Detection of Multimodal Deepfake: Datasets, Evaluation Metrics, and Approaches

Sampada Aditya Kulkarni[1,2,*], and Pooja Sapra[2]

[1]PES's Modern College of Engineering, Pune, Maharashtra, India,411005
[2] IT Department, Parul Institute of Engineering and Technology, Parul University, Vadodara, India-391760
[*]k.sampadaa2019@gmail.com

Abstract. Manipulated multimedia is a critical challenge in our day to day life. In the preceding decades, rapid progress in Artificial Intelligence, Machine Learning, and Deep Learning has resulted in new techniques and various tools for altering media formats. Advanced AI tools are used to produce fabricated videos, false images, tampered audio or staged audio-visual recordings. Altered multimedia's ability to appear and sound authentic makes them particularly unsafe. These are simply "deepfakes". Detecting these deepfakes is challenging because they are often multimodal, integrating inconsistent data sources such as video, audio, and text. This paper explores a survey of the main research in multimodal deepfake detection. In this work, we present a systematic analysis of the representative datasets and evaluation metrics used in the field, followed by a taxonomy of different deepfake detection approaches. We also compare the strengths and weaknesses of DL deepfake detection approaches. Finally, we discuss the challenges researchers face and suggest future directions to make detection systems more reliable and effective.

Keywords: Deepfake Detection, CNN, RNN, Hybrid Model, ViT Model

1 Introduction

Deepfakes are AI-generated fabricated videos, tampered images, and synthetic audio recordings. Deepfake is becoming a serious global concern, as it is hazardous to every individual who falls victim to it [1]. As altered multimedia is so realistic, they undermine trust in digital content. Day by day, AI is becoming stronger, and deepfakes are increasingly difficult to distinguish from authentic material. As a result, netizens may start to question the originality of what they watch and listen to online, creating a crisis of trust in media and communication.

J. Saxena et al. (eds.), *Proceedings of the International Conference on Sustainable Micro-Nano Materials & Innovative Technology (ICSUMMIT 2026)*, Atlantis Highlights in Engineering 44,
https://doi.org/10.2991/978-94-6239-727-9_14

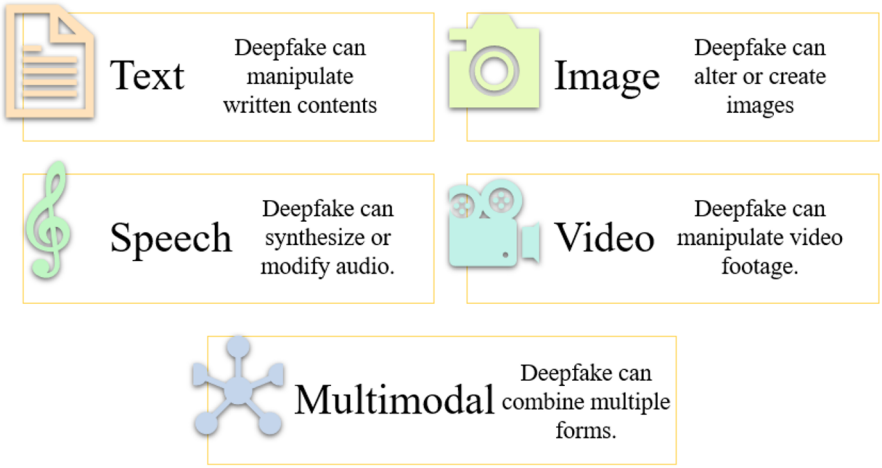

Fig. 1. Types of Deepfake

The types of deepfakes, together with their descriptions and real-world examples, are shown in fig 1 and summarized in table 1.

Table 1. Types of Deepfake with Real-world Examples

Type	Description	Real-world Examples
Text	AI-generated fake written content designed to mislead or impersonate	Fake news articles, fabricated social media posts, AI-written phishing emails
Image	Manipulated or fully generated photos using AI	Celebrity face swaps, fake profile pictures of non-existent people, morphed political campaign posters
Audio	Synthetic voice cloning to impersonate someone's speech	Fraud calls mimicking CEOs, fake audio clips of politicians, AI-generated podcasts with cloned voices
Video	Altered or fabricated moving visuals combining face, lip-sync, and gestures	Deepfake speeches of leaders, fake interviews of celebrities, manipulated adult content
Multimodal	Combination of text, audio, image, and video deepfakes for maximum realism	Fake video interviews with synthetic voice + visuals, AI-generated virtual influencers, coordinated misinformation campaigns

Ordinary individuals, as well as politicians, celebrities, and organizations such as financial institutions and government bodies, face serious repercussions from deepfakes [1], [2]. Deepfakes go beyond online frauds, they can cause real harm to finances, careers, and well-being of the victim. Each day, new cases of fraud emerge, leaving individuals vulnerable and suffering consequences in financial, emotional, and social terms. Politically, deepfakes can be exploited to spread misinformation and fabricate speeches or actions of leaders. They are also used to manipulate public opinion during elections, posing serious threats to democratic processes [80] Examples of non-consensual deepfakes include fake intimate videos, fabricated revenge pornography, and manipulated celebrity content etc. These forms of synthetic media result in reputational harm, emotional distress, and privacy loss. [82], [83].

The paper is organized as follows: Section 2 presents the background and preliminaries of deepfakes and their detection. Section 3 provides a detailed discussion of the taxonomy of deepfake detection methods. Section 4 summarizes the available datasets and benchmarks. Section 5 presents the detection methods along with a comparative analysis. Section 6 discusses the challenges and limitations. Section 7 outlines future directions, and finally, Section 8 concludes the survey.

2 Background & Preliminaries

The term ‘‘Deepfake’’ is derived from ‘‘Deep Learning (DL)’’ and ‘‘Fake,’’ and it describes specific photorealistic video or image contents created with Deep Learning. This word was named after an anonymous Reddit user in late 2017, who used DL algorithms for swapping faces in pornographic realistic videos which were fake [2].

Deepfake technology has developed at a fast pace during the past ten years, moving from basic face-swapping techniques to highly advanced generative models. Face swapping was the initial method that worked with simple autoencoders that could copy facial features from one person onto another person’s face, but the results were often unrealistic and very easy to detect too. As a developed technique, various AI-based generative models have been designed and developed in the era of deepfakes [46].

As shown in fig 2, fabricated media is created by Generative models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models [3],[4].

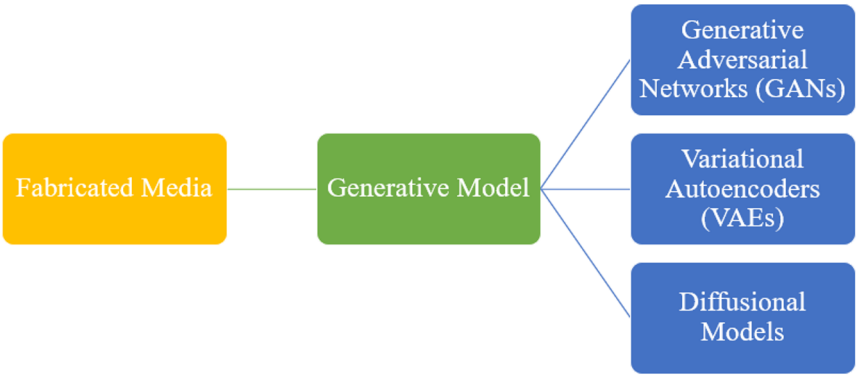

Fig. 2. List of Generative Models

The first method, GANs employ a combined framework of generator and discriminator. The generator creates synthetic images using encoder and decoder, and a discriminator evaluates the authenticity of synthetic images generated by generator [5]. The second method, VAEs, in contrast, encodes images into a latent space and reconstructs it through a decoder, which enables stable training and a variety of outputs, although with lower visual adherence compared to GANs [6]. And last but not the least method, Diffusion models, used to represent a more recent paradigm, generating data by progressively reversing a noise-adding process, and have achieved state-of-the-art performance in terms of both realism and diversity, though at the expense of computational efficiency [7]. Together, all above mentioned models show the balance between how real the outputs look, how varied they can be, and how efficiently they are produced and used in the deepfake world.

3 Taxonomy of Deepfake Detection Methods

As we discussed in the previous section, the fast advancement of generative models on one side and the increasing misuse of synthetic media on the other side, deepfake detection has manifested as a fundamental area of scholarly exploration. To study deepfake detection methods, Tolosana, Ruben, et al have proposed taxonomies that classify methods across multiple dimensions [8]. We have used these taxonomies to compare techniques, highlight strengths and weaknesses, and identify gaps in current detection strategies. Fig 3 includes the hierarchy used in deepfake detection and explained in the following explanation in this research paper.

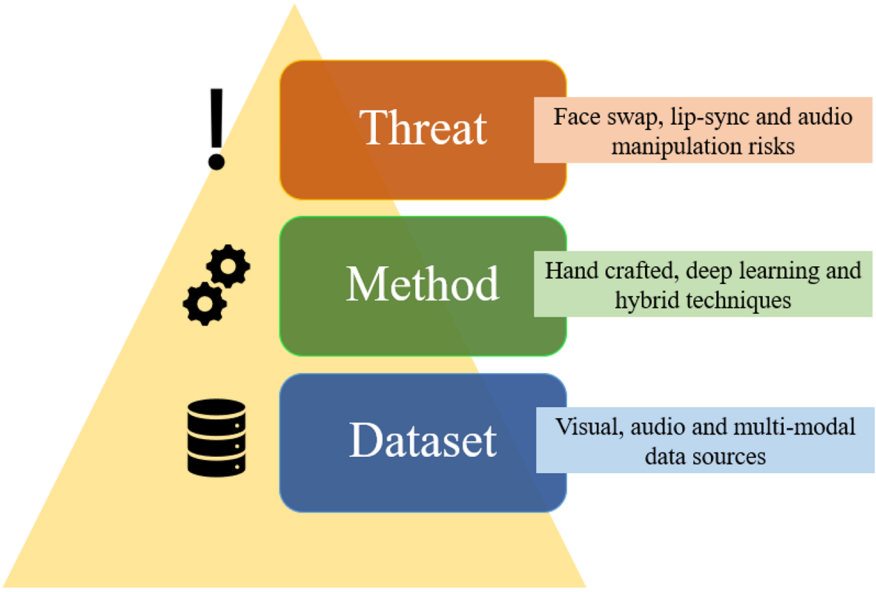

Fig. 3. Deepfake Detection Hierarchy

The foundation of deepfake detection lies in datasets that include both real and fake content. This is the first dimension of deepfake detection taxonomy. The datasets form the foundation for training and evaluation of all deepfake detection models. FaceForensics++ and Celeb-DF provide the visual datasets including manipulated images and video content for benchmarking detection models [9], [10]. To detect the audio spoofing, the audio datasets required. To design these models, Yamagishi et al. explored datasets like ASVspoof and FakeAVCeleb. They study synthetic voice detection in their research [11]. To reflect the growing complexity of deepfake attacks, Khalid et al researched the multimodal datasets. The multimodal datasets combine videos and audios to capture cross-modal manipulations [12].

Focusing on the technical approach, the next dimension categorizes detection methods. Li et al. described the research based on the previous handcrafted features like identifying artifacts such as unnatural blinking or head pose inconsistencies [13]. Lamichhane et al. and Heo et al. proposed deep learning methods, particularly CNNs, RNNs and Vision Transformers, helped the learning subtle spatial and temporal features that differentiates real from fake content [14], [15], [16] [17]. Mittal et al. proposed the new way of deepfake detection is the hybrid approach [18]. Comito et al. explored in their research that this hybrid approach improves the robustness by combining handcrafted features with deep learning. The multimodal fusion methods integrate visual and audio signals to detect inconsistencies across modalities [19].

As we discussed previously, the top dimension addresses the types of manipulations or threats. This dimension helps for detection of common categories including face swaps, lip-sync manipulations, fully synthetic identities generated by GANs or diffu-

sion models [20]. Deepfakes rely on audio that sounds like real speakers. These synthetic voices cause operational difficulties to deeffake detection systems, particularly when integrated with video manipulations [21]. While detecting deepfakes, the most challenging thing is cross-modal attacks. Here, both audio and video are altered simultaneously, and hence represent one of the most difficult scenarios for detection systems. Extensive multimodal analysis and advanced feature extraction are necessary to detect cross-modal attacks.[22], [23].

Here, we studied deepfake detection taxonomies. We showed strategies to fight new manipulations and threats. Taken together, these taxonomies focus on the fact that not a single method is sufficient on its own to detect the deepfake; rather, effective defense requires a layered, adaptive and hybrid approach. These deepfake detection approaches combine multiple detection paradigms. We tried to highlight the complexity of today's threats and points toward future research on better detection.

4 Datasets & Benchmarks

In the previous section, we have surveyed that while detecting deepfakes, the dataset plays a vital role. Various datasets are listed in previous sections that are used for different purposes, such as deepfake detection in images, videos, and audio. Table 2 covers the list of datasets, their size, modality, manipulations and availability of respective datasets. Among the listed datasets, Rossler et al. stated in his research that, FaceForensics++ (FF++) is one that supports standardized evaluation. FF++ facilitates four canonical facial manipulation methods and multiple compression settings, providing masks and consistent preprocessing that enabled cross-paper comparability [9]. The dataset DFDC introduced a large-scale, diverse, actor-driven benchmark to stress-test generalization with varied conditions, demographics, and production pipelines, catalyzing a community challenge and baselines for robust model comparison [80]. Another dataset Celeb-DF (v2) raised the bar for realism by curating high-quality YouTube source videos and synthesizing deepfakes that closely match online distribution, making many earlier detectors fail under subtle artifacts [10]. WildDeepfake shifted focus to in-the-wild content, aggregating internet-sourced deepfake videos with rich, unconstrained scenes and mixed synthesis methods, highlighting domain gaps and generalization limits of lab-trained models [25].

Table 2. Dataset Description

Dataset	Size	Modality	Manipulations	Availability
FaceForensics++ [9]	1,000 source videos; four manipulation sets with masks; multiple compression variants	Video (frame-level usable); masks enable image tasks	Deepfakes, Face2Face, FaceSwap, NeuralTextures	Public, research use via GitHub and paper page

DFDC [89]	Large-scale actor-driven video corpus; extensive augmentations	Video	Primarily face-swap and GAN-based variants; diverse capture conditions	Public challenge dataset
Celeb-DF (v2) [10]	590 real videos; 5,639 deepfake videos	Video	High-quality face-swaps designed to match online quality	Public via official page and mirrors
WildDeepfake [25]	7,314 face sequences from 707 internet deepfake videos	Video (face sequences)	Mixed, internet-sourced methods and post-processing	Public via GitHub

5 Evaluation Metrics

Performance analysis plays a vital role in the detection of deepfakes. Efficiency in deepfake detection is attained through the use of evaluation metrics. They serve as the primary means of assessing a deepfake detection model's effectiveness [84]. Yan, Zhiyuan, et al. suggested that deepfake datasets are often imbalanced, which makes the evaluation of detection models more challenging. Since manipulations can be subtle, relying on a single metric is insufficient [85]. Multiple metrics are therefore employed to capture different aspects of model performance. We surveyed numerous papers and compiled a set of evaluation metrics, along with their roles and practical examples, as presented in table 3.

Table 3. Evaluation Metrics for Deepfake Detection - Roles and Practical Examples

Metric	Role in Deepfake Detection	Example
Accuracy	**Provides an overall measure of correct predictions**, but must be interpreted carefully due to class imbalance in deepfake datasets.	A model achieves 95% accuracy on FaceForensics++ but misclassifies subtle manipulations in Celeb-DF.
Precision	**Ensures genuine content is not wrongly flagged as fake**, which is critical in applications where false positives can damage credibility.	High precision ensures genuine news multimedia contents are not wrongly flagged as fake during social media screening.
Recall	**Highlights the model's ability to correctly identify manipulated content**, reducing the risk of deepfakes being overlooked.	A recall of 92% means most fake celebrity interview clips are correctly detected, reducing overlooked manipulations.

F1-Score	**Balances precision and recall**, making it particularly valuable when datasets contain unequal numbers of real and fake samples.	A balanced F1-score shows the model performs reliably across datasets with unequal numbers of real and fake samples.
AUC (ROC Curve)	**Provides insight into the discriminative power of a model across varying thresholds**, which is essential for deployment in diverse real-world scenarios.	An AUC indicates strong separation between authentic and fake multimedia contents across varying thresholds.
Confusion Matrix	**Offers a detailed breakdown of errors**, helping researchers understand whether models are more prone to false positives or false negatives.	Analysis shows the model produces more false negatives (missed deepfakes) than false positives, highlighting a weakness in detecting subtle manipulations.

Collectively, our survey says the evaluation metrics enable fair comparison across datasets and models. Also it guides improvements in detection strategies. It also ensures that models are evaluated not only for accuracy but also for reliability, robustness, and practical applicability [86]. To address challenges such as class imbalance and subtle manipulations, researchers gain a comprehensive assessment of models by integrating evaluation metrics as accuracy, precision, recall, F1-score, AUC, and confusion matrices [87]. Abbasi, Maryam, et al. suggested in their research that a holistic evaluation framework strengthens confidence in detection systems [88]. They further noted that such a framework supports the development of resilient and generalizable models. It also enhances the suitability of these models for deployment in real-world contexts such as social media monitoring, forensic analysis, and election integrity [88].

6 Deepfake Detection Models and its Comparison

Manipulated or computer-generated media pose serious difficulties in today's digital environment. Because of this, detecting deepfakes has become a crucial focus for researchers. Altered content can fuel misinformation and inflict real damage on both individuals and society [2]. To address these risks, scholars are turning to advanced deep learning techniques, applying a range of models to improve the reliability of detection systems. Fig 4 shows the existing models of deepfake detection. Convolutional Neural Networks (CNNs) are widely applied because they can capture visual patterns in images [26]. Recurrent Neural Networks (RNNs) are useful for analyzing sequences, such as video frames or audio signals [27], [28]. Some approaches combine CNN and RNN to take advantage of both spatial and temporal features [29], [30]. More recently, Vision Transformers (ViT) have shown strong performance by modeling long-range dependencies in visual data. Hybrid models like CNN + ViT combine the strengths of convolutional layers with transformer-based attention, leading to more accurate detection results [31] [53].

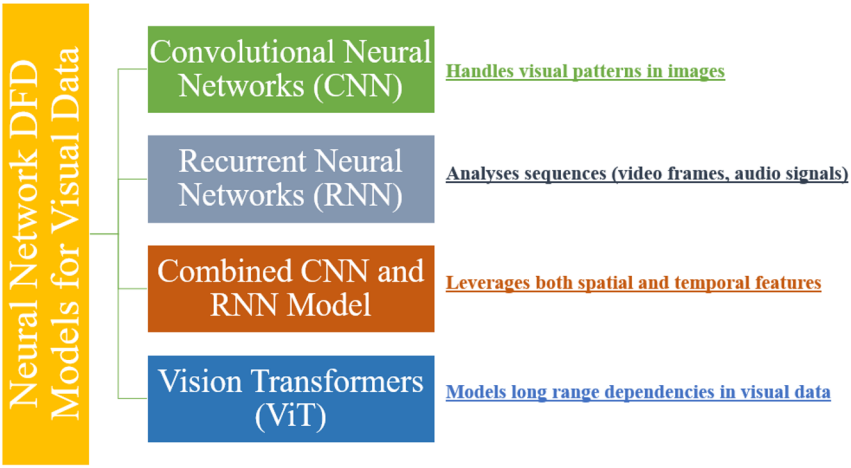

Fig. 4. Deepfake Detection DL Models

6.1 Convolutional Neural Networks (CNNs) Model

As we discussed in previous sections, deepfake detection is one of the most essential needs in the era of Artificial Intelligence (AI). Convolutional Neural Networks (CNNs) are a dedicated deep learning approach within AI. They are one of the dominant model structures used in AI, particularly for activities related to images, video, visuals and multimedia. In short, CNNs are a most powerful tool inside AI that enables machines to "see" and interpret visual information. They are also highly efficient at detecting regularities in visual information, such as edges, textures, and shapes [26]. To perform the tasks like recognition of face, number plates, analysis of images, detection of objects etc can be done with the help of CNNs. In recent decades, CNN has been the most suitable choice for researchers in detecting manipulated media. According to our survey, most researchers were convinced that CNN is the most powerful tool for deepfake detection because of its strengths. Its strength lies in automatically learning the subtle differences between real and fake content [1].

Characteristics of CNNs is mainly to focus on spatial features (details within a single frame). Detecting important features step by step after the image is divided into smaller regions is the principal working of CNN [47]. CNNs processes images through a combination of convolutional, pooling, and fully connected (dense) layers to hierarchically extract and classify various visual attributes. Initially, CNNs look at very basic attributes like edges, colors, and shapes of an input image or an individual frame of video [32]. As the layers go dense, the CNN learns more complex attributes such as facial structures, skin textures, and eye movements. This layer by layer learning makes CNNs most efficient for identifying tiny clues that separate real images or videos from fake ones.

Layers of CNNs are trained on large datasets that contain both real and fake images or videos. Trained dataset helps the researchers to detect deepfakes. Common signs of

fake image or videos are blurred boundaries around the mouth, unnatural blinking patterns, or mismatched lighting on the face. While training the model, CNN learns to identify above mentioned common signs of alteration. Once the model is trained, CNN predicts the new images or video frames to classify them as "original" or "fake" [32], [44].

Several researchers have carried out studies with advanced CNN architectures, including XceptionNet and ResNet. These models have achieved strong results on benchmark datasets, including FaceForensics++ and Celeb-DF v2. The researchers reported accuracy greater than 90% [26]. It indicates that CNNs support the identification of tiny pixel flaws in deepfakes that our eyes usually miss. Hence the CNN is best suitable for deepfake detection.

Benefits and Limitations of CNN Deepfake Model

CNNs offer multiple benefits in the context of deepfake detection. It helps to achieve high accuracy and it is fast. The most important highlighting feature is the ability to automatically learn useful visual attributes without human intervention. By learning and detecting hidden patterns automatically, CNNs prove to be powerful tools for deepfake detection [32]. As noted earlier, CNNs eliminate the need for manual feature engineering, as they autonomously extract relevant attributes from training data.

As mentioned earlier, CNNs are designed to focus mainly on spatial features, while their handling of temporal features (changes across video frames) remains limited in efficiency. This means the model may fail to record inconsistencies in sequential frames or facial expressions over time [1]. Güera, David et al addresses this limitation by making use of Recurrent Neural Networks (RNNs) as deepfake detection models for videos. The implemented model is better at analyzing sequences and time-based information [28]. RNN models can therefore provide more effective deepfake detection for temporal features with greater efficiency.

6.2 Recurrent Neural Networks (RNNs) Model

Anjaneyulu et al. stated in their research that limited efficiency is observed in CNNs for detecting deepfakes with temporal or sequential features, while superior performance is achieved by RNNs[30]. In deepfake detection, not only the spatial feature but the flow across frames within a video matters a lot [33]. Li et al. and Chintha et al. covered the importance of deepfake detection in temporal features in their research. They have explored fake videos by detecting blinking of eyes, movements of lips as well as subtle facial changes occurring over time [13], [27]. These represent temporal features, which are also referred to as sequential features. In technical language, the temporal features are frame-to-frame facial movements in a video, where the micro-changes across consecutive frames carries critical information.

Our survey indicates that RNNs are the most effective deepfake detection approach for temporal data. Unlike traditional feedforward networks, RNNs have connections that loop back. The key feature of feedforward networks is to keep information from earlier steps in memory and use it to make better predictions in later steps. This feature

makes RNNs more efficient to perform tasks like speech recognition, language modeling, and video analysis, where the sequence of data points matters [27],[28]. To address issues such as vanishing gradients, advanced variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) models were implemented by Güera et al., Anjaneyulu et al., and other researchers. As a result, RNNs were able to capture longer dependencies in sequences with high accuracy. [28], [30].

In the detection of deepfakes in videos, inconsistencies in facial expressions and subtle features are primarily focused. In real videos, the changes in faces and expressions are natural, while in manipulated videos, these patterns may appear unnatural [27]. RNN learns these disorders very efficiently and helps to detect fake videos with high accuracy [30]. As stated previously, LSTM and GRU models have been used to process sequences of frame features extracted by RNNs, making it possible to detect keen temporal irregularities in deepfake videos.

Benefits and Limitations of RNN Deepfake Model

RNN complements CNNs by adding sequence-level analysis to the deepfake detection process. RNN offers significant benefits for video data frame deepfake detection, as it captures temporal dynamics and improves accuracy [30]. However, RNNs also have some limitations. They can be slow to train the model, can require large amounts of data, and sometimes can struggle with very long sequences. In scenarios involving fine-grained spatial feature representation, RNN models tend to exhibit limitations in reliably capturing such patterns, whereas CNNs demonstrate comparatively greater effectiveness in detecting and processing the same with enhanced ease. These limitations are often overcome by CNN–RNN hybrid models. The hybrid model operates in two primary stages: CNNs extract spatial features from video frames, followed by RNNs that process these features sequentially. This hybrid approach combines both architectures to implement stronger and more reliable deepfake detection models[34].

6.3 Combined CNN and RNN Hybrid Model

Deepfake videos often contain subtle visual artifacts and unnatural movements that are hard to detect using only one type of neural network. The key feature of CNNs is high accuracy and the ability to automatically learn visual features without manual effort [30]. Also CNNs are good at analyzing individual frames and spotting spatial inconsistencies like blurred edges or mismatched lighting. Our survey lists the following subtle features. The details are gathered in table 4 with the limitations of CNN model and RNN model. Also included the solutions that are helpful to overcome the limitations by Combined CNN and RNN Hybrid Model along with real life examples for better understanding [30], [33], [34], [35], [36].

Table 4. Comparative Table - Subtle Deepfake Features vs. Model Capabilities

Subtle Feature	Why CNNs Fail	Why RNNs Fail	How combined CNN-RNN hybrid model help	Real-life Example / Explanation
Micro-expressions &	CNNs focus on spatial features within single	RNNs capture sequences but lack fine spatial	Hybrid CNN-RNN models	A genuine smile involves micro-movements in cheeks and

temporal consistency	frames, missing subtle temporal variations in facial muscles.	resolution, so tiny facial twitches are overlooked.	integrate spatial detail with sequential modeling, enabling detection of unnatural frame-to-frame consistency.	eyes; deepfakes often look "static" or frozen.
Eye movement & gaze irregularities	CNNs detect eye regions spatially but cannot track blink frequency or gaze shifts across time.	RNNs model sequences but may blur fine spatial details of eye regions, missing subtle gaze direction mismatches.	CNN-RNN hybrids track both eye dynamics and spatial accuracy across frames, flagging irregular blink rates or gaze fixation.	Humans blink naturally every few seconds; deepfakes may show long stretches without blinking or unnatural gaze fixation.
Lighting & shadow mismatches	CNNs trained on pixel intensities may fail to capture global illumination consistency across facial regions.	RNNs focus on temporal sequences, not spatial lighting cues, so mismatched shadows persist undetected.	CNN-RNN hybrids combine spatial extraction with temporal context, detecting shading inconsistencies across regions and frames.	Natural lighting casts consistent shadows (e.g., nose and cheek); deepfakes may show mismatched highlights and shadows.
Lip-sync & audio-visual mismatch	CNNs analyze mouth shapes frame-by-frame but cannot align them with spoken phonemes.	RNNs capture audio-visual sequences but lack precise spatial mapping of lip movements.	CNN-RNN hybrids jointly learn mouth region features and temporal audio cues, detecting misalignment between speech and lip motion.	In genuine speech, lip shapes match phonemes (e.g., "O" vs. "M" sounds); deepfakes may show mismatched lip motion with audio.

Previous table suggests by combining CNNs with RNNs and making it a hybrid model, the researchers can build a model that detects both static and dynamic signs of manipulation. It helps to lead more accurate deepfake detection. As stated previously, researchers combine CNNs with RNNs, which are better at analyzing sequences and time-based information. A combined CNN–RNN hybrid model can provide stronger deepfake detection by using both spatial and temporal features [9], [35].

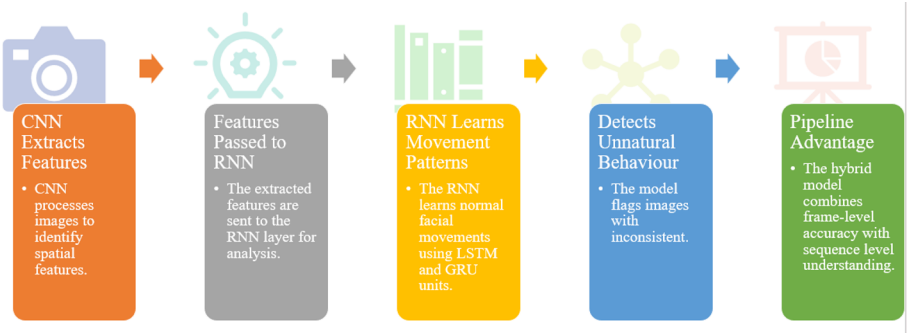

Fig. 5. Combined CNN–RNN Hybrid Model for Deepfake Detection

Based on a survey, we have outlined the steps involved in the combined CNN–RNN Hybrid Model are consolidated in fig 5 as well as in table 5 [9], [29], [30], [33], [34], [35], [36].

Table 5. Step by Step working of Combined CNN-RNN Deepfake Detection Model

Step #	Process	Details
1. CNN extracts features	Frame-level analysis	CNN looks at each video frame and identifies visual artifacts such as edges, textures, and facial landmarks.
2. Features passed to RNN	Sequence input	Extracted features from every frame are forwarded to the RNN layer for temporal modeling.
3. RNN studies movement patterns	Temporal learning	Using LSTM or GRU units, the RNN learns natural facial movements such as blinking, expressions, and motion consistency.
4. RNN detects unnatural behavior	Deepfake identification	If movements appear strange or inconsistent, the model marks the video as suspicious, helping to spot deepfakes.
5. Pipeline advantage	Combined strength	CNN ensures frame-level accuracy, while RNN adds sequence-level understanding. Together, they perform well for both short and long videos.

Benefits and Limitations of Combined CNN–RNN Hybrid Model

As we have seen, the combined CNN–RNN model performs better by offering multiple advantages. The combined model captures both spatial and temporal features, hence improves detection accuracy on video datasets, and works well across different types of manipulations [35]. Still, it also has some limitations. Training can be slowed, especially when large datasets are involved. Also, difficulties with generalization. It may be encountered when the model is tested on unseen data [33], [36]. Our study shows that Vision Transformers (ViTs) are newer models, and they work better than the older CNN-RNN hybrid models by fixing their limitations. These models use self-attention to look at images and movement sequences more flexibly [55]. The key feature of ViTs is it can process entire video sequences without relying on separate

CNN and RNN components [37]. It makes ViT faster, more scalable, efficient and better at generalizing across datasets. It also helps to overcome all limitations of the combined CNN-RNN hybrid model.

6.4 Vision Transformers (ViT) Model

Vision Transformers (ViTs) are new deep learning models that work really well for computer vision tasks [49], [50], [51]. In comparison with traditional deep learning models, ViTs use self-attention mechanisms to process images [15], [37], [42]. This feature makes them more flexible in learning both spatial and temporal patterns. In the context of deepfake detection, the survey notes that ViTs can effectively capture subtle variations in facial features and movements. These features and movements are frequently manipulated in fabricated videos [39], [41], [42]. Our survey helped us to prepare the table 6 which highlights the minute limitations of the combined model which can be overcome by the ViT model.

Table 6. Comparative Table - Subtle Deepfake Features vs. ViT Model Capabilities

Subtle Feature	Limitations of Combined CNN-RNN Model	How effectively ViT model captures
Micro-expressions & temporal consistency	Hybrids may still struggle with extremely subtle micro-movements due to limited resolution of CNN filters and vanishing gradients in RNNs.	ViTs use self-attention to capture fine-grained facial muscle variations across patches and frames, improving temporal consistency detection.
Eye movement & gaze irregularities	CNN-RNN hybrids can detect blink/gaze patterns but may miss rare or irregular eye dynamics when training data is insufficient.	ViTs attend to eye regions globally, modeling blink frequency and gaze shifts more effectively with attention heads.
Lighting & shadow mismatches	Hybrids often rely on local CNN filters, which may not capture global illumination context; temporal modeling adds little benefit here.	ViTs capture global relationships across facial regions, making them sensitive to subtle lighting inconsistencies.
Lip-sync & audio-visual mismatch	Hybrids can align mouth shapes with audio but may fail when phoneme-to-lip mapping is subtle or noisy.	Multimodal ViTs leverage cross-attention to align audio features with visual lip movements, detecting misalignment with high accuracy.

The ViT model divides each video frame into small patches. These patches are then converted into numerical tokens. The self-attention layers of the ViT analyze the relationships between these tokens. It allows the ViT model to understand how different parts of the face or background connect with each other. Once trained, the ViT can detect unnatural transitions, inconsistent blinking, and irregular facial expressions, thereby identifying potential fake videos [63], [65]. This ability to study both local details and global context makes ViTs more efficient and suitable for detecting deepfake manipulations [16], [31], [38].

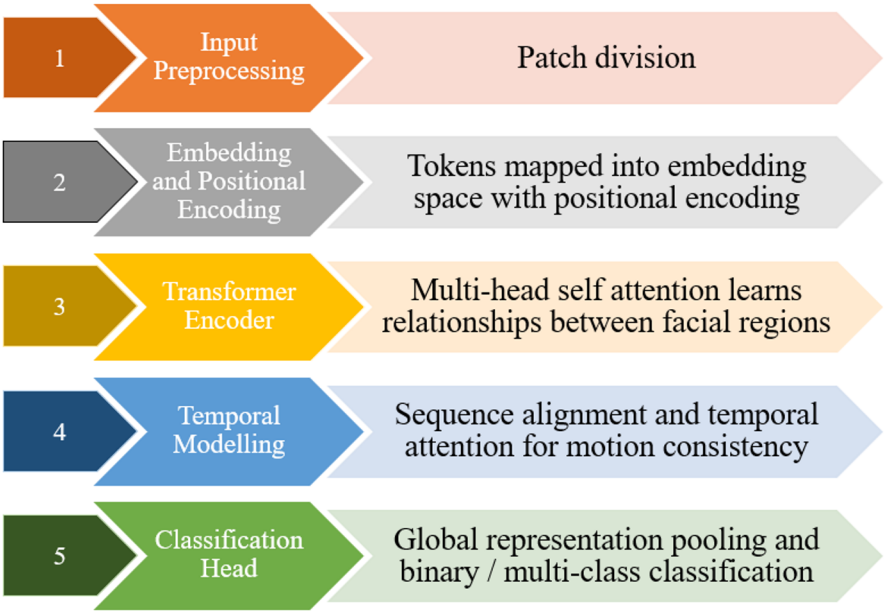

Fig. 6. Steps of Deepfake Detection using ViT Model

Based on a survey of refereed research papers, we have outlined the steps of working involved in the ViT Model for Deepfake Detection by covering very minute details. [15] [16] [31] [37] [38] [39] [41] [42]. The steps involved in deepfake detection using ViT model are depicted in fig 6 and table 7 in detail.

Table 7. Step by Step working of ViT Deepfake Detection Model

Step #	Process	Details
1. Input Preprocessing	Video frame extraction	Input video is split into individual frames.
	Patch division	Each frame is divided into fixed-size patches (e.g., 16×16 pixels).
	Tokenization	Each patch is flattened and projected into a numerical vector (token), similar to NLP word embeddings.
2. Embedding & Positional Encoding	Linear embedding	Tokens are mapped into a continuous embedding space.
	Positional encoding	Positional embeddings are added to preserve spatial order of patches.
	Sequence formation	Patches + positional encodings form a sequence representing the entire frame.
3. Transformer Encoder (Self-Attention Layers)	Multi-head self-attention	Each token attends to all others, learning relationships between facial regions and background.

	Contextual representation	Captures local micro-details (eyes, lips, textures) and global macro-structure (face dynamics, background).
	Layer normalization & feed-forward	Transformer blocks refine token representations.
4. Temporal Modeling Across Frames	Sequence alignment	Tokens from consecutive frames are analyzed together.
	Temporal attention	Learns how facial features evolve over time, detecting unnatural transitions.
	Motion consistency check	Identifies irregular blinking, lip-sync mismatches, abrupt facial changes.
5. Classification Head	Global representation pooling	Final token embeddings are aggregated.
	Binary/multi-class classification	Outputs whether the video is real or fake, sometimes with confidence scores.
	Explainability modules	Heatmaps or attention maps highlight manipulated regions (e.g., mouth, eyes).
6. Advantages in Deepfake Detection	Fine-grained feature capture	Sensitive to subtle pixel-level changes.
	Global context awareness	Understands relationships across the whole face and background.
	Temporal consistency analysis	Detects unnatural motion patterns across frames.
	Robustness to manipulation	Effective against both face-swap and facial reenactment deepfakes.

Benefits and Limitations of ViT Deepfake Detection Model

For deepfake detection, ViTs offer multiple advantages. As the survey was done by us, ViT model is highly effective at capturing fine-grained details, they generalize well to unseen data, and they can handle complex video inputs [15], [16], [31], [37]. But, ViTs also have drawbacks. For accuracy in deepfake detection, ViT requires large datasets for training, and the process can be slow and computationally expensive [61], [62].

6.5 Comparison of Deepfake Detection Models

Our survey has conducted a comparative analysis of deepfake detection models. This study helps researchers point out what each model does well and where it falls short. CNNs, RNNs, combined CNN–RNNs, and Vision Transformers (ViTs) represent the most widely adopted approaches, each designed to capture spatial, temporal, or attention-based features [56]. Our survey shows each model's strengths and weaknesses next to each other in a table. This makes it easier to see how the models work in different situations and helps choose the right one for specific detection tasks [24], [43].

Table 8. Comparison of Existing Deepfake Detection Models

Model	Basic Features	Strengths	Limitations
CNN	Extracts spatial features from individual frames (edges, textures, facial landmarks).	High frame-level accuracy; efficient feature extraction; well-established.	Limited temporal understanding; struggles with motion consistency.
RNN	Learns temporal patterns across sequences using LSTM/GRU units.	Captures facial dynamics (blinking, expressions, motion); good for video.	Weak spatial feature extraction; prone to vanishing gradients; slower.
Combined CNN - RNN	Combines CNN's spatial extraction with RNN's temporal modeling.	Balances frame-level and sequence-level analysis; effective for short/long videos.	Higher computational cost; complex training; risk of overfitting.
ViT	Uses attention mechanisms to model global context across image patches.	Strong global feature learning; robust to diverse manipulations; scalable.	Requires large datasets; computationally heavy; less efficient on small data.

Ultimately, the comparison of existing deepfake detection models is depicted in table 8. Table 8 says CNNs excel at capturing spatial features, while RNNs are better at handling temporal dynamics. When combined, they deliver more consistent results but require heavier computation [45], [60]. ViTs, by contrast, provide strong global feature learning yet depend on large datasets and significant resources [24]. Taken together, these findings suggest that no single architecture can fully address the challenges of deepfake detection. Future progress will likely rely on hybrid and multimodal frameworks that balance efficiency, temporal sensitivity, and attention-based mechanisms [40] [58], [59].

7 Challenges & Limitations

After so much research, still Deepfake detection has many challenges and limits that make it hard to fully trust. Because generative models evolve rapidly, detection systems are often left behind and existing models are circumvented [66]. Another major issue is generalization across datasets and unseen manipulations. Models trained on one dataset often fail when they face different types of forgeries [70]. Adversarial robustness is a concern, since alterations are usually crafted by attackers to mislead detection algorithms. Real-time deepfake detection is very important, but scalability has not been solved yet. The large volumes of multimedia content processed on social platforms remain a major challenge for effective deepfake detection. Beyond technical hurdles, ethical concerns like privacy, surveillance, and censorship risks must also be considered. Together, these challenges highlight the need for more adaptive, transparent, and responsible approaches to deepfake detection [64].

The challenges and limitations of deepfake detection models are shown in fig 7 and explained in the following text.

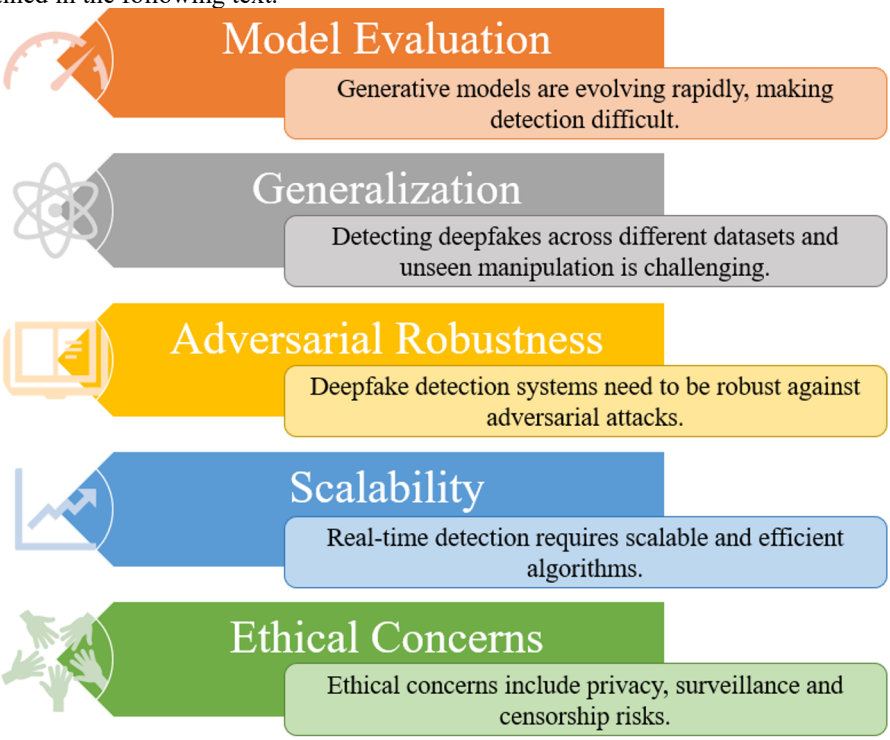

Fig. 7. Challenges and Limitations of Deepfake Detection

- **Rapid evolution of generative models:** Deepfake technology is growing fast. As detection improves, generative models also advance, making fake content more realistic. This constant race makes lasting solutions difficult [66], [67], [68].
- **Generalization across datasets and unseen manipulations:** Researchers have found that deepfake detection models usually work well only on the datasets they were trained with [69], [70]. But when tested on new data or unfamiliar manipulations, their accuracy falls sharply. This lack of generalization makes it hard to rely on these models in real-world settings, where fake content can appear in many different forms.
- **Adversarial robustness:** Deepfake detectors can be fooled by adversarial attacks. Tiny, hidden changes in images or videos may cause them to wrongly classify fake content as real. Current models are not strong enough to resist such attacks, and making them more robust is still a big challenge for researchers [71], [72], [73].
- **Scalability for real-time detection:** Detecting deepfakes in real time is very challenging. Most models are too heavy and slow for quick uses like social media monitoring or live video streams. For practical value, detection systems

must be accurate, fast, and efficient enough to process large volumes of data instantly [74], [75], [76].

- **Ethical concerns:** Deepfake detection brings important ethical concerns [77]. While it can protect people from harm, it may also be misused for surveillance or censorship. Issues such as privacy, freedom of expression, and control of these tools by governments or organizations have been raised [52], [78], [79]. The real challenge is finding a balance between safety and ethical responsibility.

By looking forward with work done while doing the survey, the challenges are included in the table 9 and the impact, research gap is highlighted. Potential directions and examples will help the researchers to move forward in the era of deepfake detection.

Table 9. Deepfake Detection Challenges

Challenge	Impact	Research Gap	Potential Direction	Examples
Rapid evolution of generative models	Newer deepfake techniques quickly outpace existing detectors.	Detection methods lag behind the pace of generative model innovation.	Continuous benchmarking, adaptive detection pipelines, and integration of generative model awareness.	GANs evolving into diffusion models; FaceSwap tools updated faster than detectors.
Generalization across datasets and unseen manipulations	Accuracy drops when tested on new datasets or manipulations.	Overfitting to specific datasets; poor cross-domain generalization.	Domain adaptation, transfer learning, and multimodal approaches.	Model trained on FaceForensics++ failing on Celeb-DF or DFDC datasets.
Adversarial robustness	Detectors can be fooled by small, invisible perturbations.	Limited resilience against adversarial attacks.	Adversarial training, robust feature extraction, and multimodal defenses.	Adding imperceptible noise to video frames causes misclassification.
Scalability for real-time detection	Real-time detection struggles with large volumes of multimedia data.	Lack of scalable architectures for high-throughput environments.	Lightweight, distributed, and hardware-optimized models.	Millions of TikTok/Instagram uploads per day overwhelming detection systems.
Ethical concerns: privacy, surveillance, censorship risks	Risk of misuse for surveillance, censorship, or privacy violations.	Absence of clear ethical frameworks and governance.	Establish ethical guidelines, transparency standards, and privacy-preserving detection methods.	Governments using detection tools for mass surveillance; platforms censoring content.

8 Future Directions

Our survey stated the comparative analysis of existing models. The subtle features which are not supported by existing models can be detected by adopting hybrid models that combine CNNs and ViTs [31]. Future work suggests that, in such hybrid models, CNNs act as efficient feature extractors, while ViTs provide powerful attention-based analysis [54]. This integration is expected to balance speed and accuracy, thereby enhancing robustness and making deepfake detection systems more practical for real-world deployment.

To overcome the challenges and limitations discussed in the previous section, the following future directions are proposed to make deepfake detection more efficient and practically useful. Looking ahead, deepfake detection models are expected to become more robust, faster, and adaptable, capable of handling real-time scenarios and unseen datasets. Combining spatial and temporal features with multimodal inputs such as audio, video, and text is likely to enhance accuracy, while lightweight architectures will be crucial for large-scale deployment. At the same time, ethical guidelines and privacy-preserving methods must evolve to ensure that technical progress is matched with social responsibility.

Apart from the implementation of a novel hybrid model to improve the accuracy, to mitigate the limitations and challenges, one more aspect that needs to be focused by the researcher is deepfake prevention. The researchers can explore training-free proactive detection schemes that use watermarking and identity-aware hashing to strengthen robustness against deepfakes [57].

9 Conclusion

Detecting deepfakes is difficult because generative models keep improving and their manipulations are increasingly subtle. Existing approaches show complementary strengths: CNNs capture fine-grained spatial features, RNNs learn temporal dynamics, and their combination improves both frame-level and sequence-level accuracy. Vision Transformers (ViTs) further enhance detection by modeling global context and attention across patches, offering robustness against diverse manipulations. However, challenges persist, including poor generalization across datasets, vulnerability to adversarial attacks, and scalability issues in real-time deployment. Ethical concerns such as privacy risks, surveillance, and censorship also limit practical adoption. Evaluation metrics such as accuracy, precision, recall, and F1-score remain critical for benchmarking these models and ensuring fair comparison across datasets. Future directions emphasize combined CNN–ViT architectures, lightweight models for large-scale platforms, and multimodal systems that integrate audio, text, and video cues. Robust adversarial defenses and domain adaptation techniques are needed to strengthen resilience. Ethical frameworks and privacy-preserving methods must also be established to ensure responsible

use. Overall, progress in deepfake detection will depend on balancing accuracy, efficiency, scalability, and ethical responsibility to make systems both technically sound and socially trustworthy.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Heidari, Arash, et al. "Deepfake detection using deep learning methods: A systematic and comprehensive review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14.2 (2024): e1520.
2. Rana, Md Shohel, et al. "Deepfake detection: A systematic literature review." IEEE access 10 (2022): 25494-25513.
3. G. Oberoi. Exploring DeepFakes. Accessed: Jan. 4, 2021. [Online]. Available: *https://goberoi.com/exploring-deepfakes-20c9947c22d9*
4. J. Hui. How Deep Learning Fakes Videos (Deepfake) and How to Detect it. Accessed: Jan. 4, 2021. [Online]. Available: https://medium. com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it c0b50fbf7cb9
5. Goodfellow, Ian J., et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
6. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
7. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
8. Tolosana, Ruben, et al. "Deepfakes and beyond: A survey of face manipulation and fake detection." Information Fusion 64 (2020): 131-148.
9. Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
10. Li, Yuezun, et al. "Celeb-df: A large-scale challenging dataset for deepfake forensics." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
11. Yamagishi, Junichi, et al. "Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan." ASV Spoof 13 (2019).
12. Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." arXiv preprint arXiv:2108.05080 (2021).
13. Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai created fake videos by detecting eye blinking." 2018 IEEE International workshop on information forensics and security (WIFS). Ieee, 2018.
14. Kaur, Arvinder, et al. "A Cross-Architecture Evaluation: CNN, LSTM, GANs, and Transformer Models for DeepFake Detection." Blockchain and Federated Learning Synergy for Privacy-Focused DeepFex Solutions. Singapore: Springer Nature Singapore, 2025. 43-63.
15. Heo, Young-Jin, Woon-Ha Yeo, and Byung-Gyu Kim. "Deepfake detection algorithm based on improved vision transformer." Applied Intelligence 53.7 (2023): 7512-7527.
16. Lamichhane, Darshan. "Advanced detection of ai-generated images through vision transformers." IEEE Access (2024).
17. Zhao, Hanqing, et al. "Multi-attentional deepfake detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

18. Mittal, Trisha, et al. "Video manipulations beyond faces: A dataset with human-machine analysis." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023.
19. Comito, Carmela, Luciano Caroprese, and Ester Zumpano. "Multimodal fake news detection on social media: a survey of deep learning techniques." Social Network Analysis and Mining 13.1 (2023): 101.
20. Korshunov, Pavel, and Sébastien Marcel. "Deepfakes: a new threat to face recognition? assessment and detection." arXiv preprint arXiv:1812.08685 (2018).
21. Dou, Zhihao, et al. "Adversarial attacks to multi-modal models." Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis. 2023.
22. Tian, Yapeng, and Chenliang Xu. "Can audio-visual integration strengthen robustness under multimodal attacks?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
23. Zhang, Liguo, et al. "Cross-modal and cross-medium adversarial attack for audio." Proceedings of the 31st ACM International Conference on Multimedia. 2023.
24. Ali, Zaryab. "A Comprehensive Overview and Comparative Analysis of CNN, RNN-LSTM and Transformer." RNN-LSTM and Transformer (December 31, 2024) (2024).
25. Zi, Bojia, et al. "Wilddeepfake: A challenging real-world dataset for deepfake detection." Proceedings of the 28th ACM international conference on multimedia. 2020.
26. Patel, Yogesh, et al. "An improved dense CNN architecture for deepfake image detection." IEEE Access 11 (2023): 22081-22095.
27. Chintha, Akash, et al. "Recurrent convolutional structures for audio spoof and video deepfake detection." IEEE Journal of Selected Topics in Signal Processing 14.5 (2020): 1024-1037.
28. Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2018.
29. Antad, Sonali, et al. "A Hybrid approach for Deepfake Detection using CNN-RNN." 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0. IEEE, 2024.
30. Anjaneyulu, Battula Prasanna, et al. "DeepFake Detection using Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN)." 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI). IEEE, 2024.
31. Ha, Hyunsoo, et al. "Robust deepfake detection method based on ensemble of vit and cnn." Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. 2023.
32. Akhtar, Zahid. "Deepfakes generation and detection: A short survey." Journal of Imaging 9.1 (2023): 18.
33. Al-Adwan, Aryaf, et al. "Detection of deepfake media using a hybrid CNN–RNN model and particle swarm optimization (PSO) algorithm." Computers 13.4 (2024): 99.
34. Al-Dhabi, Yunes, and Shuang Zhang. "Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn)." 2021 IEEE international conference on computer science, artificial intelligence and electronic engineering (CSAIEE). IEEE, 2021.
35. Dolhansky, Brian, et al. "The deepfake detection challenge (dfdc) dataset." arXiv preprint arXiv:2006.07397 (2020).
36. Aarthe, Shree, et al. "A Hybrid Approach for Deep Fake Detection Using Deep Learning Algorithm." International Conference on Power Engineering and Intelligent Systems (PEIS). Singapore: Springer Nature Singapore, 2024.

37. Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
38. Wang, Zhikan, et al. "A timely survey on vision transformer for deepfake detection." arXiv preprint arXiv:2405.08463 (2024).
39. Ghita, Bogdan, et al. "Deepfake image detection using vision transformer models." 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). IEEE, 2024.
40. Wodajo, Deressa, and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer." arXiv preprint arXiv:2102.11126 (2021).
41. Arshed, Muhammad Asad, et al. "Unmasking deception: empowering deepfake detection with vision transformer network." Mathematics 11.17 (2023): 3710.
42. Coccomini, Davide Alessandro, et al. "Combining efficientnet and vision transformers for video deepfake detection." International conference on image analysis and processing. Cham: Springer International Publishing, 2022.
43. Chanda, Kunal, Washef Ahmed, and Souvik Banik. "Deepfake Image Forgery Detection for Suspicious Images." International Conference on Cognitive Computing and Cyber Physical Systems. Singapore: Springer Nature Singapore, 2023.
44. Chanda, Kunal, Washef Ahmed, and Souvik Banik. "Deepfake Image Detection for Low and High Quality Images for Biometric Face Recognition." Proceedings of International Conference on Recent Advancements in Artificial Intelligence. Singapore: Springer Nature Singapore, 2023.
45. Gupta, Satendra, Tapas Saini, and Anoop Kumar. "Detection of AI Manipulated Videos Using Modern Deep Learning Algorithms." International Symposium on Intelligent Informatics. Singapore: Springer Nature Singapore, 2023.
46. Sunil, Reshma, et al. "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation." Heliyon 11.3 (2025).
47. Samal, Debasish, Prateek Agrawal, and Vishu Madaan. "Deepfake Image Detection & Classification using Conv2D Neural Networks." (2024).
48. Alsolai, Hadeel, et al. "Guardian-AI: A novel deep learning based deepfake detection model in images." Alexandria Engineering Journal 126 (2025): 507-514.
49. Altaei, Mohammed Sahib Mahdi. "A detection of deep fake in face images using deep learning." Wasit Journal of Computer and Mathematics Science 1.4 (2022): 60-71.
50. Sohail, Saud, et al. "Deepfake Detection Using Deep Learning: A Unified Forensic Approach to Detect AI-Generated Images and Videos with Fusion of Eye, Nose, and Mouth Landmarks." (2025).
51. Sohail, Saud, et al. "Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning." Information 16.4 (2025): 270.
52. Raza, Ali, Kashif Munir, and Mubarak Almutairi. "A novel deep learning approach for deepfake image detection." Applied Sciences 12.19 (2022): 9820.
53. Yadav, Uma, et al. "A Hybrid Approach for Robust Deep Fake Image Detection using Spatial and Frequency Domain Features." Engineering, Technology & Applied Science Research 15.3 (2025): 22786-22791.
54. Abir, Wahidul Hasan, et al. "Detecting deepfake images using deep learning techniques and explainable AI methods." Intelligent Automation & Soft Computing 35.2 (2023): 2151-2169.
55. Sen, Lord, and Shyamapada Mukherjee. "A Novel Unified Approach to Deepfake Detection." arXiv preprint arXiv:2601.03382 (2026).

56. Lai, Zhimao, et al. "A New Robust Training-free Proactive Deepfake Detection Scheme Using Watermarking and Identity-aware Hashing." Expert Systems with Applications (2026): 131126.
57. Joy, D. Stephy, and R. Thirumalai Selvi. "An enhanced pyramid deep belief network architecture design and implementation for deepfake spotting using deep learning framework." Journal of Ambient Intelligence and Humanized Computing (2026): 1-21.
58. Soudy, Ahmed Hatem, et al. "Deepfake detection using convolutional vision transformers and convolutional neural networks." Neural Computing and Applications 36.31 (2024): 19759-19775.
59. Siddiqui, Fazeela, et al. "Enhanced deepfake detection with DenseNet and Cross-ViT." Expert Systems with Applications 267 (2025): 126150.
60. Ameer, Noor Hayder Abdul, et al. "CNN vs. Transformer-Based Models for Deepfake Detection: A Comparative Analysis." 2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI). IEEE, 2025.
61. Usmani, Shaheen, Sunil Kumar, and Debanjan Sadhya. "Efficient deepfake detection using shallow vision transformer." Multimedia Tools and Applications 83.4 (2024): 12339-12362.
62. Nguyen, Dat, et al. "Fakeformer: Efficient vulnerability-driven transformers for generalisable deepfake detection." arXiv preprint arXiv:2410.21964 (2024).
63. Gong, Liang Yu, and Xue Jun Li. "A contemporary survey on deepfake detection: datasets, algorithms, and challenges." Electronics 13.3 (2024): 585.
64. Wang, Tianyi, et al. "Deep convolutional pooling transformer for deepfake detection." ACM transactions on multimedia computing, communications and applications 19.6 (2023): 1-20.
65. Babaei, Reza, et al. "Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis." Journal of Sensor and Actuator Networks 14.1 (2025): 17.
66. Soundarya, B. C., and H. L. Gururaj. "Deepfake detection: critical review of state-of-the-art approaches and future perspectives." (2025).
67. Lee, Hannah, et al. "The tug-of-war between deepfake generation and detection." arXiv preprint arXiv:2407.06174 (2024).
68. Yermakov, Andrii, et al. "Deepfake Detection that Generalizes Across Benchmarks." arXiv preprint arXiv:2508.06248 (2025).
69. Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. "Deepfake detection: analyzing model generalization across architectures, datasets, and pre-training paradigms." IEEE Access 12 (2023): 1880-1908.
70. Serrano, Adrian, Erwan Umlil, and Ronan Thomas. "Deepfake detectors are DUMB: A benchmark to assess adversarial training robustness under transferability constraints." arXiv preprint arXiv:2601.05986 (2026).
71. Pallavi, N., T. P. Pallavi, and R. Goutam. "Adversarial Robustness in DeepFake Detection: Enhancing Model Resilience with Defensive Strategies." 2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA). IEEE, 2024.
72. Abed, Al-din, et al. "Adversarial Attacks on Deepfake Detectors and Defence Mechanisms: A Cyber Security Model." International Journal of Intelligent Engineering & Systems 18.3 (2025).
73. Nejati, Faranak, et al. "Enhancing Deepfake Detection Generalization: A Scalable and Real-Time Detection." International Conference on Smart Computing and Informatics. Cham: Springer Nature Switzerland, 2025.
74. Lanzino, Romeo, et al. "Faster than lies: Real-time deepfake detection using binary neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

75. Hussain, Tarak, et al. "Decoding deception: state-of-the-art approaches to deep fake detection." Frontiers in Big Data 8 (2025): 1670833.
76. Sarkar, Diya, and Sudipta De Sarkar. "Combatting Deep-Fakes in India-An Analysis of the Evolving Legal Paradigm and Its Challenges." Indian JL & Just. 15 (2024): 346.
77. Verma, Karishma. "Digital Deception: The Impact of Deepfakes on Privacy Rights." Lex Scientia Law Review 8.2 (2024): 859-896.
78. Van der Sloot, Bart, and Yvette Wagensveld. "Deepfakes: regulatory challenges for the synthetic society." Computer Law & Security Review 46 (2022): 105716.
79. Al-Khazraji, Samer Hussain, et al. "Impact of deepfake technology on social media: Detection, misinformation and societal implications." The Eurasia Proceedings of Science Technology Engineering and Mathematics 23.429-441 (2023): 2.
80. Folorunsho, Femi, and Benedicta Frema Boamah. "Deepfake technology and its impact: ethical considerations, societal disruptions, and security threats in ai-generated media." International journal of information technology and management information systems 16.1 (2025): 1060-1080.
81. Wazid, Mohammad, et al. "A secure deepfake mitigation framework: Architecture, issues, challenges, and societal impact." Cyber Security and Applications 2 (2024): 100040.
82. Deng, Jingyi, et al. "Towards benchmarking and evaluating deepfake detection." IEEE Transactions on Dependable and Secure Computing 21.6 (2024): 5112-5127.
83. Yan, Zhiyuan, et al. "Deepfakebench: A comprehensive benchmark of deepfake detection." arXiv preprint arXiv:2307.01426 (2023).
84. Wang, Tianyi, et al. "Deepfake detection: A comprehensive survey from the reliability perspective." ACM Computing Surveys 57.3 (2024): 1-35.
85. Saeed, Najaf, et al. "Improving DeepFake Detection: A Comprehensive Review of Adversarial Robustness, Real-Time Processing and Evaluation Metrics." Journal of Computing & Biomedical Informatics 7.02 (2024).
86. Abbasi, Maryam, et al. "Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks." Applied Sciences 15.3 (2025): 1225.
87. Dolhansky, Brian, et al. "The deepfake detection challenge (dfdc) dataset." arXiv preprint arXiv:2006.07397 (2020).