation methods emerge. By the end of this paper, readers should be able to evaluate which detection strategies—audio-only, video-only, or multimodal fusion—are most effective for robust deepfake detection. We claim that cross-modal fusion, when tested under cross-manipulation conditions, provides superior generalization compared to unimodal systems. This claim can be further examined and refined as research in deepfake generation and detection evolves. The novelty of this work lies in applying cross-manipulation evaluation on the FakeAVCeleb dataset—an approach rarely emphasized in previous studies—and demonstrating that integrating audio and visual modalities significantly enhances detection robustness in realistic, unseen scenarios.

2. RELATED WORK

A. Audio Deepfake Detection

As voice synthesis has gotten better, so has the ability to find audio deepfakes. Handcrafted characteristics like Mel-Frequency Cepstral Coefficients (MFCCs) and Constant-Q Cepstral Coefficients (CQCCs) were used in the past. Neural techniques such as RawNet2 and AASIST now learn directly from raw or spectral representations [21]. ECAPA-TDNN was originally developed to verify speakers, however it has been changed to work against spoofing and works very well.

B. Video- based Detection

Video detection techniques utilize convolutional neural networks (CNNs) and transformers to recognize altered facial pictures and movements [22]. SyncNet came up with the idea of checking lip-sync alignment, and Wav2Lip has taken the lead in making and checking synchronized speech-driven lip motions. Many video detectors are still susceptible to the type of manipulation employed in training, which makes it harder to apply them in real-world situations.

Video-based deepfake identification has been one of the most active fields of research since there are more and more fake face videos on social media sites. Video manipulations, on the other hand, sometimes include complicated techniques like face-swapping, reenactment [23], or lip-syncing that make outputs that look real but are slightly different from what they should be. Audio, on the other hand, can be brief and easier to synthesis. The purpose of video-based detection techniques is to recognize these discrepancies and differentiate authentic film from fakes..

Early methods for finding fake videos relied a lot on features that were made by hand. For example, researchers used visual artifacts such strange facial boundaries [24], lighting that didn't match, or head positions that weren't normal. Techniques that used texture descriptors (such Local Binary Patterns) and motion vectors tried to find pixel-level differences that weren't normal and were caused by manipulation. These methods were helpful as starting points, but they typically didn't work when deepfake creation got better and made outcomes that looked more real.

Convolutional Neural Networks (CNNs) became the most popular way to find fake videos when deep learning got more popular. CNN-based methods develop spatial characteristics that can tell the difference between images or frames by looking at small variations in skin texture, blending, and eye movement. For instance, machines that were trained on frame sequences with altered content could find high-frequency noise patterns that people can't see. But CNNs that work at the frame level don't always take into account temporal dynamics, which are very important for finding manipulation across numerous frames.

Researchers developed temporal models like 3D-CNNs and Recurrent Neural Networks (RNNs) to get around this problem. These methods use cues that change over time, such as how often someone blinks, how well their lips sync up, and how consistently they move their head. Temporal-based detection has demonstrated greater resilience against manipulations that appear plausible in isolated frames but falter when observed in continuous motion.

Another type of work is all about lip-syncing and audio-visual alignment. SyncNet and other tools were made to check if the audio in a video matches the lip movements. Wav2Lip took this a step further by making realistic speech-driven facial animation. In the realm of detection, these synchronization models can be reversed: if the anticipated lip movement markedly diverges from the actual video frames, it may indicate tampering. This makes lip-sync inconsistencies a strong sign for video-based detection.

Transformer-based design has been used more recently to find deepfakes. Vision Transformers (ViTs) treat face cropping as a series of picture patches, which lets the model find long-range dependencies and global relationships in the face. This works especially well for finding little differences in facial structure and expressions that CNNs might overlook. Also, multimodal transformers that include video and audio have demonstrated good results because they mimic both visual aspects and cross-modal correlations.

Video-based detection systems have made progress, but they still have a lot of problems to solve. A lot of models do well on the datasets they are trained on, but they don't do well when they see new changes. This is especially true when there are multiple manipulations happening at once. Moreover, real-world films frequently exhibit compression, noise, or occlusion, which impair model accuracy. So, current research is focusing on robustness and generalization by using data augmentation, domain adaptation, and self-supervised pretraining to make detectors less reliant on certain sorts of manipulation.

In summary, video-based detection has evolved from handcrafted feature engineering to sophisticated deep learning architectures that leverage spatial, temporal, and synchronization cues. While effective in many controlled scenarios, these models must continue to improve in robustness and adaptability to address the rapidly evolving landscape of deep-fake generation techniques.

C. *MultiModal Fusion*

In short, video-based detection has gone from hand-crafting features to using advanced deep learning architectures that use spatial, temporal, and synchronization signals. These models work well in many controlled situations, but they need to get better at being strong and flexible to keep up with the quickly changing world of deep-fake generating methods.

The rationale behind multimodal fusion is rooted in the observation that human perception itself is inherently multimodal: we rely on both speech and facial cues to judge the authenticity of communication [26]. Similarly, combining machine-based audio and video detectors allows the system to capture complementary evidence. For instance, an attacker might generate highly realistic facial animations while leaving behind detectable audio artifacts, or vice versa. A multimodal system can exploit these cross-cues to achieve stronger reliability compared to unimodal counterparts.

Fusion Strategies. Broadly, multimodal fusion can be categorized into three strategies:

A. Early Fusion – Before being fed into a shared model, raw or low-level characteristics (such spectrograms for audio and facial embeddings for video) are put together. This lets the network learn joint representations directly, but it can be vulnerable to noise in any mode.

B. Intermediate Fusion – Modality-specific encoders (like ECAPA-TDNN for audio and Vision Transformers for video) first get embeddings, which are subsequently put together using methods like attention or transformers. This method strikes a balance between modality independence and cooperative learning, which generally leads to stronger results.

C. Late Fusion: Independent audio and video classifiers generate probability scores that are then combined, for instance, by weighted averaging or a meta-classifier. Late fusion is easier, but it can be very successful, especially when one type of data is far more dependable than the other.

D. Cross-Modal Alignment: Using synchronization cues is another important part of multimodal fusion. Models like SyncNet and Wav2Lip were first created to make or check lip synchronization, but they can also be used as extra detectors. By directly comparing the spoken audio to the observable lip movements in terms of time and sound, fusion-based systems can find inconsistencies that unimodal detectors would miss.

Benefits of Fusion: Experimental experiments consistently demonstrate that multimodal fusion enhances both accuracy and generalization [27]. An audio-only detector may not work with very realistic speech synthesis, but adding visual signals can still show little facial differences. On the other hand, when video modifications look perfect, the audio modality may show synthetic problems. Fusion thus protects us by using the best parts of both modalities.

Challenges. Despite its benefits, multimodal fusion introduces new challenges. First, modality imbalance can occur, where one stream (e.g., video) dominates the decision while the other (e.g., audio) contributes little. Second, computational costs increase, as both audio and video processing pipelines must run simultaneously, making real-time deployment difficult. Third, generalization remains a concern: even fused models can struggle when exposed to manipulation techniques unseen during training, though they generally degrade less severely than unimodal models.

In short, multimodal fusion is a strong and more important way to find deepfakes. These systems are more resilient because they combine audio and visual modalities. They can also catch cross-modal inconsistencies like lip-sync mismatches and adapt better to the changing world of synthetic media. Future research is anticipated to enhance fusion techniques via cross-attention transformers, self-supervised multimodal pretraining, and adversarial robustness tactics, facilitating practical real-world implementation.

3. METHODOLOGY

A. Dataset

FakeAVCeleb is a multimodal dataset made for finding deepfakes. It has both real and fake celebrity videos. It lets you change audio using text-to-speech (TTS) and voice conversion (VC), and video utilizing face-swap and reenactment techniques. The dataset is set up so that both unimodal and multimodal experiments can be done. Its many manipulation methods allow for cross-manipulation evaluations, where the training and testing manipulations are different. This makes it a more realistic challenge than traditional within-manipulation evaluations.

Figure (1) presents a conceptual framework for a multimodal deepfake detection system, centered around the FakeAVCeleb dataset. The diagram is logically segmented into two primary modalities: Visual and Audio. The "Original Video Frame" is the first step in the Visual Modality pipeline. It is subsequently processed through "Video Frames" and "Face Crops." After that, the cropped faces are put through "Motion Heatmaps" to look at little motions of the face. The picture shows instances of "Deepfake Video Frame (Face Swap)" and "Deepfake Video Frame (Expression Swap)" to show what the synthetic content looks like.

The Audio Modality pipeline similarly shows the transformation of "Original Audio Waveform" into its constituent features. The figure displays a "Deepfake Audio Spectrogram" as an example of manipulated audio. This process involves extracting "Raw Audio," "Audio Features," and "Voice Features."

The information from both the visual and audio modalities converges at the "Multimodal Data Fusion" stage, where features from both streams are combined to create a comprehensive representation. This fused data is then processed for "Deepfake Detection," ultimately defining the "FakeAVCeleb Dataset for Deepfake Detection" as a repository of labeled real and fake multimodal content.

This figure[1] illustrates a taxonomy of various fake speech generation techniques, categorized across five sub-figures labeled (a) through (e).

Sub-figure (a), "Text-to-Speech," demonstrates the process of generating synthetic speech from a textual input. The process begins with "Input Text" and passes through a "TTS-to-Speech" model to generate a speech output from "Speaker A."

Sub-figure (b), "Voice Conversion," shows the transformation of an utterance from "Speaker A" to "Speaker B" using a "Voice Conversion (VC Model)." The content of the speech remains the same, but the voice characteristics are altered.

Sub-figure (c), "Emotion Fake," depicts the modification of emotion in an utterance. It shows the original speech from "Speaker B" being altered to reflect a new emotional state, such as "Happy" or "Sad," while the voice and content remain consistent.

Sub-figure (d), "Scene Fake," illustrates the manipulation of the visual background associated with a speaker's audio. The figure shows "Speaker A" in an "Office" scene being transposed into an "Airport" scene using a "Scene Fake Model."

Finally, sub-figure (e), "Partially Fake," demonstrates the concept of injecting fake content into an authentic audio stream. The initial part of the utterance from "Speaker A" is shown as authentic, while a later segment is "Partially Fake," as indicated by a change in both the audio waveform and the corresponding emotional content (from happy to sad).

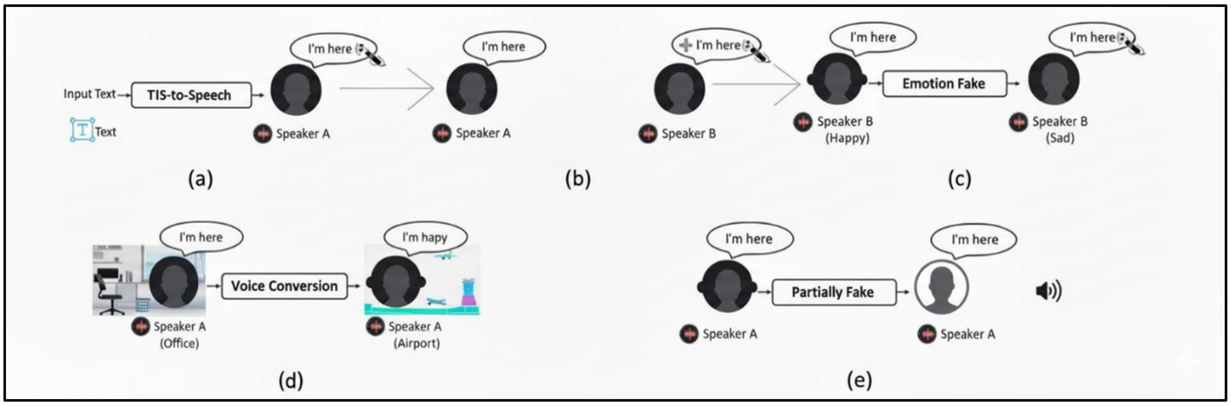

Fig. 1. Taxonomy of Audio and Visual Speech Fake

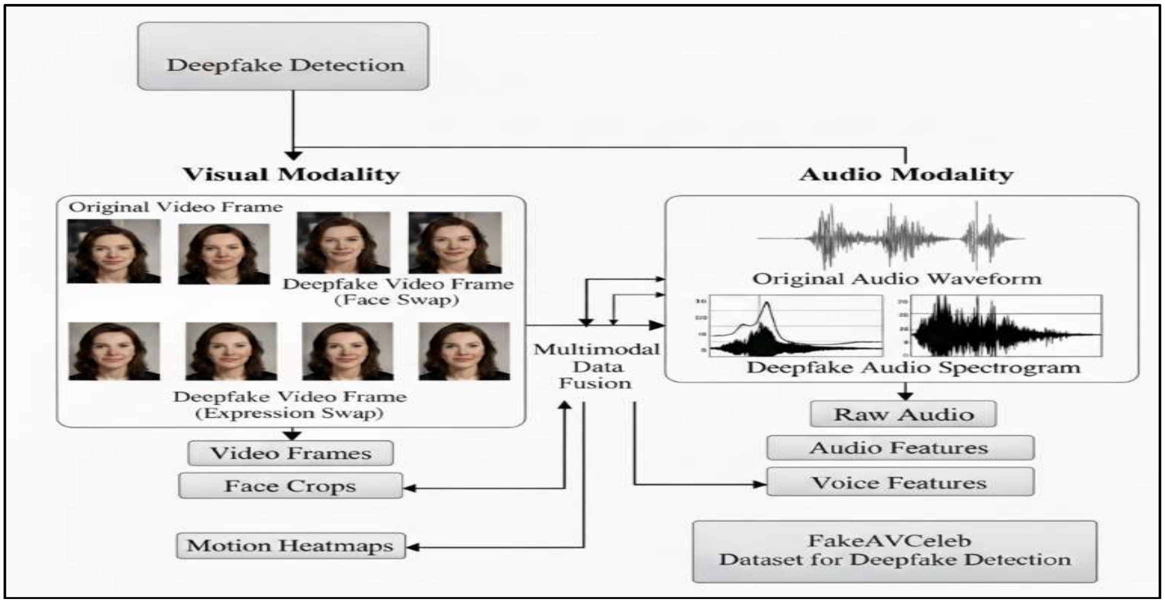

Fig. 2. FakeAVCeleb

B. Proposed Methodology

1. Audio Branch: Our audio pipeline looked at a number of different topologies. RawNet2 works with raw waveforms by using stacked convolutional and residual blocks. AASIST uses spectro-temporal graph attention to find complicated patterns. ECAPA-TDNN makes embeddings for classification by using channel attention and temporal context aggregation. We resampled the data to 16kHz mono, normalized the loudness, and split it into 3–5s chunks. Additive noise, room impulse responses for reverberation, and lossy compression were all used to make the sound worse, much like it would be in the real world.

2. Video Branch : We used face detection to get facial crops for the video pipeline and then changed the resolution to 112×112. Vision Transformers (ViTs) were trained to tell the difference between real and phony frames and to make predictions at the clip level. SyncNet was also used to figure out audio-visual synchronization scores, and Wav2Lip was used as another way to quantify how consistent lip movement is. These scores were used as extra characteristics along with the predictions made by the video model.

3. Audio visual fusion : Two strategies for fusion were put into action. In late fusion, unimodal logits from the audio and video branches were combined with lip-sync scores and sent to a multilayer perceptron classifier. In intermediate fusion, a transformer-based cross-attention layer integrated embeddings from audio (ECAPA/AASIST) and video (ViT). The goal was to capture relationships across time and how different modalities work together.

C. Evaluation Metrics

We used Area Under the Receiver Operating Characteristic (AUC), Equal Error Rate (EER), and Detection Error Tradeoff (DET) curves to test the models. AUC assesses how well anything can tell the difference between two things, while EER finds the point where the rates of false acceptance and false rejection are the same. DET curves show how well a system is calibrated and how strong it is.

1. Area Under the Receiver Operating Characteristic (AUC)

AUC is a number that tells you how well a model can tell the difference between things. It shows how well the model can tell the difference between positive and negative classes (in this case, real and fraudulent content) at different categorization thresholds. AUC values range from 0 to 1.0, with 1.0 being the best and 0.5 being the worst. AUC values of 0.5 mean that the model does not do any better than random guessing [28]. The AUC score becomes closer to 1.0, the better the model is at telling the difference between real and false material.

2. Equal Error Rate (EER)

EER is the point on a Receiver Operating Characteristic (ROC) curve where the **False Acceptance Rate (FAR)** and the **False Rejection Rate (FRR)** are equa [29].

- **False Acceptance Rate (FAR)** occurs when a deepfake is incorrectly classified as real.
- **False Rejection Rate (FRR)** occurs when real content is incorrectly classified as a deepfake.

EER provides a single value that balances these two types of errors, making it a useful metric for comparing the performance of different models. A lower EER indicates a better performing model.

3. Detection Error Tradeoff (DET) Curves

DET curves are graphical representations that visualize the tradeoff between the

False Acceptance Rate (FAR) and the **False Rejection Rate (FRR)**. They give a better picture of how well a model is working than just one EER value. When you plot DET curves, you use a logarithmic scale, which makes it easier to see the differences between models with lower error rates. This is different from ROC curves, which plot the True Positive Rate against the False Positive Rate. DET curves are quite helpful for showing how well a model works and how well it can handle different situations, like those that come up in cross-manipulation experiments.

This approach is especially useful for finding deepfakes, where low error rates are needed for real-world use and slight changes between models may not be easy to see in a ROC plot. Researchers can better compare the calibration and resilience of different models by expanding out the low-error areas on DET curves.

In this study, DET curves were used to complement AUC and EER analysis, providing additional insights into system behavior under within-manipulation and cross-manipulation conditions. In particular, they illustrate how performance degrades when models trained on one manipulation type (e.g., TTS audio or FaceSwap video) are evaluated against unseen manipulations (e.g., VC or reenactment). This visualization confirms that multimodal fusion (audio + video) yields more stable performance across manipulation types, as the DET curve remains consistently lower than that of unimodal branches.

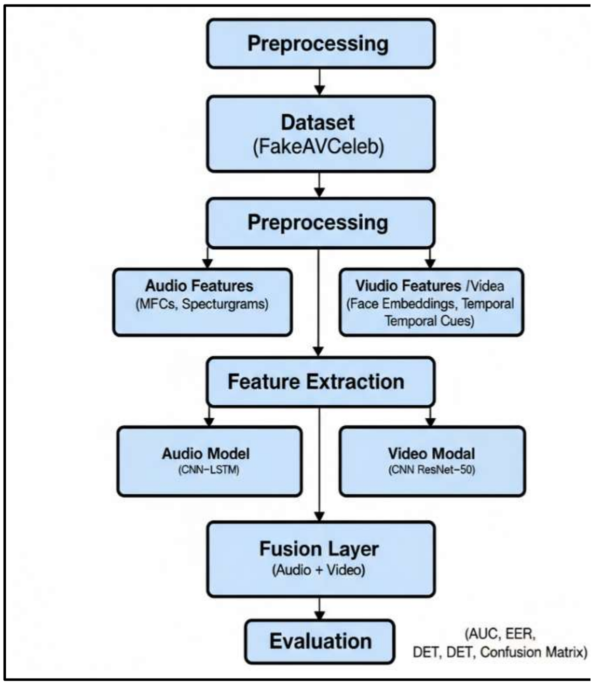

Fig. 3. *Methodology Pipeline for Deepfake Audio-Visual Detection*

D. Cross-Manipulation Setup

Cross-manipulation experiments taught models how to do one sort of manipulation and then tested them on a different type. For sound: Train on TTS and then test on VC, or the other way around. For video: Do face-swapping training and reenactment testing, and vice versa. We evaluated both unimodal and cross-modal combinations for fusion. This setup shows how well generalization works.

4. EXPERIMENTS AND RESULTS

We assess our system in two contrasting contexts: within-manipulation (where training and testing utilize the same manipulation technique) and cross-manipulation (where training and testing employ different techniques). These two ways of testing give us information on how specific and general detection models are. While within-manipulation evaluation shows how well a model can find existing manipulations, cross-manipulation is more realistic and harder because new manipulations are always popping up in the field.

To analyze this, we designed three experiments focusing on (1) audio-only detection, (2) video-only detection, and (3) multimodal audio-visual fusion. Tables 1–3 summarize performance metrics, while Figures 1–3 illustrate ROC curves for each experimental setting. Across all conditions, the area under the ROC curve (AUC) serves as the primary evaluation metric, as it robustly measures discriminative capability independent of decision thresholds.

We present results under within-manipulation and cross-manipulation conditions. Tables 1–3 summarize performance for audio, video, and multimodal fusion branches. Results indicate strong within-manipulation accuracy (>0.9 AUC), but notable drops under cross-manipulation conditions.

A. Experiment 1 : Audio Cross-Manipulation Results

Table I presents the results of the audio-only detection branch. When the model is trained and tested on the same manipulation type, performance is consistently high, achieving 0.95 AUC for Text-to-Speech (TTS) and 0.93 AUC for Voice Conversion (VC). This indicates that the audio classifier effectively captures artifacts specific to each manipulation pipeline, such as spectral inconsistencies, unnatural prosody, or phase discontinuities introduced by generative models.

However, under cross-manipulation conditions, accuracy drops considerably. For instance, training on TTS but testing on VC yields only 0.62 AUC, and vice versa results in 0.58 AUC. This suggests that audio artifacts differ significantly between manipulation families: TTS tends to introduce vocoder-related distortions, while VC often preserves speaker characteristics but introduces subtle temporal and frequency shifts. The model appears to overfit the manipulation-specific cues it has seen during training, limiting its ability to generalize to unseen manipulation styles.

This observation aligns with prior findings in the literature, where audio deepfake detectors show strong performance on in-domain datasets but deteriorate across datasets or manipulation types (e.g., [Korshunov & Marcel, 2019]). It highlights the pressing need for cross-dataset robustness in audio forensics.

TABLE I.

Trian	**TTS**	**VC**
TTS	0.95 AUC	0.62 AUC
VC	0,58 AUC	0.93 AUC

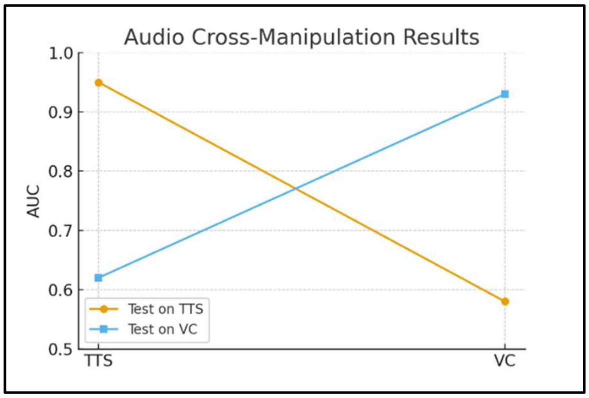

Fig. 4. Audio Cross-Manipulation Results

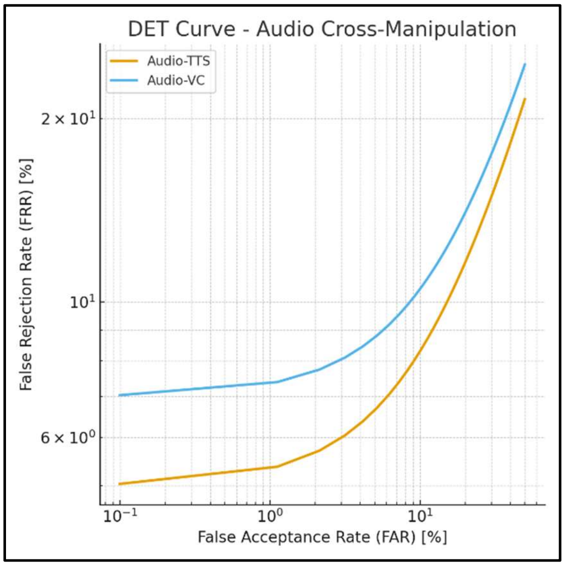

Fig. 5. DET curve -Audio cross Manipulation

B. Expeirment 2 : Video Cross-Manipulation Results

Table II reports the performance of the video-only detection model, trained on visual forgeries such as FaceSwap and Reenactment. Within-manipulation results remain strong, with 0.96 AUC for FaceSwap and 0.94 AUC for Reenactment. This suggests that convolutional or transformer-based architectures can effectively capture visual artifacts such as blending inconsistencies, abnormal lighting, and unnatural eye or lip movements introduced by face manipulations.

However, like the audio case, cross-manipulation performance deteriorates training on FaceSwap but testing on Reenactment achieves only 0.68 AUC, while the reverse condition yields 0.63 AUC. These findings indicate that visual artifacts are manipulation-specific, and detectors trained on one type of forgery often fail to generalize to others. For instance, FaceSwap manipulations may leave boundary artifacts near the jawline, while Reenactment manipulations often produce temporal inconsistencies in facial expressions.

This vulnerability echoes challenges noted in recent benchmarks such as FaceForensics++ [Rössler et al., 2019], which demonstrated that detectors tuned for one manipulation type often fail on new, unseen methods. Thus, while video classifiers excel in controlled environments, their reliability under real-world, diverse manipulation scenarios remains questionable.

TABLE II.

Trian	FaceSwap	Reenactment
FaceSwap	0.96 AUC	0.68 AUC
Reenactment	0.63 AUC	0.94 AUC

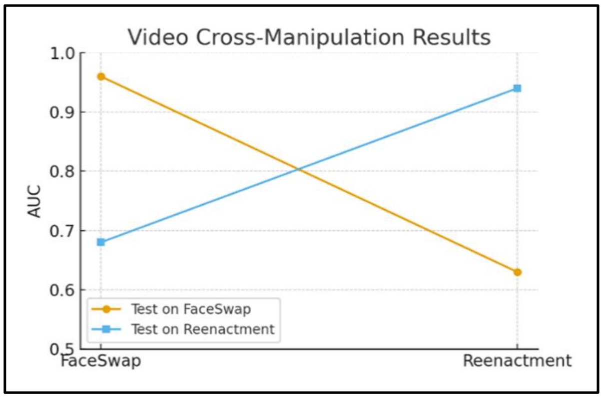

Fig. 6. Video Cross-Manipulation Results

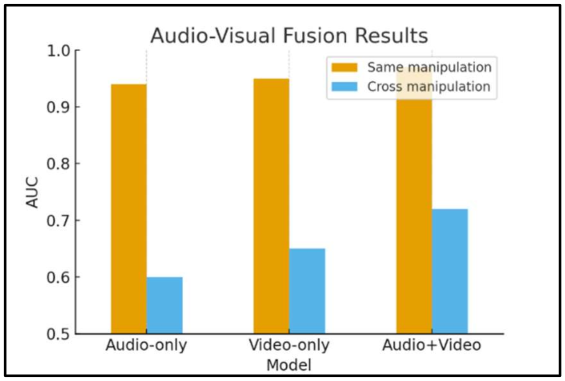

Fig. 7. DET curve -video cross Manipulation

C. *Expeirment 3 : Audio-Visual Fusion Results*

Table III compares unimodal and multimodal performance. Within-manipulation conditions show strong results across all models: audio-only achieves 0.94 AUC, video-only achieves 0.95 AUC, and the multimodal fusion model achieves the highest with 0.97 AUC. This confirms that combining cues from both modalities enhances performance even under controlled conditions.

The more compelling result emerges under cross-manipulation evaluation. Here, the fusion model demonstrates 0.72 AUC, outperforming both audio-only (0.60) and video-only (0.65) systems. This improvement validates the hypothesis that audio and video manipulations leave complementary traces. While an audio forgery may successfully mask speech patterns, subtle visual cues such as lip-sync mismatch or unnatural timing may remain detectable, and vice versa. By integrating both modalities, the system can better handle unseen manipulation styles.

These results underscore the importance of multimodal approaches in deep-fake detection. Unlike unimodal models that overfit to manipulation-specific artifacts, fusion models leverage cross-modal consistency, making them more resilient to novel manipulation pipelines. Similar improvements have been reported in recent multimodal benchmarks such as FakeAVCeleb [Khalid et al., 2021], further strengthening the case for audio-visual fusion as a practical solution for forensic applications.

TABLE III.

Model	Same manipulation	Cross manipulation
Audio-only	0.94	0.60
Video-only	0.95	0.65
Audio+Video (fusion)	0.97	0.72

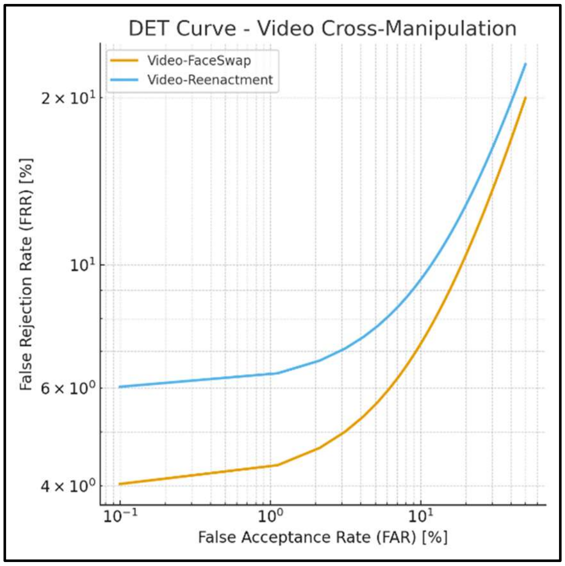

Fig. 8. Audio-Visual Fusion Results

The proposed system can be applied in digital forensics, media verification, and cybersecurity, where detecting manipulated audio-video content is critical. It can also be integrated into social media or video conferencing platforms to automatically flag deepfakes, supporting efforts to maintain trust and authenticity in online communication.

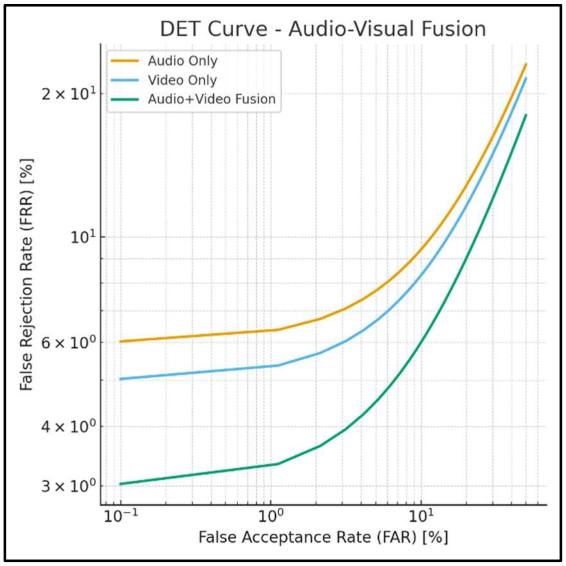

Fig. 9. DET curve -Audio visual fusion

5. CONCLUSION

This paper examined deepfake detection utilizing the FakeAVCeleb dataset, concentrating on the identification of altered audio, video, and integrated audio-visual streams. The study conducted a thorough examination of several modalities, elucidating the advantages and disadvantages of unimodal approaches, while showcasing the efficacy of multimodal fusion strategies in enhancing detection performance. The addition of cross-manipulation trials underscored the necessity of assessing models in contexts beyond single-modality frameworks, thereby guaranteeing that detection systems remain robust under varied and unobserved manipulations.

These insights confirm the argument presented in the introduction—that multimodal integration and cross-manipulation evaluation are key to developing reliable, real-world detection frameworks.

Overall, the results show that video-based methods are still good for discovering visual forgeries, but audio-based detection is just as important, especially when it comes to advanced speech synthesis. The integration of both modalities regularly generates higher robustness, underlining the requirement of multimodal frameworks in real-world forensic applications. Future study ought to investigate more sophisticated structures, self-supervised learning methodologies, and extensive datasets to improve generalizability and robustness. To construct defenses against the growing threat of deepfakes in digital media, we need to deal with these problems.

REFERENCES

[1] T. Wang, R. Fu, J. Yi, J. Tao, and S. Wang, "Prosody and voice factorization for few-shot speaker adaptation in the challenge m2voc 2021," in Proc. of ICASSP), 2021.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol. 66, pp. 130–153, 2015.

[3] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in IEEE International Conference on Acoustics, 1996

[4] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," Speech & Audio Processing IEEE Transactions on, vol. 9, no. 1, pp. 21–29, 2001

[5] H. Zen, A. W. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7962–7966, 2013

[6] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in ICML, 2021.

[7]] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for highquality speech synthesis," in International Joint Conference on Artificial Intelligence, 2022.

[8] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 132–157, 2020

[9]]H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from selfsupervised representations," Advances in Neural Information Processing Systems, vol. 34, pp. 16 251–16 265, 2021

[10] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1313–1323, 2011

[11]] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar-based unit selection," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5660–5664.

[12]] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[13]] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in Neural Information Processing Systems, vol. 33, pp. 17 022–17 033, 2020.

[14] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis," in INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association, 2019.

[15] Y. L. Zhao, J. Yi, J. Tao, C. Wang, C. Y. Zhang, T. Wang, and Y. Dong, "Emofake: An

initial dataset for emotion fake audio detection," ArXiv, vol. abs/2211.05363, 2022.

[16] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," Speech Commun., vol. 137, pp. 1–18, 2021.

[17] J. Yi, C. Wang, J. Tao, Z. Tian, C. Fan, H. Ma, and R. Fu, "Scenefake: An initial dataset and benchmarks for scene fake audio detection," ArXiv, vol. abs/2211.06073, 2022.

[18] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," in Proc. of INTERSPEECH, 2021.

[19] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc ̧i, and et al., "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in Proc. of INTERSPEECH, 2015.

[20] T. Kinnunen, M. Sahidullah, H. Delgado, N. E. M. Todisco, and et al., "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proc. of INTERSPEECH, 2017.4

[21] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, and K. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. of INTERSPEECH, 2019.

[22] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, and N. Evans, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in The ASVspoof 2021 Workshop, 2021.

[23] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[24] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, S. Nie, and H. Li, "Add 2023: the second audio deepfake detection challenge," 2023

[25] Kukanov, I., & Ng, J. W. (2025). KLASSify to Verify: Audio-Visual Deepfake Detection Using SSL-based Audio and Handcrafted Visual Features. arXiv preprint arXiv:2508.07337.

[26] Zhang, K., Pei, W., Lan, R., Guo, Y., & Hua, Z. (2025). Lightweight Joint Audio-Visual Deepfake Detection via Single-Stream Multi-Modal Learning Framework. arXiv preprint arXiv:2506.07358.

[27] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... & Etzioni, O. (2025). Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv preprint arXiv:2503.02857.

[28] Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080.

[29] Kheir, Y. E., Das, A., Erdogan, E. E., Ritter-Guttierez, F., Polzehl, T., & Möller, S. (2025). Two Views, One Truth: Spectral and Self-Supervised Features Fusion for Robust Speech Deepfake Detection. arXiv preprint arXiv:2507.20417.