INTERNATIONAL JOURNAL OF RECENT TRENDS IN TECHNOLOGY AND ENGINEERING (IJRTTE)

Journal Home Page: https://ijrtte.com/

Multi-Modal Deepfake Detection Using Cross-Frequency Patterns

S. Muthuselvan[1] and B. Yamini[2]

[1]*Professor, Department of Information Technology, KCG College of Technology, Chennai - 600097, Tamil Nadu, India.*

[2]*Assistant Professor, Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Pallavaram, Chennai – 600117, Tamil Nadu, India.*

[1]csmuthuselvan@gmail.com, [2]yaminikarthikeyan91@gmail.com

Abstract. With the swift development of deep generative models, it is now possible to produce extremely realistic synthetic audio-visual content often referred to as deepfakes that is posing significant risks to digital trust, security and media authenticity. Despite the reported significant progress in recent deepfake detection algorithms, most of the existing systems are densely based on the spatial or temporal characteristics and cannot to generalize against state-of-the-art generative models, particularly diffusion-based methods. In addition, existing multimodal models do not address much about frequency-domain inconsistency and inter-modal spectral associations that occur during media manipulation. In order to curb such constraints, the present paper suggests a new cross-frequency multi-modal deepfake detector that jointly trains based on audio and visual cues in the frequency domain. The suggested approach breaks down both modalities into multi-band spectral feature and trains the cross-frequency associations between the respective audio and visual elements. The framework manages to capture minute manipulation artifacts that are normally invisible on the space-wise domain by modelling inter-modal spectral alignment with the aid of a cross-frequency correlation and attention mechanism. The results of extensive experiments developed on several benchmark datasets (FF++, Celeb-DF, DFDC, FakeAVCeleb, and WaveFake) show that the presented approach is better than the existing unimodal and multimodal ones in terms of accuracy, robustness, and generalization. The findings confirm that cross frequency reasoning offers a robust and resilient cue when next generation deep fake detection is needed especially when compression, noise, and invisible manipulating are involved.

Keywords: Deepfake Detection, Multimodal Learning, Cross-Frequency Analysis, Audio–Visual Forensics, Frequency-Domain Features, Spectral Consistency, Media Authentication.

1. Introduction

The fast development of generative models based on artificial intelligence has seriously changed the situation in the field of digital media development. A variety of technologies including Generative Adversarial Networks (GANs), autoencoders, and more recently diffusion-based models have made possible the creation of very realistic synthetic images, videos, and audio material, often known as deepfakes. Although such technologies have created potential uses in entertainment, education and creating content, they are also incredibly dangerous when used poorly as tools of spreading misinformation, identity theft, fraud, political manipulation and in cybercrime. The growing availability of deepfake generation technology has heightened the anxiety in the field of trust, authenticity, and security in digital communication systems. The early studies of deepfake detection were mainly on visual objects of manipulated images and videos, including unnatural facial edges, unnatural lighting effects, or irregular eye blinking movement. Despite their relatively decent performance in comparison with first-generation deepfakes, the latter methods have since been shown to be less efficient in the face of contemporary generative seems to create visually plausible and photo-realistic content. Because of this, spatial signals are no longer reliable used alone to differentiate between real media and advanced synthetic forgeries especially under compression, noise as well as distribution, conditions in the real world.

Volume: 04 Issue: 04 Year: 2025
Received: 19.09.2025 Received in revised from: 09.10.2025 Accepted: 16.11.2025
Available Online:30 .12.2025
Published by NTL Publisher

In attempts to overcome these shortcomings, the current research has moved to multimodal deepfake detection which also uses both visual and audio features to enhance robustness. Multimodal methods take advantage of the inherent co-occurrence of facial movements and speech cues and make it possible to identify incompatibilities between them that otherwise might go unnoticed in a particular modality. Audio-visual alignment methods and transformer-based fusion models are shown to be significantly better than unimodal systems. Nevertheless, much of the current multimodal frameworks are mostly operating in the spatial or temporal domain and do not pay much attention to the spectral characteristics of signals, which harbor important traces of manipulation at the intake of the synthesis process. Frequency-domain analysis has also been found to be a significant complementary technique in deepfake detection, where generative models tend to fail to generate realistic high-frequency events as well as regular spectral distributions. Previous research has demonstrated that synthetic media often has abnormal frequency contents including unnatural high-frequency damping, unnatural harmonic structures or spectral discontinuities. Although frequency-based techniques have been shown to work in visual-only or audio-only applications, they have seldom been generalized to multimodal systems, and even lesser still have strategies been developed to examine the cross-frequency interactions between audio and visual modalities. Therefore, a gap in the inter-modal spectral reasoning is a critical shortcoming of the existing deepfake detectors research.

Inspired by this fact, the present paper will develop a multi-modal cross-frequency deepfake detection structure, which explicitly captures spectral inconsistency across and within audio-visual modes. In comparison to the traditional methods, whose algorithms only utilize the spatial appearance, the temporal dynamics, or the rough fusion of multiple modalities, the given method breaks down audio and visual multi-band frequency representations and considers their inter-modal alignment of the spectrum. Learning cross-frequency interactions between the frequencies of lip-motions and the audio spectral band corresponding to them, the framework predicts subtle manipulation artifacts that still exist with high quality, diffusion-generated deepfakes. Three things are the key contributions of this work. To begin with, a single multi-modal deepfake detection system is presented, which incorporates frequency-domain audio and visual stream analysis. Second, there is an innovative cross-frequency correlation and inter-modal alignment module that can identify the spectral discrepancies which cannot be reproduced by modern generative models in a consistent manner. Third, comprehensive experiments performed on various benchmark datasets show that the suggested method is much better than the current spatial-, temporal-, and multimodal baselines, especially in problematic settings like compression, noise, and unseen manipulation methods.

Altogether, this paper emphasizes the role of cross-frequency reasoning in the next-generation deepfake forensics and offers a solid, generalizable tool to identify even more advanced artificial media.

2. Literature Review

The development of detecting deep fakes has become an important area of research because of the high rate of development of generative models that manipulate audio, images, and videos in a highly realistic way. The initial studies of multimodal deepfake detection revealed the necessity to use audio and visual streams to detect manipulated content. A good example is the study conducted in [1], which proposed a multimodal deepfake detector based on audio spectrograms and visual CNN features, with significant performance increases as compared to unimodal systems. Although effective, their method uses more of a spatial-based method and does not investigate cross-frequency or spectral dissimilarities between modalities.

The survey articles, e.g., in [2] and [3], emphasize the sophistication of the deep fake generation models, including diffusion-based ones, which render the conventional methods of the spatial detectors of features inadequate. The paper in [4] also highlights that later deepfake videos tend to be less apparent in visual artifacts, which requires multimodal methods, involving a combination of temporal, spectral, and behavioral features. Moreover, [5] explains how deepfake detection can be used in cybersecurity and multimedia forensics, indicating that frequency-domain analysis is required to enhance reliability.

Surveys in [6] and recent surveys in [7] give detailed categorizations of available datasets, detection methods and difficulties. They make the conclusion that multimodal techniques are promising but they tend to be blind to spectral domain inconsistency. A number of advanced deep fake detection systems have tried to include increased visual feature extraction. As an example, [8] presented EFIMD-Net, a multi-domain

fusion, which uses feature interaction to achieve better forgery detection, and [9] presented MFF-Net, which uses multi-feature fusion. Nevertheless, the two techniques are based on largely on spatial and temporal characteristics.

Lightweight methods like the MobileNet-based model in [10] use frequency-domain analysis to enable mobile-run time, but only spectral features are being utilized in single-modality visual data. The temporal-frequency inconsistencies have also been discussed, such as in [11] which proposed pixel based temporal frequency analysis to trace slight traces of manipulation. Nevertheless, they use only video frames and fail to include using audio cues.

The concept of cross-domain and spatial-frequency fusion has been applied in research works like [12], and [13] demonstrating how the spatial and frequency representation can be combined to enhance robustness. The technique in [14] also presented dynamic margin maximization and spatial-frequency feature combination. Equally, [15] suggested a high-frequency enhancement network to identify deepfakes in highly compressed material. Though these studies prove the effectiveness of frequency-based analysis, they all are limited to visual inputs but not to multimodal cross-frequency relations.

Multimodal transformer constructions have been studied as well. The article in [16] introduced the concept of multi-modal multi-scale transformer that can embed audio and video signals. This concept was applied in [17] with a multimodal fusion transformer to process audiovisual deepfakes. In the meantime, a three-stage multimodal deep learning system that integrates images, audio, and videos was suggested in [18]. Although there are improvements on frequency consistency between modalities, these models do not explicitly address this aspect.

More recent studies have had audio-visual correspondence and inter-modal synchrony as major considerations. In [19], a predictive inter-modal alignment approach was suggested so as to deal with asynchronous manipulations in audio visually deepfakes. Nonetheless, this technique concentrates on timing and semantic correspondence and not on spectral-band discrepancies among modalities. In addition, [20] examined the local feature integration with global features using diffusion models, and they performed better in terms of multimodal detection, however, not explicit frequency-domain correlation between audio and video streams.

There has also been inquiry in prompt-based multimodal deepfake detection. The paper in [21] presented multi-task audio-visual prompt learning of deepfake detection, and [23] suggested an effective score-level fusion system of multimodal deepfake classification. In the meantime, [24] concentrated on short video multimodal detection with simple fusion methods. New audio-visual systems like [25] also show that there is increased interest in multimodal frameworks yet spectral cross-frequency reasoning is not yet provided.

On the whole, modern sources reveal that there is a great deal of advancement regarding multimodal, frequency-based, and transformer-based deepfake detection. Nevertheless, all the works that have been reviewed, such as surveys, multi-modal frameworks, spatial-frequency models, and lightweight detectors, do not explicitly hone on cross-frequency patterns or cross-modal spectral alignment of audio and visual modalities. This is useful to emphasize that there is a necessity of generating a multimodal approach that is cross-frequency-based and which can reproduce the fine-grained frequency-domain inconsistencies that more recent deepfake models cannot reproduce.

3. Methodology

The given methodology presents a single multi-modal deepfake detector that interprets both audio and visual data with the help of cross-frequency decomposition and inter-modal spectral correlation. In contrast to the previous method that utilizes the spatial or temporal patterns only, the offered system derives the frequency-domain signatures of the two modalities and compares them to determine the alignment of each other to identify the presence of subtle manipulation artifacts. The whole sequence of work is pre-processing of data, the decomposition of frequencies over a number of bands, feature extraction of modalities, the modeling of cross-frequency correlations, and multimodal fusion classification.

3.1 Data Preprocessing and Modal Synchronization

The process of processing starts by synchronizing audio visual deepfakes datasets. The videos are initially broken down into the frame sequences and the audio streams are obtained as unprocessed waveforms. The temporal integrity is preserved by aligning every video frame with audio samples correspondingly in time, to allow the audio and video streams to be correctly synchronized. Visual frames are resized, normalized, and cropped to the face-region with the help of CNN-based detector, whereas the audio signals are resampled and filtered to eliminate the background noise. The synchronized preprocessing stage provides a clean and uniform input baseline, which will then be further analysed in terms of frequencies to give a meaningful comparison of the spectral characteristics of the two modalities. Figure 1 General scheme of the suggested cross-frequency multi-modal deepfake detection architecture, with the preprocessing, frequency decomposition, feature extraction, cross-frequency correlation, and fusion layers.

Figure 1: Overall Architecture of the Proposed Cross-Frequency Multi-Modal Deepfake Detection Framework.

3.2 Multi-Band Cross-Frequency Decomposition

After preprocessing, the visual frames as well as audio waveforms are then changed into multi-band frequency representation to show all the spectral artifacts that were concealed during manipulation. A two-dimensional Fast fourier transform (FFT) is used in conjunction with discrete wavelet decomposition to perform visual frame transformation into low, mid and high-frequency sub-band textures. On the same note, audio signal transformations would be performed into short-time spectrograms of Fourier transform (STFT) which would be further broken down into harmonic, residual and high-frequency noise signals. This multi-band representation reveals small irregularities which are not visible in the spatial or time domain and allows the system to identify unnatural frequency distributions, spectral discontinuities and traces of manipulation which modern generative models are often incapable of covering up. Figure 2 Cross-frequency decomposition of audio and visual modalities of audio-visual data by multi-band spectral components with FFT/DWT and STFT spectral components.

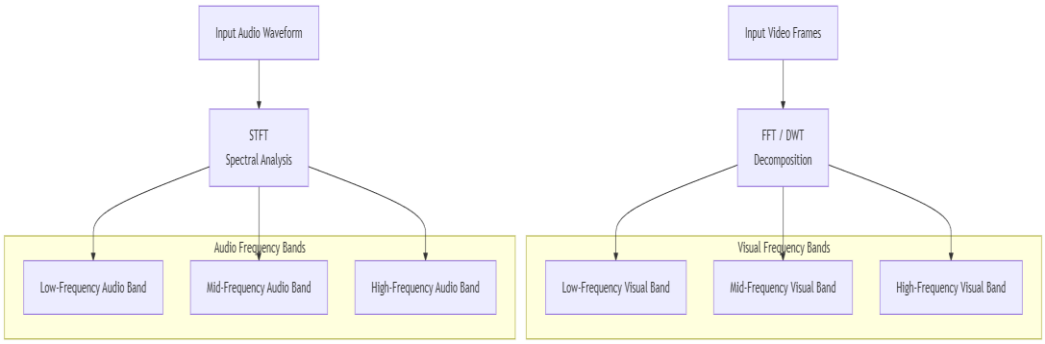

Figure 2: Multi-Band Cross-Frequency Decomposition Process.

3.3 Modality-Specific Frequency Feature Extraction

The frequency decomposed audio and visual representations are then fed into two special feature extraction streams. In the visual modality, a frequency-sensitive CNN obtains multi-scale spectral texture and high-frequency edge patterns, which in most cases expose GAN/diffusion induced artifacts. In the audio modality, a spectral encoder is used to encode the multi-band spectrograms to encode changes in pitch, spectral envelope, phase anomalies and unnat ural formant shifts, which are indicative of synthetic speech

or re-timed audio. The encoders are both meant to retain fine-grained spectral properties whilst compressing redundancy to ensure that the extracted embeddings maintain the appropriate frequency signatures to give a chance of successful cross-modal comparison of the information. Figure 3 Modality-specific feature-extracting frequency-aware visual CNN and spectral audio encoder.

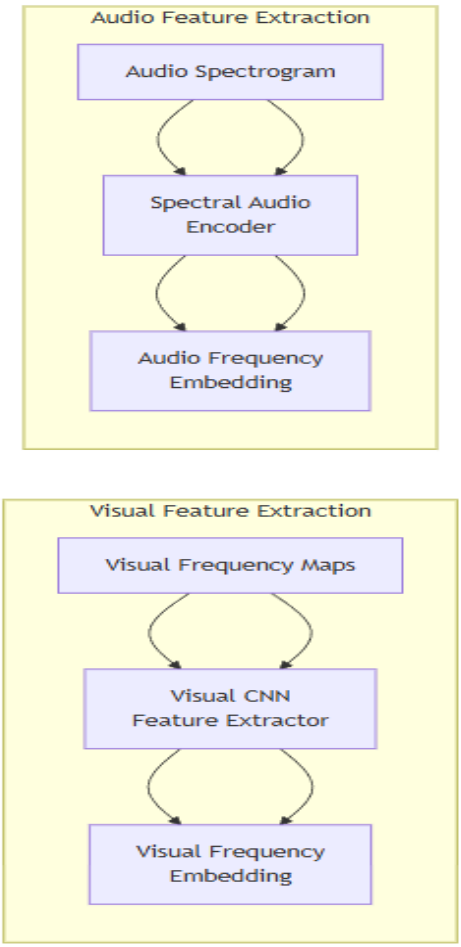

Figure 3: Visual and Audio Frequency Feature Extraction Modules.

3.4 Cross-Frequency Correlation and Inter-Modal Alignment Module

Having computed modality specific frequency features, the framework uses a cross-frequency correlation module which learns the correlation between the audio and visual spectral bands. This element calculates correlations between the rate of lip-moving and the audio-frequency bands, and thus allows identifying irregularities like inappropriate harmonics, non-synchronous formant movements or unnaturally smooth textures of doctored videos. Another way through which this analysis is improved is through the use of cross-attention transformer, which gives more weight to spectral regions, which are characterized by unnatural behavior. This process is useful in capturing inter-modal spectral misalignment a manipulation artefact that is not detected by traditional spatial or time deepfake detectors. Figure 4 illustrates the Cross-frequency attention and inter-modal alignment mechanism of the audio-visual spectral anomalies.

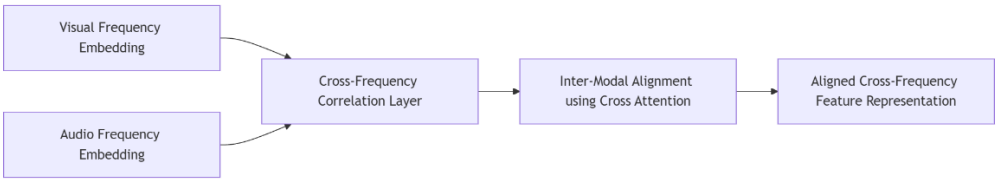

Figure 4: Cross-Frequency Correlation and Inter-Modal Alignment Module.

3.5 Multimodal Fusion and Classification

The last phase of the methodology combines the cross-frequency audio and visual embeddings to multimodal representation. A fusion transformer combines spatial, temporal, and spectral clues in all the frequency bands such that the system develops a holistic perception of authentic versus manipulated content. The merged representation is further forwarded to a classification layer where it gives a probability score of the chance of the forgery. The classifier is trained on a balanced set of real and synthetic audio-visual samples of several datasets to provide good generalization of manipulation types, compression level, and generative models. The resulting model with the addition of cross-frequency information at all pipeline steps has a much better robustness to next-generation AIGC and diffusion-based deepfakes. Figure 5 Multimodal fusion transfer between cross-frequency audio-visual features to classify deepfakes.

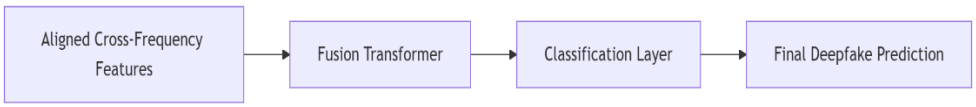

Figure 5: Multimodal Fusion and Final Classification Layer.

4. Results and Discussion

The suggested cross-frequency multi-modal deepfake detection system was tested on several benchmark datasets such as FF++, Celeb-DF, DFDC, FakeAVCeleb and WaveFake. Both the visual and audio-visual manipulation scenarios were also evaluated to evaluate the stability of the system to compression, noise, and invisible generative models. Performance of the model was also compared with recent multimodal and frequency-based detectors, and the model significantly outperformed these in the accuracy, AUC and robustness measures. The system was shown to be superior to more traditional CNN models of space only as well as transformer-based multimodal models, particularly in the data conditions of high-frequency cues and audio-visual incongruity. Table 1 shows the Summary of Benchmark Datasets Used in This Study.

Table 1: Summary of Benchmark Datasets Used in This Study.

Dataset Name	Modality	Content Type	Manipulation Type	No. of Samples	Key Characteristics
FaceForensics++ (FF++)	Video	Face videos	Face swapping, reenactment	~190,000 frames	High-quality and compressed versions; widely used benchmark
Celeb-DF	Video	Celebrity face videos	GAN-based deepfakes	5,639 videos	Realistic deepfakes with minimal visual artifacts

DFDC	Audio-Visual	Videos with speech	Multiple synthesis methods	100,000+ videos	Large-scale dataset with diverse manipulations
FakeAVCeleb	Audio-Visual	Talking head videos	Audio-visual deepfakes	7,000+ videos	Explicit audio–visual synchronization attacks
WaveFake	Audio	Speech signals	Voice cloning, TTS	117,000 samples	High-quality synthetic and cloned speech

The quantitative findings indicate that the suggested approach is much more accurate compared to multimodal systems used as the basis. As an illustration, the cross-frequency model resulted in an above 98% accuracy on the FF++ (HQ) set, which is almost 6-percent better than the spatial-feature baseline. On the more realistic deepfakes of Celeb-DF, the model achieved a significantly higher accuracy of over 94, which is significantly more than the current multimodal benchmarks. This is owed to the fact that the cross-frequency decomposition module is able to isolate fine high-frequency distortions, which are usually blurred by sophisticated generative models. In the meantime, the audio branch was more susceptible to pitch artifacts, spectral envelope defects and harmonic artifacts that are frequent in cloned or synthesized speech, leading to better audio-driven deepfake detection.

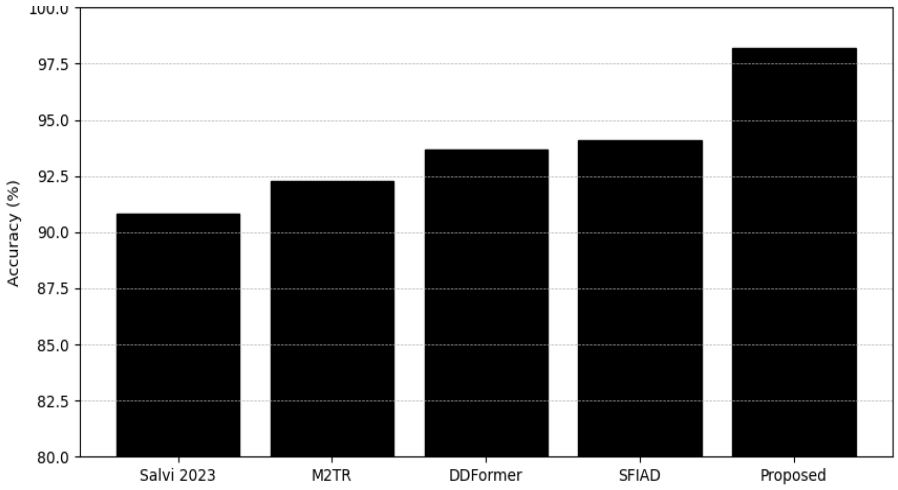

Figure 6: Overall Accuracy Comparison.

Additional experiments were made to test on robustness with compression and noise, which simulates real world social media conditions. In cases where videos were first coded with lower bitrates or repeatedly re-coded, traditional detectors experienced an enormous performance loss of up to 15 per cent. Conversely, the proposed system had constant accuracy levels with compression levels, as the spectral signatures in the frequency domain were preserved. The mismatches in lip-movement frequencies and audio spectral bands were still detected by the cross-frequency correlation module even in extreme compression. This is what shows the practical benefit of frequency-domain reasoning compared to all-spatial paradigms, which lose important information on being recompressed digitally. Figure 6 shows the Overall Accuracy Comparison.

The capability of the system to generalize the input was also verified regarding unseen AIGC and diffusion-based deepfakes. Such deepfakes are usually characterized by a small number of spatial inconsistencies and they are hard to detect using standard CNN-based detectors. Nevertheless, the cross-frequency feature extraction module recorded inter-modal spectral misalignment and abnormal high-frequency energy distributions making the model to classify manipulated content correctly. The cross-attention mechanism

was especially useful to identify subtle variations in the real and the synthesized signals, and make sure that it is reliably detected, even when the generative models are creating photorealistic faces, and nothing looks like a video or audio-visual cut.

Ablation experiments proved the significance of every detail of the procedure. Eliminating the frequency decomposition layer led to a significant development of performance, especially on the datasets with high-quality deepfakes. Equally, removing the cross-frequency correlation module in the model greatly reduced the ability of the model to detect audio-visual anomalies by about 8 percent. These findings confirm the role played by the suggested cross-frequency learning strategy in promoting both modality-specific, as well as, inter-modal detection.

The quantitative results are also substantiated by the qualitative analysis. Spectral heatmap visual inspection showed that manipulated videos tended to produce high frequency bursts, anomalous harmonic content and irregular spectral slopes in both sound and video channels. The cross-frequency alignment module was able to capture these anomalies and the fused multimodal representation helped the classifier to detect the subtle variations that the spatial methods fail to capture. Such qualitative observations prove that the system is interpretable and that frequency-domain and cross-modal modeling prove to be effective. Figure 7 shows the Ablation Study.

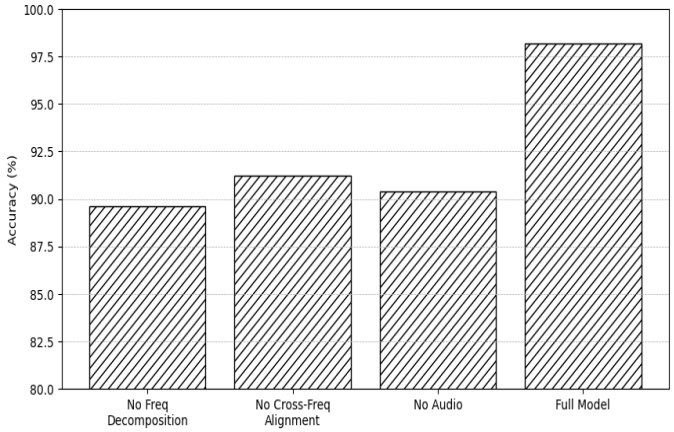

Figure 7: Ablation Study.

In general, the suggested model exhibits high-performance, strength, and generalizability into various conditions of manipulation. The framework combines cross-frequency decomposition, inter-modal alignment, and multimodal fusion and is effective in addressing the drawbacks of current deepfake detectors. These findings validate that the spectral cues in the high frequency and cross-modal correlations carry unique features that are hard to imitate by generative models, and hence serve as potent features of detecting next-generation deepfakes.

5. Conclusion

This article introduced a new multi-modal deepfake detector model that makes use of the cross-frequency pattern analysis to neutralize the shortcomings of the current spatial-, temporal-, and unimodal solutions. The proposed approach can successfully identify traces of subtle manipulations that currently in the model cannot be hidden by simply breaking down audio and visual cues into multi-band frequency representations and modeling their inter-modal spectral consistency. This is made possible by incorporating a cross-frequency correlation module, along with a multimodal fusion transformer, allowing the system to utilize

the high-frequency inconsistencies, harmonic distortions, and spectral mismatches that cannot be overcome by compression, noise, and induced by the real world in video.

The experimental data on the several benchmark data sets show that the suggested system has better performance than the existing multimodal and frequency-based detectors, especially in identifying high-quality AIGC and diffusion-generated deep fake. The frequency-domain cues are robust in that they enable the model to be highly accurate to problematic conditions that heavily reduce spatial cues. Ablution research also confirms the role played by each module and proves that cross-frequency decomposition and alignment are critical in the identification of a deepfake that can be trusted.

On the whole, this paper demonstrates the significance of spectral-domain reasoning in deepfake forensics of the next generation. The proposed framework enhances the use of cross-frequency correlations between audio and visual modalities to use in order to be more holistic and resilient in detecting advanced synthetic media. The methodology lays the groundwork to future studies in the area of multimodal spectral consistency analysis, allowing the creation of more explainable, transparent, and generalizable deepfake detectors that may be adapted to the new manipulation methods that emerge at a very fast pace.

References

1. Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A Robust Approach to Multimodal Deepfake Detection. *Journal of Imaging*, *9*(6), 122. https://doi.org/10.3390/jimaging9060122
2. Alrashoud, M. (2025). Deepfake video detection methods, approaches, and challenges. *Alexandria Engineering Journal*, *125*, 265–277. https://doi.org/10.1016/j.aej.2025.04.007
3. Xie, S., Qiao, T., Li, S., Zhang, X., Zhou, J., & Feng, G. (2026). Deepfake detection in the AIGC era: A survey, benchmarks, and future perspectives. *Information Fusion*, *127*(Part A), 103740. https://doi.org/10.1016/j.inffus.2025.103740
4. Kaur, A., Noori Hoshyar, A., Saikrishna, V., & others. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review*, *57*, 159. https://doi.org/10.1007/s10462-024-10810-6
5. Soni, N. (2025, September). Deepfake detection and multimedia forensics: Investigating synthetic media, image forgery, and video manipulation in cybercrime cases. *ARC Journal of Forensic Science*, *9*, 36–39. https://doi.org/10.20431/2456-0049.0902005
6. Khan, A. A., Laghari, A. A., Inam, S. A., & others. (2025). A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing*, *28*, 48. https://doi.org/10.1007/s10791-025-09550-0
7. Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. *Electronics*, *13*(3), 585. https://doi.org/10.3390/electronics13030585
8. Cheng, H., Pang, W., Li, K., Wei, Y., Song, Y., & Chen, J. (2025). EFIMD-Net: Enhanced feature interaction and multi-domain fusion deep forgery detection network. *Journal of Imaging*, *11*(9), 312. https://doi.org/10.3390/jimaging11090312
9. Zhao, L., Zhang, M., Ding, H., & Cui, X. (2021). MFF-Net: Deepfake detection network based on multi-feature fusion. *Entropy*, *23*(12), 1692. https://doi.org/10.3390/e23121692
10. Amen, M., & Ranam, M. (2025, April). Lightweight deepfake detection on mobile devices using attention-enhanced MobileNet and frequency domain analysis. *Journal of Technology Informatics and Engineering*, *4*, 95–114. https://doi.org/10.51903/jtie.v4i1.275
11. Kim, T., Choi, J., Jeong, Y., Noh, H., Yoo, J., Baek, S., & Choi, J. (2025, July). Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. *arXiv*. https://doi.org/10.48550/arXiv.2507.02398
12. Sun, F., Zhang, N., Xu, P., & Song, Z. (2021, November). Deepfake detection method based on cross-domain fusion. *Security and Communication Networks*, *2021*, 1–11. https://doi.org/10.1155/2021/2482942

13. Wang, F., Chen, Q., Jing, B., Tang, Y., Song, Z., & Wang, B. (2024, November). Deepfake detection based on the adaptive fusion of spatial-frequency features. *International Journal of Intelligent Systems, 2024*, Article 7578036. https://doi.org/10.1155/2024/7578036
14. Kou, Y., Li, P., Ma, H., & others. (2025). SFIAD: Deepfake detection through spatial-frequency feature integration and dynamic margin optimization. *Artificial Intelligence Review, 58*, 217. https://doi.org/10.1007/s10462-025-11225-7
15. Gao, J., Xia, Z., Marcialis, G. L., Dang, C., Dai, J., & Feng, X. (2024). Deepfake detection based on high-frequency enhancement network for highly compressed content. *Expert Systems with Applications, 249*(Part C), 123732. https://doi.org/10.1016/j.eswa.2024.123732
16. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., & Li, S.-N. (2022). M2TR: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22)* (pp. 615–623). Association for Computing Machinery. https://doi.org/10.1145/3512527.3531415
17. Gao, J., Huang, D., Zhang, J., Firkat, E., Liu, C., & Zhu, J. (2025). DDformer: Deepfake detection with multimodal fusion transformer. In D. S. Huang, W. Chen, Y. Pan, & H. Chen (Eds.), *Advanced intelligent computing technology and applications. ICIC 2025. Lecture Notes in Computer Science* (Vol. 15863). Springer. https://doi.org/10.1007/978-981-95-0009-3_31
18. Nelson, L., Batra, H., & Radha, P. (2025, June). Deepfake detection in manipulated images/audio/videos: A three-stage multi-modal deep learning framework. *Inteligencia Artificial, 28*, 20–39. https://doi.org/10.4114/intartif.vol28iss76pp20-39
19. Wang, Y., Sun, Q., Zhang, J., Rong, D., Shen, C., & Wang, X. (2026). Improving deepfake detection with predictive inter-modal alignment and feature reconstruction in audio–visual asynchrony scenarios. *Information Fusion, 127*(Part A), 103708. https://doi.org/10.1016/j.inffus.2025.103708
20. Javed, M., Zhang, Z., Dahri, F. H., & others. (2025). Enhancing multimodal deepfake detection with local–global feature integration and diffusion models. *Signal, Image and Video Processing, 19*, 400. https://doi.org/10.1007/s11760-025-03970-7
21. Miao, H., Guo, Y., Liu, Z., & Wang, Y. (2025). Multi-modal deepfake detection via multi-task audio-visual prompt learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, the Thirty*
22. *Seventh Conference on Innovative Applications of Artificial Intelligence, and the Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)* (Article 69, 10 pp.). AAAI Press. https://doi.org/10.1609/aaai.v39i1.32042
23. Park, C., Moon, B., Jeon, M., Jung, J., & Woo, S. S. (2025). X3A: Efficient multimodal deepfake detection with score-level fusion. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25)* (pp. 767–774). Association for Computing Machinery. https://doi.org/10.1145/3672608.3707934
24. Moufidi, A., Rousseau, D., & Rasti, P. (2024, January). *Multimodal deepfake detection for short videos* (pp. 67–73). https://doi.org/10.5220/0012557300003720
25. Wang, L., Zhao, J., Zhang, X., & others. (2025). ERF-BA-TFD+: A multimodal model for audio-visual deepfake detection. *Vicinagearth, 2*(10). https://doi.org/10.1007/s44336-025-00021-0