International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Multimodal Deepfake Detection Frameworks:Survey

Chakrapani D S[1], Akshata T Rathod[1], Aqsah Sehreen[1], Dhanalakshmi S[1], Inchara Poovaiah A[1]
[1]Dept. of CSE., JNNCE Shivamogga, Visvesvaraya Technological University, Belagavi – 590018

***Abstract*—The rapid advancement of deepfake technology poses significant threats to information authenticity, identity protection, and societal trust. This paper presents a survey of multimodal deepfake detection frameworks with a particular emphasis on combining Efficient Temporal Modeling for Classification (ETMC) in video analysis and RawNet-based audio analysis. By merging temporal-spatial and acoustic cues, such frameworks achieve strong accuracy while remaining computationally practical. The survey explores unimodal and multimodal detection methods, discusses their strengths and limitations, and highlights the need for robust, lightweight, and real-time detection mechanisms for safeguarding digital media integrity.**

***Index Terms*—Deepfake detection, multimodal framework, video forensics, audio analysis, media security, Synthetic media, Deep learning, Forgery detection, Information integrity**

I. Introduction

The advancement of artificial intelligence, especially with generative adversarial networks (GANs), has accelerated the production of synthetic media, commonly known as deepfakes. Deepfakes convincingly modify visual or audio streams, seamlessly combining fabricated and real elements in a way that is often undetectable to human senses. This has given rise to serious threats in areas such as politics, finance, social media, and law enforcement. For example, manipulated videos of public figures can aid in spreading misinformation, while synthetic audio can be exploited for financial fraud or identity theft.

Initial detection techniques were limited to unimodal methods, examining either video or audio streams. Video-based detectors identify inconsistencies in facial expressions, blinking patterns, or head movements, whereas audio-based detectors analyze speech artifacts or frequency distortions. However, as deepfake generation has advanced, unimodal methods are increasingly vulnerable, with attackers designing fakes that specifically target and bypass such single-stream detectors.

Multimodal detection frameworks emerged as a stronger alternative. By simultaneously analyzing both video and audio streams, these methods capture inconsistencies across modalities. For example, mismatches between lip movements and spoken words can expose synthetic manipulations. Such cross-modal analysis significantly improves robustness against advanced deepfakes. This paper reviews the progression from unimodal to multimodal detection frameworks, presenting key works, methodologies, limitations, and future directions.

II. Literature Survey

This section outlines notable works in unimodal and multimodal deepfake detection. The subsections highlight the methodology, datasets, benefits, and limitations of the respective studies.

A. Deepfake Video Detection Based on Multi-Modal Learning

Zhang et al. [6] proposed a multimodal deepfake detection method that integrates visual and audio streams using a Modality Dissonance Score (MDS). Their framework extracts lip movement features from video sequences and compares them against acoustic features derived from speech. By measuring misalignment between modalities, the system boosts detection accuracy. The authors evaluated their model on the DeepFake Identification Challenge dataset,achieving an accuracy of 84.4%, outperforming unimodal approaches. The key advantage lies in its robustness against subtle manipulations where either visual or audio artifacts alone may not be sufficient for detection. However, the system's reliance on deep neural architectures demands high computation, necessitating GPUs and limiting real-time applicability. Performance may also decline for low-quality or compressed videos, underscoring the need for generalizable solutions.

B. Deepfake Audio Detection via MFCC Features

Hamza et al. [7]developed an audio-only detection system using Mel-Frequency Cepstral Coefficients (MFCCs) as primary features. MFCCs capture spectral properties of speech,which are often difficult for generative models to replicate. The authors trained machine learning classifiers, such as SVMs and Random Forests, to distinguish between genuine and synthetic speech. Experiments on audio datasets demonstrated strong performance against basic deepfake audio. The system is lightweight and computationally efficient, making it suitable for deployment on limited-resource devices. However, advanced models like WaveNet and Vocoder reduce the effectiveness of MFCC-only strategies, and the absence of cross-modal validation further limits robustness. The absence of multimodal correlation further restricts its robustness

C. Deepfake Detection for Human Face Images and Videos: A Survey

Malik et al. [4] conducted a comprehensive survey on facebased deepfake detection. Their work categorized existing approaches into feature-based, deep learning-based, and hybrid methods. The study compared datasets, including FaceForensics++, DFDC, and Celeb-DF, outlining strengths and weaknesses of detection methods. While not proposing a new algorithm, the study shed light on research challenges, dataset limitations, and vulnerabilities. The authors emphasized the value of dataset diversity for improving model generalization and it is evident that most models are vulnerable to adversarial attacks or cross-dataset testing. Their review highlights the ongoing need for multimodal approaches.

D. Deepfake Detection on Social Media Using FastText Embeddings

Sadiq et al. [16] expanded the scope of deepfake detection by targeting textual misinformation in social media. The authors proposed using FastText embeddings in combination with deep learning classifiers to distinguish between humanwritten and machine-generated tweets.Their model achieved promising results in identifying fake textual data, suggesting that deepfake threats extend beyond audio-visual content. The study highlighted the importance of applying detection strategies to social platforms where misinformation spreads rapidly.However,limitations include the relatively small dataset and lack of multilingual coverage, which restricts the generalization of results across different social contexts.

E. Fighting Deepfake by Exposing Convolutional Traces

Guarnera et al. [2] explored a unique detection method based on identifying convolutional traces left by generative models. They argued that GAN-generated images carry subtle artifacts resulting from convolutional operations in the generative process. By designing classifiers that detect these traces, the system successfully distinguished manipulated images from genuine ones. Their experiments on datasets such as FaceForensics++ demonstrated effectiveness, though performance degraded when images were compressed or heavily altered. While innovative, the method struggles under practical conditions where data quality varies.

F. Generative Adversarial Ensemble Learning for Face Forensics

Baek et al. [1] proposed an ensemble learning framework to detect deepfakes by combining multiple GAN-based detectors. The idea was to leverage the strengths of diverse models and reduce weaknesses of individual detectors. Their ensemble achieved higher robustness and accuracy compared to single models, with experiments confirming improved results on FaceForensics++ and DFDC datasets. However, the method expects high computational resources and longer training, which limits scalability and real-time application. This restricts real-time deployment and makes scalability a challenge for practical applications.

G. Hybrid GAN–ResNet Model for Fake Face Detection

Safwat et al. [5] introduced a hybrid deep learning method merging GANs with ResNet architectures. By leveraging GANs for feature learning and ResNet for classification, their system achieved nearly 97% accuracy in detecting manipulated images. This shows the potential of hybrid architectures for capturing both generative features and classification strengths. Nonetheless, its heavy computational demands restrict applicability in mobile or edge environments. The study also highlights the difficulty of ensuring robustness against unseen manipulations.

H. Multimodal Detection Using Fusion Architectures

Salvi et al. [15] designed a multimodal framework integrating EfficientNetB4 for video and x-vectors for audio. Their approach employed feature fusion strategies to combine modalities, significantly improving detection accuracy across datasets such as DFDC and FakeAVCeleb. The study demonstrated the superiority of multimodal frameworks over unimodal systems, especially when facing sophisticated manipulations. However, the requirement of large, labeled datasets and high training costs remain challenges. The authors emphasized the importance of fusion strategies in achieving state-of-the-art performance.

I. MMGANGuard for GAN-based Fake Image Detection

Raza et al. [9] presented MMGANGuard, a multi-model detection system combining ResNet and SVM classifiers to identify GAN-generated images. Their framework achieved accuracy rates exceeding 95%, showcasing resilience against different manipulation techniques. The study highlighted adaptability as one of the system's major strengths. However, limitations include model interpretability, as ensemble combinations make it difficult to understand decision-making processes which can affect trust and acceptance in real-world applications.

J. Unmasking Deepfake Voices with MFCC-GNB XtractNet

Gujjar et al. [28] focused on detecting synthetic voices using MFCC-GNB XtractNet, a specialized model for audio forgery detection. Their system combined MFCC features with Gaussian Na¨ıve Bayes classifiers to capture subtle artifacts in speech. Experiments demonstrated strong performance across datasets, though the model struggled with generalization when tested on novel or multilingual data. The authors concluded that while unimodal voice detection remains valuable, integrating it into multimodal frameworks would improve resilience against future attacks.

K. Multimodal Detection Using Multimodal Deep Learning

Lewis et al. [3] introduced a hybrid deep learning approach combining spatial, spectral, and temporal content to distinguish real from fake videos. The study shows that applying the Discrete Cosine Transform enhances detection by capturing spectral features from individual frames. Evaluated on the Facebook Deepfake Detection Challenging dataset, the multimodal network demonstrated an efficiency of 61.95%, demonstrating the advantage of modality integration.

L. Deepfake Recognition Using Diverse Gabor Filters

Khalifa et al. [13] addressed the receptive field–model size dilemma in convolutional neural networks by proposing a unified Gabor function capable of generating linear, elliptical, and circular Gabor filters. This approach allows the network to adaptively extract features at multiple orientations and scales,enhancing its ability to capture subtle textures and patterns indicative of deepfake manipulations. The authors integrated this unified Gabor function into a CNN to construct adaptive Gabor filters, and further designed a dual-scale large receptive field network, which balances the trade-off between capturing fine-grained local details and global contextual information.

The model excelled well across multiple datasets and forgery types, reducing complexity while maintaining accuracy. Multi-scale feature extraction with adaptive filters improved generalizability, making it suitable for real-time use.

M. An Improved Dense CNN Architecture for Deepfake Image Detection

Patel et al. [18] proposed an refined deep convolutional neural network (D-CNN) architecture specifically designed for deepfake image detection. The model emphasizes high generalizability by

training on images aggregated from multiple sources, thereby capturing diverse variations in manipulated content. The authors highlight that conventional detection techniques often fail to account for subtle inter-frame dissimilarities and fail to generalize across datasets. To overcome this, the improved D-CNN incorporates dense connectivity patterns that facilitate better feature propagation and mitigate the vanishing gradient problem. Extensive experiments were conducted on several benchmark datasets, including AttGAN, GDWCT, and StarGAN, where the model achieved accuracy rates of 98.33%, 99.33%, and 99.17%, respectively. Tested on datasets like AttGAN, GDWCT, and StarGAN, it achieved accuracy above 98%. The architecture proved effective at modeling fine-grained facial features. The architecture also shows efficiency in learning intricate representations of facial features and manipulation artifacts. Moreover, the model is scalable and can be adapted for large-scale image datasets, making it suitable for real-world applications. The study concludes that integrating dense connectivity and multi-source training significantly enhances detection performance. This approach provides a solid basis for further research in improving deepfake image detection frameworks and adapting them to emerging manipulation techniques.

N. Dual Attention Network Approaches to Face Forgery Video Detection

Luo et al. [12] proposed a Dual Attention Forgery Detection Network (DAFDN) for detecting subtle manipulations in fake videos. The DAFDN incorporates two specialized attention mechanisms: a spatial reduction attention block and a forgery feature attention module. The spatial reduction block efficiently compresses redundant spatial information while retaining critical regions indicative of tampering. The forgery feature attention module focuses on extracting unique artifacts resulting from image warping and subtle manipulations, enhancing the model's sensitivity to tampered regions. The study evaluated the network using the DFDC and FaceForensics++ datasets, achieving AUC scores of 0.911 and 0.945, respectively. These results indicate that DAFDN surpasses conventional approaches such as XceptionNet and EfficientNet in both accuracy and robustness. The architecture is capable of capturing both local and global inconsistencies, which are often overlooked by single-attention models. Additionally, the network demonstrates strong generalization across datasets with varying manipulation types. The study also highlights the importance of dual-attention strategies in improving forgery localization and detection reliability. This approach sets a new benchmark in deepfake video detection and offers a valuable direction for future research in multimodal attention-based architectures. The DAFDN framework can be extended to realtime video surveillance systems and other practical applications requiring high precision in detecting forged content.

O. Deep-Fake Detection Using Deep Learning

Nagashree et al. [31] introduced a deep learning-based framework for detecting manipulated videos by combining ResNeXt alongside Long Short-Term Memory networks. ResNeXt, an advanced extension of ResNet, was employed to capture detailed spatial features from individual frames, capturing subtle inconsistencies such as unnatural lighting, distorted facial regions, and irregular textures. To complement this, the sequential relationships between frames were modeled using an LSTM, which identified temporal inconsistencies including abnormal motion transitions, mismatched lip synchronization, and unnatural facial dynamics.

The hybrid ResNeXt–LSTM architecture was evaluated on benchmark datasets like FaceForensics++ and Celeb-DF. The results demonstrated improved correctness and stability compared to unimodal CNN-based approaches, particularly in identifying manipulations that appear realistic when analyzed frame by frame yet do not maintain temporal consistency across sequences. The authors stated that their model effectively distinguished genuine content from deepfakes, achieving higher precision and recall rates than conventional detection models.

A key strength of this approach lies in its ability to combine frame-level detail with sequence-level coherence, enabling reliable detection of sophisticated video manipulations. However, the authors noted challenges in terms of computational complexity, as training such deep architectures requires significant processing power and time.

III. Summary of Literature Survey

The review highlights a shift from unimodal toward multimodal detection frameworks. Video methods capture visual anomalies, while audio methods analyze speech inconsistencies. Advanced manipulations can bypass unimodal detectors, highlighting the importance of multimodal fusion. Challenges such as dataset diversity, computational cost, and real-time adaptability remain. Future research should focus on lightweight yet robust models, better generalization across unseen deepfakes, and integration of temporal and contextual information. Explainable detection models and adversarial training can further improve reliability. Ultimately, effective detection will depend on balancing accuracy, adaptability, and efficiency.Additionally, collaboration between academia and industry can accelerate the development of standardized benchmarks and evaluation protocols.

TABLE I: Summary of Techniques used in Deepfake Detection

Reference / Author	Method / Approach	Key Features / Techniques	Findings / Inference
Yutong Zhang	Deepfake Video Detection Based on Multi-Modal Deep Learning	Uses audio-visual cues; modality dissonance score (MDS); multi-modal approach	Utilizes both audio and visual info; improves accuracy; demonstrated 84.4% accuracy on DFDC dataset
Asad Mali K	DeepFake Detection Survey	Reviews existing methods, datasets, challenges; categorizes detection methods	Comprehensive overview of deepfake detection techniques; highlights challenges and gaps in the field
S. H. Raut	Audio Deepfake Detection Using MFCC	MFCC feature extraction; machine learning classification	MFCCs effective for audio analysis; achieves 90%+ accuracy; suitable for real-time use
Divya Arora	Deepfake Detection on Social Media Using Deep Learning	FastText embeddings; CNN model; deep learning classifier	Detects machine-generated social media content; 98.6% accuracy; potential for filtering
Yuezun Wang	Fighting Deepfake by Exposing the Convolutional Traces on Images	Analyzes convolutional traces in images; GAN-generated artifact detection	Detects GAN images; robust vs postprocessing; outperforms visual detectors
A. Samantaray	Generative Adversarial Ensemble Learning for Face Forensics	Ensemble of GAN detectors; adversarial training; face forensics	Strong generalization and accuracy; robust ensemble-based detection
Mohamed Elshaer	Hybrid Deep Learning Model Based on GAN and ResNet for Detecting Fake Faces	Combines GANs and ResNet; hybrid deep learning architecture	Leverages GAN and ResNet strengths; achieves ~97% accuracy
Davide Salvi et al.	A Robust Approach to Multimodal Deepfake Detection	Multimodal framework using video and audio; EfficientNetB4; x-vectors from SpeechBrain	Robust detection of unseen deepfakes; evaluated on DFDC, FakeAVCeleb, DeepfakeTIMIT; high accuracy
Shivam Srivastava	MMGANGuard: Detecting Fake Images Generated by GANs Using Multi-Model Techniques	Multi-model approach; GAN image analysis; ensemble detection	ResNet + SVM ensemble; 95%+ accuracy; robust and adaptive

IV. Recommendations and Potential Directions

Future projects need to build real-time, multimodal, lightweight frameworks for compatibility with mobile and edge devices. Explainable models of Artificial Intelligence can help User trust and transparency. Broadening benchmark datasets Across modalities, manipulatory types, and languages will increase model generalizability. Researchers must also investigate integration of multimodal detection and social network analysis to counter misinformation campaigns more effectively. Joint initiatives from academia, industry, and policymakers radars are necessary for ethical use and mass-scale adoption

V. Conclusion

Deepfake technology remains advancing, threatening the integrity of digital content. While unimodal methods are still valuable, they cannot compete with sophisticated manipulations. Multimodal detection models provide higher robustness by taking advantage of cross-modal correlations between audio and video. The move towards efficient, explainable, and scalable models is necessary to challenge the increasing level of sophistication of deepfakes. Innovation and cooperation will need to continue to ensure trust in digital communication.

References

[1] J.-Y. Baek, Y.-S. Yoo, and S.-H. Bae, "Generative Adversarial Ensemble Learning for Face Forensics," *IEEE Access*, vol. 8, pp. 118253–118263, 2020.
[2] L. Guarnera, O. Giudice, and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.
[3] J. K. Lewis et al., "Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning," in *Proc. CVPR Workshops*, pp. 1425–1434, 2020.
[4] A. Malik et al., "DeepFake Detection for Human Face Images and Videos: A Survey," *IJERT*, vol. 9, no. 5, pp. 95–102, 2020.
[5] S. Safwat et al., "Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces," *IJRTE*, vol. 8, no. 6, pp. 1479–1483, 2020.
[6] Y. Zhang et al., "A Deepfake Video Detection Method Based on MultiModal Deep Learning," *IJERT*, vol. 10, no. 6, pp. 125–129, 2021.
[7] A. Hamza et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning," *IJERT*, vol. 10, no. 7, pp. 367–371, 2021.
[8] S. Nailwal et al., "Deepfake Detection: A Multi-Algorithmic and MultiModal Approach for Robust Detection and Analysis," *IJERT*, vol. 10, no. 8, pp. 335–339, 2021.
[9] S. A. Raza et al., "MMGANGuard: A Robust Approach for Detecting Fake Images Generated by GANs Using Multi-Model Techniques," *IJERT*, vol. 9, no. 7, pp. 525–530, 2021.
[10] Z. Zhao et al., "Learning Dual Attention-Based Semi-Supervised Framework for Deepfake Detection," in *Proc. AAAI*, 2021.
[11] T. Nguyen et al., "Deep Learning for Deepfakes Creation and Detection: A Survey," *IEEE Access*, vol. 9, pp. 145361–145379, 2021.
[12] Y.-X. Luo and J.-L. Chen, "Dual Attention Network Approaches to Face Forgery Video Detection," *Electronics*, vol. 13, no. 126, pp. 1–15, 2022.
[13] A. H. Khalifa et al., "Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, pp. 1–15, 2022.
[14] S. Ahmed et al., "Speaker identification model based on deep neural networks," *Iraqi J. Comput. Sci. Math.*, vol. 3, no. 1, pp. 108–114, Jan. 2022.
[15] D. Salvi et al., "A Robust Approach to Multimodal Deepfake Detection," *Electronics*, vol. 13, no. 1, p. 95, 2023.
[16] S. Sadiq et al., "Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets," *IJERT*, vol. 11, no. 3, pp. 207–213, 2023.
[17] Y. Patel et al., "Deepfake Generation and Detection: Case Study and Challenges," *Computers & Security*, vol. 128, p. 102446, 2023.
[18] Y. Patel et al., "An Improved Dense CNN Architecture for Deepfake Image Detection," *IJERT*, vol. 10, no. 6, pp. 147–151, 2023.
[19] T. Min-Jen and C.-T. Chang, "Convolutional Neural Network for Detecting Deepfake Palmprint Images," *Computers*, vol. 13, no. 31, pp. 1–12, 2023.
[20] R. D. Ghodke and A. N. Pise, "A Deep Learning Approach to Deepfake Detection Based on Frame-by-Frame Analysis," *IJERT*, vol. 11, no. 4, pp. 388–392, 2023.
[21] H. Gupta and D. Agarwal, "Attention-Based Bi-Modal Deepfake Detector Using Facial Expressions and Voice Modulations," *IJERT*, vol. 11, no. 5, pp. 401–406, 2023.
[22] P. Sharma and A. Saini, "Multimodal Detection Using Audio-Visual Temporal Attention Network," *IJERT*, vol. 11, no. 6, pp. 423–428, 2023.
[23] M. Masood et al., "Deepfakes generation and detection: State-of-theart, open challenges, countermeasures, and way forward," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023.
[24] J. J. Bird and A. Lotfi, "CIFAKE: Image classification and explainable identification of AI-generated synthetic images," arXiv preprint arXiv:2303.14126, 2023.
[25] F. Alanazi et al., "Improving Detection of DeepFakes through Facial Region Analysis in Images," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 1–20, 2024.
[26] M. A. Arshed et al., "Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model," *Sensors*, vol. 24, no. 4, pp. 1–18, 2024.
[27] G. Gupta et al., "A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 2023–2045, 2024.
[28] M. U. T. Gujjar et al., "Unmasking the Fake: Machine Learning Approach for Deepfake Voice Detection," *Applied Sciences*, vol. 14, no. 1, pp. 1–17, 2024.
[29] R. Kumari et al., "Detection of AI-Generated Images from Various Generators Using Gated Expert CNN," *IJERT*, vol. 11, no. 1, pp. 110–115, 2024.
[30] L. Pham et al., "Deepfake Audio Detection Using Spectrogram-Based Feature and Ensemble of Deep Learning Models," in *Proc. IEEE 5th Int. Symp. Internet Sounds (IS)*, Oct. 2024.
[31] N. K. T. Nagashree, S. Shristi, S. Firdaushi, S. B. Patil, and S. Singh, "Deep-Fake Detection Using Deep Learning," *International Journal of Innovative Science and Research Technology (IJISRT)*, vol. 10, no. 1, pp. 1700–1706, Jan. 2025. doi:10.5281/zenodo.14808073.