IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141065

A Comprehensive Review and Prototype Implementation for Deepfake Detection System using Multi-Modal

Adesh Borude[1], Nikam Abhishek[2], Waghmode Vaibhav[3], Mayur Gavhane[4], Prof. B.Y. Baravkar [5], Prof. R. S. Gandhi[6]

Information Technology, Dattakala group of Institute, Pune Maharastra India[1-4]

Guide, Informatiom Technology, Dattakala group of Institute, Pune Maharastra India[5]

Co-Guide, Informatiom Technology, Dattakala group of Institute, Pune Maharastra India[6]

Abstract: The advancements in deepfake technology have come swiftly, allowing for the creation of extremely realistic altered images, videos, and audio material. Although there has been considerable progress in unimodal detection in current research, most approaches tend to concentrate on a single modality. This paper analyses more than 20 cutting-edge studies on deepfake detection and pinpoints significant research shortcomings, including the absence of multi-modal frameworks, limitations in datasets, lack of robustness, and insufficient interpretability. To address these issues, we built a prototype detection system based solely on single-modality images that employs two models: a custom Convolutional Neural Network (CNN) and Exception CNN. Our findings underscore the necessity for solutions that incorporate multiple modalities. We suggest an integrated framework for multi-modal detection encompassing images, videos, and audio, which represents the next advancement toward reliable and effective detection systems.

Keywords: Deepfake detection, CNN, Captioned, multi-modal system, Video and Audio Forensics etc.

I. INTRODUCTION

The rapid growth of artificial intelligence has accelerated the development of deepfakes—super realistic altered images, videos and audio created using techniques such as generative adversarial networks (GANs), variational autoencoders (VAEs) and diffusion models. These pose a big risk to digital trust, cybersecurity and society as a whole as they can be used for misinformation campaigns, political manipulation, identity fraud and financial crime. Deepfakes are being used in advanced scams including impersonating individuals during live video calls to get signing bonuses or sensitive data as highlighted by recent studies and reports from industry experts [1].

The Traditional detection systems have mainly focused on a single modality, such as only for images, video or audio. Image-based anomaly detection methods typically use CNNs or transformers to discover spatial characteristics in images and detect anomalies. For video, RNN and Long Short Term Memory (LSTM) based models are utilized to exploit temporal inconsistencies in eye, head and lip movements [2]. As for audio deepfake detection, the solutions are in capturing spectrogram and frequency information [3], using MFCC features[4] and adversarial training that identifies a scenario with generated speech. Although these unimodal approaches work well in controlled datasets, they struggle in real world applications due to challenges like compression, channel noise, unseen manipulation techniques or lack of diversity in training datasets.

Recent research has started to explore multi- modal deepfake detection systems to overcome these challenges. For instance, SIDA [5] integrates features from images with metadata to determine, find, and clarify altered content, whereas Vision- Language Models [6] utilize cross-modal reasoning to enhance interpretability. In the same vein, cross-modal attention-based transformers [7] combine audio and visual components to enhance detection, albeit at the price of computational resources and issues with data alignment. These multi-modal approaches show that combining visual, temporal and auditory information can greatly improve robustness and generalization compared to methods that rely on a single modality.

Moreover, state-of-the-art methods have explored alternative signals beyond standard feature extraction. For instance, Intel's FakeCatcher detects deepfakes in real-time by utilizing subtle pixel-level variations reflective of

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141065

blood flow in videos, offering a physiological viewpoint for verification [8]. Other solutions like Reality Defender and GetReal Labs emphasize live video detection, addressing the escalating problem of real-time scams [1], [9], [10].

These issues, however, are merely the beginning of the challenges that remain to be solved regarding the scaling of detection systems for real-world applications. Many models are trained on limited datasets (Celeb-DF, FaceForensics++, VoxCeleb) and don't generalize well to different environments or unknown generative models. Adversarial attacks and compression artifacts further degrade detection accuracy and interpretation and explanation of complex models are not good enough for deployment in sensitive scenarios. In audio deepfake detection, multi-speaker context, background noise and variability in channels are the challenges and we need robust multi-modal architectures.

In summary, although considerable progress has been made in this domain, the majority of current methodologies continue to be either unimodal or only partially multi-modal, which restricts their effectiveness in realistic scenarios where fabricated content is present across images, videos, and audio simultaneously. This underscores the necessity for comprehensive multi- modal detection frameworks that amalgamate spatial, temporal, and auditory signals, enhanced by attention- based interpretability, to create dependable and scalable systems for deepfake detection. Building upon these insights, this study reviews over 20 recent works, implements a prototype image-based system, and suggests a future multi-modal detection framework to enhance the prevailing state-of-the-art.

II. RELATED WORK

Recent research on deepfake detection has primarily concentrated on unimodal methods, examining altered content separately for images, videos, or audio. Though successful in specific areas, these techniques frequently encounter difficulties when applied to real-world scenarios and diverse manipulation methods.

a. Image-based Detection:
Convolutional Neural Networks (CNNs) and Vision Transformers continue to be essential for image-based detection. Models such as XceptionNet and ResNet effectively identify spatial variations in facial features and textures [11]. Huang et al. [5] developed SIDA, which improves localization and interpretability for forensic analysis of social media images. In the same manner, Guo et al. [6] integrated visionlanguage modeling to enhance detection performance and explainability. Even with impressive outcomes on datasets such as Celeb-DF and FaceForensics++, systems that rely solely on images do not possess temporal or auditory comprehension.

b. Video-based Detection:
Video detection utilizes temporal signals like unnatural blinking, head movement, or lip- sync inaccuracies with RNNs, LSTMs, and 3D CNNs [2], [12]. Video models based on transformers [13] similarly capture long- range dependencies. Nonetheless, these techniques are very sensitive to compression artifacts, noise, and are resource-demanding, restricting real-time scalability.

c. Audio-based Detection: Audio detection emphasizes recognizing synthetic or altered speech through the use of spectrogram-based CNNs coupled with MFCC features and LSTM networks [3], [4]. Even though adversarial training enhances resilience, models continue to have difficulties with background noise, overlapping voices, and synchronizing with visual information.

d. Limitations and Multi-Modal Approaches: Unimodal systems, while performing well on benchmarks, exhibit a lack of robustness to unfamiliar manipulations [14]. Recent advancements, including vision-language integration [6] and cross-modal attention transformers [15], show promise by merging image, video, and audio modalities.

III. RESEARCH GAPS

Recent research in deepfake detection has investigated various methodologies aimed at enhancing robustness and generalization across images, videos, and audio. Katamneni and Rattani [16] introduced a cross-modal attention mechanism for audio-visual detection, which resulted in improved accuracy and localization; however, it encounters difficulties when audio and visual data are misaligned or contain noise. In a similar vein, Wang et al.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141065

[17] employed dual transformers with dynamic weight fusion for audio-visual detection, demonstrating strong performance across different datasets, although the computational demands and the necessity for high- quality synchronized data continue to be a limitation. Koutlis and Papadopoulos [18] presented DiMoDif, which utilizes feature pyramids and local attention to identify small fake intervals, but achieving precise audio-visual alignment is essential for its effectiveness.

The work by Koutlis and Papadopoulos [18] presented a framework (DiMoDif) that exploits feature pyramids and local attention mechanism to capture faint fake intervals. However, the successful application of this approach heavily depended upon accurate audio–visual synchrony. Katamneni and Rattani [16] also investigated modality- invariant and modality-specific representations for improving the detection efficiency, but it was still sensitive to mismatch among modalities, with high costs in the training phase. Oorloff et al. [19] proposed an audio– visual fusion model by self-supervisely pretrained approach, achieving high accuracy and AUC values [21]. However, performance decreased when one or both modality types were absent or not continuous with each other.

Vision Transformer-based methods [22] have shown remarkable accuracy in identifying manipulated faces within controlled datasets; however, their effectiveness diminishes in real-world or 'wild' environments, especially when faced with compression or variations in artifacts.

In order to identify subtle forgeries, Gao et al. [5] concentrated on forecasting temporal variables for bimodal detection. It is still challenging to identify forgeries that are extremely brief or misplaced.

A comprehensive analysis of machine learning and fusion-based detection techniques was presented by Gupta et al. [23]. They emphasized promising multimodal strategies, gaps, and trends. They also pointed out the paucity of empirical validation and the absence of real-world testing.

These investigations highlight persistent issues such limited multi-modal integration, high processing requirements, sensitivity to noise and compression, and dependence on unimodal datasets. These problems highlight the necessity of unified frameworks that efficiently integrate audio, video, and picture data to enhance interpretability, scalability, and strength in practical deepfake detection.

IV. METHODOLOGY (PROTOTYPE BASED)

To take forward the noted research insufficiency in unimodal deepfake detection, we created a epitome based on a single modality image as a preparatory measure initial to writing a comprehensive multimodal system. The purpose of this prototype was to verify the model's efficacy and to reveal limitations that involve the integration of multiple modalities.

1. Dataset Preparation

For the early version of the prototype, we are working with the Celeb-DFv2 dataset, which is made up of high-resolution facial video frames taken from both authentic and manipulated videos. The dataset encounters the following preprocessing steps:

a. **Frame Extraction:** The videos were converted to separate specific frames at 30 fps.
b. **Normalization of the Data:** Pixel values were regularized in range [0, 1] to normalize input data.

c. **Resizing:** The frames were re-sized to 224×2 pixels representing the input of the CNN architectures.

d. **Image Augmentation:** We made use of the random horizontal flipping, rotation and brightness alteration to augment data for better generalization ability.

The dataset was divided into 80% for training, 10% for validation, and 10% for testing, ensuring an equitable representation of both real and fake samples.

2. Model Architectures

Two Convolutional Neural Network (CNN) architectures were developed and assessed:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141065

a. Custom CNN Architecture

A tailored CNN was created as a baseline model. The architecture includes:

i. **Convolutional Layers:** Several layers featuring 3×3 kernels and ReLU activation functions to capture spatial hierarchies.

ii. **Max-Pooling Layers:** 2×2 pooling layers designed to decrease spatial dimensions and lower computational demand.

iii. **Fully Connected Layers:** Dense Layers are the that leading to a give us a finalizing output layer with a sigmoid initiating function which helps us for binary classification.

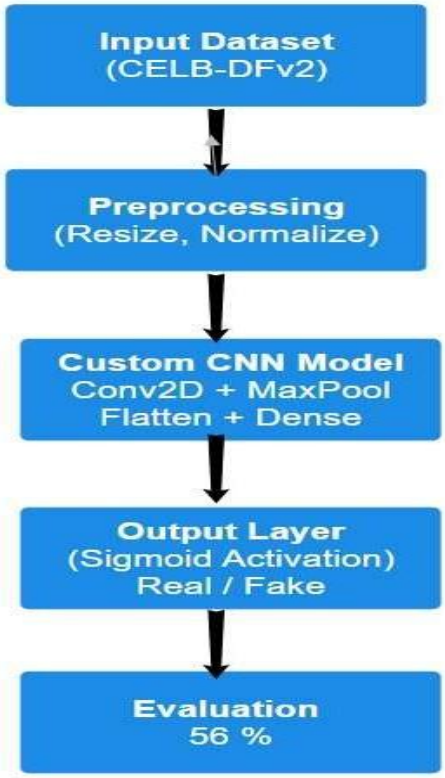

Figure 1: Custom CNN Based Model Workflow This model was chosen for its simplicity and successfulness, which was allowing us for a quick prototype designing and initial performance evaluation of that prototye.

b. Xception CNN Architecture

We used the Xception model, which is based on an Inception style architecture known for a high level of accuracy in image classification contests. The some Notable attributes includes :

i. **Depthwise Separable Convolutions:** These convolutions distinguishes the filtering and combining process for a better model efficiency and performance improvement of our prototype model.

ii. **Residual Connections:** The shortcut connections between layers that makes easier the flow of gradients and help to mitigate and minimize the risk of disappearing gradients.

iii. **Pre-trained Weights:** The model was created using the weights that were pre-trained on the ImageNet dataset, which was allowing us to earn benefit from the already learned features.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141065

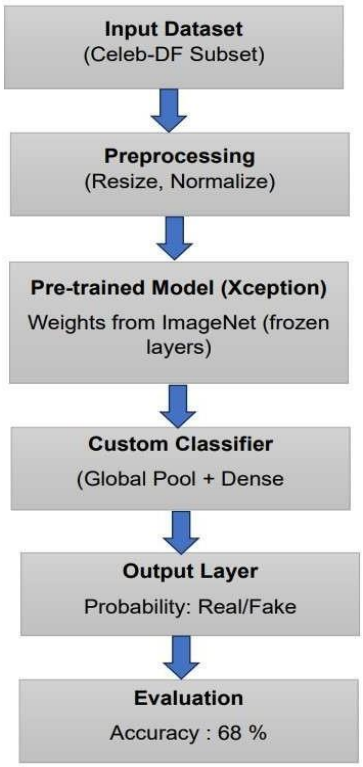

Figure 2: Xception CNN Based Model Workflow

The Xception architecture was chosen to test its efficiency in deepfake detection tasks after giving a practical exhibition and explanation of exceptional performance in a number of picture classification tests.

3. Training and Evaluation

The following conditions were used to train models:

i. **Optimizer:** We had applied 0.0001 learning rate to the Adam optimiser.

ii. **Loss Function:** The binary classification problem was addressed using binary crossentropy loss.

iii. **Batch Size:** There were 32 samples in each batch.

iv. **Epochs:** To avoid the overfitting problem, we have implemented the early halting based on validation loss during a total of 50 epochs.

From all the evaluation metrics were:

I accuracy: Accuracy is proportion of samples that were correctly classified in the model.

II Precision: Precision is the percentage of actual positive forecasts from all predictions with a positive label.

III Recall: The number of positive things out of all And every were really right that came from a true place.

IV F1-Score: The average of recall and precision.

Confusion matrices were generated in order to study types of misclassification, and provide information regarding the performance of the model.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141065

4. Observations and Limitations

a. Custom CNN: This model performed just fairly well and showed that it is not feasible for simpler architectures to efficiently discern the complexities of deepfake images.

b. Xception CNN: Showed an improvement, however, faced the same issue of not generalising well to other synthesis techniques highlighting that more robust models were required.

These results underscore the need for advanced networks capable of learning complex patterns from different modalities and pave the way to a multimodal detection tool,

V. PROPOSED MULTI-MODEL FRAMEWORK

i. **Overview of the System** The recent works attempting to detect deepfakes have been largely observed for unimodal approaches separately analyzing modalities like visual, audio or text. Works like [24], [2] and [25] showed promising accuracy in image-level detection with convolutional or transformer- based networks. Yet, unimodal methods break down when the manipulations involve several modalities : voicecloned videos or lip-synched speech [26], [27].

In order to solve these problems, we are developing a new mult-modal deepfake detection approach that uses the intermodal correlation of images, videos and audios. Leveraging multimodal fusion methods proposed in [5], [6], [12] our approach considers multiple data streams, utilizes dedicated neural networks for feature extraction, and combines them using a fusion layer to improve the reliability of detection and generalization.

The framework can be expected to better realize the explainability, robustness and adaptability to practical applications as [3] and [28]. The framework consists of segments for encoding data, classifying, and evaluation.

ii. Data Flow Diagram (System Architecture The Data Flow Diagram (DFD) in Figure 3 illustrates the multi-level data movement within the system, inspired by hierarchical designs from [16], [17], [29], [30].

a. Level 0: Defines a top-level process that accepts multimodal input (video, audio, or text) and generates classification results.

b. Level 1: Expands the structure into preprocessing, model training, and prediction modules.

c. Level 2: Implements modality-specific feature extraction pipelines — a CNN for visual data [20], spectrogram-based feature learning for audio [4], and transformer embeddings for text [15].

d. Level 3: The mutual fusion and ensemble prediction are integrated to improve the interpretability and model explainability, which combines refinements from [13] and [14].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141065

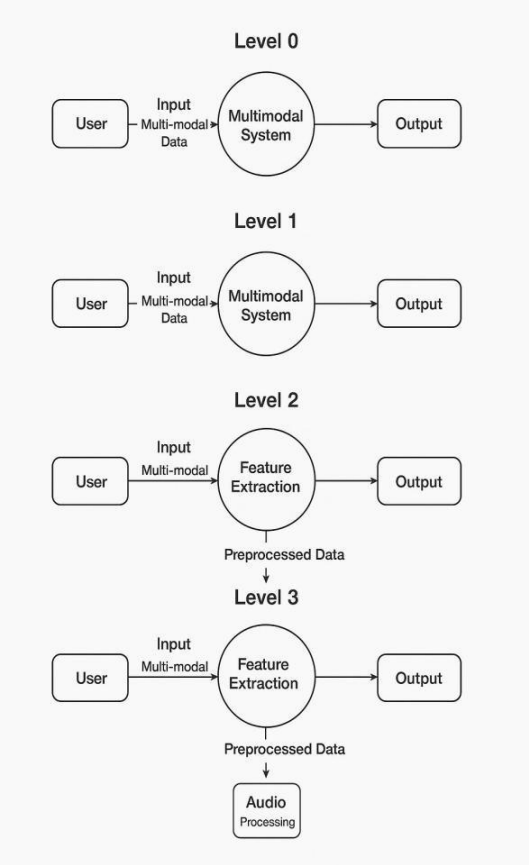

Figure 3: Data Flow Diagram of the Proposed Multimodal Deepfake Detection Framework.

This architecture is hierarchical and modular, making this framework easy to incorporate new models or datasets as demonstrated in [11], [21].

iii. Functional Description (Use Case Diagram) Fig.4 shows the Use Case Diagram, which describes the operational framework and user interactions of the system that refers to the functional architecture proposed by [31] and [32]. Two key players—the Admin and the Researcher/User—interact with key modules:

a. Admin: Responsible for user management, updating datasets and configuring the system.

b. Researcher/User: responsible for user management, updating datasets and configuring the system. These scenarios were designed based on the modular development methodology proposed in [1,33] that allows for scalability and continual model improvement.

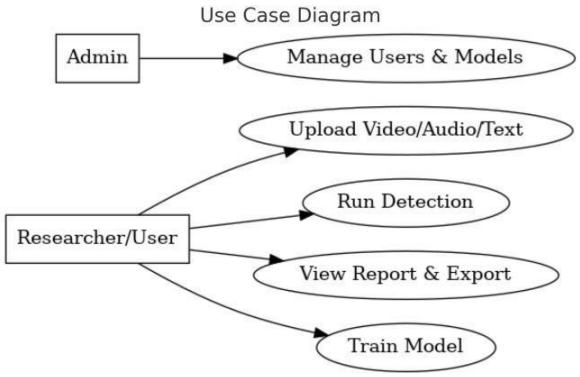

Figure 3: Use Case Diagram Representing System Functionality and User Interaction.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141065

iv. Workflow Integration

The DFD, as well as Use Case Diagram, demonstrates a complete technical and an operational aspect of the new system. DFD shows the internal structure, data exchange and flow of information while Use Case Diagram exhibits how users interact with one another and with the system. Taken together, they form a combined end-to-end framework for manipulated content detection across all modalities (visual, auditory and text), filling the gaps in existing unimodal frameworks [2], [4], [5], [6], [24]-
[28],[13]-[15],[34].

VI. CONCLUSION

This study reviewed over 20 papers related to deepfake detection in images, videos and audios. Unimodal methods like CNNs for image forensics, temporal models for videos and audio spectrogramLSTMs perform well on clean datasets but fail when dealing with noisy/compressed/multi-modal data.

Our prototype implementation with handcrafted and Xception based CNN also validates these bounds. Multi-modal approaches, which fuse image, video and audio through attention mechanism, can be used to improve robustness, interpretability and generalization; but there remain issues of computation and dataset alignment. It's important we integrate multiple modalities to be able to have effective, scalable and robust deepfake detection so that the digital media is verifiable and public trust is enhanced."

Acknowledgements

The authors would like to thank Prof. Sunil Yadav Sir, of the Department of Computer Engineering at Dr. D Y Patil College of Engineering & Innovation in Varale, Talegaon, Pune, for his priceless guidance and precious tips, wholehearted endorsement, and perceptive recommendations during this research.

For their encouragement, academic contributions, and assistance during the creation of this work, we are also incredibly grateful to the entire faculty of the Department of Computer Engineering at Dr. D Y Patil College of Engineering & Innovation, Varale, Talegaon, Pune.

Our experimental approach has advanced thanks to the efforts of open-source research communities and public datasets like FaceForensics++ and Celeb-DF which the authors recognize.

REFERENCES

[1]. Reality Defender. (2025). Reality Defender Launches Public API and Free Tier to Bring Enterprise-Grade Deepfake Detection to Every Developer. PR Newswire.
[2]. Li, Y., Chang, M., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. IEEE International Workshop on Information Forensics and Security (WIFS).
[3]. Elsevier. (2025). Advances in Audio Deepfake Detection. Information Fusion.
[4]. ArXiv. (2025). Audio Deepfake Detection Using Spectrogram and LSTM.
[5]. Huang, Z., Li, Y., Chen, X., Wu, S., & Lin, H. (2025). SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Multi-Modality. CVPR.
[6]. Guo, M., Li, S., Chen, Y., Wang, X., & Zhang, Q. (2025).
Rethinking Vision-Language Model in Face Forensics: MultiModal Interpretable Forged Face Detection. CVPR.
[7]. Zhang, K., Pei, W., Lan, R., Guo, Y., & Hua, Z. (2025). Lightweight Joint Audio-Visual Deepfake Detection via Single-Stream Multi-Modal Learning Framework. ArXiv
[8]. Intel. (2025). FakeCatcher: Detecting Deepfakes Using Subtle Blood Flow Signals in Video. Lifewire.
[9]. GetReal Labs. (2025). Real-Time Deepfake Detection Technology.
[10]. Wired. (2025). The Rise of Live Video Deepfake Detection Tools.
[11]. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. arXiv.
[12]. AAAI. (2020). Deepfake Video Detection Using Temporal Features. AAAI Conference on Artificial Intelligence.
[13]. ArXiv. (2025). Transformer-based Video Forensics.
[14]. Springer. (2025). Hybrid CNN-RNN for Deepfake Detection.
[15]. ArXiv. (2025). Cross-Modal Attention for Forgery Detection.
[16]. Katamneni, V. S., & Rattani, A. (2024). Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization. arXiv.
[17]. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Lim,S. N., & Jiang, Y.-G. (2024). Dual Transformers with

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141065

Dynamic Weight Fusion for Audio-Visual Deepfake Detection. arXiv.
[18]. Koutlis, C., & Papadopoulos, S. (2024). DiMoDif: Audio-Visual Deepfake Detection using Feature Pyramids and Local Attention. arXiv.
[19]. Oorloff, R., et al. (2024). AVFF: Self-Supervised AudioVisual Feature Fusion for Deepfake Detection. arXiv.
[20]. ScienceDirect. (2025). Fusion Model on Audio-Video Deepfake Detection.
[21]. Springer. (2025). Video Transformer-based Deepfake Detection.
[22]. MDPI. (2024). Vision Transformer-Based Approaches for Deepfake Detection: Performance and Challenges.
[23]. Gupta, A., et al. (2024). A Comprehensive Review of Machine Learning and Fusion-Based Deepfake Detection Techniques. arXiv.
[24]. Ahmed, N. R. (2024). Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects. IEEE.
[25]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. arXiv.[26]. Nguyen, H., Yadav, R., Nguyen, T., & Hoang, T. (2019). Capsule-Forensics: Using Capsule Networks for Detecting Forged Images and Videos. arXiv.
[27]. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. arXiv.
[28]. Springer. (2024). Survey on Multi-Modal Deepfake Detection. Multimedia Tools and Applications.
[29]. Khan, N., Nguyen, T., Bermak, A., & Issa, K. (2025). CAMME: Adaptive Deepfake Image Detection with MultiModal Cross-Attention. arXiv.
[30]. Springer. (2024). Image Forensics with Deep Learning.
[31]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[32]. TensorFlow/Keras. (2025). Convolutional Neural Network Guide. Keras Documentation.
[33]. Francois, C. (2017). Xception Model in Keras for Image Classification. GitHub Repository.
[34]. AAAI. (2020). Deepfake Video Detection Using Temporal Features. AAAI Conference on Artificial Intelligence.
[35]. Ahmed, N. R. (2021). DeepFake Video Detection: A Review. Semantic Scholar.
[36]. Durall, R., Keuper, M., et al. (2020). Unmasking DeepFakes with Simple Features. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
[37]. OJS AAAI. (2019). GAN Artifact Detection via Patch Analysis. AAAI Conference on Artificial Intelligence.
[38]. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. IEEE International Workshop on Information Forensics and Security (WIFS).
[39]. Ahmed, N. R. (2021). DeepFake Video Detection: A Review. Semantic Scholar.
[40]. ScienceDirect. (2025). Fusion Model on Audio-Video Deepfake Detection. Information Fusion.