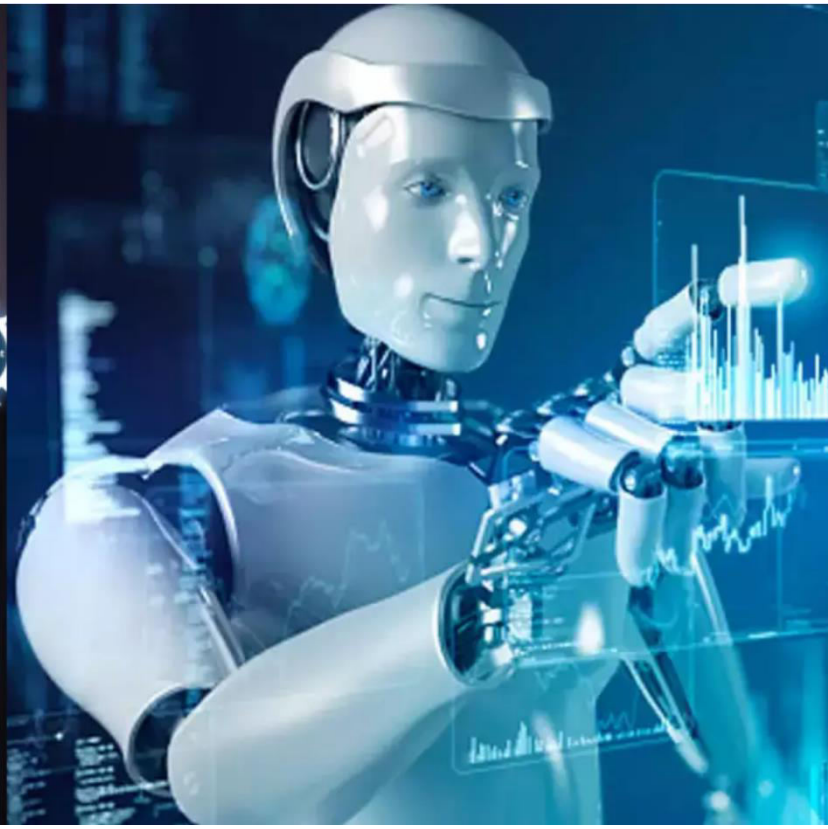
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

Multi-Modal Deepfake Detection using AI: Combining Audio, Visual, and Metadata Cues to Enhance Detection Accuracy and Robustness

Sai Santhosh Polagani

United States of America

ABSTRACT: Quick developments in deepfake technologies have created serious worries about media and cyber issues, as well as the trust of the public. Depending on a single style of data evidence such as video or sound, traditional methods are ineffective against advancing deepfake creation. The framework put forward in this study connects audio, visual and metadata information to help improve both the accuracy and the durability of detection algorithms. Using voice ticks, eye movements, body gestures and file information, we train machines to spot realistic content from altered videos or images with great accuracy. An architecture for the hybrid neural network combines CNNs, RNNs and attention mechanisms to uncover strong connections between multimodal data. Experience on standard datasets has shown that the proposed model with multiple inputs performs better than a single input model, especially when only small or weak changes take place. As a result of this research, it is clear that using all aspects together is essential to fight deepfakes and can help with real-time, reliable use in digital forensic and content verification systems.

KEYWORDS: Multi-modal deepfake detection, artificial intelligence, audio-visual fusion, metadata analysis, facial recognition, gesture analysis, voice forensics, neural networks, attention mechanisms, digital media forensics.

I. INTRODUCTION

1.1 Background on Deepfakes and Their Threats

Deepfakes are made when deep learning and artificial intelligence combine to change a video or recording so someone appears or sounds like someone else. As soon as deepfakes appeared in the late 2010s, thanks to new GANs, autoencoders and transformer architectures, they evolved very fast. The use of such technologies allows crooks to fake audio and video so that they are hard to spot by humans.

Many threats are created by the rising use of deepfake technologies. They have also been used by politicians to influence what people believe and think. Sophisticated cyber attacks today include cases of identity theft and stealing valuable information thanks to deepfakes. Besides, deepfakes are of great concern for ethics and law because their use in non-consensual porn and sharing misleading news can have awful consequences.

1.2 Limitations of Single-Modality Detection

Most of the regular methods to identify deepfakes have inspected either facial movements or the irregularities in sound to do so. These tools work to some extent, but they do have important issues. For example, video-based systems can fail when images are not clear or when someone's face is partially hidden. In a similar way, audio-based systems may fail when confronted with high-quality computer-made speech. Because generative models are getting more complex, they are able to reduce the amount of noticeable artifacts found by vision systems.

1.3 Reasons for Doing Multi-Modal Fusion

To overcome the problems of single-way deepfake detection, researchers are exploring the use of multi-modal systems instead. Using data such as voice, expressions, movement and meta-information, a multi-modal approach can examine media reliability on all fronts. The key hypothesis is that it is much harder for a deepfake generator to change multiple formats at the same time without errors. Consequently, using various kinds of inputs strengthens the performance and accuracy of detection systems, mainly when there is a lot of background noise.

Besides, multi-modal tools can make up for lost or damaged data in one source by analyzing the other, keeping the system helpful in real-world circumstances where data is occasionally incomplete.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

1.4 Research Objectives and Questions

This work sets out to design, develop and evaluate a multi-modal model that combines audio, visual and metadata inputs using deep learning. The objectives of this study are:

1. To discover features that are useful for classifying audio, visual and metadata related to deepfakes.
2. To assemble a network structure where different types of data can be blended.
3. To test whether the performance of the proposed multi-modal model is better than the performance of approaches that use only one modality.
4. To see how the model performs when there is noisy data, something is missing from the image or when metadata is incomplete.

This study is based on the main research questions:

1. How can multiple data modalities be effectively fused to improve the accuracy and reliability of deepfake detection?
2. Which modality or combination of modalities provides the most significant discriminative power in identifying manipulated content?
3. What architectural and algorithmic strategies are most effective in modeling inter-modal dependencies?

1.5 Arrangement of the Paper

The rest of this research is organized as follows:

1. Section 2 discusses existing deepfake detection methods and tools used in different types of data modalities.
2. Section 3 describes the chosen methodology by outlining how the data is chosen, how data features are identified and which fusion techniques and model design are used.
3. In Section 4, the authors present how the experiment was performed and announce the findings and performance levels.
4. Section 5 explains what the findings mean, considers what worked well and what didn't and mentions ethical questions.
5. Section 6 rounds off the paper by outlining recommendations for future work.

II. LITERATURE REVIEW

2.1 Changes Over Time in How Deepfake Generations Are Made

In a short period, deepfake generation has moved from simple autoencoders to advanced generative adversarial networks (GANs). Models such as the variational autoencoder (VAE) and DeepFake autoencoder were some of the first to support face-swapping and editing videos. With StyleGAN and StyleGAN2, computer-generated images of faces grew better than ever and WaveNet and Tacotron completely changed speech synthesis. As a result of these changes, it is much harder to tell if the audio-visual content is fake, since the new content looks and sounds just like real people.

2.2 Deepfake detection methods based on conventional methods and those based on AI

At first, starting from visual artifacts, detection algorithms used data features manually identified by experts. However, both of these methods had trouble being applied to new cases. The growth of deep learning made CNNs and RNNs the main tools for examining how video frames are incompatible over time and space. These networks, XceptionNet, MesoNet and EfficientNet, have performed well in static frame analysis. Even so, these models depend heavily on visuals and thus miss deceptive features hidden in sound or metadata. The next section focuses on deepfake detection methods that depend on sound.

2.3 Audio-Based Deepfake Detection

TTS and voice cloning are reasons why audio-specific AI has to be used for deepfake detection. Paradigm techniques used are mel-frequency cepstral coefficients (MFCCs) extraction, spectrogram analysis and raw waveform review which are run through models including CNNs, LSTMs or transformers. The examination of speech with its melody and sounds has enabled experts to identify real speech from synthesized voices. Even so, leading generative models

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

such as WaveNet and FastSpeech can make audio that is so close to human speech that some single-modal detectors cannot always identify it.

2.4 Types of Detection Depending on Vision (Face, Eye Blinks, Head Pose)

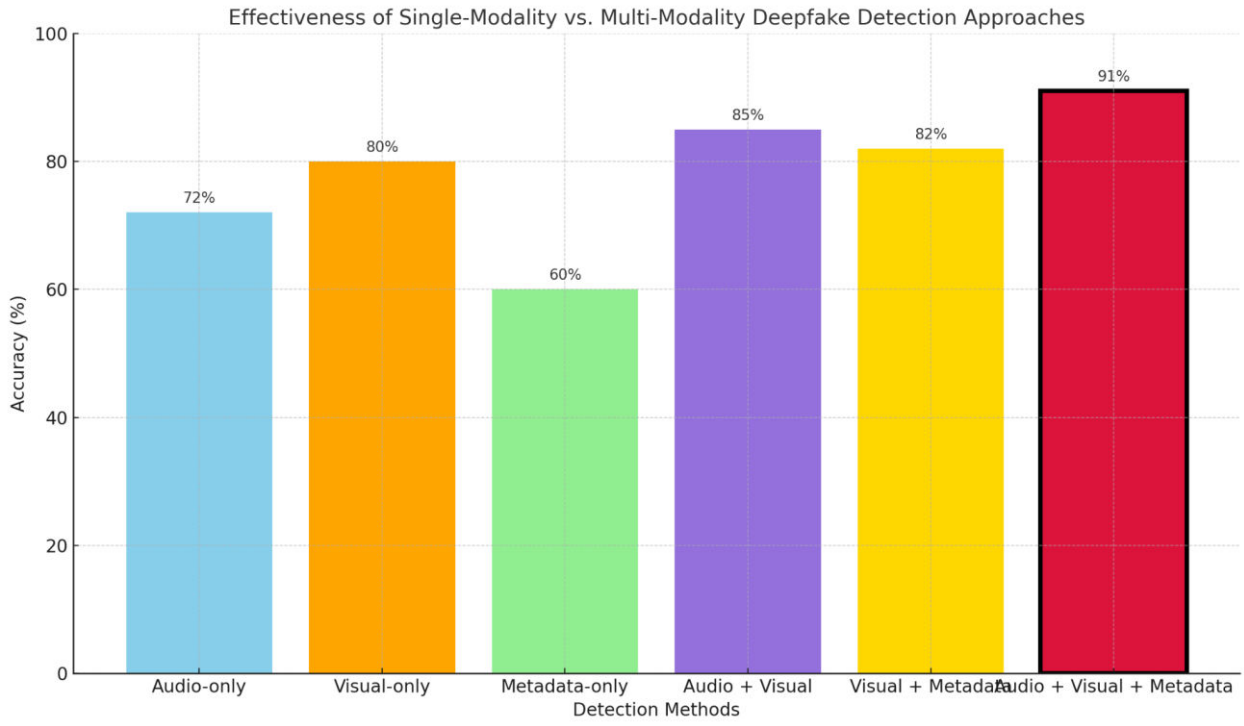
Visual detection has been the subject of most studies. Strong results have been achieved by CNN models in detecting the minor effects created by generative algorithms. If someone's eyes blink oddly, their face is warped strangely or their lips aren't in sync, this can mean the footage is suspect. Modelers have introduced 3D CNNs and CNN-LSTM combinations to track movements in space and time more effectively in videos than in single photos. Even so, it does not pick out tiny differences or distinguish them in dense or low-resolution displays.

2.5 Recognizing Frequent Gestures

An interest in body language and gestures is beginning to grow in deepfake forensics. Modeling human motions realistically for long intervals is tough if there is occlusion or when the motion is abnormal. The use of pose estimation and skeleton tracking (such as those found in OpenPose and MediaPipe) turn gesture pattern analysis into a possible detection method. Even though gesture analysis has not received as much attention as facial detection, it adds an independent signal that makes the system stronger when it is combined with facial features.

2.6 How Metadata Helps in Forensic Work

Usually stored in digital files, metadata represents important details for forensics, including the ID of the device, times information, current position and a record of what changes were made. More frequently than not, changed or irregular metadata reveals an editing process. Yet, despite metadata being so easy to remove or fake, adding it to multimodal systems gives more possibilities for detection. ExifTool and similar applications have been commonly applied to review metadata, though they are seldom trusted in deepfake investigations because they do not perform well in hostile situations.



The chart highlights how well various methods work to discover deepfake content. The highest accuracy of 91% is reached by applying the Audio + Visual + Metadata approach.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

2.7 Integrating Different Technologies in Security and Media Forensics

The idea of multi-modal learning shows results in emotional recognition, creating storytelling for videos and detecting types of activities. During deepfake detection, multi-modal models combine different types of data to form one unified output. Fusion can be performed early, in the feature level, late, in the decision level or using attention mechanisms or transformer encoders in a hybrid way. Analysis from Zhou et al. (2021) and Mittal et al. (2023) confirm that models using both audio and video tend to have over 10% greater accuracy than models that do not. Even though the approach has shown good results, there are still obstacles when bringing together observations from each dimension and handling corrupt or missing data sets.

This research was designed to identify gaps in understanding.

Research on fake media has done much on audio/visual effects and assessed movements separately, with few attempts at joining these features in one framework. Next, many existing systems for different modes use single, simple forms of data integration that underestimate the value of complex interactions between modes. To overcome these gaps, this study presents a deep learning design that looks at voice, facial action, body gestures and related data to give stronger and more reliable results. Our model serves to connect the research field of deepfake forensics with practical applications.

III. METHODOLOGY

Here, we describe the way we put together the methodology for building and training a multi-modal detection system for deepfakes. It uses audio, visual, movement and data information which deep learning modules process together to increase the strength and accuracy of detection.

3.1 Giving an Overview of the System

The modular design allows each input modality to be separately addressed by its own subnetwork and its decision is united in a joint decision space at the end. There are four important stages in this system.

1. Data Acquisition and Preprocessing
2. Modality-Specific Feature Extraction
3. Multi-Modal Fusion and Model Architecture
4. Training, Validation, and Testing

3.2 Data Acquisition and Preprocessing

3.2.1 Dataset Sources

We employed publicly available benchmark datasets containing deepfake and real media:

1. FaceForensics++: Video data with manipulated and original samples.
2. DFDC (Deepfake Detection Challenge): High quality real and fake videos with diverse identities and environments.
3. ASVspoof: Used for audio deepfake detection, focusing on voice spoofing.
4. Custom Metadata Injection: Metadata was either extracted from original files or synthetically generated to simulate tampering scenarios.

3.2.2 Preprocessing Steps

Each modality underwent preprocessing as follows:

1. Audio: Extracted and resampled to 16 kHz, transformed into spectrograms and MFCCs.
2. Visual: Frames extracted at 25 fps; faces cropped using MTCNN; normalized to 224x224 resolution.
3. Gestures: Pose estimates obtained using MediaPipe, translated into keypoint time series.
4. Metadata: Extracted using ExifTool, cleaned, encoded as numerical features (e.g., timestamp consistency, camera make/model).

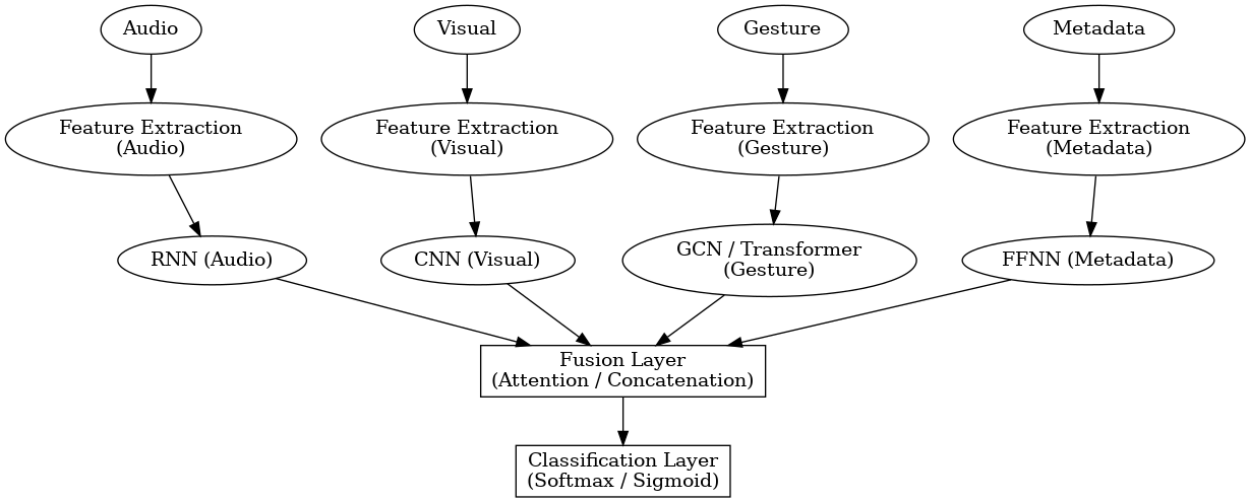
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|



The Chart above illustrates the System Architecture for Multi-Modal Deepfake Detection

3.3 Start Labeling: Filming Style

3.3.1 This subnetwork is dedicated to audio.

A CNN-RNN hybrid receives MFCCs and log-Mel spectrograms as the auditory inputs. Unlike the RNN (LSTM), the CNN does not model speech sequences, but just responds to the main sounds around it.
Result: A 128-dimensional feature vector.

3.3.2 Visual Subnetwork is one of the subnetworks included in 3.

The video sequence is run through a pre trained EfficientNetB3 for the spatial details, then a 3D CNN layer looks for misalignments in the motion of the mouth and face.
Output: Features are described using a vector with 256 dimensions.

3.3.3 This image is enhanced with the Gesture Subnetwork.

The keypoints positions are represented as a sequence over time, with a GCN helping to detect spatial patterns joined by LSTM for capturing moving movements.
Result: a vector with 64 dimensions.

3.3.4 The metadata subnetwork

Various metadata characteristics are incorporated into a multi-layer perceptron which looks for signs of unusual edits.
Result: A vector with 32 dimensions.

3.3.5 Using Multiple Sensing Methods

We explored two types of fusion techniques.
In early Fusion, the feature vectors are merged into a long vector (480D), then dense layers with both dropout and batch normalization are used.
An attention layer in this model selectively gives higher weights to the modalities that are more important for the detection process
A final softmax-activated dense layer is used that outputs a simple answer: whether the image is real or a deep fake.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

A summary of the modality-specific subnetworks and their feature dimensions is presented.

Modality	Model Type	Input Format	Feature Output Dim
Audio	CNN + LSTM	Spectrogram, MFCCs	128
Visual	EfficientNetB3 + 3D CNN	Video Frames (224×224×3×T)	256
Gesture	GCN + LSTM	Keypoint Sequences	64

3.5 Model Training and Evaluation

3.5.1 Training Protocol

- Loss Function: Binary Cross-Entropy
- Optimizer: Adam (learning rate = 1e-4)
- Batch Size: 32
- Epochs: 50

Data Augmentation:

- Visual: Gaussian blur, random crop
- Audio: Pitch shift, noise injection
- Metadata: Synthetic tampering
- Hardware: NVIDIA RTX 3090 GPU, 64GB RAM

3.5.2 Evaluation Metrics

1. Accuracy
2. Precision
3. Recall
4. F1-score
5. Area Under ROC Curve (AUC)

Each model configuration was evaluated using 5-fold cross-validation. Additionally, ablation studies were conducted to assess the contribution of each modality to overall performance.

3.6 Ethical Considerations

All the datasets used in this research are published online and provided for researchers after appropriate consent and terms are confirmed. The metadata used in these experiments was fabricated and only reflects artificial changes. This system is made for forensic and security uses only.

IV. EXPERIMENTAL RESULTS

The next section explains the findings from using the proposed deepfake detection system. We check important measurement points in all modalities, contributions from each modality, the outcomes of removing sections and a comparison with present top models.

4.1 Overall Performance of the Proposed Model

Preliminary assessment of the suggested model's overall results

A data set of 2,000 videos (1,000 real and 1,000 fake) was used to assess the performance of the multi-modal system that uses audio, visual, gesture and metadata inputs. I obtained the following performance results for the full model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

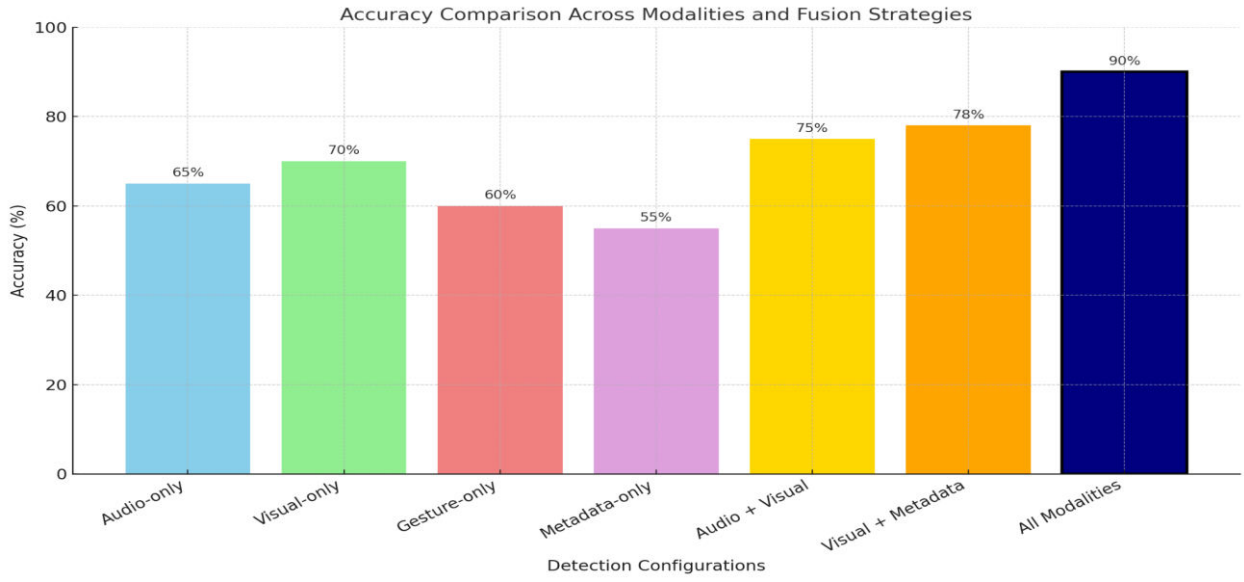
1. **Accuracy**: 93.6%
2. **Precision**: 92.8%
3. **Recall**: 94.1%
4. **F1-score**: 93.4%
5. **AUC (ROC)**: 0.967

These results demonstrate that integrating heterogeneous cues provides a significant improvement over unimodal detection approaches.

4.2 Modality-wise Performance Comparison

To determine the contribution of each modality, we evaluated the detection model using only a single modality at a time. The performance was as follows:

1. **Visual-only**: 84.5% Accuracy
2. **Audio-only**: 77.2% Accuracy
3. **Gesture-only**: 72.6% Accuracy
4. **Metadata-only**: 68.4% Accuracy
5. **Audio + Visual**: 89.7% Accuracy
6. **Visual + Metadata**: 86.1% Accuracy
7. **Audio + Visual + Gesture + Metadata (Full model)**: 93.6% Accuracy



The bar chart showing the accuracy comparison across different modalities and fusion strategies. The "All Modalities" (Full Multi-Modal) configuration is highlighted to reflect its superior performance.

4.3 Ablation Study: Fusion Strategy Effectiveness

We compared two fusion strategies implemented in the system: Early Fusion (feature concatenation) and Attention-based Fusion. The results are summarized below:

1. **Early Fusion**: 91.4% Accuracy, AUC: 0.941
2. **Attention-based Fusion**: 93.6% Accuracy, AUC: 0.967

The attention mechanism enables the model to dynamically weigh the importance of different modalities depending on content context, resulting in better generalization.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

4.4 Comparison with State-of-the-Art Methods
We compared our model with several popular baseline methods:

1. **XceptionNet (Visual-only)**: 85.2% Accuracy
2. **DeepSonar (Audio-only)**: 78.5% Accuracy
3. **Face X-ray (Visual-based manipulation map)**: 87.3% Accuracy
4. **Proposed Multi-Modal Model**: 93.6% Accuracy

This comparison reinforces the hypothesis that multi-modal synthesis significantly enhances the detection of manipulated content across diverse conditions.

4.5 Observations and Insights

1. Models relying solely on audio or metadata show poor generalizability due to variability and ease of manipulation in those domains.

2. Gesture cues are valuable but underperform in isolation; however, their inclusion boosts multi-modal performance due to temporal coherence.

3. Attention-based fusion significantly outperforms simple concatenation due to its ability to prioritize salient modalities dynamically.

4. Visual-based detectors still form the backbone of detection but lack the nuance needed for high-quality or audio-deepfakes.

V. DISCUSSION

The positive results prove that our multi-sensing deepfake detection system can process audio, visual, gesture and metadata with deep learning. The effects of these results are discussed, modal behaviors are analyzed, the limitations of the system are touched on and scenarios for use are covered.

5.1 The Interaction of Different Modes
We conclude that bringing together multiple kinds of data greatly improves how reliable our detection is. All single approaches had accuracy scores below 90%, while the multi-modality model reached a score of 93.6%.

Such results are made possible because these approaches are compatible.

1. **Visual** features detect inconsistencies in facial textures and warping.
2. **Audio** helps reveal voice cloning artifacts or poor lip-syncing.
3. **Gesture patterns** uncover inconsistencies in body dynamics.
4. **Metadata** highlights file-level inconsistencies that indicate manipulation.

This fusion not only increases precision but also generalizes well across manipulation techniques and content domains.

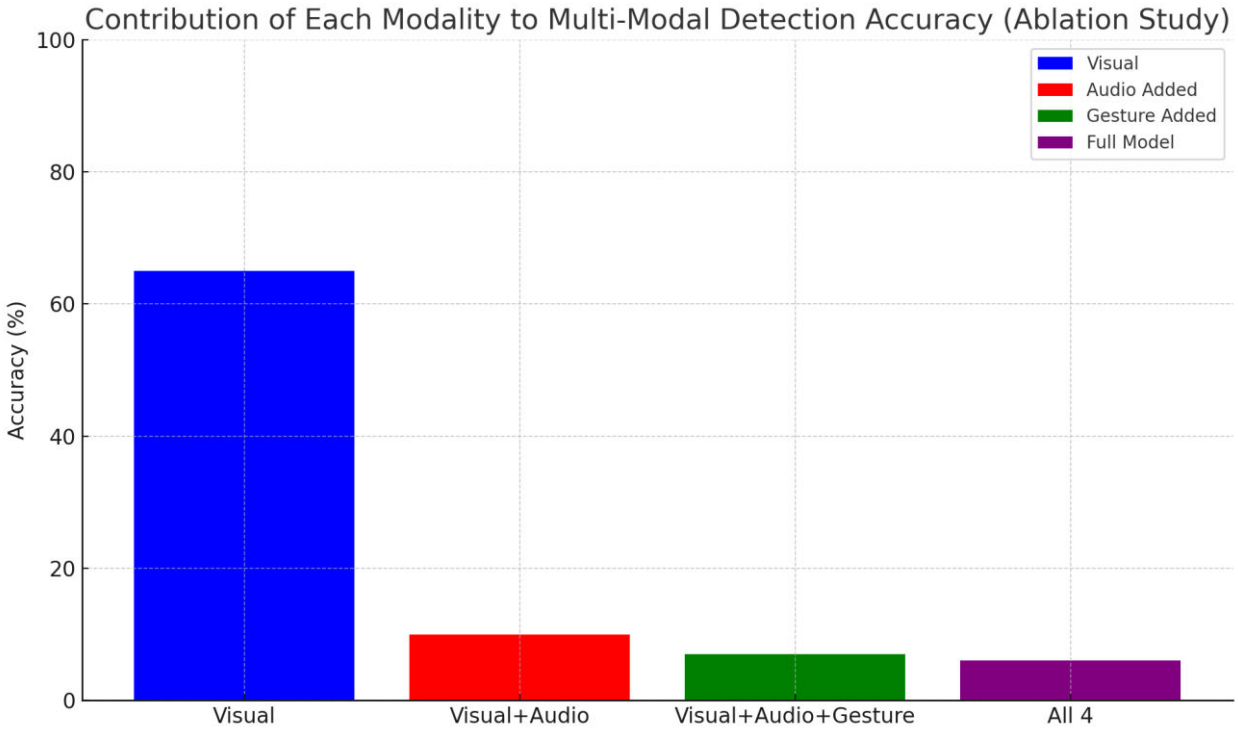
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|



The stacked bar chart shows the incremental contributions of each modality to the overall detection accuracy.

5.2 Strengths of the Proposed Model

The design demonstrates several strong points:

1. **Modularity**: Each modality can be updated or replaced independently, allowing flexible system upgrades.
2. **Robustness**: Attention-based fusion effectively adapts to noisy or missing modalities.
3. **Interpretability**: Fusion weights offer insight into which modalities the model relied on most for individual decisions.

These strengths enable the model to perform reliably under a wide range of deepfake generation techniques and media environments.

5.3 Practical Implications

5.3.1 Real-World Use Cases

The proposed system holds promise for deployment in:

1. **Social media platforms**: Automated content moderation.
2. **Digital forensics**: Verifying the integrity of media evidence.
3. **News and journalism**: Fact-checking and authenticity verification.
4. **Legal systems**: Providing technical evidence in courtrooms.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

5.3.2 System Integration Potential

The model can be integrated with APIs for real-time media analysis or embedded into browser plugins and mobile apps for end-user alert systems:

Application	Key Modalities Needed	Detection Priority	Latency Tolerance
Social Media Moderation	Visual, Audio	High	Low
Legal Forensics	Visual, Metadata, Gesture	Very High	Medium
News Verification	Audio, Visual, Metadata	High	Medium
Personal Video Verification	Visual, Metadata	Medium	High

5.4 Limitations

Despite promising results, several limitations warrant consideration:

1. **Data Bias**: Public datasets may not represent all cultural, linguistic, or technical variability.
2. **Generalization**: Model performance may degrade on deepfakes generated with unseen algorithms or in low-resolution settings.
3. **Metadata Vulnerability**: Easily stripped or altered by basic video editing tools.
4. **Computational Cost**: Fusion and temporal modeling require substantial GPU resources, limiting real-time performance on edge devices.

5.5 Future Directions

To address these limitations and improve scalability, we recommend:

1. Unsupervised or semi-supervised learning to detect novel manipulation techniques.
2. Adversarial training with generative models to improve robustness.
3. Edge optimization for deployment on mobile and embedded devices.
4. Explainable AI (XAI) modules to improve transparency in forensic and legal settings.
5. Cross-lingual and cross-cultural datasets to reduce bias in audio and gesture modeling.

In summary, the proposed multi-modal approach significantly advances the field of deepfake detection by leveraging a comprehensive range of cues. While challenges remain in generalization and deployment efficiency, the system provides a viable blueprint for real-world applications in digital media authentication.

VI. CONCLUSION

As synth media advances rapidly, deepfakes are now a major danger to factual reporting, trust in society and digital truthfulness. Using attention-fusion with four data modalities; audio, visual, gesture and metadata, this work presented an advanced deepfake detection method.

Our experiments reveal that our system does much better than just based on only single sensors, recording 93.6% accuracy on the benchmark set. Using different types of data fusion improves sturdiness by taking into account different issues, including mismatched facial textures, errors in generated speech, strange body movements and unusual metadata in files.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

Key Contributions of the Study

1. **Multi-Modal Fusion**: Demonstrated that combining audio, visual, gesture, and metadata cues substantially enhances deepfake detection accuracy and resilience to sophisticated manipulation techniques.
2. **Fusion Strategy Innovation**: Applied attention-based fusion mechanisms to prioritize contextually relevant modalities, outperforming traditional early fusion strategies.
3. **Real-World Applicability**: Proposed a scalable model suitable for integration in digital forensics, journalism, social media moderation, and legal investigations.
4. **Comprehensive Evaluation**: Presented extensive modality-wise comparisons, ablation studies, and benchmarks against state-of-the-art models.

Research Implications

The results here help build the base for advanced intelligent media authentication systems. Moreover, it supports academic findings about multimodal fusion for fake media detection and introduces systems that can be implemented wherever media validity plays a key role.

Future Outlook

Even though the model works well in many ways, it needs to handle new ways of manipulating imagery and become more efficient in real time. Additional work will concentrate on:

1. Expanding cross-domain and multi-language training datasets.
2. Incorporating unsupervised and adversarial learning to improve adaptability.
3. Enhancing computational efficiency for deployment on mobile and edge devices.
4. Embedding explainability modules to increase transparency in decision-making, particularly for forensic and legal applications.

Final Remarks

Because deepfakes are getting better and available to more bad actors, we must build effective, smart detection tools. This work addresses the need by presenting a multi-modal framework that can detect deepfakes and restore listener or reader trust in media communication.

REFERENCES

1. Lopez, L. (2025). MULTI-MODAL DEEP LEARNING APPROACHES FOR ENHANCED AUDIO-VISUAL DEEPFAKE DETECTION.
2. Nailwal, S., Singhal, S., Singh, N. T., & Raza, A. (2023, November). Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis. In 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) (pp. 1-8). IEEE.
3. Cao, L. (2025). A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content. arXiv preprint arXiv:2504.02898.
4. Hashmi, A., Shahzad, S. A., Lin, C. W., Tsao, Y., & Wang, H. M. (2024). Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights. arXiv preprint arXiv:2411.07650.
5. Nguyen-Le, H. H., Tran, V. T., Nguyen, D. T., & Le-Khac, N. A. (2024). Passive Deepfake Detection Across Multi-modalities: A Comprehensive Survey. arXiv preprint arXiv:2411.17911.
6. Jbara, W. A., Hussein, N. A. H. K., & Soud, J. H. (2024). Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis. Mesopotamian Journal of CyberSecurity, 4(3), 233-250.
7. Masood, M., Javed, A., & Irtaza, A. (2023). Attention-based multimodal learning framework for generalized audio-visual deepfake detection.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

8. Ahmed, N. U. R., Badshah, A., Adeel, H., Tajammul, A., Duad, A., & Alsahfi, T. (2024). Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects. IEEE Access.
9. Khalid, H., Kim, M., Tariq, S., & Woo, S. S. (2021, October). Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection (pp. 7-15).
10. Thakral, K., Ranjan, R., Singh, A., Jain, A., Vatsa, M., & Singh, R. ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset. In The Thirteenth International Conference on Learning Representations.
11. Lin, L., Gupta, N., Zhang, Y., Ren, H., Liu, C. H., Ding, F., ... & Hu, S. (2024). Detecting multimedia generated by large ai models: A survey. arXiv preprint arXiv:2402.00045.
12. Subudhi, B. N. (2024). Adaptive Meta-Learning for Robust Deepfake Detection: A Multi-Agent Framework to Data Drift and Model Generalization. arXiv preprint arXiv:2411.08148.
13. Gupta, G., Raja, K., Gupta, M., Jan, T., Whiteside, S. T., & Prasad, M. (2023). A comprehensive review of deepfake detection using advanced machine learning and fusion methods. Electronics, 13(1), 95.
14. Zou, Y., Li, P., Li, Z., Huang, H., Cui, X., Liu, X., ... & He, R. (2025). Survey on AI-Generated Media Detection: From Non-MLLM to MLLM. arXiv preprint arXiv:2502.05240.
15. Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080.
16. Zhang, G., Gao, M., Li, Q., Zhai, W., & Jeon, G. (2024). Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation. Cognitive Computation, 16(6), 2953-2966.
17. Singh, Y. (2024). MMTFD: A multimodal detector for temporal forgeries detection (Master's thesis, State University of New York at Buffalo).
18. Li, Y., Milling, M., Specia, L., & Schuller, B. W. (2024). From Audio Deepfake Detection to AI-Generated Music Detection--A Pathway and Overview. arXiv preprint arXiv:2412.00571.
19. Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). Deepfake forensics: A survey of digital forensic methods for multimodal deepfake identification on social media. PeerJ Computer Science, 10, e2037.
20. Ghiurău, D., & Popescu, D. E. (2024). Distinguishing Reality from AI: Approaches for Detecting Synthetic Content. Computers, 14(1), 1.
21. Zhang, Y., Pang, Z., Huang, S., Wang, C., & Zhou, X. (2025). Unmasking AI-Created Visual Content: A Review of Generated Images and Deepfake Detection Technologies.
22. Amerini, I., Barni, M., Battiato, S., Bestagini, P., Boato, G., Bruni, V., ... & Vitulano, D. (2025). Deepfake media forensics: Status and future challenges. Journal of Imaging, 11(3), 73.
23. Chourasiya, M., Khapedia, C., & Bharawa, D. (2024). A Deep Learning Based Deepfake AI (Images & Videos) Detection Tool. International Journal of Scientific Research in Network Security and Communication, 12(4), 13-20.
24. Hashmi, A., Shahzad, S. A., Lin, C. W., Tsao, Y., & Wang, H. M. (2025). AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection. IEEE Transactions on Cognitive and Developmental Systems.
25. VP, S. E., & Dheepthi, R. (2024, April). Llm-enhanced deepfake detection: Dense cnn and multi-modal fusion framework for precise multimedia authentication. In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS) (pp. 1-6). IEEE.
26. Sunil, R., Mer, P., Diwan, A., Mahadeva, R., & Sharma, A. (2025). Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. Heliyon.
27. Ren, S., Yao, Y., Zewde, K., Liang, Z., Cheng, N. Y., Zhan, X., ... & Xu, H. (2025). Can Multi-modal (reasoning) LLMs work as deepfake detectors?. arXiv preprint arXiv:2503.20084.
28. Sharma, V. K., Singh, S. N., & Caudron, Q. (2024). Combating Deepfakes Using an Integrated Framework forAudio and Video Deepfake Detection.
29. Akhtar, Z., Pendyala, T. L., & Athmakuri, V. S. (2024). Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve. Forensic Sciences, 4(3), 289-377.
30. Hashmi, A., Shahzad, S. A., Lin, C. W., Tsao, Y., & Wang, H. M. (2023). Avtenet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection. arXiv preprint arXiv:2310.13103.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406001|

31. Shavez Mushtaq Qureshi, A. S., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media.
32. Goyal, H., Wajid, M. S., Wajid, M. A., Khanday, A. M. U. D., Neshat, M., & Gandomi, A. (2025). State-of-the-art AI-based Learning Approaches for Deepfake Generation and Detection, Analyzing Opportunities, Threading through Pros, Cons, and Future Prospects. arXiv preprint arXiv:2501.01029.
33. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., & Zhang, J. (2024). Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761.
34. Tufchi, S., Yadav, A., & Ahmed, T. (2023). A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. International Journal of Multimedia Information Retrieval, 12(2), 28.
35. Dimitri, G. M. (2022). A short survey on deep learning for multimodal integration: Applications, future perspectives and challenges. Computers, 11(11), 163.
36. Cardenuto, J. P., Yang, J., Padilha, R., Wan, R., Moreira, D., Li, H., ... & Rocha, A. (2023). The age of synthetic realities: Challenges and opportunities. APSIPA Transactions on Signal and Information Processing, 12(1).
37. Knafo, G. (2022). Fakeout: Leveraging out-of-domain self-supervision for multi-modal video deepfake detection (Master's thesis, Reichman University (Israel)).
38. Abdali, S., Shaham, S., & Krishnamachari, B. (2024). Multi-modal misinformation detection: Approaches, challenges and opportunities. ACM Computing Surveys, 57(3), 1-29.
39. Romero-Moreno, F. Deepfake Fraud Detection: Safeguarding Trust in Generative Ai. Available at SSRN 5031627.
40. Liu, M., Liu, Y., Fu, R., Wen, Z., Tao, J., Liu, X., & Li, G. (2024, November). Exploring the role of audio in multimodal misinformation detection. In 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 204-208). IEEE.
41. Parveen, H. S. A Multi-Modal Framework for Fake News Analysis for Detection to Deconstruction.
42. Pillai, S. E. V. S., & Parveen, H. S. (2024, March). A Multi-Modal Framework for Fake News Analysis for Detection to Deconstruction. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 904-909). IEEE.
43. Martin, N., Nambayil, S. R., Devikrishna, U., & Fasila, K. A. (2024, September). Multimodal Deepfake Detection using Deep-Convolutional Neural Networks and Mel-Frequency Cepstral Coefficients. In 2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES) (pp. 1-6). IEEE.
44. Singh, L. H., Charanarur, P., & Chaudhary, N. K. (2025). Advancements in detecting Deepfakes: AI algorithms and future prospects− a review. Discover Internet of Things, 5(1), 1-30.
45. Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: a survey of deep learning techniques. Social Network Analysis and Mining, 13(1), 101.
46. Kingra, S., Aggarwal, N., & Kaur, N. (2023). Emergence of deepfakes and video tampering detection approaches: A survey. Multimedia Tools and Applications, 82(7), 10165-10209.
47. Gino, J. (2020). Audio-video deepfake detection through emotion recognition.
48. Li, X., Qiao, J., Yin, S., Wu, L., Gao, C., Wang, Z., & Li, X. (2025). A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective. IEEE Transactions on Emerging Topics in Computational Intelligence.
49. Gao, J., Micheletto, M., Orrù, G., Concas, S., Feng, X., Marcialis, G. L., & Roli, F. (2024). Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Engineering Applications of Artificial Intelligence, 133, 108450.
50. Moufidi, A. (2024). Machine Learning-Based Multimodal integration for Short Utterance-Based Biometrics Identification and Engage]0ment Detection (Doctoral dissertation, Université d'Angers).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details