International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
www.computersciencejournals.com/ijcai
IJCAI 2025; 6(1): 246-256
Received: 19-03-2025
Accepted: 21-04-2025

Jaspreet Singh
Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab, India

Madan Lal
Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab, India

Kanwal Preet Singh Attwal
Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab, India

Corresponding Author:
Jaspreet Singh
Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab, India

Deep learning techniques for deep fake identification: A review

Jaspreet Singh, Madan Lal and Kanwal Preet Singh Attwal

DOI: https://www.doi.org/10.33545/27076571.2025.v6.i1d.159

Abstract
Deep learning has been a hugely successful approach which is utilized in a wide diversity of domains such as Natural Language Processing, Computer Vision, and Machine Learning One of the most controversial applications of Deep Learning is the creation of Deep fakes-synthetically generated videos that closely mimic real individuals, often making the transition from genuine to fabricated footage virtually undetectable to the human eye. To deal with Deep fakes, numerous deep learning-based techniques have been developed to detect such manipulations. This paper presents a comprehensive analysis of deep fake generation and detection systems, emphasizing learning-based approaches. The study provides an in-depth review of the latest detection methods, highlighting their respective strengths and limitations. Furthermore, we examine state-of-the-art techniques used to identify Deep fakes in social media content, positioning our analysis as a valuable resource for academic research. By detailing the most recent methodologies and datasets, this work facilitates meaningful comparisons with prior studies and advances understanding in the rapidly evolving field of deep fake detection.

Keywords: Deep learning, machine learning, deep fakes, fake detection, social media, fake videos

Introduction
Deep fake videos have become much more widespread on social media due to the fact that deep fake technology is now more accessible. Deep fakes are pieces of digital media that have been altered so that another person's face is inserted into them instead of the original one. This is usually the case with photos or films. The fear of this technology is rapidly growing in today's society. A lot of politicians are targeted by Deep fakes when rumors are spread about them or their face is swapped in nude photos or videos [1, 2, 3]. In 2018, a fake video of Barack Obama circulated that included statements he never made [4]. During the 2020 U.S. election, Deep fakes of Joe Biden were circulated in which his jaw was made to look like it was sticking out. By lying especially on social networks, this type of Deep fake, which is intended for malevolent purposes, dangers greatly to be a societal sickness, caused by the disjoint reality that is their truth.

Growing deep learning techniques that are known as generative adversarial networks (GANs) [5] can provide society with the manipulation to blend the authentic and fake slipped together in surface textures and shadows, which are undetectable by humans. The models are brought to the task of creating synthetic media after being trained on datasets. Consequently, Deep fake photos and videos that have been generated are more real-life-like and have bigger chances of persuading the bigger dataset. Through streaming or posting online, presidents and Hollywood celebrities' videos are too easily shared on social media, making fake footage so real-looking that its spread can have a significant influence on society in the form of rumors and misinformation. Deep fake images and videos are now increasingly found on social media platforms; hence, it is quite imperative to identify them. The companies involved include Facebook, Google, and the U.S. Defense Advanced Research Projects Agency (DARPA), which have initiated projects toward deep fake detection and prevention in order to help the study in this area [6, 7]. Deep learning methods, one of which are LSTMs, the other RNNs, and even a combination of them, have been developed primarily to identify the fake imagery generated and spur research in the field of Deep fake development. From the perspective of recent studies, deep neural networks have shown their amazing

ability of detecting social media posts that are either rumors or unauthorized information [8, 9, 10, 11].

This paper uncovers a multifaceted examination of deep fake detection applying deep learning techniques including long short-term memory (LSTM), convolutional neural networks (CNN), and recurrent neural networks (RNN). By means of: 1) a painless context of cutting-edge science, 2) a database review in this discipline, and 3) the exposure of limitations prevailing methodologies along with certain possible solutions, this survey is expected to be the best in the hand of the researchers. The following are the summary of the contributions of this survey:

- This paper offers a complete analysis of the limitations of the artificial intelligence methods currently utilized for Deep fake detection.
- We go through various obstacles researchers face in this field and suggest possible line of investigations.
- We accumulate and demonstrate all the annotated datasets that are available for the research on Deep fake detection.

The remainder of the paper is arranged like this: First, a related work review follows in Section 2, then advanced Deep fake production and detection processes are explained in Section 3, afterwards public datasets especially used for the Deep fake industry are demonstrated in Section 4. The study comes to an end in Section 6 after a review of the issues and challenges faced by this area is expressed in Section 5.

2. Overview of Existing Research

2.1 Basic Structure and Function of ANNs

The basic concept of artificial neural networks (ANNs) is derived, among other things, from the way our brain works. The structure of artificial networks is illustrated in Figure 1. Neural networks, whose structure is multistoreyed, consist of one input layer, at least one hidden layer, and an output layer. A neural network is a tool that is used to turn a group of input values into an input [12]. The skill of neural networks is the prediction and classification of the information into already given types of values.

The input layer, being the first in a neural network, outputs input data to the next layer [13]. The input parameters in our case are x1, x2, x3, and x4. A collection of interconnected components called the second layer's hidden layers constitute artificial neurons (or nodes). The neurons' edges illustrate their connections with each other and ways of sending and receiving signals at different levels. Each connection is assigned a specific weight, which determines the degree of relation between the two nodes. Three neurons form the initial layer of our network, whereas four neurons constitute the second one. Each neuron gets inputs and a bias value from the previous layer. A bias value of 1 is also added as an extra value. The learning equations (Equations (1) and (2)) can be interpreted as the n weight values that a neuron will have if it has n inputs:

$$z = x_1 w_1 + x_2 w_2 + x_3 w_3 + \cdots + x_n w_n + b * 1$$

$$z = \sum_{n=1} x_n w_n + b$$

Output layer, which is the third layer, is responsible for predicting the output values y1, y2, and y3 by using the information obtained from the previous layer. In order to reduce the error and predict the output values, the learning and forecasting process is focused on adapting the connection weights between those units. Neural networks are using activation functions to determine the output values $z = \sum_{n=1} x_n w_n + b$ of the model. The objective of an activation function is to minimize the range of value. In neural networks, activation functions (3), (4), and (5) are mostly the ones that are the most used.

$$sigmoid(z) = \frac{1}{1+exp(-z)}$$

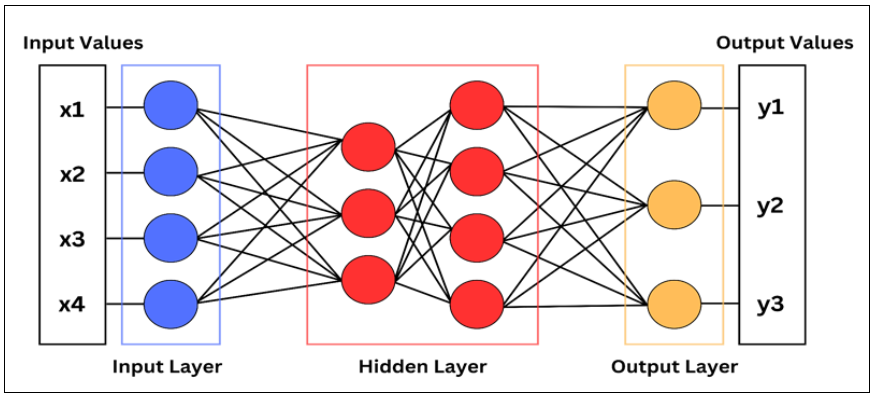

Fig 1: Artificial neural networks architecture (ANNS)

$$tanh(z) = \frac{e^z - e^{-z}}{e^z - e^{-z}}$$

$$ReLU(z) = Max\ (0, z)$$

The sigmoid function (Equation (3) is commonly implemented because it transforms input values to a range in between [0, 1]. As the tanh function (Equation (4)) exhibits this behavior, where the output falls between [-1, 1] rather than [0, 1], it is zero-centered. If the input values are positive, the non-linear Rectified Linear Unit (ReLU) function (Equation (5)) will output the values directly, otherwise it outputs zero. Due to its computational simplicity and faster training time, ReLU has replaced other activation functions as a default in many neural network applications.

A neural network receives the input values just to be able to anticipate the desired output. We call this technique of introspection, forward propagation. Each hidden layer in forward propagation is an instance where an input representation of the previous layer is delivered to it. The damaged layer processes the input, exploits the transfer function, and produces the output. Besides, back propagation is the inverse method, which can be implemented to refine the network's parameters with the result that a model can accurately predict the output. Stochastic gradient descent is a technique used to minimize the error or cost to achieve this.

2.2 Deep Learning

A machine learning technique called deep learning is predicated on the neural network concept [13, 14]."Deep," the expression in deep learning here is the use of many hidden layers in the network. Deep learning architectures, which borrow their patterns from artificial neural networks, make use of altogether finite fixed-size hidden layers to pull out deeper information from raw input data. The number of hidden layers depends on the complexity of the training data [6]. More hidden layers are added for covering more intricate data which is the source of proper results. In the last few years, deep learning has been successfully used in many different areas, such as natural language processing, computer vision, audio processing, and machine translation. Deep learning overcomes traditional machine learning algorithms in many fields that produce research and development using these methods. Deep fakes are also one of the areas in which it has shown a lot of potentials. The literature has demonstrated several techniques based on deep learning, among which are 1) CNN's, 2) RNN's, and 3) LSTM, the Long Short-Term Memory (LSTM). These methods will be the foundation of deep fake detection and we will give a brief explanation in the sections that follow.

2.2.1 Convolutional Neural Network (CNN)

Convolutional neural networks (CNN) are the most preferred deep neural network model. CNN, like neural networks, has one or more hidden layers besides input and output layers.The data that come to a CNN's [15] input layer go first through the hidden layers. Then, these layers perform a convolution operation on this data. This convolution operation is carried out by matrix multiplication or some similar dot product. The CNN after obtaining the result of the matrix multiplication makes use of a non-linear activation function, such as the Rectified Linear Unit (RELU). Further, other layers that follow such as pooling layers are added. Pooling layers are mainly used to calculate outputs by applying such functions as max pooling and average pooling which remove the redundancy of the data.

2.2.2. Recurrent Neural Network (RNN): A Recurrent Neural Network (RNN) is another type of artificial neural network engineered to extract features from sequential input [16]. An artificial neural network (ANN) has multiple hidden layers, each layer having its own weights and biases just like other neural networks. Thanks to the directed cycle created by the connections between the nodes in an RNN, sequential data processing becomes a reality. However, the availability of temporal dynamic behavior by RNNs is one of the pros. In contrast with FFNs, RNNs internal memory is also a component through which they store data relevant to previous inputs. As a result, RNNs are widely used in speech recognition and natural language processing. A recurrent hidden state that records the dependencies of the sequences at different time scales is integrated into RNNs for the temporal sequence [17].

2.2.3. Long Short-Term Memory (LSTM): A recurrent neural network (RNN) of the type LSTM [18, 19] (Long Short-Term Memory) is designed to be effective in dealing with long-term dependencies. The feedback that the unit can make use of enables it to remember sequences of data completely. The LSTMs are widely used in many time series applications including processing, prediction, and classification. The conventional LSTM design has three main components: The action of input, forget, and output gates are the first three things that are happening. It is the earlier Information that is being retained by the cell state that acts as long-term memory. The input gate allows only those values to be fed into the cell state that are managed. The forget gate is a binary classifier that marks the input data for removal with the application of the sigmoid function which has a range of [0, 1]. The output gate decides which data from the current state should be forwarded to the next stage.

3. Deep fake Generation and Detection: Deep fake technology with the use of Generative Adversarial Networks (GANs) makes it possible to create photos or videos that look real but are fake. At first, this section gives a short description of the current instruments and applications for generating the illusion of being real and the art of fabricating videos and images. Next, to cope with and reduce this problem we consider different deep learning-based deception methods.

3.1 Deep fake Generation: A deep neural network that is commonly used in the creation of Deep fakes is called a Generative Adversarial Network (GAN). One of the advantages of GANs lies in their ability to be trained on a dataset and later on, they can present totally new images with similar characteristics and analogies. The GANs, for instance, can construct a "fake" image or video of a person “occupying” a “real” space [1]. GAN's basic architecture consists of an encoder of two parts and a decoder. An encoder is a part of the process which means first, it is trained on a few of the exemplary datasets. This will lead to the generation of the non-genuine data. Afterward, the decoder then can tell between false and real data. Nevertheless, a considerable number of data (different images and videos) are to be used for this model to be able to generate outputs that are not fake. In Figure 2, the GAN architecture is shown. In order to produce a fake sample, the encoder is fed random pixels as inputs. The fakes are presented to a decoder to train it as well. Through the use of a differentiation SoftMax function, the decoder plays a classifier that distinguishes cubes of green and white ones to the given input besides real and fake ones.

Over the past couple of years, many deep fake applications have been popping up. FakeApp was the first and most successful one so far for face-swapping with Deep fakes. The autoencoder-decoder model is employed in this user-generated software on Reddit to interchange faces in videos [20, 21]. FakeApp processes latent features of the human face photographs via an autoencoder and generates face-like features that the decoder then needs to make recognizable, in a way that is essentially the same luckily as in GANs. It is

a simple method but is very effective and often the produced fake videos are so lifelike that it is a real challenge for the viewers to tell the difference. VGGFace is a product of the deep fake method that gains a lot of popularity, it is developed on a GAN basis.Adversarial loss and perceptual loss are the two additional elements of the VGGFace [22] network design. To obtain the potential impact of latent factors of face images such as eye movements for the synthesis of more believable fake images, these layers are added to the autoencoder-decoder structure.

CycleGAN [23] is a Deep fake technique that employs the GAN architecture to obtain the features of one image and then construct another one with the same characteristics. This approach, by employing a cycle loss function, facilitates learning the latent features of the model. In comparison to FakeApp, CycleGAN is an unsupervised method that can do image-to-image translation without the need for paired examples. The model, in a different manner, learns the characteristics of the set of source and target images that are not always interconnected.

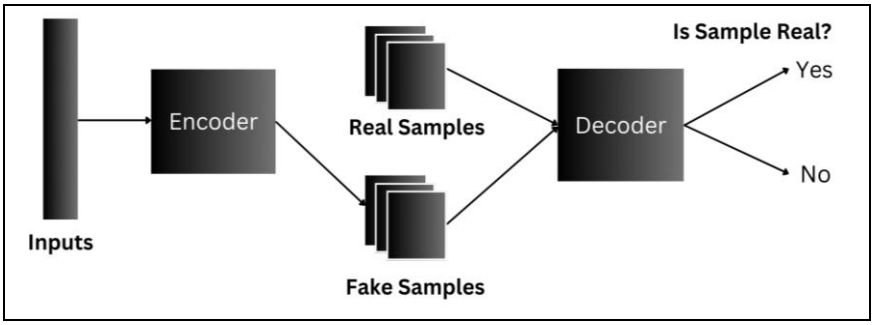

Fig 2: GNA Design

3.2 Deep fake Detection

The contributions of deep learning to deep fake detection have been tremendous to date. The next section talks about video detection models after analyzing image detection algorithms that are based on deep learning techniques.

3.2.1 Image Detection Models

Deep learning networks have been exploited in several ways to develop techniques used in the detection of GAN images. Tariq *et al.* [24] developed and implemented neural network-based methods for fake GAN movies detection. Moreover, this model lessens the probability of human-created fake face photos by processing the images statistically using preprocessing techniques [25]. Correspondingly, Nhu *et al.* [26] proposed the use of a deep convolutional neural network-based approach to identify fake images, which are the creation of GANs. The model starts with the acquisition of the potential features related to real/fake image detection from a deep learning network in the first instance and then these features are fine-tuned to perform image authentication. A few strategies have obtained good results with the data from competition validation.

However, the problem of the lack of generalizability of the forensic models is only considered as a minor issue in many studies to date. In other words, they usually do the same thing, which is they train and test with the same type of dataset. A solution was presented by Xuan *et al.* [27] who developed a forensic convolutional neural network (CNN) to solve such problems that detect fake human pictures through the combination of two image preprocessing techniques; Gaussian Blur and Gaussian Noise. The idea of this model lies in either how to solve high-standard pixel noise at the statistics level or eliminate errors in GAN-generated images for these preprocessing techniques. It tells the difference between a real photograph and a fake one because the fake one has pixel-level indicators that are real in low stats but give too much-constant noise in a high-frequency domain. The data acquired during actual imagery operations is one of the best relevant training sets that allow the classifier to see and recognize not only realistic but also fake surfaces appearing similar to real examples. The model's capability of identifying fake images was verified via the uncompromising results of the tests. Moreover, the technology is advanced further by the creation of hybrid methods besides the traditional deep fake detection models which makes it easier to identify fraudulent images [28, 29, 30]. Thus, a two-stream network for the detection of face manipulation was suggested by Zhou and colleagues. *et al.* [29] (see Figure 3). In the face classification stream, the network uses the GoogleNet [31] as its algorithm. It provides the learners with both real and fake images to train the model in face classification. In addition, the patch triplet stream captures local noise leftovers and the camera's basic properties, by examining the features with a steganalysis module.

The findings prove that this approach can effectively work the network to be able to tell a real photo from a fake one.A couple of other methods have been investigated as a hybrid technique of detecting deep fake images [30]. Generating the fake images first and then capturing the variation between real and fake images, this method is based on a paired learning model of the quite popular fake feature network (CFFN) which was invented by the GANs.

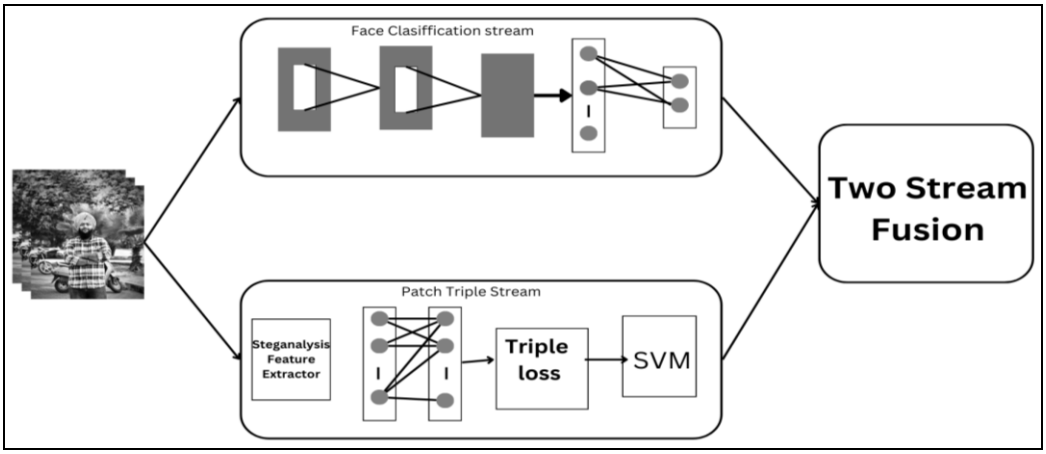

Fig 3: Two-stream neural networks.

Through the evaluation process, it becomes evident that this method is the best one above the currently available fake image detection models.

3.2.2 Video Detection Models

Deep learning techniques have been shown to be effective in detecting fake photos in recent years. On the other hand, due to video compression, this leads to a huge loss of frame information and thus, the techniques cannot be directly applied to the detection of deep fake videos [32, 33]. The earlier work of detecting deep fake videos is split into two major categories in the subsections below: biological signal analysis and the study of temporal and spatial elements.

1) Biological Singles Analysis

Yuezun Li [34] presented a brand-new method for identifying films with false faces that uses neural networks. This strategy, in contrast to other ones, concentrates on eye blinking, a crucial physical characteristic that can aid in identifying phony videos. This is accomplished by combining a recurrent neural network (RNN) and a convolutional neural network (CNN) to record physiological data like blinking and eye movement. After that, the model uses a binary classifier to identify whether the eyes are open or closed. An eye-blinking dataset, especially collected from the internet, was used to test this method. The eye-blinking dataset, created especially for eye-blinking detection, is the first of its type. The outcomes of the experiment show how well this method works to detect phony videos.

In addition, other bodily cues, such as heartbeats, have been identified as reliable indicators for distinguishing between artifacts. A Generative Adversarial Network (GAN)-based algorithm, for instance, was put forward by Ciftci *et al.* [35], which is applicable to the analysis of Deep fake video origins caused by the along-initially 'heartbeat.' In the presence of the authentic video as the input, the model proposed by the authors consists of several detector networks. Subsequently, a pair of real and false films is the input for the registration layer. It makes use of biometric signals and face regions of interest (ROIs) in order to generate photoplethysmogram (PPG) images. A face detector was also used to obtain multiple faces from these spatiotemporal windows, which are referred to as PPG cells. At the last layer of the architecture, the model makes the decision whether the obtained video is genuine or manipulated. The accuracy of the model to distinguish between Deep fakes according to the results of several publicly accessible datasets was 97.3%.

Some earlier studies have demonstrated that there is a significant association between the different audio-visual modalities of the same sample together with biological signals. [36, 37, 38. 39, 40]. Mittal *et al.* have designed a deep learning framework for detecting Deep fakes in multimedia content. [41]. The focus of the model is on the analysis and comprehension of the audio and video modalities and their mutual interaction. To do this, the model at once obtains speech and facial modalities via a Siamese network-based architecture. For the audio data, pyAudioAnalysis and for the visual data, Open Face are the modality representation networks that are used for the vector representation to come to a conclusion of whether the video is real or hoax. Finally, the triplet loss function is used to measure the degree of similarity between the real and fake videos. The Deep fake TIMIT dataset [42] together with the DFDC [43], the two deep fake detection benchmark datasets, was used as test sets for this technique. In the DF-TIMIT dataset, the accuracy was 96.6% and in the DFDC dataset, it was 84.4%.

2) Spatial and Temporal Features Analysis

Deep fake detection is mainly based on analyzing face images from a video, which is not the most suitable method for detecting fake ones, but only a single still image. The frame-level attributes are the possible areas in which video manipulation can take effect. The key fact that has been proven by research is that temporal changes between frames have proved to be an efficient method of distinguishing Deep fakes from real videos. This paper [8] described a time-perception based approach the authors took in order to do deep fake detection. Firstly, the model derived features from face frames using a CNN then added them to LSTM so that the controlling aspect of time is guaranteed and face alteration detection is conducted between frames. Finally, a softmax function is applied to discover whether the video is genuine or not. The overview of the model is displayed in Figure 4. A total of 600 videos were taken from a variety of websites for final evaluation. The display of the model's deep fake detection ability is evidenced by the test results.

Bansal *et al.* [45] unveiled the Recycle-GAN method, which brings about the conversion between spatial and temporal data on the conditional generative adversarial networks and makes it based on the previous Cycle-GAN version [44]. The results from the assessment show that using spatiotemporal restrictions result in better performance. Sabir *et al.* [46]

came up with a novel concept of a convolutional network with recurrent capabilities and applied it in this experiment. The process has two stages: first, face processing and second, face manipulation detection. A Spatial Transformer Network (STN) is enlisted to perform face cropping and alignment in the processing step. The recurrent convolutional network not only detects manipulations on the face through a single frame but also tracks temporal information across frames. See Figure 5 for reference.

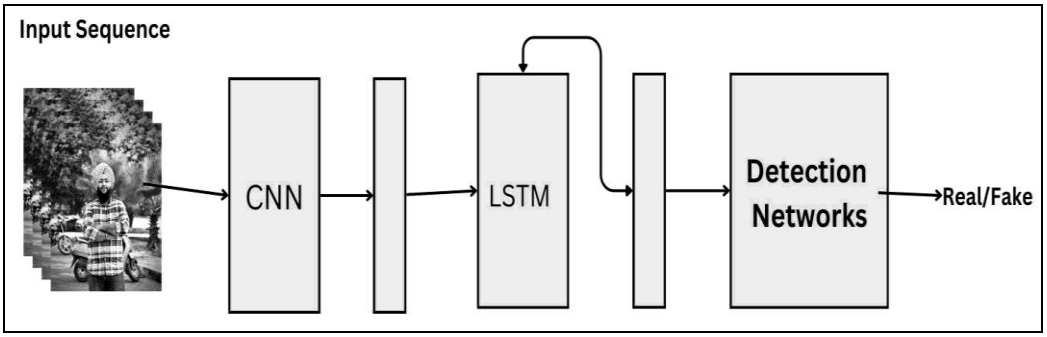

Fig 4: Convolutional neural network for analysis of temporal and spatial features.

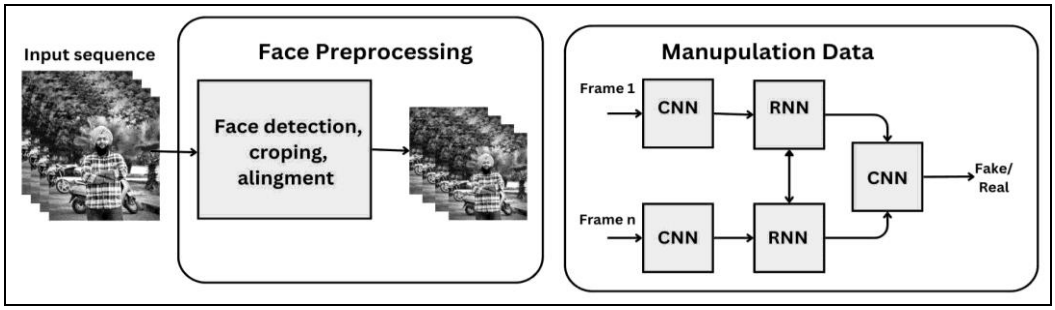

Fig 5: The suggested approach consists of two steps. Face detection, cropping, and alignment are the initial steps. The detection of manipulation is the second step.

Face Forensics++ dataset of public accessibility is employed to evaluate the proposed methodology [47]. In comparison, the outcomes indicate that the model with respect to the model has state-of-the-art performance.

4. Public Datasets

That Are Available besides the above four, there are five additional deep fake detection datasets that are publicly available: 1) DeeperForensics-1.0, 2) Face Forensics++ in version 2; 3) DF40; 4) Celeb-DF; 5) The Eye-Blinking Dataset; 6) 100K-Faces. The relevant information about these datasets, such as their properties and the download URLs, are aggregated in Table 2 in the form of a brief description of the

4.1 DeeperForensics-1.0

DeeperForensics-1.0, made by Jiang *et al.* [48] represents genuine deep fake utilization by giving a huge video dataset with controlled conditions such as the illumination, compression, and adversarial attacks. This dataset was specifically used to test the detection systems' compliance in different episodes.

4.2 Face Forensics++

However, Face Forensics++ is a database of movies that have been changed through different face modification techniques. It was Rössler and others [49] that introduced the dataset to the public. It is common for deep fake detecting methods to be compared to the above fakes, which include both real and manipulated videos. YouTube videos were collected and processed using algorithms such as Face Swap and Deep fakes that modified the faces.

4.3 DF40

One of the most comprehensive sets of deep fake generation methods, namely, the DF40 dataset, consists of 40 different datasets [51]. This system was designed to replicate several types of forgery including full-frame synthetic pictorial synthesis and facial exchange work. This has been one of the main factors in the development of systems for useful detection.

4.4 Celeb-DF

Celeb-DF has been enhanced by including a wider range of and more authentic examples, thus to fix the weaknesses of the old Deep fake datasets [52]. With more than five thousand videos of different and very good-quality material, the established dataset incorporates better face alignment and smoother blending than its predecessors.

4.5 The Eye-Blinking Dataset

The eye-blinking detection has not been addressed by the dataset that is currently available. The eye-blinking datasets, created specifically for this purpose, were made public by Li *et al.* [34]. This dataset includes 50 films and conversations with each individual that run roughly 30 seconds and include at least one instance of eye blinking. The author then marks each video clip's left and right eye states using their own tools.

Dataset	Type	Link
DeeperForensics-1.0	Videos	https://github.com/EndlessSora/DeeperForensics-1.0
Face Forensics++	Videos	https://github.com/ondyari/Face Forensics
DF40	Images and Videos	https://ar5iv.labs.arxiv.org/html/2406.13495
Celeb-DF	Videos	https://github.com/yuezunli/celeb-Deep fake forensics
The Eye-Blinking Dataset	Videos	http://www.cs.albany.edu/%E2%88%BClsw/downloads.html
100K-Faces	Images	https://generated.photos/

4.6 100K-Faces: The 100K-Faces [53], a well-known publicly available dataset, is composed of 100,000 automatic human photos created with StyleGAN [50]. Also, StyleGAN was used to create pictures with a flat background from a big set that included more than 20,000 images from 69 different models.

5. Comparative Analysis of Prior Studies

5.1 Overview of Related Works in Deep fake Detection

Similar approaches are used in several recent surveys that concentrate on how resilient detection methods are to possible attacks. The effectiveness of various detection technologies is unclear, though, particularly in terms of robustness and computing complexity. Only a few surveys explore how detection techniques are actually used in real-world situations. Furthermore, only a small number of these polls deal with the detection of deep fake movies, whereas the bulk focus on the detection of false photos. Since these detectors frequently do not attain comparable performance in real-time applications, the accuracy results provided in many studies are generally unduly optimistic.

We can see below several main findings that are supplementary to the research of studies on various methodologies and repositories for detecting Deep fakes.

Sr. No.	Author(s)	Method	Dataset(s)	Dataset Size	Media Type	Strengths	Limitations	Result Accuracy
1	Nataraj *et al.* (2019) [1]	Co-occurrence matrices for texture analysis	Not specified (synthetic GAN images)	Not specified	Images	Simple and effective for detecting GAN-generated images	Limited to specific GAN architectures; less effective on diverse datasets	Not specified
2	Wang *et al.* (2020) [2]	CNN-based classification with data augmentatio-n	CelebA, custom GAN-generated dataset	1M images	Images	Robust to various GAN models; high generalizability	May fail with advanced GANs; requires large training data	99.5%
3	Hsu *et al.* (2018) [3]	CNN for fake face detection	Custom wild dataset	Not specified	Images	Effective in real-world scenarios	Limited dataset diversity; sensitive to image quality	85%
4	Güera & Delp (2018) [8]	Recurrent Neural Networks (RNN) with CNN features	Custom Deep fake videos	Not specified	Videos	Captures temporal inconsistencies effectively	High computational cost; limited to specific video types	Not specified
5	Li & Lyu (2018) [9]	Face warping artifact detection	Custom Deep fake dataset	Not specified	Video	Detects subtle spatial artifacts	Fails with high-quality Deep fakessensitive to compress- ion	Not specified
6	Yang *et al.* (2019) [10]	Inconsistent head pose analysis	Custom Deep fake dataset	Not specified	Video	Robust to facial manipulatio-ns	Requires clear head pose visibility; less effective on low-resolution videos	89%
7	Marra *et al.* (2018) [11]	CNN-based detection of GAN images	Social media dataset (GAN-generated)	Not specified	Images	Effective for social media contexts	Limited to specific GAN models; sensitive to image compress-ion	90%
8	Tariq *et al.* (2018) [24]	CNN for machine and human-created fakes	Custom wild dataset	Not specified	Image	Handles both synthetic and human-edited fakes	Limited dataset size; sensitive to noise	92%
9	Hsu *et al.* (2020) [30]	Pairwise learning with CNN	Custom Deep fake dataset	Not specified	Images	Improved generalization through pairwise comparison	Computationally intensive; requires paired data	87%
10	Afchar *et al.* (2018) [33]	MesoNet (compact CNN)	Custom Deep fake video dataset	Not specified	Video	Lightweight and fast; suitable for real-time detection	Less effective on high-quality Deep fakes	95%
11	Li *et al.* (2018) [34]	Eye blinking detection	Custom Deep fake dataset	Not specified	Video	Highly specific to biological cues	Fails if eye region is obscured or manipulat-ed	92%
12	Ciftci *et al.* (2020) [35]	Biological signal analysis (heart rate residuals)	Custom Deep fake dataset	Not specified	Video	Novel use of physiologic-al signals	Requires high-quality video; sensitive to noise	90%
13	Mittal *et al.* (2020) [41]	Audio-visual Deep fake detection	Custom audio-visual dataset	Not specified	Audio + Video	Multimodal approach improves robustness	High computational complexity; limited dataset diversity	93%
14	Sabir *et al.* (2019) [46]	Recurrent convolutional networks	Face Forensics++	~1M frames	Video	Effective for face manipulation detection	Sensitive to compress-ion; requires large	96%

							datasets	
15	Rossler *et al.* (2019) [47]	CNN-based detection (Face Forensics++)	Face Forensics++	~1M frames	Images + Videos	Comprehensive dataset; robust to multiple manipulation techniques	Limited to known manipulat-ion methods	98%
16	Li *et al.* (2023) [54]	CNN + RNN (Eye Blink)	Celeb-DF, FF++	~8,000 videos	Video	Detects temporal eye patterns	Fails with high-quality fakes	89.5%
17	Masi *et al.* (2022) [55]	Lip-sync + Audio-Visual Transformer	DFDC, VoxCeleb	100,000+ videos	Audio + Video	Works on audio-video mismatch	Needs clear voice track	92.3%
18	Chugh *et al.* (2022) [56]	Vision Transformer (ViT)	FF++, DFD	10,000+ videos	Video	Captures spatial + temporal features	High computati-on	94.8%
19	Verdoliva (2021) [57]	Image Forensics + Frequency Analysis	FF++	~5,000 videos	Image/Video	Good for compressed videos	Limited to visual-only data	86.7%
20	Sabir *et al.* (2023) [46]	Multi-stream CNN-LSTM	Celeb-DF, FF++	12,000 videos	Video	Uses temporal sequence	Training time is high	91.2%
21	Tharindu Fernando *et al.* (2025) [58]	Comprehensive review of face Deep fake generation and detection methods	Various public datasets	Not specified	Image, Video	Provides a structured analysis of state-of-the-art methods and their implications on face biometrics	Lacks experime-ntal evaluation of methods	Not applicable
22	Ping Liu *et al.* (2024) [59]	Survey on evolution from single-modal to multi-modal facial Deep fake detection	Multiple datasets	Not specified	Image, Video	Highlights the transition to multi-modal detection approaches	Does not provide quantitate-ve performa-nce metrics	Not applicable
23	Binh M. Le *et al.* (2024) [60]	Systematization of Knowledge (SoK) on facial Deep fake detectors	16 detectors evaluated	Not specified	Image, Video	Offers a unified framework categorizin--g detectors and assessing their generalizabi-lity	Focuses primarily on facial Deep fakes	Not specified
24	Florinel-Alin Croitoru *et al.* (2024) [61]	Review of detection in generative AI (GANs, NERFs)	Multiple datasets	Not specified	Image, Video, Audio	Highlights emerging threats and methods	Existing models fail on new types	Not specified
25	Xixi Hu (2025) [62]	Experimental comparison of detection models	Various datasets	Not specified	Image, Video	Compares cross-dataset generalization	Low performance on unseen forgeries	Not specified
26	Liang Yu Gong & Xue Jun Li (2024) [63]	Survey on datasets and detection algorithms	Multiple datasets	Not specified	Image, Video	Summarizes existing datasets and model types	Lacks analysis depth on detection technique-s	Not specified
27	Dennis Lucky Tuanwi Bale *et al.* (2024) [64]	Review of image-based detection from video	Various datasets	Not specified	Image, Video	Discusses techniques applied to extracted video frames	Does not cover full video-based detection	Not specified
28	Singh Chauhan *et al.* (2025) [65]	Review of detection techniques and challenges	Multiple datasets	Not specified	Image, Video	Covers trends, methods, and research gaps	No model testing or performa-nce metrics	Not applicable
29	Gourav Gupta *et al.* (2024) [66]	Review of ML and fusion-based detection	Various datasets	Not specified	Image, Video	Highlights machine learning and hybrid systems	No accuracy values reported	Not specified
30	Nuria Alina Chandra *et al.* (2025) [67]	Evaluation on Deep fake-Eval-2024 benchmark	Deep fake-Eval-2024	45h video, 56.5h audio,1,975 images	Image, Video, Audio	Real-world benchmark showing generalization limits	Detectors drop ~50% performa--nce	50%

6. Challenges and Open Issues

There is an enormous amount of deep fake photos and videos produced daily due to the wide availability of tools and applications for doing so. To those who are students who are into the examination and analysis of Deep fakes, such a jump is a big deal. A key factor is the absence of extensive and high-quality datasets. Severely limited scalability faces the current deep learning techniques, which means although models are often trained on fragmented datasets that are unique to face-swap tasks, the real-life efficacy of these datasets is low, and therefore, they are not that useful.

To efficiently process vast, high-quality databases, scientists must as well produce models that can be easily extended and are resistant to failure. Another difficulty connected to deep learning model training is the huge dataset need that comes out of it; these datasets mostly have a limited scope or are not accessible publicly due to the permissions requested by social media networks.

Moreover, the constant innovation of deep fake creation techniques, especially by the advanced GAN models, allows the appearance of new types of synthetic media that could be missed by the current detection Algorithm. As people keep on using deep fake techniques, the issue of detecting and removing altered content in information becomes even more challenging. In the end, the challenge of remaining perfectly in tune with detection tech regulations becomes unbearable with the introduction of synthetic speech and video Deep fakes and the addition of algorithmic AI to real-time applications.

Thus, a corporate requirement for more comprehensively agile and scalable deep learning is structured and the

development of these models is needed to catch up with the exponential development of deep fake technology and remain intact on various data sources and kinds of data.

7. Conclusion and Future Directions: In conclusion, the rapid rise of deep fake technology is reshaping how we understand truth in the digital world. Social media, in particular, has become a space where real and fake content often blend together, making it harder than ever to tell what's genuine. This study examined how Deep fakes are created and detected, focusing on the role of deep learning techniques like CNNs, RNNs, and LSTMs. While GAN-based models have made encouraging progress in spotting fake media, the challenge remains ongoing. With tools like FakeApp and CycleGAN becoming more accessible, creating realistic Deep fakes is easier than before, fuelling a continuous battle between creators and detectors. Our review of datasets such as Face Forensics++ and Celeb-DF highlighted the importance of having diverse, high-quality data to train reliable detection models. Yet, challenges remain—large data demands, limited dataset variety, and the constant evolution of deep fake techniques mean detection methods must continually adapt. Ultimately, deep fake detection is not a one-time fix but an evolving effort that must keep pace with the technology it seeks to counter. Continued innovation and vigilance will be key to staying ahead in this ongoing digital arms race.

Deep fakes have become a real challenge for how we trust what we see and hear online. They blur the lines between what's real and what's fake, sometimes with serious consequences for public trust and even democracy. While deep learning has made good progress in spotting these tricks—catching subtle signs like unnatural blinking or odd timing—it's still tough to make these tools work perfectly in the messy, unpredictable world outside the lab. Some newer methods that combine different approaches look promising, but they still struggle with things like compressed videos or differences in people's appearance. Tests show a wide range of success rates, which reminds us that real-world detection is far from a solved problem.

Looking ahead, the key to staying ahead will be flexible, adaptable detection systems trained on rich, diverse data that reflects real-life conditions—different lighting, video quality, and cultural contexts. Projects like Deeper Forensics-1.0 are great starting points, but we need more open access to datasets so the whole community can contribute and improve. Detection models also can't stay static. They need to learn and evolve quickly as new deep fake techniques appear. Lightweight designs will help make detection tools usable even on devices with limited power. And combining multiple types of information—like video, audio, and context—seems to be the smartest way forward, as seen in models that merge these signals effectively.

But this fight isn't just about technology. It requires everyone—from researchers and industry to governments and communities—to work together: sharing data, setting standards, and thinking carefully about ethics. Open, community-driven platforms can speed progress and make things more transparent. Early detection systems that can stop Deep fakes before they go viral are especially important. Finally, as we develop these tools, we have to balance security with creativity and fairness. The goal is to protect truth without limiting expression or introducing new biases.

In the end, Deep fakes aren't just a technical problem—they touch the very heart of how we share and trust information in society. As the technology keeps improving, our commitment to defending authenticity must grow right alongside it. With smarter models, better data, and global cooperation, we can help people and institutions face this new reality with confidence.

References

1. Nataraj L, *et al*. Detecting GAN generated fake images using co-occurrence matrices. Electron Imaging. 2019;2019(5):532-1-532-537. https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532
2. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot. for now. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2020 Jun 13-19; Seattle. 2020. p. 8695-8704. https://doi.org/10.1109/CVPR42600.2020.00872
3. Hsu C-C, Lee C-Y, Zhuang Y-X. Learning to detect fake face images in the wild. In: 2018 IEEE Int Symp Comput, Consum Control (IS3C); 2018 Dec 6-8; Taichung. 2018. p. 388-391. https://doi.org/10.1109/IS3C.2018.00104
4. Vaccari C, Chadwick A. Deep fakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Soc Media Soc. 2020;6(1):1-13. https://doi.org/10.1177/2056305120903408
5. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv. 2014. Available from: https://arxiv.org/abs/1411.1784
6. Kwok AO, Koh SG. Deep fake: a social construction of technology perspective. Curr Issues Tour. 2020;23(20):1-5. https://doi.org/10.1080/13683500.2020.1738357
7. Westerlund M. The emergence of Deep fake technology: a review. Technol Innov Manag Rev. 2019;9(11):40-53. https://doi.org/10.22215/timreview/1282
8. Güera D, Delp EJ. Deep fake video detection using recurrent neural networks. In: 2018 15th IEEE Int Conf Adv Video Signal Based Surveill (AVSS); 2018 Nov 27-30; Auckland. 2018. p. 1-6. https://doi.org/10.1109/AVSS.2018.8639163
9. Li Y, Lyu S. Exposing Deep fake videos by detecting face warping artifacts. arXiv. 2018. Available from: https://arxiv.org/abs/1811.00656
10. Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses. In: 2019 IEEE Int Conf Acoust Speech Signal Process (ICASSP); 2019 May 12-17; Brighton. 2019. p. 8261-8265. https://doi.org/10.1109/ICASSP.2019.8683164
11. Marra F, Gragnaniello D, Cozzolino D, Verdoliva L. Detection of GAN-generated fake images over social networks. In: 2018 IEEE Conf Multimedia Inf Process Retrieval (MIPR); 2018 Apr 10-12; Miami. 2018. p. 384-389. https://doi.org/10.1109/MIPR.2018.00084
12. Grekousis G. Artificial neural networks and deep learning in urban geography: a systematic review and meta-analysis. Comput Environ Urban Syst. 2019;74:244-256. https://doi.org/10.1016/j.compenvurbsys.2018.10.008

13. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A. 1982;79(8):2554-2558. https://doi.org/10.1073/pnas.79.8.2554
14. Pouyanfar S, *et al*. A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv. 2018;51(3):1-36. https://doi.org/10.1145/3234150
15. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge: MIT Press; 2016.
16. Elman JL. Finding structure in time. Cogn Sci. 1990;14(2):179-211. https://doi.org/10.1207/s15516709cog1402_1
17. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157-166. https://doi.org/10.1109/72.279181
18. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
19. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 1997;45(11):2673-2681. https://doi.org/10.1109/78.650093
20. Faceswap: Deep fakes software for all [Internet]. GitHub; [cited 2025 May 30]. Available from: https://github.com/Deep fakes/faceswap
21. FakeApp 2.2.0 [Internet]. Malavida; https://www.malavida.com/en/soft/fakeapp
22. Keras-VGGFace: VGGFace implementation with Keras framework [Internet]. GitHub; Available from: https://github.com/rcmalli/keras-vggface
23. CycleGAN [Internet]. Jun-Yan Zhu; https://junyanz.github.io/CycleGAN/
24. Tariq S, Lee S, Kim H, Shin Y, Woo SS. Detecting both machine and human created fake face images in the wild. In: Proc 2nd Int Workshop Multimedia Privacy Security; 2018 Oct 15; Toronto. 2018. p. 81-87. https://doi.org/10.1145/3267357.3267367
25. Li H, Li B, Tan S, Huang J. Detection of deep network generated images using disparities in color components. arXiv. 2018. Available from: https://arxiv.org/abs/1808.07276
26. Do N-T, Na I-S, Kim S-H. Forensics face detection from GANs using convolutional neural network. ISITC. 2018;:1-5.
27. Xuan X, Peng B, Wang W, Dong J. On the generalization of GAN image forensics. In: Chinese Conference on Biometric Recognition. Berlin: Springer; 2019. p. 134-41. https://doi.org/10.1007/978-3-030-31456-9_15
28. Liu F, Jiao L, Tang X. Task-oriented GAN for PolSAR image classification and clustering. IEEE Trans Neural Netw Learn Syst. 2019;30:2707-19. https://doi.org/10.1109/TNNLS.2018.2885799
29. Zhou P, Han X, Morariu VI, Davis LS. Two-stream neural networks for tampered face detection. Proc IEEE Conf Comput Vis Pattern Recognit Workshops. 2017;:1831-9. https://doi.org/10.1109/CVPRW.2017.229
30. Hsu C-C, Zhuang Y-X, Lee C-Y. Deep fake image detection based on pairwise learning. Appl Sci. 2020;10:370. https://doi.org/10.3390/app10010370
31. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proc IEEE Conf Comput Vis Pattern Recognit. 2016;:2818-26. https://doi.org/10.1109/CVPR.2016.308
32. Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S. Deep learning for Deep fakes creation and detection. 2019;1:1-10.
33. Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: A compact facial video forgery detection network. 2018 IEEE Int Workshop Inf Forensics Secur (WIFS). 2018;:1-7. https://doi.org/10.1109/WIFS.2018.8630761
34. Li Y, Chang M-C, Lyu S. In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking. 2018 IEEE Int Workshop Inf Forensics Secur (WIFS). 2018;:1-7. https://doi.org/10.1109/WIFS.2018.8630787
35. Ciftci UA, Demir I, Yin L. How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. 2020 IEEE Int Joint Conf Biometrics (IJCB). 2020;:1-10. https://doi.org/10.1109/IJCB48548.2020.9304909
36. Ciftci UA, Demir I, Yin L. FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell. 2020;:1. https://doi.org/10.1109/TPAMI.2020.3009287
37. Shan C, Gong S, McOwan PW. Beyond facial expressions: Learning human emotion from body gestures. Proc Br Mach Vis Conf. 2007;:1-10. https://doi.org/10.5244/C.21.43
38. Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41:423-43. https://doi.org/10.1109/TPAMI.2018.2798607
39. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex. 2007;17:1147-53. https://doi.org/10.1093/cercor/bhl024
40. Sanders DA, Goodrich SJ. The relative contribution of visual and auditory components of speech to speech intelligibility as a function of three conditions of frequency distortion. J Speech Hear Res. 1971;14:154-9. https://doi.org/10.1044/jshr.1401.154
41. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D. Emotions don't lie: An audio-visual Deep fake detection method using affective cues. Proc 28th ACM Int Conf Multimedia. 2020;:2823-32. https://doi.org/10.1145/3394171.3413570
42. Korshunov P, Marcel S. Deep fakes: A new threat to face recognition? Assessment and detection. 2018;:1-8.
43. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The Deep fake Detection Challenge (DFDC) preview dataset. 2019;:1-5.
44. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf Comput Vis. 2017;:2223-32. https://doi.org/10.1109/ICCV.2017.244
45. Bansal A, Ma S, Ramanan D, Sheikh Y. Recycle-GAN: Unsupervised video retargeting. Proc Eur Conf Comput Vis (ECCV). 2018;:119-35. https://doi.org/10.1007/978-3-030-01228-1_8
46. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. 2019;:1-10.
47. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Face Forensics++: Learning to detect

manipulated facial images. Proc IEEE/CVF Int Conf Comput Vis. 2019;:1-11. https://doi.org/10.1109/ICCV.2019.00009
48. Jiang L, Yang H, Li C, Sun X, Qian Y, *et al*. DeeperForensics-1.0: A large-scale dataset for realistic Deep fake detection. 2020;:1-10.
49. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, *et al*. Face Forensics++: Learning to detect manipulated facial images. 2019;:1-11.
50. Yan Z, Ding Y, Zhang T, Wang S, Yu N, *et al*. DF40: Toward next-generation Deep fake detection. 2023;:1-9.
51. Li Y, Chang MC, Shi H, Lyu S, *et al*. Celeb-DF: A large-scale challenging dataset for Deep fake forensics.
52. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 15-20 June 2019, 4401-4410. https://doi.org/10.1109/CVPR.2019.00453
53. Generated Photos. 100,000 faces generated by AI. 2018. Available from: https://generated.photos
54. Li Y, Chang MC, Lyu S. In ictu oculi: exposing AI created fake videos by detecting eye blinking. IEEE International Workshop on Information Forensics and Security (WIFS), 2018. Available from: https://arxiv.org/abs/1806.02877
55. Masi I, Tran AT, Hassner T, Medioni G. Audio-visual synchrony for Deep fake detection. CVPR Workshops, 2022.
56. Chugh K, Jain A, Jalal AS. A multi-scale vision transformer for Deep fake video detection. Pattern Recognition Letters. 2022;162:38-45. https://arxiv.org/abs/2110.00036
57. Verdoliva L. Media forensics and Deep fakes: an overview. IEEE Journal of Selected Topics in Signal Processing. 2020;14(5):910-932. https://ieeexplore.ieee.org/document/9109282
58. Fernando T, Priyasad D, Sridharan S, Ross A, Fookes C. Face Deep fakes - a comprehensive review. arXiv preprint arXiv:2502.09812. 2025.
59. Liu P, Tao Q, Zhou JT. Evolving from single-modal to multi-modal facial Deep fake detection: a survey. arXiv preprint arXiv:2406.06965. 2024.
60. Le BM, Kim J, Tariq S, Moore K, Abuadbba A, Woo SS. SoK: facial Deep fake detectors. arXiv preprint arXiv:2401.04364. 2024.
61. Croitoru FA, Hiji AI, Hondru V, Ristea NC, Irofti P, Popescu M, *et al*. Deep fake media generation and detection in the generative AI era: a survey and outlook. arXiv preprint arXiv:2411.19537. 2024.
62. Hu X. A comprehensive evaluation of Deep fake detection methods: approaches, challenges and future prospects. ITM Web of Conferences. 2025;73:03002.
63. Gong LY, Li XJ. A contemporary survey on Deep fake detection: datasets, algorithms, and challenges. Electronics. 2024;13(3):585.
64. Bale DLT, Ochei LC, Ugwu C. Deep fake detection and classification of images from video: a review of features, techniques, and challenges. International Journal of Intelligent Information Systems. 2024;13(2):20-28.
65. Chauhan SS, Shieh CS, Horng MF. A comprehensive review of Deep fake detection techniques: challenges, methodologies, and future directions. Journal of Neonatal Surgery. 2025;14(18S):323-3.
66. Gupta G, Raja K, Gupta M, Jan T, Whiteside ST, Prasad M. A comprehensive review of Deep fake detection using advanced machine learning and fusion methods. Electronics. 2024;13(1):95.
67. Chandra NA, Murtfeldt R, Qiu L, Karmakar A, Lee H, Tanumihardja E, *et al*. Deep fake-Eval-2024: a multi-modal in-the-wild benchmark of Deep fakes circulated in 2024. arXiv preprint arXiv:2503.02857. 2025.