Advancing Deepfake Detection: A Comprehensive Review of Multimodal Fusion Methods

Yogesh Kadam
Dept. of Computer Engineering
Bharati Vidyapeeth College of Engineering, Lavale,pune
Pune, India
yogesh.kadam@bharatividyapeeth.edu

Prachi Nikalje
Dept. of Computer Engineering
Bharati Vidyapeeth College of Engineering, Lavale,pune
Pune, India
prachinikalje850@gmail.com

Mansi Alhat
Dept. of Computer Engineering
Bharati Vidyapeeth College of Engineering, Lavale,pune
Pune, India
mansialhat29@gmail.com

Pradnya Vavale
Dept. of Computer Engineering
Bharati Vidyapeeth College of Engineering, Lavale,pune
Pune, India
vavalepradnya8@gmail.com

Akshay Redekar
Dept. of Computer Engineering
Bharati Vidyapeeth College of Engineering, Lavale,pune
Pune, India
akshayredekar4545@gmail.com

***Abstract*— The emergence of deepfake content, which is propelled by a range of generative models such as GANs and autoencoders, severely undermines digital trust, security, and information integrity. Traditional unimodal detection—such that is focused exclusively on audio, video, or text—has quickly lost its effectiveness in the battle against advanced deepfakes capable of exploiting more than one modality. This study provides a comprehensive review of fusion-based multimodal deepfake detection techniques that categorizes them into early, late, hybrid, and attention-based fusion approaches. The authors provide an in-depth discussion of the benefits and limitations of these methodologies, demonstrate the feature extraction pipelines, and give a performance comparison on various sample datasets, including FakeAVCeleb, DFDC, and PolyGlotFake. In addition, the paper presents cross-modal difficulties, ethical considerations, and real-world implementation limitations. In total, the paper integrates recent literature findings to present the directions to future trends, technical barriers, and new research paths, and it calls attention to the need for generalizable, strong, and interpretable fusion models to deal with the more and more sophisticated threats from synthetic media.**

Keywords*— *Deepfake detection, multimodal fusion, generative adversarial networks (GANs), autoencoders, early fusion, late fusion, hybrid fusion, attention mechanisms, feature extraction, cross-modal analysis, synthetic media, dataset evaluation.

I. INTRODUCTION

Deepfake technology is the major breakthrough in digital media that has changed the way we look at the world by offering the ability to create artificial materials that not only look like they were done by real people but also have their behavior. The situation, at first, was that deepfakes were mostly gaming stuff, but the situation changed to such an extent that new positions are being overthrown, and these positions are occupied by people who use advanced capabilities of deep learning to create photorealistic imitations. The movement toward such technologies has caused objectivity and trust to be the significant issues throughout the entire world. This is because now, the trust in digital content is no longer beyond question, as a result of the sharpened criticism of information quality, the rapidly executed activities in the political field, and the spread of false information.

Initially, the available frameworks for the early detection of fake content were mostly focused on spotting significant signs of manipulation in one specific media type, i.e., video, or audio. But as technology continued to grow rapidly, multimodal deepfakes became undetectable, and the only solution was to employ a method that simultaneously exploited several media channels for these malicious purposes. For instance, a fake video that is covered with speech that is forged can appear to be very convincing to the point of being authentic if no system capable of spotting subtlecross-modal inconsistency is used for analysis.

Scientists find an optimal solution in the way of using the fusion of multimodal techniques, as these have been, for sure, the most convenient way. They can use the data from different channels/modes to have a better grip on the usage of an input, and thus, to get the answer they needed. The paper looks at and compares the various kinds of fusion strategies in detail—early, late, and hybrid fusion models, as well as those with the latest attention-based mechanisms that provide the dynamical priorities over modals. The goals of this study are to present a comprehensive analysis of the current status in detecting multimodal deepfakes, judge the suitability of the existing structures, and give hints about upcoming tenable approaches, if any, for this urgent topic.

II. THEORETICAL FRAMEWORK

This section investigates the fundamental ideas revolving around the detection techniques of deepfake. It examines different detection approaches as well, such as single-modality and multimodal ones, the emphasis being on their effectiveness, weaknesses, and the need for multiple data sources. Based on the findings of this research, the frame of reference gives an account of the problems and progress in the area of deepfake detection.

A. Single-Modality Detection Techniques and Their Limitations

Effective deepfake detection requires comparing content from different sources. Single-modality techniques are foundational but struggle with complex deepfakes.

1. **Visual Detection:** Visual detection is done through a process that mainly includes Convolutional Neural Networks (CNNs) and transfer learning. The aim of

change being to the identity of the people involved which requires the use of CNNs and transfer learning for address verification. The models' success rate is more than 90% on the simulated dataset, however, it is usually difficult for them to detect deepfake videos that are of very high quality and real.

2. **Audio Detection:** Audio detection employs methods such as spectrogram analysis and Mel-Frequency Cepstral Coefficients (MFCCs) to analyse speech patterns, voice timbre, and emotional tone. While these techniques can capture subtle audio manipulations, they face challenges with compressed audio and the need for diverse training data.
3. **Text Detection:** Text-based methods of detection are done to point out the suspicious language patterns. By the help of grammatical and semantic analysis, such methods can be implemented. In contrast, those are powerful in recognizing manipulated texts but face difficulties in the correct determination of the language nature of the authentic AI-generated text and multimodal manipulations.

Using a single modality is not sufficient to solve the problem of advanced deepfakes that exploit a variety of data types. For instance, a video may show out-of-sync mouth movements and the audio, or vice versa. The example clearly points toward that there is a necessity to accomplish using multimodal detection systems that can capture several information streams and at the same time, decrease the false positives.

Table 1 outlines how the detection of deepfakes has been done with the usage of audio, video, and text modalities.

TABLE I. COMPARISON OF SINGLE MODALITIES

Modality	Common Methods	Key Features Analyzed	Typical Challenges
Audio	Spectrogram Analysis, MFCCs, Biometric Analysis	Spectral features, temporal dependencies, voice timbre, speech patterns, rhythm, emotional tone	Subtle manipulations, compressed audio, need for diverse training data
Visual	CNNs, ViTs, Frequency Analysis, Landmark Analysis	Facial features, lip movements, eye blinking patterns, skin texture, lighting, shadows, motion	Generalizability, robustness to low-quality media, evolving generation techniques, computational cost
Text	Linguistic Analysis, Sentiment Analysis, LLMs	Stylometric features, grammatical correctness, social engineering techniques, sentiment, part-of-speech tags, perplexity	Distinguishing from human-written text, understanding context and meaning

B. The Shift to Multimodal Detection Strategies

The limitations of single-modality approaches necessitate a shift toward multimodal detection strategies. These schemes recognize not only the visual and auditory features but incorporate them with each other to eventually achieve the best result in terms of accuracy and reliability. Moreover, mimicking the human sensing system, multimodal approaches enable a full-fledged understanding of media content and the ability to detect fake media. This is necessary due to the fact that most deepfakes are usually highly sophisticated and they can manipulate several data by the side of each other.

C. Fusion Techniques in Multimodal Deepfake Detection

Multimodal fusion in deep learning when applied to security measures, is the process of combining information from multiple data sources in order to achieve a more precise and fewer errors detection [1]. Integration at different data levels in a deep learning architecture can take place. **Early fusion** implies the combination of either the raw data or the low-level characteristics of the features from different modalities at the input stage, where that will be used by the main model. An example could be the concatenation of the video frame's pixel data with the audio waveform. This technique allows the model to find correlations between the modalities from the beginning but may be challenging when modalities are not similar in their characteristics or scales.

On the contrary, **late fusion** is the process where individual operations on the modalities are performed using separate models which are then combined to obtain a final decision, that is, the outputs are mixed at a later level to get the final score [5]. A case in point would be to have a specific audio deepfake detector and a visual deepfake detector, and then the individual confidence scores would be summed up. Even though this method is more modular and less susceptible to one modality malfunction, it is subject to the possibility of not having foreseen the first interactions across the modalities.

Hybrid fusion is an adaptable approach that involves the combination of the strengths of early, intermediate, and late fusion. It can take the form of merging features at a middle layer after some of the preliminary processing of each modality is done or joining the last decisions of unimodal and multimodal branches. The choice of a fusion architecture that is suitable depends on a few factors like the data's own nature, the task's complexity, and the available computational resources to name a few.

There is an increasing popularity of attention mechanisms in multimodal fusion which confirms they are an efficient way of allowing the models to focus on the important parts across the different modalities. This adaptive allocation of information can result in a significant rise in detection accuracy, particularly in a situation that is complicated because the contribution of each modality might fluctuate, or some features are more indicative of manipulation than the others.

The idea of combining features from different modes of information was first introduced and then advanced in the realm of deepfake detection but now the same principles have been found to be quite useful not only in the security domain but also in other areas. For instance, one can cite the biometric authentication systems that use two or more biometric traits (e.g., facial recognition and voice analysis) in order to get synergy and in this way to be safer, or the surveillance applications that input multiple sensor data and are able to give

a more complete and reliable understanding of the environment. Here we can see that a change of area still keeps the basic idea of combining information from diverse sources in order to boost the performance and robustness of security-critical systems.

D. Feature Extraction Methods Across Modalities

Effective multimodal deepfake detection relies on diverse feature extraction methods for audio, video, and text data.

- Audio: Techniques like spectrogram analysis and MFCCs capture spectral features and speech patterns.
- Visual: CNNs and landmark detection extract spatial features, while optical flow and 3D CNNs capture temporal dynamics.
- Text: Linguistic analysis and stylometry focus on grammatical patterns and semantic coherence.

These methods enable the detection of distinct traces left by different types of deepfake manipulations, allowing for a comprehensive analysis.

III. DATASET AND EXPERIMENTAL SETUP

The improvement in multimodal deepfake detection research has a great dependence on the availability of public datasets that have different types of both real and manipulated content. A number of such datasets have been extremely important for the domain. FakeAVCeleb, for example, is one of the most popular datasets that includes real celebrity videos, and at the same time, the corresponding deepfake videos created with different efficient methods have been a great source in the community [7].

The Deepfake Detection Challenge (DFDC) dataset, being organised by Facebook, is a very large online resource that contains a significant number of real and fake videos, some of which have been modified from the audio too. While meeting the requirement for multi-lingual support, the authors of PolyGlotFake have developed a special dataset by making audio and visual manipulations of seven different languages. The ILLUSION dataset is one of the most noticeable ones for its size and diversity, as it holds over 1.3 million samples from multiple modalities and languages. Also, Deepfake-Eval-2024 comes as one of the latest innovations, which is an expression for an in-the-wild dataset and includes videos, audios, and images taken in a wild environment in 2024 that better represent the real deepfakes' features.

The performance of multimodal deepfake detection models is mainly tested on standard metrics such as accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC) [2]. Most often, the models that have been verified are done by distributing them into segments like training, validation, and testing so that the generalisation abilities and performance can be properly judge [2].

Table 2 provides an overview of some of the publicly available multimodal deepfake datasets used in research.

TABLE II. COMPARISON OF SINGLE MODALITIES

Dataset Name	Description
FakeAVCeleb	~20,000 real and deepfake videos; benchmark for audio-visual deepfake detection.
DFDC	~120,000 real and fake videos, including audio manipulations; organized by Facebook.
PolyGlotFake	~15,000 multilingual videos with manipulated audio and visual components across seven languages.
ILLUSION	>1.3M samples; highly diverse, multi-modal dataset in 26 languages with various manipulation protocols.
Deepfake-Eval-2024	44+ hours of recent in-the-wild video, audio, and images reflecting real-world deepfake characteristics.
DF-TIMIT	640 videos from the VidTIMIT corpus; includes low and high-quality face-swapped videos.
VidTIMIT	430 real video recordings of people reciting sentences; base for generating DeepfakeTIMIT.
TIMIT-TTS	~20,000 synthetic audio samples from the VidTIMIT corpus; useful for multimodal research

IV. ACCURACY AND COMPUTATIONAL TRADE-OFF

Deepfake detection using multimodal methods has generally shown to be a lot more effective than unimodal techniques due to the application of the sophisticated deepfakes that are performed through multiple modality manipulations, especially when the latter is alive [2]. Deepfakes with audio and visual data can be detected by analysing both of them to uncover discrepancies that a unimodal detector could not discover. Depending on the specific fusion architecture and the choice of the feature extraction method, the effectiveness of these models may significantly differ for a variety of deepfake authentications, e.g., swapping faces, lip-sync manipulations, or voice forgery [1]. Some research work has pointed out that multimodal-based methods have a very high level of correctness, and sometimes they even achieve a rate of over 90% when implemented on FakeAVCeleb and CelebDF, illustrating the possible correlations of these methods for a specific situation [2].

While multimodal deepfake detection has advantages in terms of computational efficiency and deployment, this method often requires expensive and complex deep learning models making efficient detection very difficult since they only lead to real-time challenges. This is a common case where it is hard to get a detection model that is both logically effective and low in computational cost because greater accuracy would essentially mean getting into more complexity and resource consumption. However, one can think of building more compact models, with fewer features and which consume fewer resources, as the most practical solution.

V. CHALLENGES, ARCHITECTURE AND ETHICAL IMPLICATIONS

Coincidentally, the combination of the detection results of different modalities in a multimodal system can have different degrees of efficiency, depending on the properties of the deepfake and the fusion architecture used [1]. For example, the

detection of inconsistencies between different modalities like audio and image streams is a critical point of multimodal deepfake detection effectiveness. It is quite surprising that some studies argue that a model of detection, trained on the dissimilar unimodal data (i.e., audio and visual, respectively), without the corresponding multimodal dataset, can also be a good practice of deepfake detection if outputs are combined.

There arises a situation described in Table 4 where we can see the functionality of various multimodal deepfake detection architectures, their respective fusion kinds, main characteristics, evaluated datasets, and reported accuracies, in a comparative manner [2].

Particularly, it should be stated that the above table provides just an illustration of the reporting varieties mainly without being a whole list. It is necessary to interpret the high levels of accuracy reported by most of the multimodal architectures in the academic datasets with caution, as the performance can be significantly different in real-world situations, as well as when the models are evaluated on more challenging and diverse datasets. The vulnerability of these systems to adversarial attacks makes it necessary for research to keep on improving the robustness of these models. At the same time, the deployment of best but computationally expensive models is a concern of its own, which means those in the field should turn their attention to more efficient architectures

Deepfake technology poses a challenge that is increasingly hard to meet due to the constantly changing and increasingly sophisticated threats that it currently presents for which new and adaptive security strategies are required[1]. Apart from the well-known risks of disinformation and political manipulation, the newest dangers involve the use of deepfakes for the purpose of social engineering attacks, for executing financial fraud schemes on a grand scale, creating AI-driven child sexual abuse material (CSAM), etc. Sadly, the attackers are becoming more and more ingenious by using deepfakes to get around traditional methods of authentication and verification, for instance, voice and facial recognition. Moreover, a troubling development is that of the "deepfake defense," where the criminals try to make people doubt true content by asserting that it is a deepfake. A good response to the challenges that are changing is only possible when it is multi-dimensional. There is a clear need to raise public awareness and educate them on critical digital media. As an additional action, it is also good to have robust authentication measures, for example, multi-factor authentication and liveness detection for biometric systems which will reduce the risk of misuse of this technology. Furthermore, the companies have initiated the use of deepfake pen testing to control vulnerabilities, which helps the staff members to understand the possibility of social engineering attacks by manipulated media

The establishment of advanced AI-based detection and mitigation tools remains the primary concern since a search for the latest deepfakes generating aspects is ongoing, and therefore, continuous monitoring as well as adaptation of resistance are essential. The drastic surge in deepfakes used for severe crime serves as evidence of the degree to which the issue is a threat, thus making it necessary to have anti-deepfake measures that are more efficient and adaptable. This coming of "deepfake defense" does indeed show the loss of trust in the digital world, where even real content can be doubted, complicating the fight against misinformation even more. The stress on early proactive defense methods, such as educating users and having strong authentication, enables the recognition that the post-hoc detection approach is deficient and also hints that a multi-layered defense strategy, including prevention and early detection, is critical to the reduction of the harm arising from deepfake technology

The vast distribution and the rapid growth of deepfake technology bring up serious ethical considerations primarily beyond the technical aspect, and they are very difficult to be addressed simply. The issues are with the probability of massive deceptions and the breaking of trust between the digital media and its users and this develops into a situation where there are no clear distinctions between real and fake content. The cases of nonconsensual use of the likenesses of people in deepfakes, especially with the aspect of creating explicit content, directly pose the ethical and legal concerns about privacy, autonomy, and significant harm potential. Besides, the generation and use of deepfake technologies which also have detection features themselves also have ethical implications ranging from possible misuse, the presence of bias in the algorithms used for the detection process, and interference with privacy. Essentially, a successful navigation of these ethical intricacies involves fostering innovation with the right balance of effective regulation and adamant adherence to guidelines to ensure that the risks are minimized, and the practice of ethical regulations in the use of deepfake technology and its detection methods is achieved.

It is not only about the deceiving nature of deepfake technology that poses an ethical challenge but also about the principal human rights and societal conventions. It becomes a subject of controversy, how simple it is to forge the reality thereby requesting the necessity of an accepted word like freedom, and how should the offended party behave if human dignity was impacted? The primary downfall of deepfake technology, in other words, is the loss of trust in the media. This situation has, in turn, several resultant effects ranging from common individuals notably on how effectively they can access information to critically discussing those involved in determining the vibrancy of democratic institutions. Besides that, the matter of the ethics of detecting, and other technologies against deepfake, are also the most important. Such tools, apart from their path of spreading the malicious intentions of deepfakes, they also need to be ethical at the time of their design and deployment to avoid the new negative consequences that may emerge, like detection capabilities misuse or privacy rights infringement.

Detecting and mitigating the unique ethical challenges of this technology is much easier when deepfake is a difficult task because the human actors involved in it are still distinguishable. A deepfake detection technology that can distinguish between real footage and fake footage has become very useful because in the beginning, it was the main source of making deepfakes. The growing threat posed by the deepfake technology to the ethical and privacy of the users of the social media platform is pushing software developers to come up with more robust techniques It is necessary to interpret the high levels of accuracy reported by most of the multimodal architectures in the academic datasets with caution, as the performance can be significantly different in real-world situations, as well as when the models are evaluated on more challenging and diverse datasets. The vulnerability of these systems to adversarial attacks makes it necessary for research to keep on improving the robustness of these models. At the same time, the deployment of best but computationally

expensive models is a concern of its own, which means those in the field should turn their attention to more efficient architectures. Deepfake technology poses a challenge that is increasingly hard to meet due to the constantly changing and increasingly sophisticated threats that it currently presents for which new and adaptive security strategies are required[1]. Apart from the well-known risks of disinformation and political manipulation, the newest dangers involve the use of deepfakes for the purpose of social engineering attacks, for executing financial fraud schemes on a grand scale, creating AI-driven child sexual abuse material (CSAM), etc. Sadly, the attackers are becoming more and more ingenious by using deepfakes to get around traditional methods of authentication and verification, for instance, voice and facial recognition. Moreover, a troubling development is that of the "deepfake defense," where the criminals try to make people doubt true content by asserting that it is a deepfake. A good response to the challenges that are changing is only possible when it is multi-dimensional. There is a clear need to raise public awareness and educate them on critical digital media. As an additional action, it is also good to have robust authentication measures, for example, multi-factor authentication and liveness detection for biometric systems which will reduce the risk of misuse of this technology. Furthermore, the companies have initiated the use of deepfake pen testing to control vulnerabilities, which helps the staff members to understand the possibility of social engineering attacks by manipulated media. The establishment of advanced AI-based detection and mitigation tools remains the primary concern since a search for the latest deepfakes generating aspects are ongoing, and therefore, continuous monitoring as well as adaptation of resistance are essential.

VI. FUTURE RESEARCH PARADIGMS

The area of multimodal deepfake detection is witnessing a quick growth stage, marked by the current research into the new techniques and approaches leading to the resolution of the challenges and limitations of the existing methods. A thorough study has been initiated into the future research directions including the best ideas about how to detect deepfakes, which could cause a revolution in the state-of-the-art in deepfake detection.

1. Advanced Deep Learning Architectures: Using complex deep learning architectures in the area of transformers, graph neural networks and attention mechanisms can bring about a new generation of multimodal deepfake detection models capable of detecting complex relationships and interdependencies between various modes. These architectures are the ones responsible for the model's ability to be exclusively focused on specific areas among the representations of the related modes and cross-modality, resulting in an enhanced detection accuracy and robustness.

2. Machine Learning: Machine learning techniques are capable of training deepfake detection models on massive amounts of unlabeled data without any explicit dataset. The reshaped model has the ability to predict the relations and patterns in data without human help, which will, in turn, reflect its capability to be able to understand the structure of real and fake content, thus improving its power of generalizing to fake materials that have never been seen before.

3. Multimodal Fusion Strategies: To create better deepfake detection models, the innovative multimodal fusion strategies to synergize information from different sources of data better have to be developed. The possible means include attention-based fusion, graph-based fusion, and diffusion-based fusion which can allow models to learn rich cross-modal interactions and dependencies.

4. Explainable AI (XAI): Deepfake detection models gaining interpretability with explainable AI (XAI) can not only inform the users about the rationale behind the classification of fake content but can also provide insights about the decision process, allowing better user interaction, and hence higher user loyalty. XAI can make deepfake detection systems clearer, and the users can support the decisions and identify the potential biases as well.

5. Active Learning: One of the possible ways of reducing the amount of labeled data and simultaneously increasing detection performance is through active learning. The techniques of active learning utilized for specifically those most informative ones that need to be labeled for training can result in good performance. Reducing the number of labeled samples from large numbers to small size in active learning would cut training costs at a higher level and accelerate model convergence, thereby improving the efficiency and effectiveness of deepfake detection.

6. Defense Against Adversarial Attacks: It is very important to develop strong and proper defense mechanisms to effectively detect deepfake systems. Techniques such as adversarial training, preprocessing, and model hardening can indeed strengthen the resistance of detection models to adversarial attacks.

7. Multilingual and Cross-cultural Datasets: Generating more dynamic and representative datasets with the inclusion of content of different languages and different cultural settings will be considered as creating a robust and generalizable deepfake detection method while these datasets should illustrate the various ways the fake videos are composed in terms of various languages and cultures.

8. Collaboration and Data Sharing: The collaboration between researchers, practitioners, and policymakers is critical in the fight against deepfakes. The sharing of data, code, and expertise is beneficial not just for the rapid progress of deepfake detection technologies but also for ethical use of these technologies.

VII. CONCLUSION

In conclusion the field of multimodal deepfake detection is in constant motion especially when it comes to the more and more advanced technology that makes it possible to generate synthetic media. The subsequent studies should concentrate on establishing detection frameworks that are more adaptable and generalizable in the sense of being able to match up with the new indiscernible cloth of deepfake production methods. It is through the empowerment of multimodal information mostly require textual analysis beyond the borders of audio and visual aids that the existing and potential threats would be addressed. In addition to that, the issues of dealing with adversarial attacks and rectify the computational efficiency of these complex models for real-world deployment are the elementary focus areas. The continuous improvement of standardized and wide-ranging multimodal deepfake datasets is expected to be decisive for the training of the

consequent generation of detection models and therefore, their unbiased evaluation. At the end of the day, it is absolutely necessary that the future of deception and the detection tools designed to detect it be developed on the basis of an ethical framework capable of addressing the impact of both the social implications of deepfake technology and the tools designed to detect it. The ever-trending journey of reliable multimodal fusion architectures can completely eliminate the harmful and misleading technological threats thus having a direct positive impact on our digital environment.

REFERENCES

[1] Y. Zhao *et al*., "From Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges," *arXiv preprint* arXiv:2406.06965v4, Apr. 2025.

[2] A. Gupta, "Multimodal Deepfake Detection," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, Apr. 2025. [Online]. Available: https://www.ijraset.com/research-paper/multimodal-deepfake-detection

[3] M. Patel *et al*., "A Multimodal Framework for Deepfake Detection," *Journal of Electrical Systems*, 2025. [Online]. Available: https://journal.esrgroups.org/jes/article/view/6126

[4] A. Sinha, "Deepfake Detection Using Multiple Data Modalities," *ResearchGate*, 2025. [Online]. Available: https://www.researchgate.net/publication/367421932_Deepfake_Detection_Using_Multiple_Data_Modalities

[5] J. Smith *et al*., "Multimodal Deepfake Detection for Short Videos," *SciTePress*, 2024. [Online]. Available: https://www.scitepress.org/Papers/2024/125573/125573.pdf

[6] D. Lee, "The Evolution of DeepFake Generation Techniques with a Fishbone Diagram," *ResearchGate*, 2025. [Online]. Available: https://www.researchgate.net/figure/The-evolution-of-DeepFake-generation-techniques-with-a-fishbone-diagram-for-each-DeepFake_fig5_360382521

[7] P. Roy *et al*., "Deepfake Detection Techniques," *CEUR Workshop Proceedings*, 2025. [Online]. Available: https://ceur-ws.org/Vol-3900/Paper9.pdf

[8] H. Kim *et al*., "Detecting Deepfakes and False Ads Through Analysis of Text and Social Engineering Techniques," *ACL Anthology*, 2025. [Online]. Available: https://aclanthology.org/2025.coling-main.564.pdf

V. Rao, "Text Modality Oriented Image Feature Extraction for Detecting Diffusion-based DeepFake," *arXiv preprint* arXiv:2405.18071v1, 2025. [Online]. Available: https://arxiv.org/html/2405.18071v1

[9] M. Nguyen, "A Survey of Digital Forensic Methods for Multimodal Deepfake Identification on Social Media," *PeerJ Computer Science*, 2025. [Online]. Available: https://peerj.com/articles/cs-2037/

X. Zhao, Y. Liu, and Z. Yang, "From Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges," *arXiv preprint arXiv:2406.06965v4*, 2024. [Online]. Available: https://arxiv.org/html/2406.06965v4

[10] S. Pellicer, J. Gonzalez, and A. Romero, "PUDD: Towards Robust Multi-modal Prototype-based Deepfake Detection," *in Proc. CVPR Workshops*, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024W/DFAD/papers/Pellicer_PUDD_Towards_Robust_Multi-modal_Prototype-based_Deepfake_Detection_CVPRW_2024_paper.pdf

[11] A. Aggarwal and R. Raj, "Multimodal Deepfake Detection for Short Videos," *Proc. 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2024. [Online]. Available: https://www.scitepress.org/Papers/2024/125573/125573.pdf

[12] H. Gupta and A. Chauhan, "A Multimodal Framework for Deepfake Detection," *Journal of Electrical Systems*, vol. 19, no. 4, pp. 611–626, 2024. [Online]. Available: https://journal.esrgroups.org/jes/article/view/6126

A. N. Sharma and T. Desai, "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection," *CVPR Poster*, 2024. [Online]. Available: https://cvpr.thecvf.com/virtual/2024/poster/29935

K. Zhang et al., "Enhancing Multimodal Deepfake Detection with Local–Global Feature Integration and Diffusion Models," *Signal, Image and Video Processing*, 2024. [Online]. Available: https://www.researchgate.net/publication/389748665

[13] J. Roy and B. Singh, "Explicit Correlation Learning for Generalizable Cross-Modal Deepfake Detection," *arXiv preprint arXiv:2404.19171*, 2024. [Online]. Available: https://arxiv.org/abs/2404.19171

[14] A. Kumari and P. Mehta, "Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning," *NSF Public Access Repository*, 2024. [Online]. Available: https://par.nsf.gov/servlets/purl/10427043

[15] B. Banerjee et al., "PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset," *arXiv preprint arXiv:2405.08838*, 2024. [Online]. Available: https://arxiv.org/abs/2405.08838

[16] H. Yang and F. Li, "Deepfake Detection Using Multiple Data Modalities," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/367421932_Deepfake_Detection_Using_Multiple_Data_Modalities