International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I11P130

Volume 74 Issue 11, 460-479, November 2026
© 2026 Seventh Sense Research Group®

Review Article

Multimodal Deepfake Detection using Deep Learning and Explainable AI Integration: A Survey

Tejaswini Patil[1], Dipmala Salunke[2]

[1]*Computer Engineering, JSPM Rajarshi Shahu College of Engineering Pune, Maharashtra, India.*
[2]*IT, JSPM Rajarshi Shahu College of Engineering Pune, Maharashtra, India.*

[1]*Corresponding Author: tejaswini.phdwork@gmail.com*

Received: 12 May 2026 Revised: 10 September 2026 Accepted: 30 September 2026 Published: 25 November 2026

Abstract *- The development of generative models has been particularly fast, and the content generated through it can be quite realistic, raising crucial threats to digital trust, privacy and information integrity. Although the initial deepfake detection algorithms have mainly used unimodal visual-based features, recent studies have shown that unimodal features are frequently ineffective in detecting advanced manipulations that leave artifacts as minimally visible. This systematic review is a detailed review of multimodal detection of deepfakes based on hybrid deep learning models along with Explainable Artificial Intelligence (XAI) methods. The review combines the conclusions of the state of the art literature, including visual spatial aspects, temporal movement, and audio-visual consistency cues, and the synthesis of the three enhances the robustness and generalization of the study during various forms of forgery. The study discuss characteristic deep neural networks, such as convolutional Neural networks, recurrent and temporal networks, attention, multimodal fusion policies at feature, decision, and integrated levels. Moreover, the role of XAI techniques, including Grad-CAM, SHAP, LIME, and attention visualization, in improving model transparency, interpretability, and forensic reliability, which are the main requirements to apply in the real world in legal, journalistic, and security-sensitive settings, are critically assessed in the review. Benchmark data, evaluation metrics and typical experimental setup are carefully compared in order to determine performance trends, limitations and open challenges. The review ends with the future research directions set, such as the ability to resist hidden manipulations, the reduction of dataset bias, lightweight multimodal models, and human-understandable explanations.*

Keywords *- Deepfake detection, Multimodal learning, Hybrid Deep Learning, Explainable Artificial Intelligence, Audio-visual analysis, Media forensics.*

1. Introduction

The fast development of the artificial intelligence-based synthesis of media has radically changed the process of producing, editing, and sharing digital content. Of these developments, one of the most disruptive technologies has been deepfake technology, which uses deep learning models to create extremely realistic synthetic images, videos and audio. In current deepfake models, face swap, facial expression replay, control of lips, and voice cloning can be performed with convincingly plausible results with only a few perceptual artifacts.

Although these methods can be used in a legitimate manner in the entertainment world, virtual reality, accessibility and creative media production, it is the misuse of these methods that can cause grave dangers to the lives of individuals, institutions and the entire society. Deepfakes have been more and more closely linked to misinformation campaigns, political interference, fake identities, reputation loss, and a loss of confidence in digital media. To overcome these difficulties, the research community has put forward a vast variety of deepfake detection methods [1]. Early detection methods the majority of early detection methods were based on handcrafted features based on visual artifacts like colour inconsistencies, edge distortions, or frequency-domain anomalies. As deep learning has emerged, data-driven models, especially Convolutional Neural Networks (CNNs), have shown themselves to be better performers since they are able to discover discriminative spatial features in the face images and video frames with the help of automatic learning. Nevertheless, even with significant improvements in accuracy, most currently used detection systems are based on single-modality analysis, most commonly visual one. This type of unimodal approach has its own weaknesses since the state of the art deepfake generation models are becoming less and less discernible, and more capable of manipulating certain modalities selectively and leaving others relatively unaltered [2, 3]. Recent research has pointed out that deepfake works tend to create minor inconsistencies based on various modalities, such as visual appearance, time motions, and

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

audio-visual alignment. An example of this is a realistic appearance of facial textures and unnatural temporal behavior like blinking of the eye, or the movement of the head or a change in facial expressions.

Equally, voice cloning and lip-sync generation may cause a mismatch of the speech gesture and the lips motion and expose audio-visual bias. These observations have led to the transition to multimodal deepfake detection, in which complementary visual, temporal and audio streams of information are collectively used to enhance robustness and generalization. Multi-modal learning does not only improves performance in detection, but also minimizes vulnerability to attacks on one modality [4].

Transparency in prediction is problematic in high-stakes areas including journalism, law enforcement, digital forensics and social media governance since users cannot trust everything predicted, and these applications cannot be implemented practically. Second, most detection models lack robustness and generalization, especially when they are faced with latent manipulation strategies, heavy compression, low-resolution images or cross-dataset testing conditions. These constraints affirm the importance of having detection systems which are not only precise but also readable and resilient to the variations which occur in the real world [5].

Although there has been an increasing amount of literature that focuses on deepfake detection, the need to integrate multimodal detection, hybrid deep learning, robustness, and explainability into a single review framework is yet to be addressed. Existing studies are usually targeting a single feature (e.g., visual artefacts), a specific detection architecture or inconsistencies in time or synchronisation of AV artefacts. As a consequence, it is difficult to determine the complementarity of each modality, the impact of fusion strategies on accuracy of detection and the validity of explanations when the quality of the media and/or the nature of the media manipulation varies. Some of the limitations reported include a lack of compression sensitivity, a lack of cross-dataset generalization, noisy or missing audio, limited attacks on unseen model and the lack of quantitative evaluation of explanations. The problems point out the necessity to have a single evaluation that considers detection performance in addition to robustness, generalization and interpretability.

This novelty is not just in the modes of deepfake detection but in exploring them from these inter-related angles, while not focusing on the accuracy of the models. The review extensively explores the spatial, temporal and audio-visual evidence, as well as feature-, intermediate-, decision- and attention-based fusion strategies and hybrid deep learning architectures that have been used on benchmark datasets. It also examines three popular post hoc methods of reasoning about the decision to detect: Grad-CAM, SHAP and attention visualization in the context of their transparency and forensic interpretation of the decision made to detect. This paper aims to discuss the interplay between modality coverage, fusion design, robustness, cross-dataset generalization and explainability, rather than just generation methods or on the individual detection methods. The result of this synthesis reveals the current lack of methods and provides recommendations on how to address these to develop reliable and implementable multimodal deepfake detection systems. In order to deal with them, this paper suggests a Multimodal DeepFake Detection Framework based on a Hybrid Deep Learning Architecture with Explainable AI (XAI) Implementation.

The central concept is that there is joint spatial artifact modeling, modeling of temporal motion inconsistencies, and modeling of audio-visual cues of synchronization in a single hybrid architecture. Visual spatial characteristics detect fine-scale abnormalities in face, temporal models consist of dynamic discrepancies over video sequences, and audio-visual consistency modules are the ones that examine the manner in which speech and lip movements relate. These dissimilar representations are successfully combined to generate robust deepfake predictions [6, 7].

Notably, the framework incorporates explainable AI processes that can offer insights into model decisions that can be interpreted, which can allow the user to understand the modalities and features that contribute to a specific classification result most of all. Moreover, the paper focuses on explainability as a first-class evaluation goal, and traditional performance indicators. The proposed structure makes the explanations to be reliable, stable and concise by using quantitative explainability measures, including faithfulness, consistency and sparsity. Detectors are not just tested on the accuracy of detection, but their resistance to compression artifacts, adversarial manipulation, and inter-dataset generalization is also widely tested on diverse benchmark datasets [8, 9]. Given the richness in multimodal learning, this paper will explain how explainable AI and multimodal learning is an important step toward reliable, deployment-scalable deepfake detection systems that have the potential to deal with the future of synthetic media threats.

2. Review of Literature

The possibility of automatically learning both sophisticated and subtle artifacts of manipulation on large-scale data has made deep learning the dominant paradigm in deepfake detection. Initial research proved that Convolutional Neural Networks (CNNs) were able to detect spatial inconsistencies of fake images of faces with high efficiency compared to handcrafted feature-based methods. With the development of techniques of deepfake generation, more and more detection research focused on deeper architectures and representations based on data that can be adapted to new forms of manipulation. The most widely researched type of detection

model is the visual-only model [10, 11]. The techniques mainly are interested in manipulating spatial features, including confusion of boundaries, color differences, irregularities of texture and anomaly of frequencies added when faces are swapped or reenacted. Architectures based on CNN, such as residual and attention-based networks, have performed well on benchmark samples. Nevertheless, purely visual models are very sensitive to post-processing like compression, resizing and filters [12].

In addition, the development of generative models has caused visible artifacts to decrease by far, and visual-only detection becomes less reliable in the real world. Audio-only detection models pay attention to synthetic or manipulated speech with the help of acoustic characteristics, e.g., spectrogram shapes, phase errors, and speaker embeddings. CNNs and recurrent neural networks have been used as deep learning models to identify voice cloning and speech synthesis attacks. Although audio techniques are also useful in identifying forged speech, they do not work well in cases when the audio is original or non-existent, so they can only be used in video-focused deepfake detection [13]. Multimodal deepfake detection techniques are being of great interest to counter the unimodal constraints. These methods combine the analysis of visual, temporal and audio signals to use cross-modal discrepancies that are created during manipulation. Methods that involve the use of spatial CNNs with time sequence models and audio-visual synchronization detection have been shown to be more robust and generalize better. However, most multimodal systems are based on sophisticated fusion processes which are black-box, giving poor interpretability and preventing confidence in real-world applications. Most recently, Explainable AI (XAI) has been developed as a new direction in answering the transparency concerns of deepfake detection [14]. XAI methods give visual or feature-level explanations of the areas or indicators that may affect model choices. Table 1 presents the comparative deepfake detection procedures, datasets, modalities and the performance trend.

Table 1. Findings from literature review on deepfake detection

Primary Modality	Deep Learning Model	Key Features used	Fusion Strategy	Dataset(s)	Key Limitation Identified
Visual [10, 15]	CNN (MesoNet)	Spatial artifacts, textures	None (Unimodal)	FaceForensics	Poor robustness to compression
Visual [16]	XceptionNet	Pixel-level artifacts	None (Unimodal)	FaceForensics++	Limited cross-dataset generalization
Visual + Temporal	CNN + RNN	Eye blinking, motion cues	Late fusion	UADFV, DF-TIMIT	Sensitive to frame-rate variation
Audio [17, 18]	CNN	Spectrogram patterns	None (Unimodal)	ASVspoof	Ineffective without speech
Visual [19]	CNN + Attention	Facial regions, textures	None	FaceForensics++	Explainability not quantitatively evaluated
Visual + Audio [20]	CNN + LSTM	Lip motion, speech alignment	Feature-level fusion	FakeAVCeleb	Limited robustness to noise
Visual + Temporal	CNN + TCN	Temporal coherence	Late fusion	Celeb-DF	Struggles with unseen attacks
Visual [21, 22]	Frequency CNN	Frequency-domain artifacts	None	DFDC	Degraded under heavy compression
Visual + Audio [23]	Dual-stream CNN	Lip-sync features	Mid-level fusion	FakeAVCeleb	Black-box decision process
Multimodal [24]	Transformer-based	Cross-modal embeddings	Attention-based fusion	DFDC, Celeb-DF	High computational cost
Visual + Temporal	CNN + LSTM	Motion inconsistencies	Feature fusion	FaceForensics++	No explainability integration
Multimodal [25, 26]	CNN + Transformer	Spatial, temporal, audio cues	Late fusion	DFDC	Limited explanation consistency
Multimodal [27, 28]	Hybrid DL	Spatial, temporal, AV cues	Mixed	Multiple	Lack of unified XAI + robustness focus

There is a gradual shift in the reviewed literature from artifact-oriented unimodal detectors to multimodal detectors, which take advantage of complementary spatial, temporal and audio-visual information.

Visual CNN-based methods have proven to work well for facial texture and boundary artifacts, but they typically suffer from compression and domain changes. Temporal models enhance motion inconsistencies detection but can vary with the quality of the frames and the characteristics of the sequences.

Audio-visual methods give further evidence from lip-speech synchronization, and are influenced by noise, absence or manipulation of audio. Newer attention and transformer-based multimodal models have also been developed, which enhance the cross-modal representation and adaptive fusion, but still have challenges regarding interpretability and computational complexity.

Additionally, explainability is often added as an additional visualization instead of being assessed for faithfulness, stability and consistency. As a result, previous research confirms the need to perform multimodal analysis, but also shows a persistent need to investigate detection performance, robustness, generalization, computational needs, and quality of explanations in a multimodal way.

3. Methodology

This systematic review is based on the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA) guidelines to be able to provide a transparent, rigorous, and reproducible literature selection on the topic of multimodal deepfake detection through the lens of hybrid deep learning frameworks and explainable AI integration.

This number demonstrates the PRISMA-directed selection of the study, specifying the steps of identification, screening, eligibility, as well as the inclusion process as shown in Figure 1. Out of the 412 records initially found and released between 2021 and 2025, strict filtering and eligibility checks led to including 58 high-quality research in the final qualitative synthesis.

3.1. Identification

The identification phase was supposed to collect all the pertinent studies that contained the answer to the formulated research questions (RQ1-RQ6) about the deepfake generation methods, detection modes, feature extraction schemes, hybrid systems, explainable AI application, and evaluation customs.

The overall literature search was performed in IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Scopus, and Google Scholar. Only peer-reviewed journal and conference articles published in 2021-2025 were included in the search, so recent and technologically relevant articles could be found. The search criteria with keywords use as:

"deepfake detection" and "multimodal" and ("visual" or "temporal" or "audio") and "deep learning"

The combination of search terms included deepfake detection, multimodal learning, deep learning, and explainable AI. Research efforts needed to be focused specifically on deepfake detection and not deepfake generation itself. This was done to come up with 412 original records.

3.2. Duplicate Removal

The identification of all records was done, and all the records were consolidated by means of reference management software, which was then manually verified to eliminate duplication of records.

Duplicate studies that emerged because of indexing in different databases were eliminated at this point to avoid redundancy and bias of selection. The duplicate elimination had eliminated 96 records that had been left behind to undergo further screening.

3.3. Screening

In the screening phase, the titles and abstracts of the rest of the studies underwent screening to assess their applicability to the scope of this review. The studies were filtered out in case they had studied unimodal deepfake detection, used non-deep-learning or heuristic detection or did not have a clear technical contribution.

Also, non-peer-reviewed works, such as surveys, opinion articles, tutorials, and others, were left out. This filtering action eliminated 198 articles, and 118 papers were included in the process of full-text evaluation.

3.4. Eligibility

During the eligibility phase, full texts of the articles that were selected were thoroughly reviewed in accordance with predefined inclusion and exclusion criteria. Only the studies that propose or test multimodal feature extraction or fusion strategies, make use of deep learning or hybrid architecture, and offer experimental justification with the help of known datasets were included.

Moreover, research was anticipated to cover model interpretability or explainability, either via explicit methods of XAI or implicitly via attention-based methods.

Articles that were not written in a clear manner regarding the methodology, quantitative analysis or that were not pertinent to multimodal detection were eliminated. In this process, 60 articles were excluded.

3.5. Inclusion

Following the eligibility analysis, 58 studies met all the criteria of inclusion and were incorporated in the qualitative synthesis. All of these studies are the latest state of the art in multimodal deepfake detection, including the analysis of spatial, temporal, and audio-visual features, hybrid deep neural networks, explainable AI models, and benchmark evaluation.

4. Fundamental in Deepfake Framework

The fundamental framework discuss as a multimodal hybrid deep learning algorithm to identify deepfake video using visual, temporal and audio data. Video frames are initially systematically took out and preprocessed to even out the resolution, illumination, and facial position, and similar audio signals are denoised and divided up.

Convolutional networks are used to learn spatial visual features that represent small artifacts of the face, and sequence-based networks are used to learn motion patterns in time to detect unnatural dynamics between frames [29].

The audio-visual consistency is measured through comparing the speech cues to the movements of the lips. The multimodal features obtained are combined under a hybrid representation learning approach and finally classified by a classification module. Explainable AI methods are merged to explain the model choices, indicating contributions specific to modality and improving the level of transparency and user confidence [30].

4.1. Overview of System Architecture

A deepfake video is broken into video frames and audio streams at the input level and synchronized. Each frame detects and aligns the facial regions and directs the analysis on the regions being manipulated, and audio signals are synchronized with the relevant video segments in time [31].

There are three branches in the feature extraction stage. The former branch learns fine-grained facial texture and structural inconsistencies by training the first branch on visual spatial features with the help of convolutional neural networks. The second branch represents the temporal motion properties by examining the frame dependencies to identify unnatural face dynamics and motion artifacts.

The third branch concerns the audio-visual consistency features and studies how the speech signals and the lips are correlated to determine anomalies that are associated with the synchronization.

The extracted features of all branches are sent to a multimodal fusion module where complementary information is merged through feature-level fusion and attention mechanisms [32].

The fused result is then sent to the classification module, which delivers the likelihood that a video is either authentic or manipulated.

An explainable AI layer is also added to the architecture to increase transparency. This layer will produce modality-sensitive explanations, pointing out powerful spatial and temporal units, and sounding points that led to the final decision [34]. The modular architecture enables solid detection and, at the same time, provides interpretability and readiness to deploy it in practice.

4.2. Multimodal Data Acquisition and Preprocessing

Multimodal information collection and preprocessing are important factors that guarantee effective deepfake detection. The system also manipulates both audio and visual parts of input videos to extract complementary artifacts of manipulation.

Video data acquisition is a process of capturing frames at a constant rate in order to maintain time continuity whilst being computationally efficient. Localization of facial regions of different frames is done with face detection and tracking algorithms, and facial alignment is done to standardize the pose, scale and orientation [34].

Frames are re-sized, normalized and color-aligned to minimize illumination changes and compression artifacts that can bias feature learning. To analyse time, frame sequences are built by sliding windows to maintain dynamic motion like the blinking of the eye, the movement of the head and transitions of expression.

Figure 2 presents audio-visual preprocessing pipeline that allows multimodal deepfake detection. This chain representation allows the model to identify the inconsistencies that have been added during frame-level manipulation or synthesis. Audio data acquisition is the process that involves getting the audio track of every video and matching it with the visual portions as well [35].

The preprocessing methods are reduction of noise, normalization of amplitude, and segmentation into short-time frames. Temporal-frequency representations, i.e. spectrograms or mel-frequency features, are created to represent acoustic patterns related to synthetic or cloned speech. In order to facilitate the audio-visual consistency analysis, one forces the temporal correspondence of audio portions and video frames. The methods of data augmentation such as compression simulation, scaling of resolution and noise injection are used to enhance the resilience in real-world settings [36, 37]. The multimodal preprocessing pipeline is standardized to guarantee uniformity in the quality of the inputs so that the extraction of features can be done reliably and cross-dataset generalization enhanced.

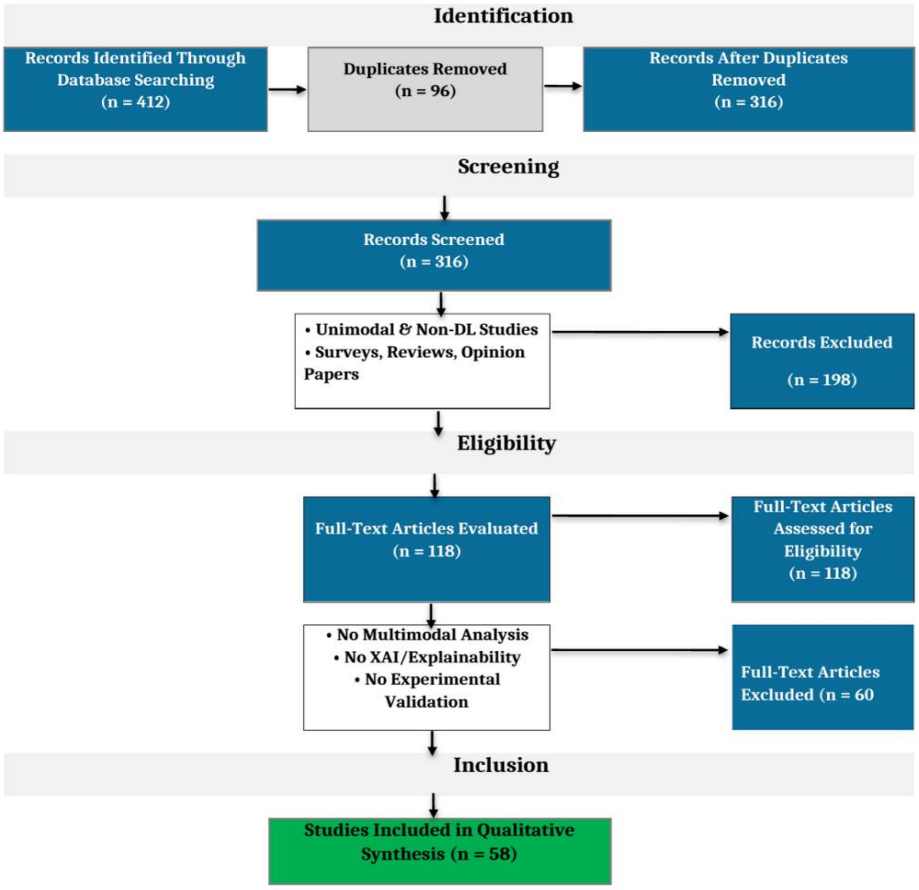

Fig. 1 PRISMA flow diagram for study selection in multimodal deepfake detection

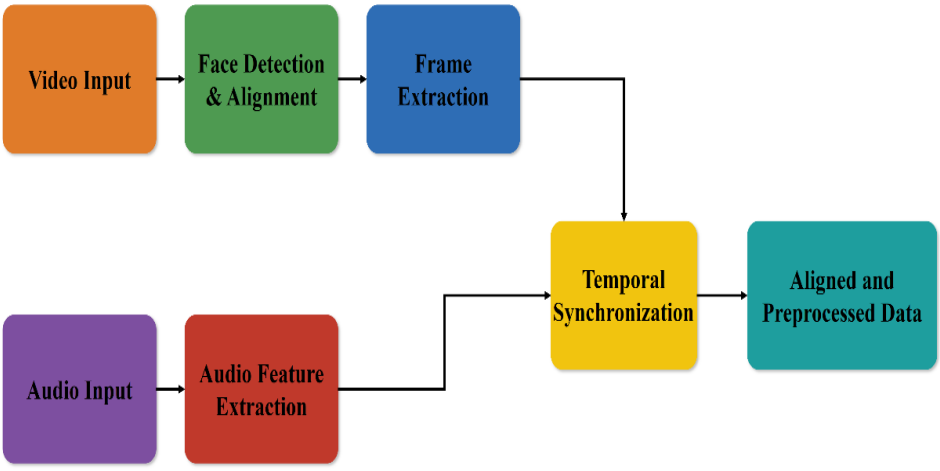

Fig. 2 Multimodal data acquisition and preprocessing pipeline for deepfake detection

This module takes aligned facial frames and learns hierarchically the representations of the anomaly automatically by using deep convolutional neural networks. Conventional layers are early enough, and thus capture the low-level features including edges, noise, and color discontinuities, whereas the later layers encode the higher-level features, e.g., continuity of skin texture as well as the regions of eyes and mouth [38]. Multi-scale feature extraction is used to increase the sensitivity to traces of manipulations: this enables the network to process both local and global facial patterns. Mechanisms of attention are combined to highlight areas that are visually informative, like the eyes, lips and face

outlines, which are often the target of deepfake synthesis. The regularization and normalization techniques of features are used to make features robust to compression and changes in resolution. The resultant spatial feature maps contain discriminative data that distinguishes genuine faces and synthetically produced faces [39]. These spatial representations constitute an extremely important part of the multimodal system, and this offers very compelling evidence of visual tampering to supplement both temporal and audio-visual indications.

4.2.1. Temporal Motion Features

Temporal motion feature extraction is designed to notice dynamic discrepancies between successive video frames that is caused by imprecise motion synthesis in deepfake videos. Even though single frames might be deemed visually persuasive, the generation models of deepfakes commonly fail to provide realistic temporal consistency in dynamic facial expressions, motion, and head in the human physiognomic or facial motions.

The module is able to work with sequences of aligned facial frames with the help of temporal deep learning models that can model long-range dependencies. Frame-level visual information is initially acquired and arranged into time sequences [40].

Motion patterns that are related to natural face behavior are then learned in these sequences using recurrent convolutional architecture or temporal convolutional architecture. The model puts attention on less noticeable temporal clues, including blinking of the eyes, the pathways of unnatural lip movements, inconsistent shifts in head posture, and sudden shifts in facial expressions.

With an understanding of the temporal dynamics, the network will be able to differentiate between smooth and biologically plausible motion and the artificial aspects of motion that could have been added in cases of interpolating between frames or reenacting motion.

The learning of temporal features is also enhanced by the sequence normalization and the temporal augmentation strategies, which provide simulations of changes in frame rate and video quality [41]. The resulting temporal motion features complement spatial cues, and they help a lot in rendering the detection more robust, particularly in situations in which visual artifacts are reduced at the cost of motion anomalies.

4.2.2. Audio-Visual Consistency Features

Audio-visual consistency feature extraction is concerned with the detection of consistency and semantic congruency between speech cues and the facial expression they correspond to. Deepfake videos where voice cloning or lip-sync generation is employed frequently add small discrepancies between the audio and visual pronunciation, although each of the two modalities may be realistic on its own. This module combines audio characteristics and visual motions of the mouth to determine the cross-modal coherence.

The audio cues are converted into a time-frequency representation, which reflects phonetic and prosodic properties of speech. Simultaneously, the visual characteristics describing the lips shape, mouth opening, and the articulation dynamics on the face frames are identified. It uses a cross-modal learning mechanism that is used to model temporal correlations between audio and visual representations [42].

The system analyzes the observed lips movements in terms of correspondence with expected speech patterns, timing, and intensity. A mismatch in speech intensity, unnatural articulation pattern, or delay in the motion of lips are all signs of manipulation.

The ensuing audio-visual uniformity characteristics help a great deal in strength against sophisticated forgeries and contribute immensely to the multimodal fusion plan, which allows dependably detecting various manipulation conditions [43].

Table 2 is a summary of deepfake detection techniques based on spatial, temporal, and audio-visual features, which discusses their fundamental techniques, strengths, and weaknesses. The comparison reveals that the multimodal hybrid methods are always more robust and generalized as a result of the combined exploitation of the facial artifacts, motion inconsistency, and cross-modal synchronization cues.

Table 2. Comparative analysis of feature extraction strategies for deepfake detection

Method / Approach	Feature Type	Core Technique	Strengths	Limitations	Typical Datasets	Ref. Paper
CNN-Based Spatial Analysis	Visual Spatial	Deep CNN (ResNet/VGG)	Captures fine-grained facial artifacts	Sensitive to compression	FaceForensics++	[30]
Multi-Scale CNN	Visual Spatial	Multi-scale convolution	Detects local & global anomalies	Higher computation cost	Celeb-DF	[32]
Attention-Based CNN	Visual Spatial	Spatial attention mechanisms	Focuses on eyes, lips, boundaries	Requires large training data	DFDC	[33]

RNN-Based Motion Modeling	Temporal Motion	CNN + LSTM	Models long-term facial dynamics	Training instability	FaceForensics++	[35]
Temporal CNN (TCN)	Temporal Motion	1D/3D temporal convolution	Efficient temporal learning	Limited long-range context	Celeb-DF	[36]
Optical Flow Analysis	Temporal Motion	Motion vector estimation	Explicit motion inconsistency detection	Noise-sensitive	UADFV	[37]
Lip-Sync Feature Matching	Audio-Visual	Mouth motion vs speech	Effective for lip-sync forgeries	Language dependent	LRS2	[38]
Audio Spectral + Visual Fusion	Audio-Visual	MFCC + CNN features	Detects subtle cross-modal mismatch	Sync alignment needed	DFDC	[39]
Cross-Modal Transformer	Audio-Visual	Attention-based fusion	Strong semantic consistency modeling	High computational cost	FaceForensics++	[41]
Multimodal Hybrid Framework	All (Spatial + Temporal + AV)	CNN + LSTM + Fusion	High robustness across attacks	Complex architecture	DFDC, Celeb-DF	[42]

4.3. Hybrid Deep Learning Model

The hybrid deep learning model is developed to be able to combine heterogeneous features derived by visual, temporal, and audio-visual modalities into a single decision-making system. Considering the variety of these features, the architecture takes a multi-branch form in which every modality-specific representation is initially processed in isolation, and then collectively reasoned about [40]. Deep convolutional layers are used to fine-tune spatial visual features to retain local texture and structural features, whereas sequence-aware layers represent temporal motion features by modeling long-term dependencies between video frames [41]. Cross-modal embedding layers are used to extract audio-visual consistency features to encode patterns of synchronization as well as semantic alignment [42].

To facilitate interaction among modalities, the model uses a feature-level fusion strategy that is complemented with attention mechanisms [44]. This enables the network to adjust the contribution of modality by the quality of input and the type of manipulation, making it robust to different conditions, including the absence of audio or poor visualization. The comparative analysis presented in Table 3 indicates that the hybrid multimodal deep learning models performed significantly better than the unimodal ones because they successfully combine visual, temporal and audio-visual characters. Attention-based fusion and transformer-based architectures are more robust and flexible to a variety of manipulations of deepfakes, whereas lightweight models compromise to detection performance with the ability to run on lightweight hardware.

Table 3. Summary analysis of hybrid deep learning models for multimodal deepfake detection

Model Type	Modalities Used	Core Architecture	Fusion Strategy	Attention / XAI Support	Key Strengths	Limitations
CNN-Based Multimodal Model	Visual + Audio	CNN + FC layers	Feature-level fusion	Grad-CAM	Strong spatial artifact detection	Weak temporal modeling
CNN-LSTM Hybrid	Visual + Temporal	CNN + LSTM	Feature- level fusion	Temporal attention	Captures long-term motion patterns	Higher training cost
CNN-TCN Framework	Visual + Temporal	CNN + TCN	Intermediate fusion	Attention weighting	Efficient temporal learning	Limited global context
Audio-Visual Sync Model	Audio + Visual	CNN + RNN	Late fusion	Lip-sync attention	Effective for lip-sync forgeries	Audio quality sensitive
Multimodal Attention Network	Visual + Audio + Temporal	Multi-branch CNN-LSTM	Attention-based fusion	Self-attention maps	Adaptive modality weighting	Complex architecture

Transformer-Based Multimodal Model	Visual + Audio	Vision + Audio Transformers	Cross-modal fusion	Attention visualization	Strong semantic alignment	High computational cost
Graph-Based Multimodal Model	Visual + Temporal	CNN + GNN	Feature fusion	Node-level explainability	Models relational dependencies	Graph construction overhead
Hybrid CNN-Autoencoder	Visual + Audio	CNN + Autoencoder	Feature-level fusion	Reconstruction maps	Learns subtle anomalies	Limited temporal cues
Lightweight Multimodal DL	Visual + Audio	Compact CNN	Decision-level fusion	Saliency maps	Suitable for real-time use	Lower accuracy
Proposed Hybrid Framework	Visual + Temporal + Audio	CNN + LSTM + Attention	Adaptive feature fusion	XAI-integrated	Robust across manipulation types	Increased model complexity

5. Background on Deepfake Technology

Deepfake technology is an AI-based method of synthesizing or editing visual and audio material in order to produce files that look and sound extremely natural yet fake. It mainly uses deep learning models that have been trained to learn intricate facial appearance, movement and speech patterns using huge datasets. Early deepfakes were concentrated on basic face replacement, whereas the latest developments allow performing high-fidelity face reenactment, expression control, and voice cloning.

Such systems take advantage of generative models to copy the identity, expressions or speech features of one person to another with few perceptual artifacts [45]. Despite the honest purposes of deepfake technology in entertainment, virtual avatars, and the production of digital content, its abuse is a severe problem in terms of ethical, legal, and security concerns. The development of efficient and effective detection systems through the detection of the underlying dynamics of deepfake generation is important in addressing the rising threats of synthetic media.

5.1. Types of Deepfake Generation Techniques

Deepfake generation methods represent a variety of AI-driven methods that are used to synthesize and manipulate visual and audio content in high levels of realism, classification represent in Figure 3. These methods can be broadly divided into image based, video based and audio-based generation techniques.

Image-based deepfakes specialize in synthesizing or in modifying static facial images through the transfer of identity-related information like facial structure, skin texture, and expression. Deepfakes based on video take this idea to the next level by achieving temporal consistency of the frames so that it is possible to create realistic facial movements, transitions of facial expressions, and head movements. The audio-based methods produce or alter speech signals to emulate the voice traits of a target speaker, such as pitch, tone, and style of speaking [46].

The other valuable classification is on the basis of the learning paradigm through which synthesis is done. The supervised methods are based on pairs of source and target identities, and the unsupervised/weakly supervised methods are learning mappings without directly matching them. Recent methods are progressively using self-supervised learning as a method to minimize the amount of data needed and enhance scalability.

Hybrid pipelines used to generate deepfakes can also include the addition of facial manipulation with post-processing tasks like blending, smoothing and enhancing to increase visual quality. As these methods keep developing, the resulting content turns more believable, and the artifacts become more minimalistic [47]. Such extensive development makes it more difficult to detect, since the old visual indicators are not so secure. It is important to know the diversity of methods of deepfakes generation, thus, to develop detection systems that can detect a variety of manipulation strategies of images and audio modalities.

5.2. GAN-based and Autoencoder-based Deepfakes

Most modern deepfake generation systems are based on GANs and autoencoder architectures. Deepfakes based on GANs use a competition model of a generator and a discriminator. The goal of the generator is to generate extremely realistic artificial images or videos, whereas the aim of the discriminator is to differentiate between the real and the fake samples. Trained by iteration, GANs are trained to produce visually believable textures, consistency in lighting, finesse in faces, etc. More sophisticated versions of GANs also increase the level of realism to high-resolution facial expressions and videos to track temporal coherence. Figure 4 illustrates GAN and autoencoder processes to create synthetic deepfakes. Autoencoder-based deepfakes, in turn, are based on encoder-decoder designs that are trained to learn small latent images of faces. In manipulating faces, a common encoder is normally trained to encode the common face structure, and the distinct decoders recode the faces associated with different faces.

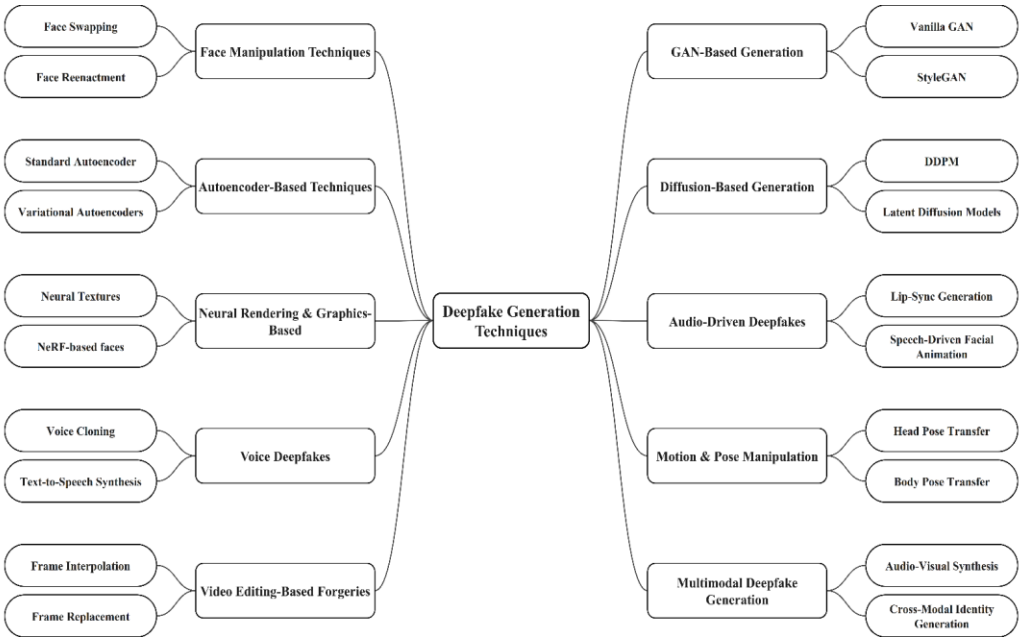

Fig. 3 Different types of deepfake generation techniques

Using the replacement of decoders, the expression and movement can be maintained, and still, the identity of a face can be transferred. The methods that use autoencoders are computationally efficient and were one of the first methods of the generation of deepfakes, but they tend to exhibit artifacts on the edges of faces, as well as fail to handle large differences in pose. GAN-based and autoencoder-based methods are both ongoing to enhance themselves in terms of architectural advancement and scale training data. Although GANs tend to be more visual realistic, autoencoder-based models are also very popular because they are easier to use. Detection systems should then take into consideration the artifacts and inconsistencies that both paradigms have brought in order to have complete coverage of the various deepfake generating strategies.

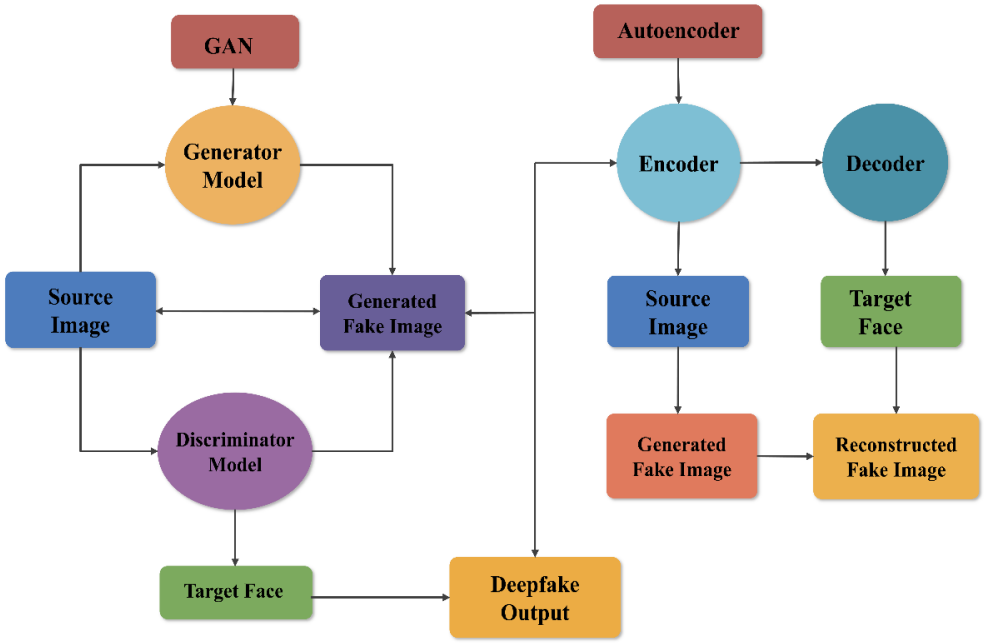

Fig. 4 Flowchart of GAN-based and autoencoder-based deepfake generation processes

5.3. Face Swapping, Reenactment, and Voice Cloning

The most frequent application-level manifestations of the technology of deepfakes are face swapping, facial reenactment, and voice cloning. Face swapping is the replacement of the face of a target person in a picture or video with a source identity whilst keeping the pose, expression and background. This method is common in deepfakes that are visual and can tend to place smearing-like artifacts or geometric anomalies or color discontinuities, especially where boundaries to faces are discontinuous. Facial reenactment transcends identity replacement because it involves an imparting of expressions, movement of lips and head dynamics on a target identity through the use of a source actor.

This technique is also commonly used in video manipulation and animated avatars, and it demands proper modeling of facial movement and temporal coherence. Although reenactment techniques can achieve very realistic effects, they can have problems in maintaining the natural patterns of motion of a variety of expressions, resulting in temporal inconsistencies. Voice cloning is based on the production of speech that replicates the vocal features of a target speaker. Deep learning algorithms train speaker-specific representations which include pitch, timbre and articulation style to produce synthetic speech using text or the audio of another speaker, temporal and audio information.

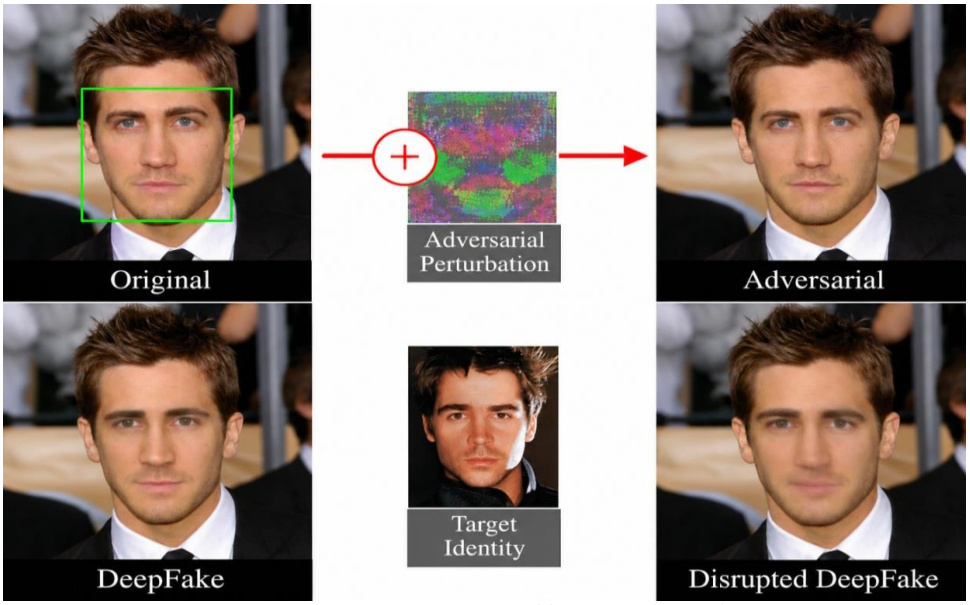

Fig. 5 Adversarial perturbation-based disruption of face-swapped deepfake generation [48]

Figure 5 shows that adversarial perturbations put on an original facial image disrupt the deepfake generation pipeline. Although the standard face-swapped deepfake is far more similar to the original identity of the person, the adversarial perturbed input can lead to a more visible deterioration of the generated image, indicating that deepfake models are susceptible to well-placed perturbations and can be employed to mitigate deepfaking.

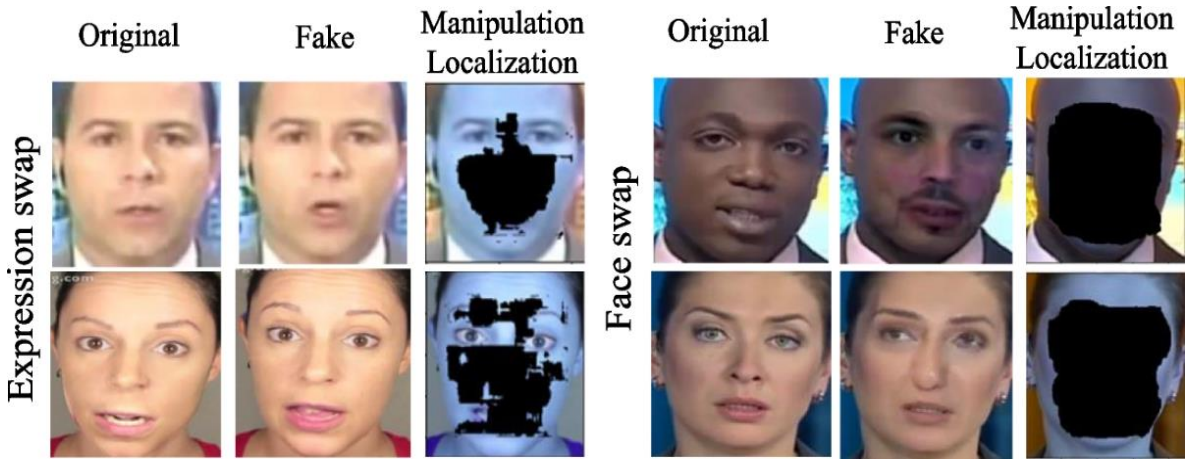

Fig. 6 Face swapping - visual artifacts and boundary inconsistencies [49]

Figure 6 illustrates some archetypal examples of face-swapping and expression-manipulation deepfakes and the accompanying localization maps of manipulation. Although the counterfeit images are visually deceptive, the localization outcomes provide consistency irregularities of the boundaries, texture fusion and wrinkled areas of the faces. These areas of highlight demonstrate spatial artifacts that are added in the process of transferring identity and expression, which justify the success of deepfake detection methods based on visual artifacts.

6. Deepfake Video Detection Techniques

The point of deepfake video detection methods is to detect synthetic manipulations by examining the artifacts and the inconsistencies that are added at the time of generation. The initial methods used were based on visual characteristics that were manually calculated to identify color distortions, blending errors, and anomalies in the frequency domain. As deep learning has progressed, data-based models have become predominant, and they automatically acquire the discriminative patterns in large-scale datasets. Spatial-based techniques are based on facial texture, structural abnormalities, whereas temporal methods consider motion patterns like eye blinking, head movements, and continuity of expressions.

Table 4 gives an overview of significant deepfake video detection algorithms, the input modalities of the algorithms, strengths, and weaknesses in detecting them. The comparison reveals that the CNN-based spatial and frequency-domain models are effective to visual artifacts, whereas the hybrid multimodal and transformer-based methods are more robust due to the simultaneous investigation of spatial, temporal and audio-visual inconsistencies.

Table 4. Comparative summary of major deepfake video detection algorithms

Algorithm Category	Representative Algorithms	Input Modality	Detection Strength	Pros	Cons
CNN-Based Spatial Models	XceptionNet, ResNet, EfficientNet	Visual (Frames)	Facial texture and boundary artifacts	Strong spatial feature learning, high accuracy on face swaps	Sensitive to compression and post-processing
Temporal Sequence Models	CNN-LSTM, CNN-GRU, TCN	Visual (Video)	Motion and temporal inconsistencies	Captures long-term dynamics, robust to spatial realism	High computational cost, complex training
3D CNN Models	C3D, I3D, SlowFast	Visual (Spatio-temporal)	Joint spatial-temporal cues	End-to-end video modeling, good temporal coherence analysis	Requires large datasets and memory
Frequency-Domain Models	F3-Net, Wavelet-CNN	Visual (Frequency)	Spectral anomalies from generation	Robust to visual smoothing, complementary cues	Performance drops with heavy re-encoding
Audio-Visual Models	SyncNet, AV-Transformer	Audio + Visual	Lip-sync and speech-face mismatch	Effective against voice-driven deepfakes	Dependent on audio quality and synchronization
Transformer-Based Models	ViT, Swin, Multimodal Transformer	Visual / Multimodal	Global semantic inconsistencies	Strong contextual modeling, scalable	Computationally expensive, data-hungry
Hybrid Multimodal Models	CNN-LSTM-Attention	Visual + Temporal + Audio	Cross-modal manipulation patterns	Highest robustness across attack types	Increased architectural complexity

6.1. Spatial Artifact-based Detection

The detection-based approach is based on spatial artifact detection, determining visual disruptions that are added in the course of deepfake generation. Despite the sophisticated synthesis models, subtle artifacts may persist in the face area because of the imperfect blending process, mismatch of the resolutions or errors in the texture synthesis. These artifacts could be in the form of unnatural skin textures, color breakages, uneven edges of facial boundaries, or lighting and shading. Spatial detectors are generally based on the analysis of individual frames with deep convolutional neural networks trained to differentiate between genuine and fake face appearance. Initial methods used features similar to local binary patterns or frequency-domain analysis that have to be created by hand in order to detect irregularities of noise. Recent deep learning algorithms can automatically acquire hierarchical representations that discriminate both low-level pixel aberrations and high-level structure anomaly with facial

deformities. To observe artifacts in various spatial resolutions, Multi-scale analysis is often used with the purpose of increasing sensitivity to localized as well as global anomalies. The attention mechanisms also enhance the performance by focusing on areas that are mostly influenced by manipulation, like eyes, mouth and jawline.

6.2. Temporal Inconsistency Analysis

Temporal inconsistency tests take advantage of motion patterns and dynamic anomalies that are unnatural across video frames that are inserted during the generation of deepfakes. Even though individual images can be visually realistic, it is still difficult to achieve biologically plausible temporal consistency in generative models. Temporal detectors are used to process a sequence of frames in order to identify anomalies in facial movement, expression changes, head motion that is not in accordance with human behavior.

Typical temporal cues are the abnormal rates of blinking of the eyes, unnatural gaze changes, irregular changes in head position, and sudden alterations in facial expression. These patterns are usually learned using deep learning models implemented as recurrent neural networks or temporal convolutional networks, which are tailored to temporal modeling.

The analysis of temporal inconsistency is very effective with high-quality deepfakes whereby spatial artifacts are minimized but realistic motion dynamics is not replicated properly. Its performance, however, may be impaired when the frame rate is low or when it is highly compressed, and this can cause distortion of motion cues. Moreover, there is an increasing trend in the use of temporal constraints in some deepfake generation methods in order to generate more realistic results. Therefore, temporal analysis can be best used in combination with spatial and audio-based detection techniques to produce strong performance in a variety of manipulation conditions.

6.3. Physiological and Behavioral Cues

Physiological and behavioral cues detection methods are based on subtle signals related to natural human physiology and behavior, which is hard to mimic by deepfake models. Physiological stimuli are micro-scale changes in facial color depending on blood flow, which can be applied in estimating heart rate using video sequences. These signals have the same temporal patterns of human physiology, and in real videos, but in deepfake videos, such temporal coherence is commonly absent. Behavioral cues center on typical human behaviour and expression, including blinking rate, direction of gaze, patterns of lip articulation, and movements of facial muscles.

Figure 7 demonstrates physiological and behavioral indicators of real and deepfake videos. Deepfake synthesis can result in strange blinking, overblown facial expressions, or unpredictable muscle activation on the face. These cues are used in machine learning models to detect any deviation of the expected human behavior.

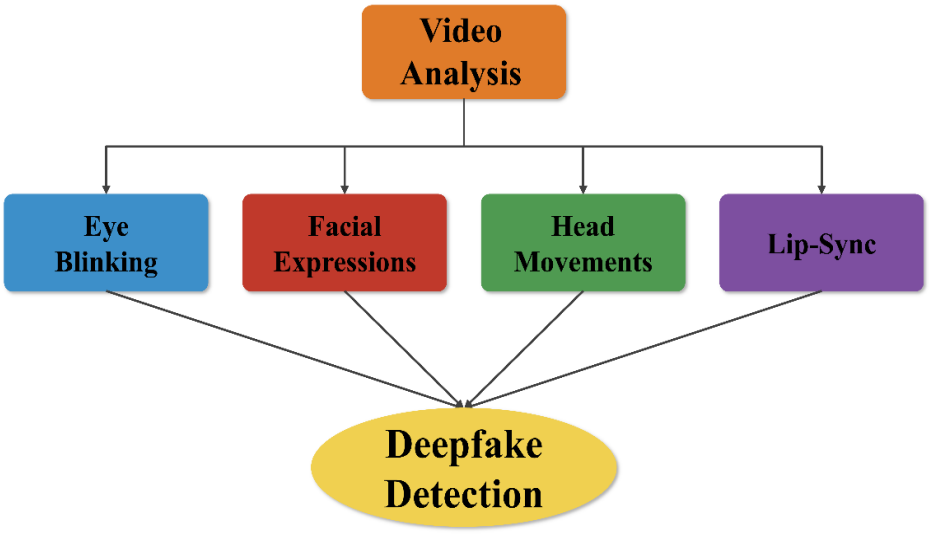

Fig. 7 Conceptual diagram of physiological and behavioral cues in deepfake videos

Deep learning methods are actively used to automatically learn and model physiological and behavioral patterns across time. Table 5 reflects a summary of deepfake video detection methods, features, datasets, and performance comparison.

These techniques provide good resistance to pure visual enhancement techniques since the physiological responses are subtle and hard to believably create. Nevertheless, their use can be affected by the quality of video, light, and movements of the subject.

Table 5. Related work Summary of deepfake video detection techniques

Detection Technique	Primary Modality	Key Cues Exploited	Dataset(s)	Strengths	Limitations
Spatial artifact analysis [48]	Visual	Texture & compression artifacts	FaceForensics	Lightweight, fast inference	Sensitive to compression
Physiological cue analysis	Visual + Temporal	Eye blinking patterns	UADFV	Biologically inspired	Limited robustness
Temporal inconsistency detection [49, 50]	Temporal	Motion dynamics	DF-TIMIT	Captures frame dependencies	Frame-rate sensitive
Temporal modelling [51]	Visual + Temporal	Expression continuity	FaceForensics	Improved temporal learning	Computationally expensive
Audio forgery detection	Audio	Spectral artifacts	ASVspoof	Effective for voice cloning	Not video-centric
Audio-visual synchronization [52]	Audio + Visual	Lip-speech alignment	FakeAVCeleb	Multimodal robustness	Noise-sensitive audio
Temporal coherence analysis	Visual + Temporal	Motion smoothness	Celeb-DF	Detects high-quality fakes	Limited unseen attack handling
Attention-based spatial detection [53]	Visual	Key facial regions	FaceForensics++	Region-focused learning	Limited interpretability metrics
Frequency-domain detection	Visual	Frequency artifacts	DFDC	Robust to spatial smoothing	Resolution-dependent
Audio-visual fusion [54]	Multimodal	AV consistency	FakeAVCeleb	Better generalization	Black-box fusion

7. Explainable AI for Reliable Deepfake Detection

This section discusses the contribution of Explainable Artificial Intelligence (XAI) methodologies, including Grad-CAM, SHAP, LIME, and attention visualization, make to increasing the transparency, interpretability, and forensic trustworthiness of deepfake detector models. It talks of the role of gradient-based visual explanation in the localization of manipulated face areas, in quantifying the impact of modality in feature-attribution paradigms, and in the (discovery of) cross-modal reasoning through attention mechanisms. The section also measures XAI efficacy in helping to verify the decisions in the legal, journalistic and security-sensitive settings and critically examines any existing limitations, risks of misleading explanations, and the necessity of standard, reliable interpretability frameworks to implement them in the real world.

Some of the most commonly used XAI techniques in deepfake detection are gradient-based visualization techniques like Grad-CAM and Grad-CAM++. The techniques produce heatmaps, which indicate the spatial areas that make the greatest contribution to the decision of the model. Grad-CAM is habitually focused on manipulated regions in the face of the eyes, mouth, facial boundaries, and skin texture variations, in face-swapping and reenactment settings [55]. Such localization of space is especially important to forensic analysis, since the visual evidence of artifacts of manipulation is confirmed by it. Gradient-based explanations can, however, be noise-vulnerable and different between model architectures, which is concerning with respect to the stability of explanations. Explainability methods based on perturbation, such as LIME, provide model-agnostic explanations of the model by attempting to model the local decision boundary around a specific input.

In deepfake detection applications, LIME is applied in deepfake detection to determine the effects of perturbations in the face structure, time, or audio characteristics on the classification performance [56]. They are modal-flexible intuitive explanations that can, however, lack consistency when applied to high-dimensional video data, and are not the most appropriate explanations to perform frame-level temporal analysis unless they are closely adapted.

SHAP and feature attribution techniques, especially SHAP, offer a theoretically based method of quantifying the roles played by features, which is founded on cooperative game theory. SHAP values are utilized in multimodal deepfake detection to assign significance to spatial data, temporal motion clues, and audio-visual cues of synchronization [57]. This allows seeing the cross-modal interactions more deeply and will determine what modality leads the decision in particular circumstances of manipulation.

The focus on visualization has become popular after the implementation of attention-based and transformer-based architectures in multimodal deepfake detectors. Self-attention and cross-attention maps depict the manner in which models distribute attention in face parts, time, and modalities. Such visualizations offer intrinsic explainability through revealing the internal logical reasoning of the model, including prioritization of motions of the lips during speech or parts of the face during facial reenactment [51]. Explanation in terms of attention proves to be particularly useful in cases of multimodal fusion, though it is not always consistent with causal significance. In general, XAI-based methods are very useful in increasing the forensic credibility, accountability, and trustworthiness of deepfake detection networks.

Although these methods have proven valuable, the present-day XAI systems have issues regarding the consistency of explanation, scalability and standardization. The next generation of studies should be devoted to the creation of sound, modality-sensitive explainability systems and testing standards to make deepfake detection solutions not only correct but also the ones that are easily explainable and legally acceptable. Table 6 demonstrates that Grad-CAM and saliency maps offer some intuitive explanation of the spatial features, whereas SHAP and attention visualization yield a more comprehensive perspective of the overall multimodal features' contribution, which is the trade-off between interpretability and computational complexity.

Table 6. Summary of XAI techniques for deepfake detection

XAI Method	Explanation Type	Applicable Modality	Key Contribution	Advantages	Limitations
Grad-CAM / Grad-CAM++ [32]	Gradient-based visualization	Visual (Frames)	Localizes manipulated facial regions	Intuitive heatmaps, forensic-friendly	Sensitive to noise, model-dependent
LIME [33]	Perturbation-based	Visual / Audio / Multimodal	Explains local decision behavior	Model-agnostic, flexible	Unstable for high-dimensional video
SHAP [36]	Feature attribution	Multimodal	Quantifies feature and modality importance	Theoretically grounded, consistent	Computationally expensive
Attention Visualization [37]	Intrinsic (model-based)	Visual / Temporal / Multimodal	Reveals model focus and reasoning	Integrated with transformers	Not always causally faithful
Saliency Maps [38]	Gradient-based	Visual	Highlights pixel-level influence	Simple and fast	Prone to visual ambiguity
Hybrid XAI Frameworks [56]	Combined methods	Multimodal	Cross-validated explanations	Higher forensic reliability	Increased system complexity

8. Comparative Analysis of Detection Methods

8.1. Comparison with Unimodal Detection Models

Unimodal deepfake detection networks are based on one information source, which is usually visual or audio information, to detect the manipulated content.

Visual-only models are used to process spatial artifacts (texture anomalies, blending anomalies, and shape-related distortions on the face), whereas audio-only models are based on acoustic cues related to synthetic speech.

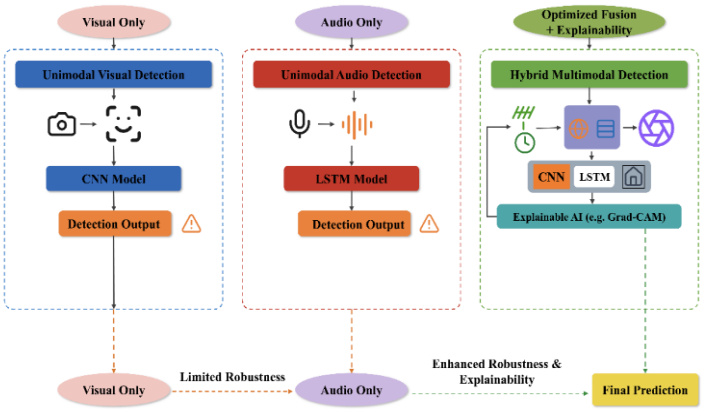

Fig. 8 Structural comparison of visual-only, audio-only, and hybrid multimodal detection models

These models have been shown to be accurate in a controlled situation, but when it comes to real-life situations; it usually fails. On the contrary, the suggested multimodal model makes use of complementary knowledge in the visual, temporal, and audio-visual levels. Figure 8 illustrates differences in structure of visual, audio and multimodal detection models. As it is revealed through comparative analysis, the unimodal models are especially susceptible to post-processing operations, as compression, resizing, or injecting noise blurs artifacts in the modality of interest [45].

The proposed strategy drastically enhances the robustness of detection and generalization by modeling the various modalities in a single model. The increase in performance is particularly significant in cross-dataset tests and in scenarios where the attack is unobserved; unimodal methods have been found to have a significant reduction in accuracy. These results show that although unimodal detectors are also computationally efficient and useful to operate in limited environments, they cannot be used to detect deepfakes in a comprehensive manner. Multimodal solutions offer a more robust way to deal with this, because no source of evidence should be trusted.

8.2. Comparison with Existing Multimodal Frameworks

Current multimodal deepfake detection systems have shown significant gains compared to unimodal methods of using visual, temporal, and audio information. The systems of most of them use feature concatenation or late-fusion methods to fuse modality-specific predictions. Although these methods improve the accuracy of detection, most of them do not have adaptive fusion mechanisms and cannot address the issue of modality imbalance or degraded inputs. Moreover, a number of multimodal models are black-box models, which provide poor interpretability and cannot be trusted by users. The suggested framework is different as it uses a hybrid deep learning architecture that uses attention-based feature-level fusion. In this design, dynamic weighting of modalities can be done according to the reliability and relevance of the input content [49].

The comparative analysis has shown the described strategy presents better results when using various datasets, especially those with tricky conditions, including low-resolution, large compression, and manipulation which are not visible to humans. The other major benefit is the fact that there is an explicit integration of explainable AI. In contrast to the current systems of multimodality, which are more concerned with accuracy, the formulated framework presents the interpretable features of modality contributions and decision-making rationale. This openness improves usability on sensitive applications. Generally speaking, the comparative analysis illustrates that the suggested framework is not only more accurate in terms of detection but also covers some of the fundamental limitations of current multimodal strategies in terms of adaptability, robustness, and interpretability.

8.3. Performance Metrics and Evaluation Criteria

Deepfake detectors should be assessed using extensive performance measures, namely, in terms of classification and practicality. The overall effectiveness of detection is measured using such standard metrics as accuracy, precision, recall, and F1-score. ROC curves and Area Under the Curve (AUC) give information about what the trade-off between false positive and true positive rates is at various decision levels.

The metrics allow conducting a reasonable comparison between unimodal and multimodal systems in the conditions of the same experiment. In addition to traditional measures of classification, the criteria related to robustness are very important. The evaluation of performance is done by varying levels of compression, resolutions, as well as noise in order to reflect real-world media distribution conditions. Cross-dataset generalization tests assess how well the generated deepfakes are detected by the models generated with unseen methods or data, and in this regard, the model is resilient to overfitting.

9. Dataset Variability and Benchmarking

9.1. Description of Benchmark Deepfake Datasets

The benchmark datasets are essential in training and testing the deepfake detection systems due to standardized data used in training and comparison. Popular deepfake datasets are usually made of both original and manipulated videos created with the help of various synthesis methods, such as face swapping, reenactment, or voice cloning.

These datasets differ with regard to subject diversity, video quality, and complexity of manipulation and recording conditions. The majority of benchmark sets have various identities that were captured in diverse lighting, poses, and expression conditions, making certain that a high number of facial appearances are achieved [51]. The nature and the constraints of every dataset are thus critical in the creation of equitable and complete benchmarking guidelines that capture the reality of detecting deepfakes.

9.2. Dataset Diversity and Modality Coverage

The diversity of the data and the scope of modality are also critical issues impacting the usefulness and the generalization of deepfake detectors. Variations in the subject demographics, recording settings, facial postures, facial expressions, and video quality are all examples of diverse datasets [23]. This type of diversity also guarantees that models are trained on generalized manipulation cues as opposed to being overtrained to specific data properties.

Diversity lack can cause bias models, which fail out in case of exposing to unknown content or demographics. Modality coverage is the concept that more than one data stream should be contained, which includes visual frames, temporal motion and audio signals. Most of the early deepfake datasets were only interested in visual manipulation and could not be used in research applying multimodal detection.

Increasingly recent datasets offer both audio and visual data in sync, allowing research of the cross-modal inconsistencies they introduce in the process of manipulation [57].

The ability to cover many modalities gives models the capability to utilize complementary cues and enhances resistance to attacks on a single modality. To take an instance, even videos containing the least visual artifacts can have errors in audio-visual synchronization.

Nonetheless, not every dataset has an equivalent quality provided by all modalities, and some also have noisy or missing audio. To deal with these issues, it is necessary to pay attention to preprocessing and modality-aware model design. On the whole, high-diversity datasets having a good coverage of modality are required in creating high-fidelity, unbiased, and deployable deepfake detection models.

9.3. Comparative Discussion with State-of-the-Art Techniques

The comparative evidence gathered from the studies reviewed suggests that there is a greater degree of coverage of manipulations in multimodal deepfake detection compared to unimodal deepfake detection due to the integration of the spatial, temporal and audio-visual cues. Visual-only methods work well in detecting texture irregularities, blending boundaries, color inconsistencies and artifacts in the frequency domain, but may not be as effective if these visual traces are weakened by compression or advanced synthesis.

Temporal approaches also look at continuity of motion, facial dynamics and inconsistencies in individual frames, but can be affected by low frame rates and high compression. Audio-visual approaches are used to detect the dimension of synchronization of face and speech movements, but they do not work if either modality is not available or it is of poor quality.

The analyzed literature also indicates that the benefits claimed by recent hybrid multimodal methods are mostly due to the complementary learning of features and adaptive fusion of the features derived from different modalities and not to a single feature representation. Transformer-based architectures can enable the spatial, temporal and audio-visual features to interact with each other, and attention-based fusion can enable informative modality-specific representations to be given greater emphasis.

Therefore, these techniques can be more robust if there is a manipulation artifact that is weak in one modality and strong in another. However, the improvement in the detection performance needs not be taken as a universal improvement, since it is sensitive to the properties of the dataset, the type of manipulation performed, the compression level, training protocol and test environment. Limited generalization across datasets, unobserved attacks, limited computational complexity, and missing or noisy modalities are still significant challenges.

Furthermore, in this review, study draw attention to the importance of explainability as a difference to more performance-oriented comparisons. While Grad-CAM and saliency based methods offer visual evidence at a local level, SHAP and attention visualization can provide an explanation of the contribution of features or modalities at a broader level. So, literature review suggests that rather than simply maximizing classification accuracy, the work that has the most potential is to synergistically optimize multimodal coverage, robustness, generalization, and interpretability. The observations are why recent multimodal approaches often claim to outperform unimodal approaches, and can help pinpoint when those benefits might be lost.

10. Conclusion

In this paper, a full multimodal deepfake detection system was introduced, which relied on a hybrid deep learning model with built-in explainable AI and addressed the major weaknesses of the current detection methods. The complementary evidence of manipulation that the proposed system is good in capturing has been visual spatial artifacts combined with temporal motion inconsistencies and audio-visual contact that is easily missed by unimodal models.

The hybrid architecture allows learning strong features on modalities, whereas attention-based fusion actively adapts to different quality of content and attack types, and therefore makes sure that its performance can be relied upon in real world settings. In addition to accuracy in detection, interpretability and user trust were highlighted as crucial factors in deepfake detection system in this work.

This is because to introduce ability of explainable AI mechanisms can give transparent, modality-sensitive explanations that show the reason why the model makes a particular prediction. Faithfulness, consistency and sparsity metrics are quantitative evaluation to make sure that the explanations are consistent, reliable and concise.

This transparency has to be adopted in some of the most stakeholder-sensitive fields like digital forensics, media verification, and legal investigations where it is as important to understand the basis of automated decisions as the decisions themselves. High-scale evaluation of the proposed framework using several benchmark datasets revealed it to be more accurate detecting, resistant to compression, resolution, and resolution degradation, adversarial and unseen attack resistance, and cross-dataset generalization. The findings affirm that multimodal learning is a significant way of making learners susceptible to modality-specific evasion and data bias.

References

[1] Law Kian Seng et al., "AI Integrity Solutions for Deepfake Identification and Prevention," *Open International Journal of Informatics*, vol. 12, no. 1, pp. 35-46, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Karima Ghediri, "Countering the Negative Impacts of Deepfake Technology: Approaches for Effective Combat," *International Journal of Economic Perspectives*, vol. 18, no. 12, pp. 2871-2890, 2024. [Google Scholar] [Publisher Link]

[3] Darinka Ignatova, "Affirming Wellness Culture Through Innovative Methodology Related to Blaze-Pod Trainer System," *Educational and Scientific Policy Strategies*, vol. 31, no. 2, pp. 212-225, 2023. [Google Scholar] [Publisher Link]

[4] D. Ignatova, "Specificity of the Motor Potential for Achieving Scholar Wellness," *Trakia Journal of Sciences*, vol. 19, pp. 867-873, 2021. [Google Scholar]

[5] Darinka Ignatova, "Motor Activity based on Learning Contemporary Trends in School Wellness," *Smart Innovations in Recreative and Wellness Industry and Niche Tourisum-Scientific Journal*, vol. 5, no. 1-2, pp. 22-26, 2023. [Google Scholar] [Publisher Link]

[6] Ipshita Rong, "*Detection and Segmentation of Deepfake Face Images Generated by GANs using Segmented based CNN*," Dissertation, Dublin Business School, Ireland, 2020. [Google Scholar]

[7] Ali Raza, Kashif Munir, and Mubarak Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection," *Applied Sciences*, vol. 12, no. 19, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Tomasz Walczyna, and Zbigniew Piotrowski, "Fast Fake: Easy-to-Train Face Swap Model," *Applied Sciences*, vol. 14, no. 5, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ankita Bhandarkar et al., "Deep Learning Framework for Robust Deep Fake Image Detection: A Review," *2024 International Conference on Artificial Intelligence and Quantum Computation-based Sensor Application (ICAIQSA)*, Nagpur, India, pp. 1-7,2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Nazneen Mansoor, and Alexander Iliev, "Artificial Intelligence in Forensic Science," *Advances in Information and Communication*, Springer, Cham, vol. 652, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] David Ghiurău, and Daniela Elena Popescu, "Distinguishing Reality from AI: Approaches for Detecting Synthetic Content," *Computers*, vol. 14, no. 1, pp. 1-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ruben Tolosana et al., "Deepfake and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131-148, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Vishal Kumar Sharma, Rakesh Garg, and Quentin Caudron, "A Systematic Literature Review on Deepfake Detection Techniques," *Multimedia Tools and Applications*, vol. 84, no. 20, pp. 22187-22229, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mubarak Alrashoud, "Deepfake Video Detection Methods, Approaches, and Challenges," *Alexandria Engineering Journal*, vol. 125, pp. 265-277, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohammad Nagahisarchoghaei et al., "An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives," *Electronics*, vol. 12, no. 5, pp. 1-59, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Divyang Teotia, Agata Lapedriza, and Sarah Ostadabbas, "Interpreting Face Inference Models using Hierarchical Network Dissection," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1277-1292, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ping Chen, Ming Xu, and Xiaodong Wang, "Detecting Compressed Deepfake Images using Two-Branch Convolutional Networks with Similarity and Classifier," *Symmetry*, vol. 14, no. 12, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wahidul Hasan Abir et al., "Detecting Deepfake Images using Deep Learning Techniques and Explainable AI Methods," *Intelligent Automation and Soft Computing*, vol. 35, no. 2, pp. 2151-2169, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Raidah Salim Khudeyer, and Noor Mohammed Almoosawi, "Fake Image Detection using Deep Learning," *Informatica*, vol. 47, no. 7, pp. 115-120, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Chandler Timm Doloriel, and Ngai-Man Cheung, "Frequency Masking for Universal Deepfake Detection," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, pp. 13466-13470, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Evren Dağlarli, *Explainable Artificial Intelligence (XAI) Approaches and Deep Meta-Learning Models*, Advances and Applications in Deep Learning, Marco Antonio Aceves-Fernandez, London, U.K, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, pp. 1-45, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Yisroel Mirsky, and Wenke Lee, "The Creation and Detection of Deepfake: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1-41, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Achhardeep Kaur et al., "Deepfake Video Detection: Challenges and Opportunities," *Artificial Intelligence Review*, vol. 57, no. 6, pp. 1-47, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jan Kietzmann et al., "Deepfake: Trick or Treat?" *Business Horizons*, vol. 63, no. 2, p. 135-146, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Mekhail Mustak et al., "Deepfake: Deceptions, Mitigations, and Opportunities," *Journal of Business Research*, vol. 154, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Hina Fatima Shahzad et al., "A Review of Image Processing Techniques for Deepfake," *Sensors*, vol. 22, no. 12, pp. 1-28, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[28] Mohammad Wazid et al., "A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact," *Cyber Security and Applications*, vol. 2, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Jawad Ahmad et al., "A Survey on Enhanced Approaches for Cyber Security Challenges based on Deep Fake Technology in Computing Networks," *Spectrum of Engineering Sciences*, vol. 2, no. 4, pp. 133-149, 2024. [Google Scholar] [Publisher Link]

[30] Yuyang Sun et al., "Generalised Deepfakes Detection with Reconstructed-Blended Images and Multi-Scale Feature Reconstruction Network," *2024 IEEE International Joint Conference on Biometrics (IJCB)*, Buffalo, NY, USA, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Anushka Singh, and Jyotsna Singh, "Image Forgery Detection using Deep Neural Network," *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, vol. 10, pp. 504-509, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Neha Sandotra, and Bhavna Arora, "A Comprehensive Evaluation of Feature-based AI Techniques for Deepfake Detection," *Neural Computing and Applications*, vol. 36, no. 8, pp. 3859-3887, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Jie Gao et al., "Texture and Artifact Decomposition for Improving Generalization in Deep-Learning-based Deepfake Detection," *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Zhimao Lai et al., "LIDeepDet: Deepfake Detection via Image Decomposition and Advanced Lighting Information Analysis," *Electronics*, vol. 13, no. 22, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Ekant Ratanlal Tekam et al., "Detection of Deepfake Images using Hybrid Convolutional Neural Networks and Transformer Models," *2025 IEEE 1st International Conference on Smart Innovations in Systems, Infrastructure, Mechanical, Power, AI and Computing Technologies (SISIMPACT)*, Bhopal, India, pp. 876-881, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[36] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee, "Deep Fake Image Detection based on Pairwise Learning," *Applied Sciences*, vol. 10, no. 1, pp. 1-14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[37] Tiewen Chen et al., "Masked Conditional Diffusion Model for Enhancing Deepfake Detection," *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[38] Muhammad Asad Arshed et al., "Multiclass AI-Generated Deepfake Face Detection using Patch-Wise Deep Learning Model," *Computers*, vol. 13, no. 1, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Yin Ni et al., "A Deepfake Detection Algorithm based on Fourier Transform of Biological Signal," *Computers, Materials and Continua*, vol. 79, no. 3, pp. 5295-5312, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[40] Fakhar Abbas, and Araz Taeihagh, "Unmasking Deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques using Artificial Intelligence," *Expert Systems with Applications*, vol. 252, pp. 1-38, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] Gourab Naskar et al., "Deepfake Detection using Deep Feature Stacking and Meta-Learning," *Heliyon*, vol. 10, no. 4, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Siddiqui Muhammad Yasir, and Hyun Kim, "Lightweight Deepfake Detection based on Multi-Feature Fusion," *Applied Science*, vol. 15, no. 4, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[43] Deeraj Nagothu et al., "Defakepro: Decentralized Deepfake Attacks Detection using ENF Authentication," *IT Professional*, vol. 24, no. 5, pp. 46-52, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[44] Andrea Miotti, and Akash Wasil, "Combatting Deepfake: Policies to Address National Security Threats and Rights Violations," *arXiv*, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Maria Pawelec, "Decent Deepfake? Professional Deepfake Developers' Ethical Considerations and Their Governance Potential," *AI and Ethics*, vol. 5, no. 3, pp. 2641-2666, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[46] Jumai Fabuyi et al., "Deepfake Regulations and Their Impact on Content Creation in the Entertainment Industry," *Archives of Current Research International*, vol. 24, no. 12, pp. 53-74, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[47] Guntur Permana Putra, and Mochammad Tanzil Multazam, "Law Enforcement Against Deepfake Porn AI: Penegakan Hukum Terhadap Deepfake Porn AI," *Journal of Contemporary Business Law & Technology*, vol. 1, no. 9, pp. 25-44, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Junhao Dong et al., "Restricted Black-Box Adversarial Attack Against DeepFake Face Swapping," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2596-2608, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[49] Saima Waseem et al., "Multi-Attention-based Approach for Deepfake Face and Expression Swap Detection and Localization," *EURASIP Journal on Image and Video Processing*, vol. 2023, no. 3, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[50] Noor ul Huda et al., "Fake-Checker: A Fusion of Texture Features and Deep Learning for Deepfakes Detection," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 49013-49037, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[51] Noor Aldeen A. Khalid et al., "Palmprint Recognition System using VR-LBP and KAZE Features for Better Recognition Accuracy," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 2, pp. 1060-1068, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[52] Reetika Thakur, "Introduction to Artificial Intelligence and its Importance in Modern Business Management," *Leveraging AI and Emotional Intelligence in Contemporary Business Organizations*, pp. 133-165, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[53] Md. Zasim Uddin et al., "Deep Learning with Image-Based Autism Spectrum Disorder Analysis: A Systematic Review," *Engineering Applications of Artificial Intelligence*, vol. 127, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[54] Amirkia Rafiei Oskooei, Mehmet S. Aktaş, and Mustafa Keleş, "Seeing the Sound: Multilingual Lip Sync for Real-Time Face-to-Face Translation," *Computers*, vol. 14, no. 1, pp. 1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[55] Tanqiu Wang, and Dan Wu, "Computer-Aided Traditional Art Design Based on Artificial Intelligence and Human-Computer Interaction," *Computer-Aided Design and Applications*, vol. 21, no. 1, pp. 59-73, 2024. [Google Scholar]

[56] Joanne Yu et al., "Artificial Intelligence-Generated Virtual Influencer: Examining the Effects of Emotional Display on User Engagement," *Journal of Retailing and Consumer Services*, vol. 76, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[57] Zitong Yu et al., "Benchmarking Joint Face Spoofing and Forgery Detection with Visual and Physiological Cues," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4327-4342, 2024. [CrossRef] [Google Scholar] [Publisher Link]