International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Analysing the landscape of Deep Fake Detection: A Survey

Kishan Vyas[1], Preksha Pareek[1], Ruchi Jayaswal[1], Shruti Patil[1]

Submitted: 05/11/2023 **Revised**: 24/12/2023 **Accepted**: 06/01/2024

Abstract: With the rapid advancement of deep learning and generative modelling techniques, the creation of hyper-realistic synthetic media, known as deepfakes, has become a growing concern. These manipulated media assets are in various domains, including politics, entertainment, and security. As a result, the development of effective deepfake detection systems has gained significant attention. To identify Deepfakes, many researchers have developed various binary-classification-based detection techniques. This survey provides a comprehensive overview of state-of-the-art deepfake detection methods, their underlying principles, datasets used for training and evaluation, and the challenges faced in this evolving field. Then, alternative methods have been discussed in literature to address an issue raised by Deepfake. We analyse different techniques by groping them into four categories: Image-based DFDT, Audio-based DFDT, Video-based DFDT, and Multimodality based DFDT. Researchers in this area will learn from our study because it contains cutting-edge methods for detecting deep-fake videos and photos in social media. To benchmark and facilitate the advancement of deepfake detection systems, numerous datasets are examined, such as FaceForensics++, DeepFake Detection Challenge (DFDC), and Celeb-DF. Also discuss their characteristics, variations, and limitations, emphasizing the need for diverse and realistic datasets to ensure the model generalization.

Keywords: *DeepFake, Artificial Intelligence, Machine learning, Deep learning, CNN, RNN, Multimodality, Forensics, Cyber security*

1. Introduction

The proliferation of artificial intelligence and deep learning techniques has led to the emergence of highly sophisticated manipulated media, commonly referred to as "deepfakes." Deepfakes are synthetic media, including images, videos, and audio, that are convincingly altered or generated using advanced machine learning algorithms. DeepFake uses different types of machine learning methods to edit or manufacture auditory and visual content that can more easily mislead, and producing fake content is not a new task [1]. The traditional deep learning and machine learning algorithms for creating deepfakes training generative neural network designs such Autoencoders or Generative Adversarial Networks (GANs). This problem has been around for a while. Digital data played little of a role in the process of legitimizing papers in the past. It was limited to the proofing, verification, and investigation of documents. It is impossible to ignore the rapid expansion of digital data throughout the Internet and its importance in daily life, including applications for digital advertising, legal forensic imagery, health visuals, sensitive satellite image processing, and more [2]. A rise in cybercrime is also being fuelled by the manner that digital data are developing in various applications.

For a very long time, people have been interested in the alteration of image, audio, and video content. While using tools like Photoshop to modify images is quite simple. It can be very difficult to manipulate audio and video. Nowadays videos are to be manipulated frame by frame to bring unknowable effects into the film industries. Nowadays, everyone is enabled to do video editing on their personal computer. Recent developments in AI and ML algorithms have led to the introduction of a new method for editing images, audio, and video. When celebrities' faces were projected into pornographic videos using misleading Face Swap Machine Learning algorithms, it first became known to the public [3].

One of the most famous examples of this technique is a fake video of the US president where he got a warning from Deepfake techniques [3]. Deepfake is the overarching term for images, videos, and sounds that have been used deep learning techniques. Fully computer-generated photos that are almost unrecognizable from actual photographs can also be created using certain techniques. Such images and films can be produced by professionals, but there are tools available that make it simple to produce false media. Available methods are used to detect fake media but there is one more problem lack of labelled data, and we cannot do it manually. For that, some of the models can be classified the real and fake but still, it is lower accuracy with Convolutional Neural Network (CNN) and some of the pre-trained models. Some models give accuracy better but only one or two datasets with simple Deep Neural Networks (DNN) and GAN [4]. Deep Learning researchers achieved numerous related advancements in generative modeling, according to a yearly report in Deepfake. For facial re-enhancement, for instance, computer vision experts suggested a technique called Face2Face [5]. Using this technique, facial emotions from one person can be instantly transferred to an actual computer "avatar." Cycle GAN

[1] *Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune, Maharashtra, India.*
* *Corresponding Author Email: kishanvyas012@email.com*

developed by UC Berkeley academics can change the aesthetic of photos and movies [5].

Fig. 1. Understanding of DeepFake.

Figure 1 shows the general understanding of how people forage media by using different techniques in machine learning and how fake media is used in the real world. Here cybercrime is just one example of Deepfake.

According to recent news, many examples are there which introduce Deepfakes are more and more used in day-to-day life for earning, cyber-crime, etc. A Bloomberg writer using the name Maisy Kinsley on social media platforms like Twitter and LinkedIn, for instance, was most likely Deepfake. Her profile photo seems odd, possibly created by a computer. The fact that Maisy Kinsley's profile on social media kept attempting to get in touch with Tesla stock short sellers suggests that the profile was constructed for financial gain. The amount of multimedia content (such photos and videos) that are freely available has increased because of the widespread use of digital smart devices like smartphones, computers, and cameras. Secondly, the growth of social media over the past ten years has enabled individuals to instantly access amassed media content, greatly increasing output and accessibility of multimedia content [49].

For textual Deepfake, a software program dubbed Grover has been created by the Allen Institute for Artificial Intelligence to identify fake content that circulates online. This software, according to researchers, can identify deep-fake writings 92% of the time. Grover utilizes a test set created by the free and open-source web crawler and archive Common Crawl [35]. As an example, a team of academics from Harvard and the MIT-IBM Watson laboratory developed the Giant language model test room, a website which seeks to determine whether the given language produced by AI.

1.1. Motivation

Images are more spoken compared to text and in the digital world, early days every domain will use an image for authentication. Some of the areas are already using it. Also, there are machine learning techniques also used for generating fake images by using different variants of GAN. Now the world of databases human beings cannot be classifying real and fake images separately that's why the concept is to detect forage or detect fake images. Since at least 15 years ago, prominent IT corporations, governmental agencies, and research communities have all contributed to the development of multimedia forensics [31]. The DARPA (Defence advanced research projects agency) of the department of defence (US) launched a large-scale digital forensic study in 2016 to further research on media integrity, with remarkable results in terms of methodology and benchmark dataset [47]. The MediFor taxonomy states that physical, digital, and semantic integrity may be checked during digital media confirmation. Deep learning models' superiority is now beyond question; in fact, they are rapidly gaining acceptance among numerous academic institutions and significant IT firms, gradually displacing most technology.

Merging of computer vision and deep learning methods, Deepfake also known as highly realistic fake images and videos is possible to Generative adversarial networks and Auto encoders [32, 33]. Hackers or non-technical machine learning engineers can alter photos or videos by changing the information and then creating a new picture or video that cannot be distinguished by human beings or machines. In 2018, justify that how simple it is to use this innovation for immoral and malicious uses, such as spreading false information, presenting as political figures, and defaming innocent people. These "deepfakes" have significantly improved since then [48].

1.2. Evaluation of DeepFake techniques

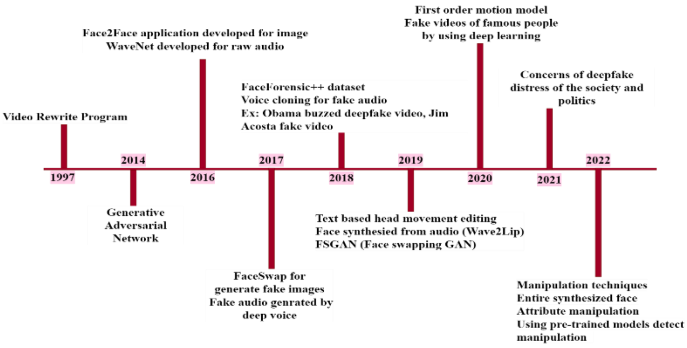
Video Rewrite Program
1997
2014
Generative Adversarial Network
Face2Face application developed for image
WaveNet developed for raw audio
2016
2017
FaceSwap for generate fake images
Fake audio genrated by deep voice
FaceForensic++ dataset
Voice cloning for fake audio
Ex: Obama buzzed deepfake video, Jim Acosta fake video
2018
2019
Text based head movement editing
Face synthesied from audio (Wave2Lip)
FSGAN (Face swapping GAN)
First order motion model
Fake videos of famous people
by using deep learning
2020
Concerns of deepfake
distress of the society and
politics
2021
2022
Manipulation techniques
Entire synthesized face
Attribute manipulation
Using pre-trained models detect
manipulation

Fig. 2. Evaluation of DeepFake detection techniques.

Figure 2 is related to the timeline of DFD techniques. An evolution of DFDT says the entire background and existing techniques of Deepfake detection. Forage or manipulated images existed for many years ago.

2. Literature Review

This section will discuss different strategies and compare them to give an understanding of how each can be used. Also discussed different types of datasets and existing surveys with their models. This section gives the idea behind what is Deepfake and how they classify with the use of artificial intelligence and different types of neural networks.

2.1. Process of DeepFake detection

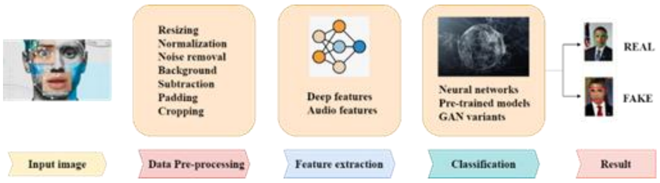

Fig. 3. Generalized diagram of DeepFake detection technique

Figure 3 shows a generalized processing pipeline for Deepfake detection. The majority of Deepfake detection techniques have either used hand-crafted features or deep learning-based feature extraction techniques. There are serval methods that concentrate on combining handcrafted and intricate features with numerous methods, such as auditory and visual information, for efficient manipulation.

2.1.1 Data Pre-processing

Working with feature variables (inputs) and target variables (outputs) for exploratory data analysis, feature variables may be person-specific data like the age of the person, height, and weight of the person; in pictorial format, flatten and normalized the image. To create models, several data pre-processing procedures are required. Then, what kind of data do we have, such as time series data, audio format, png/jpeg format, structured or unstructured, numerical or string, etc. The methods for processing photos play a crucial part in the acquisition, pre-processing, clustering, segmentation, and classification of various types of photographs, including those of fruits, medical devices, vehicles, and digital text, among others. Data preparation includes fundamental operations including feature normalization, feature impute, feature encoding, one-hot encoding, feature encoding, feature engineering, and feature selection using different methods such as dimensionality reduction approaches.

Image quality can be improved by image augmentation. This method is most frequently utilized in a variety of applications. Because of illuminations, satellite and digital camera photos can have reduced contrast, brightness, and noise. The aim of image enhancement is to sharpen and remove noise from the image and order to improve blurrily and a noisy image [30]. Here, this research provides some image-enhancing methods, like sharpening with text images and bib numbers, histogram equalization, filtering to remove salt and pepper noise, and so on [30].

Data pre-processing improves the quality of the data, so we can easily examine the data. In deepfake there are some pre-processing techniques are available which include attribute manipulation, identity, and expression swapping, etc. Figure 4 represents some of the techniques for data pre-processing,

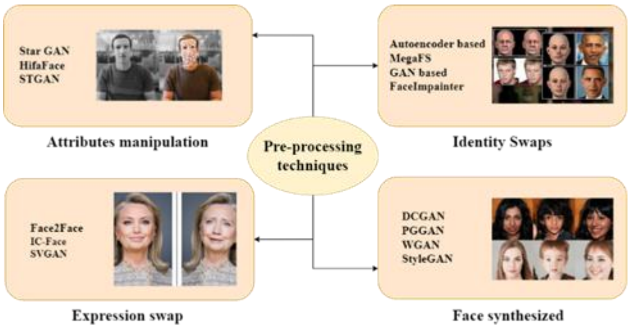

Fig. 4. Overview of pre-processing techniques

2.1.2 Feature extraction

Feature extraction means the process of dimensionality reduction, which involves splitting up and condensing a set of raw data into small groups. As the result, processing will be easier. Dimensionality reduction, or the process of breaking down a starting collection of raw data into small, simpler groups, is referred to as feature extraction. These huge datasets have a variety of different factors that make up a significant portion of their characteristics. To process variables, a lot of processing power is required. Feature extraction assists in successfully removing the best feature from such big data sets to reduce the amount of data by selecting and combining variables into features. These characteristics accurately and distinctly represent the real data set while still being simple to utilize. Figure 5 says some of the techniques for feature extraction in audio and image format,

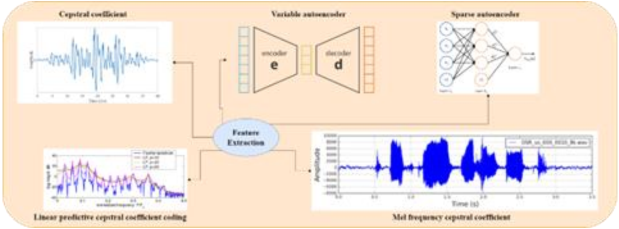

Fig. 5. Overview of feature extraction

There are many techniques for feature reduction but for the synthesized media variable autoencoder and Convolutional autoencoder. In 2013, the generative model class known as the Variational autoencoder (VAE) was defined [14]. It is a model that can produce accurate data and derive useful latent representations of the input. VAE is built on Bayesian inference to do this. By a latent variable, the VAE approximates the probability distribution of input data p(x). As was previously indicated in this paper, VAE may also be thought of as a probabilistic Principal component analysis (PCA) that is non-linear and uses neural networks to introduce non-linearity [15].

In developing a high-fidelity face-swapping approach, we consider three essential criteria to address the low visual quality issues of earlier research. For us to produce a big number of high-quality videos, it should be universal and scalable. It is necessary to address the issue of face style mismatch brought on by different variances. It is important to evaluate temporal continuity of generated videos [17].

There are two networks that make up variational auto encoder. A latent presentation of input data that includes key data properties is created by the encoder or inference network. The decoder or generative network subsequently maps this latent representation into the reconstructed data [15].

The recommended sparse autoencoder-based LSTM model is utilized to extract computer vision features from a pre-processed image, which is delivered in part. The primary auto-encoder method has drawbacks like its impotence to locate features by duplicate memory into an implied layer [19]. In an unsupervised deep learning method, the data is encoded for feature extraction using a single hidden layer. This will estimate the error and get the utterance from the hidden layer with the most relevant features [20]. With the use of an autoencoder known as a sparse autoencoder, this issue is handled using a sparsity technique (SAE) [21]. The data is encoded for feature extraction using a single hidden layer in an unsupervised deep-learning technique [22]. This will estimate the mistake and retrieve the hidden layer expressions' expressions with the most pertinent features.

An encoder is a compression space, and A decoder make up an SAE. The encoder uses hidden layers to encode the given input into the parameter space, while the decoder oversees decoding the data from the parameter space into the output layer. A sigmoid function is employed as an activation function instead of a ReLU activation function since autoencoders cannot handle negative values, which limits the network's capacity for training [22]. The variance between the input and the rebuilt data can be decreased using SAE. The sparsity limitation lowers the use of hidden layers. Regularization can be done on larger datasets and will prevent the overfitting problem in this case.

2.1.3 Detection models

Recently, Deep learning techniques have been used to identify false images with success. However, there is a large loss of pixel information following video compression, and the current deep-learning methods for images can't be used to detect false movies [6]. The related study in Deepfake video recognition is broken down into two primary sections in the part below: Temporal and spatial features analysis and biological singles analysis.

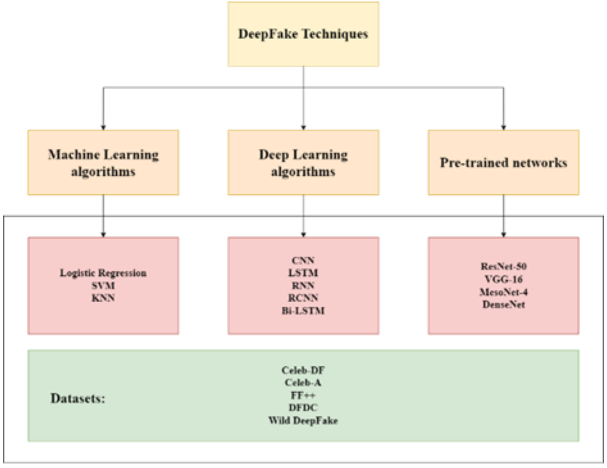

Fig. 6. Some existing detection methods of DeepFake

Figure 6 represents some existing methods that are used for detecting the deepfake. It includes machine learning algorithms, deep learning algorithms, and pre-trained networks with their datasets. A few research have been done with the use of machine learning algorithms.

Motivation behind any decision that can be described in human terms may be understood using conventional machine learning (ML) algorithms. These techniques work well in the Deepfake arena since the processes and data are well recognized. Further, it is much easier to adjust hyper-parameters and modify model designs. The decision process is displayed as a tree in a tree-based ML techniques, such as Decision Trees, Random Forest, Extremely Randomized Trees, etc [5]. Using traditional machine learning algorithms, we got the best accuracy, but the model depends only on some datasets. A model cannot be well-trained in machine learning algorithms. That's why we apply deep learning for Deepfake detection. To avoid an attack from moving through a community of voice assistants (or other IoT devices), federated, learning-based, lightweight techniques must be used [52, 53].

i) LSTM (Long short-term memory)

LSTM is one of the types of recurrent neural networks. Long-term dependencies are dealt with using the artificial recurrent neural network (LSTM) type. The full data sequence can be learned using the feedback connections in LSMT. Various fields that rely on time series data for classification and prediction have used LSTM [28]. Input gates, forget gates, and output gates make up the basic LSTM architecture. The values from earlier intervals are remembered by the cell state, which stores them in the LSTM cell. The range of the sigmoid function is [0,1] and it was decided by forget gate [29].

ii) CNN (Convolutional neural network)

Convolutional neural networks are a widely used type of deep learning model. Like neural networks, convolutional neural networks have output and input layers with hidden layers. When using this network, hidden layers read the inputs from the top layer before the convolution of the data. In this usage, convolution refers to matrix multiplication. After matrix multiplication, CNN uses Rectified Linear Unit (RELU), a nonlinearity activation function, followed by subsequent convolutions like the pooling layer [27]. The basic aim of pooling layers is to decrease the dimensionality of data by defining output using functions like average or maximum pooling.

With deep learning techniques, synthetic voice quality has substantially improved in recent years. Voice adaptation [54], one or few-shot learning [55, 56], self-awareness network [57], and semantic voice translation [58, 59] are just a few of the key improvements. It remains difficult for them to produce more naturally sounding, human-like utterances given the limited training examples they have available in a variety of environments.

A method to recognise image modification was reported in DL-based work by Guarnera et al. [60]. Prior to training the naive classifier to distinguish between real and false photos, the using the Expectation Maximization (EM) method was to extract the image features. Although this method exhibits improved deepfake recognition accuracy, it can only be used with static images.

Usual deep neural network model is a type of convolutional neural network. In CNN, the input data is first convolutional by the hidden layers after reading the input from the top layer. Convolution in this context denotes a dot product or matrix multiplication. Rectified linear unit (RELU), a nonlinearity activation function, which used by CNN after matrix multiplication, posted by additional convolutions such as pooling layers.

iii) RNN (Recurrent neural network)

It is a type of neural network; it extracts the features from the sequence data. RNNs are composed of numerous hidden layers every layer has a weight and a bias, much like neural networks. A connection among the nodes of a direct cycle graph occurs sequentially in RNN. The ability to identify temporal dynamic behavior is one advantage of RRN. In contrast, to a feed-forward network, this neural network employs private storage to store the order of data from earlier input, making it beneficial in several fields such as speech recognition and natural language processing [13]. By providing a recurrent hidden state that captures dependencies across several time scales, RNN can handle the temporal sequence.

The initial layer of the neural network is known as the input layer, and it receives the input data and passes it to the next layer. Take input data are A1, A2, A3, and A4. The next layer consists of hidden layers, which are made up of artificial neurons, a collection of connected units. The edges that link the neurons show how one neuron is related to the others and how it can send and receive information over multiple levels. Each link has a weight assigned to it that represents the links connecting two units. If neurons have N inputs, then values represented in Eq. (1) and Eq. (2),

$$Z = AW1 + A2W2 + A3W3 + \cdots + AnWn + B \quad \textbf{(1)}$$

$$Z = \sum_{i=1}^{n} i\, A * W + B \quad \textbf{(2)}$$

iv) Techniques based on deep features.

The scientific community has suggested numerous methods to recognize facial alterations by recognizing the internal GAN pipeline. Similar findings were made in [63], where the author pushed the idea that examining internal neuron activities could help distinguish between real and artificial faces. For this, combinations of layer-by-layer neuron activation provide a more representative set of visual traits. A few DL-based face recognition frameworks, such as VGG-Face [64], OpenFace [65], and FaceNet [66], were used by FakeSpoter, the suggested approach in [63], to calculate the deep features. The SVM classifier was trained to distinguish between fake and real faces using the extracted features. Although the work [63] performed well for detecting facial attribute alteration, samples with extreme light fluctuations may not respond as effectively. Using a YOLO face detector, the face areas were recovered from video frames [67]. Using transfer learning [68], we can create models that have already been trained for prediction. The capability to use the learned features for highly accurate prediction is provided by transfer learning. The fine-tuning techniques based on transfer learning use a network that has already been trained and retrains a portion of the network using the new dataset.

v) GANs (Generative Adversarial Networks)

GAN is a cutting-edge method for training generative models which are used to produce realistic examples from the distribution of data. In general, GANs are a mixture of generator and discriminator networks. Dynamic mini-max game pits these two neural networks against one another. The idea for this concept is that while D tries to differentiate between actual data and data samples, the generator tries to make fake data. If the two models are given enough time to compete, they will eventually improve. To put it another way, the aim of the generator is to represent data distribution, but the aim of the discriminator is to determine how likely it is that data came from the training examples rather than from the generator. The mathematical optimization of GAN is represented in Eq. (3),

$$G \in arg \arg minmax\, V(G, D) =$$

$$minmax\ \ Ex\,[\log D(x)] +\ Ez\,[1 - logD(G(z))]$$
(3)

Here, Generator represent as G and Discriminator represents as D. To improve designs, losses, and training techniques, various types of GANs, including deep convolutional GAN (DCGAN), Wasserstein GAN (WGAN), progressive growing GAN (PGGAN), BigGAN and Style-GAN, have recently been developed.

In addition to a GAN-based attribute modification network to implement change, the ANN model for attribute modification proposed in SAGAN [61] gave a spatial attention strategy to find and explicitly constrain editing inside a specific region.

By using an attention mask that switches from a high to a low feature level, PA-GAN [62] used a progressive attention method in GAN to gradually blend the attribute features into the encoder features constrained to a proper attribute area. Lower feature levels (better resolution) result in more precise attention masks and finer attribute manipulation. This method effectively manipulates many attributes while keeping irrelevance inside a single model.

2.2. Existing surveys with different modalities.

In this section, we have discussed the DeepFake techniques based on image, video, and hybrid datasets.

2.2.1. Image based DFDT.

Deep neural networks are used in a variety of ways to detect GANs-generated images. Deep neural network-based techniques for spotting fake images were proposed by Tariq et al [39]. This method improves the detection of fake face images made manually with the use of pre-processing analysis to detect statistical features of an image. Another method for identifying fraudulent images produced by GANs is also introduced by Nhu et al [40]. And is based on a deep convolutional neural network. This method first uses a deep learning network to extract a face attribute from the face recognition networks. The facial features are then changed to make them useful for distinguishing real photos from fake ones. Using these techniques, the data from the contest validation produce good results.

The difference is subsequently avoided by boosting the shallow layer-derived textural features, which are then combined with high-level semantic information to represent each local component. During training, the author can voluntarily some high reaction attention in blurred regions and force the network to learn from other attention regions. The author checks this technique on different datasets [8]. A mathematical formula of attention-guided data augmentation of this paper is as below:

$$i' = i * (1 - Ak') + id * Ak' \tag{4}$$

Where ‘i’ was our original image, ‘Ak’ as the weight of the original image and get the output will be our degraded image. Table 1 shows the comparison based on images data.

Table 1: Existing work based on Image data.

Year	Methodology	Advantages	Disadvantages
2021 [8]	Using the deep semantics, the author creates multi-attention heads to forecast various spatial attention maps. Detect the manipulated images as fine-grained classification. This paper proposed attention guided data augmentation mechanism.	Detect fake images as classification problem. Easy to understand.	For different datasets accuracy will be reduce, due to differ shapes.
2021 [12]	Classifies real or fake data as an unsupervised approach. Image augmentation with two augmented version as unsupervised contrastive learning. Using Backbone network learn the feature for data pre-processing.	Overcomes the problem of unlabeled dataset using unsupervised contrastive learning.	Labelling is time consuming process.

2022 [9]	This paper refers the different techniques for manipulated data like, swapping the images (image augmentation), attribute manipulation, content creation (text-to-image), etc. Detect manipulated data by using variants of CNN like, RCNN and YOLO, and StyleGAN.	The methods for forage the images are helpful to build an accurate model.	The methods which used in this paper it takes more time to regenerate manipulated images.
2022 [18]	Facial swapping techniques, like Deepfake, modify the face area to mimic the appearance of the context while changing only the face. These differences present exploitable manipulation red flags.	Increase the size of the dataset. Change the context of the images to generates fake images.	Difficult to merge the original and generated data due to different types of datatypes.
2023 [69]	With the use of Dual shot face detector extract the faces from multiple images and videos and for detect deepfake applies MesoNet, FWA, XceptionNet and Capsule methods.	Detect low- and high-resolution images.	Requires higher computational system for training.

2.2.1.1 Discussion related image based DFDT.

Generation of Deepfake images using artificial intelligence was very easy with the help of different variants of GANs. And for the detection techniques there are lots of pre-trained models are available in TensorFlow and using it we can easily be trained and classify the real or fake images in a neural network. Different models were used in image-based detection techniques, table 6 denotes model’s performance with their respective datasets.

2.2.2. Audio based DFDT.

Audio-driven talking-head animation is one cross-modal research field with a lengthy history in the computer graphics community. Previous techniques take one of two routes, depending on whether they want to create lifelike videos. These non-photorealistic techniques focus on discovering a relationship between source waveforms and facial movements, for instance by relating 3D vertex coordinates to rigging parameters or facial model parameters [41]. This technique typically requires artistic assistance for the rigging settings or high-quality 4D face capture data. With the aid of pre-trained models and audio analysis techniques, multiple methods are utilised to train audio datasets, including training audio without video first, followed by training audio without video and with sync stream [4]. Table 2 represent about the comparison of different DFDT techniques on audio data.

Table 2: Existing work based on Audio data.

Year	Methodology	Advantages	Disadvantages
2021 [4]	For the joint dataset, the author used an existing video Deepfake dataset with real audio; That audio was converted to a mel-spectrogram. Apply it on some vocoders and generate fake audio. The author uses the R(2+1)D 18 network for the video stream. To analyze 1D raw wave-form data for audio streams, the author uses a straightforward 1D convolutional network.	In this paper authors check with three different types of datasets which are audio, video and combined audio-video datasets.	Combined dataset has low accuracy with compared to separate datasets.
2022 [10]	Introduce the Mel-frequency cepstral coefficient parameter. Like a machine learning cycle in this first, we analyze the datasets with help of Bispectral analysis and NTU approaches. Trained different RNN models and feature likes, MFCCs, RMS, zero crossing, chroma frequency, and spectral roll-off.	Minimum complicated computation.	Limitations are audio clips only for 2 sec long.
2022 [11]	The author applies lots of machine learning algorithms to classify real and fake audio like SVM and KNN with around 97% accuracy. Another deep learning method is CNN gives better accuracy which is 99% with only 2% misclassification. The author used many types of CNN, which includes ResNet34, LSTM, and RNN with good accuracy.	By using these techniques detect an imitated based audio clip.	There are limited audio Deepfake detection methods with non-English languages.
2020 [43]	Overcome the problem of lip synchronization with the use of the matching concept of arbitrary identity and target speech segment. It also uses multiple consecutive frames rather than a single frame as the discriminator.	Wav2Lip model performs significantly better than the methodologies in use. Overcome the problems of lip synchronization.	Must check with more evaluation metrics to decide the model. Need to increase robustness.

2023 [70]	Use a unified approach for detect fake audio and visual modality with a novel AVFakeNet framework. Also defined novel Dense Swin Transformer Net for feature extraction.	Detect multimodality with novel approach.	Requires higher computational resources.

2.2.2.1 Discussion related Audio based DFDT.

For the detect manipulated audio mostly used technique known as the Mel-spectrum process. Using it converts the audio dataset into a raw format and then we can easily classify the audio data by applying different variants of neural networks like CNN, RNN, or variants of LSTM. Different models were used in audio-based detection techniques, table 6 denotes model's performance with their respective datasets. The motion of the lip is generated in prior work on lip-syncing [50, 51] by reselecting frames from a clip or transcript and target emotions.

2.2.3. Video based DFDT.

Deep learning techniques have been successfully used in recent years for the detection of fake data. But, due to the heavy loss the information during video compression, the ongoing deep learning methods for images cannot be used to detect fake videos. The existing work in deepfake video recognition is divided into two categories in the section below, 1) biological singles analysis, 2) temporal and spatial features analysis. For the video dataset space-time conditioning volume formulation and transfer of the source image into target images [44]. Table 3 is all about the comparison of the DFDT techniques on video data.

Table 3: Existing work based on Video data.

Year	Methodology	Advantages	Disadvantages
2018 [7]	Detects manipulated videos that are generated by Deepfake with the use of encoders and decoders. By using CNN and LSTM. But there are variants of the LSTM model. LSTM with 20 frames, 40 frames, and 80 frames.	Used to detect videos with the different variants of LSTM.	Cannot run two autoencoders separately during training phase.
2018 [38]	This adaptive technique considers eye blinking, a crucial physical characteristic that can be utilized to identify fraudulent videos. It is combining a convolutional neural network with a recursive neural network to identify physiological signals like blinking and eye movement.	The outcomes of the research showed that the proposed method is effective in identifying fake photographs.	Must try to convert into frames, and it was time consuming process.

2018 [44]	First to transfer a source actor's facial expression, eye gaze, head pose, and orientation to a target actor. The suggested technique is based on a revolutionary rendering-to-video translation network that produces photorealistic and temporally consistent video from a series of basic computer graphics renderings.	These methods first demonstrated the highly realistic re-enactment results in large variety of applications.	Time consuming approach due to video-based classification.
2018 [45]	An algorithm can synthesis face pictures of a limited size due to resource and production time constraints, and they must go through an affine warping process to match the source's facial features.	Overcome the limitations of the creation time of the dataset.	Take more time compared to the pre-trained networks.
2023 [71]	Proposed Hierarchical file system-based method to detect fake videos with grey wolf optimization and Vortex search algorithm feature selection techniques.	Proposed HFS method with higher accuracy on different datasets.	Time consuming process.

2.2.3.1 Discussion related Audio based DFDT.

The generation of Deepfake videos is a difficult task because it is a time-consuming process. And for classify these we must use neural networks like CNN, RNN. Because first, we convert the video dataset into frames and then trained using neural networks. It was a time-consuming process, but we got the best accuracy due to more training. The eye-blinking datasets are the first ones that are now accessible that were created specifically for the detection of eye-blinking. These models use binary classification either in the detection of real or fake or in eye-blinking datasets. Different models were used in video-based detection techniques, table 6 denotes model’s performance with their respective datasets.

2.2.4. Multimodality in DFDT.

We have studied and analyse unimodal datasets and their respective techniques, in this section, we discuss some related multimodality in Deepfake detection and their techniques. How we can classify the media which is real or fake the muti model data? In this, we have two different modalities like (audio + video) and (audio with images). And here there are many techniques for analyzing and training the separate datasets, like MFCC for audio, CNN, and RNN for the images, and pre-trained networks for the combined datasets [3]. Table 4 states the comparison of the multi-modality.

Table 4: Existing work based on Multimodality data.

Year	Methodology	Advantages	Disadvantages
2021 [3]	DFDC, contains mixing of real and synthesis video and audio without labelling. Model they used like Xception, Efficient Net and VGG16 for multimodal.	No need to apply labelling.	Existing detection methods are limited to videos or audio individually.
2022 [42]	Using GAN and LSTM predict audio and video. For audio detection applies SincNet operates on the unprocessed waveforms, preventing any anomalies or lost data during preprocessing and lowering pipeline overhead. For	Tries to get maximum accuracy due to step wise model.	Due to classifies individual modalities take more time to train and feature extraction process.
2022 [2]	Generates fake data with the use of StyleGan with applies techniques like face synthesis, swap attributes or expression, etc.	Generating deepfake dataset methods that can withstand view angle.	Must applies labelling process.
2022 [46]	Paper introduced a two-stream network Multi-modal Multi-scale Transformer (M2TR) for Deepfake detection	Detect hidden forgery of the images with the use of multi-scale transformer.	Must crop the faces.
2023 [72]	Extract the audio-visual features from input data with the use of time-aware neural networks on monomodal datasets.	Don’t require to train models on existing dataset.	Must train disjoint monomodal datasets.

2.3. Discussion.

This literature survey refers to what models are developed recently in DFDT. Given tables show the models and respective datasets with their accuracy. In this survey, there are separate tables for the audio, video, and image dataset. This survey says related to generalized methods for the DFDT, which includes popular steps like data pre-processing, feature extraction, and classification and classifies the binary output as media was real or fake.

Table 5: Popular datasets of forage datasets.

Year	Dataset	Images/Videos (i/v)	Details	URL
2021	Open Forensic [23]	16k (i)	Large-scale dataset for segmenting and detecting multiple face forgeries in the real world.	https://paperswithcode.com/dataset/openforensics
2021	Wild Deepfake [24]	11m (i)	Tiny dataset that tests the real-world dataset's ability to recognise deepfakes.	https://github.com/Deepfakeinthewild/Deepfake-in-the-wild
2020	Celeb-DF [25]	6k (v)	Challenging large-scale dataset for deepfake forensics. Videos from youtube.	https://github.com/danmohaha/celeb-Deepfakeforensics
2020	DFDC [16]	23k (i)	Used 8 different facial modifications for generating.	https://ai.facebook.com/datasets/dfdc/
2015	Celeb-A [26]	2m (i)	A large dataset of facial traits contains more than 2 million photos of celebrities and 10,000+ identities.	https://tinyurl.com/yc3pa85r
2018	DF-TIMIT	34k	Low- and High-quality images generated by GAN models	https://www.idiap.ch/en/dataset/Deepfaketimit
2017	VISION	34k	Combination of images and videos; application-based dataset	https://lesc.dinfo.unifi.it/VISION/dataset/

3. Conclusion and Future scope

Deep learning approaches have drawn great interest from various sectors. A few deep learning-based approaches have recently been put forth to deal with the problem as well as effectively identify phony photos and movies. To successfully recognize false films and photos, the present deep learning techniques must also be improved. Multiple attention mapping is used in the proposed background to study differential regions, and deep layer texture characteristics are enhanced to capture more subtle abnormalities. Then, using attention mapping as a guide, low-level textural features, and high-level semantic features are compiled. For a purpose of training disembarrassing multiple awareness, an independence loss function and an attention-guided data augmentation method are presented. In numerous metrics, our strategy produces good improvements. The entire study can be summed up as follows: Deep learning algorithms are broadly used in DFDT, and the largest dataset is Face forensic ++ for the experiments. A significant percentage of all the models are deep learning models, mostly CNN models. The performance metric that is most frequently used is detection performance. The alternative result shows that deep learning methods can successfully identify Deepfake. Additionally, it can be said that the deep learning model performs better than the non-deep learning model overall.

Acknowledgment

The paper named "Analysing the landscape of DeepFake Detection - A Survey" includes this study. We would like to say thank you to Prof. Shruti Patil for her insightful comments. We appreciate everyone who volunteered to take part in the study.

Author's contribution

All the authors have contributed to this work equally.

Conflict of interest

None of the researchers of this survey has any financial or personal links to any individuals or groups that would inappropriately influence or affect the content of this survey. It must be stated clearly that there are no competing interests or conflicts of interest with other people or groups that could improperly impact or skew the content of this survey.

References

[1] Shraddha Suratkar, Sayali Bhiungade, Jui Pitale, Komal Soni, Tushar Badgujar & Faruk Kazi (2022) Deep-fake video detection approaches using convolutional – recurrent neural networks, Journal of Control and Decision.

[2] A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "Deepfake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[3] Hasam Khalid, Minha Kim, Shahroz Tariq, Simon S. Woo, "Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors", 2021

[4] Y. Zhou and S. Lim, "Joint Audio-Visual Deepfake Detection," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 14780-14789.

[5] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[6] "Deepfakes Detection Techniques Using Deep Learning: A Survey" written by Abdulqader M. Almars, published by Journal of Computer and Communications, Vol.9 No.5, 2021

[7] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[8] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen and N. Yu, "Multi-attentional Deepfake Detection," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2185-2194, doi: 10.1109/CVPR46437.2021.00222.

[9] Ahmed S Abdulreda; Ahmed J Obaid. "A landscape view of Deepfake techniques and detection methods". International Journal of Nonlinear Analysis and Applications, 13, 1, 2022, 745-755. doi: 10.22075/ijnaa.2022.5580

[10] Khan, Madeeha B.; Goel, Sanjay; Katar Anandan, Jaswant; Zhao, Jersey; and Naik, Ramavath Rakesh, "Deepfake Audio Detection" (2022). AMCIS 2022 Proceedings. 23.

[11] Almutairi, Z.; Elgibreen, H. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms 2022, 15, 155. https:// doi.org/10.3390/a15050155

[12] Sheldon Fung, Xuequan Lu, Chao Zhang, Chang-Tsun Li. DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning. https://doi.org/10.48550/arXiv.2104.11507.

[13] Badhrinarayan Malolan, Ankit Parekh, Faruk Kazi, "Explainable Deep-Fake Detection Using Visual Interpretability Methods",2020 3rd International conference on Information and Computer Technologies (ICICT).

[14] D. P Kingma and M. Welling, "Auto-Encoding Variational Bayes," ArXiv e-prints, Dec. 2013. arXiv: 1312.6114 [stat.ML].

[15] Sánchez Martín, Pablo. (2018). Unsupervised Deep Learning: Research and Implementation of Variational Autoencoders.

[16] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, Cristian Canton Ferrer AI Red Team, Facebook AI; The Deepfake Detection Challenge (DFDC) Preview Dataset; 23 Oct 2019.

[17] L. Jiang, R. Li, W. Wu, C. Qian and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2886-2895, doi: 10.1109/CVPR42600.2020.00296.

[18] Nirkin Y, Wolf L, Keller Y, Hassner T. Deepfake Detection Based on Discrepancies Between Faces and Their Context. IEEE Trans Pattern Anal Mach Intell. 2022 Oct;44(10):6111-6121. doi: 10.1109/TPAMI.2021.3093446. Epub 2022 Sep 14.

PMID: 34185639.

[19] Olshausen, B., Field, D. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996).

[20] Jiewu Leng, Pingyu Jiang. A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm, Knowledge-Based Systems, 2016, ISSN 0950-7051.

[21] Hoq, Uddin & Park (2021) Hoq M, Uddin MN, Park S-B. Vocal feature extraction-based artificial intelligent model for Parkinson's disease detection. Diagnostics. 2021;11(6):1076. doi: 10.3390/diagnostics11061076.

[22] Kandasamy V, Hubálovský Š, Trojovský P. Deep fake detection using a sparse auto encoder with a graph capsule dual graph CNN. PeerJ Comput Sci. 2022 May 31;8:e953. doi: 10.7717/peerj-cs.953.

[23] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, ''Open Forensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild,'' in Proc. Int. Conf. Computer. Vis., Oct. 2021, pp. 10117–10127.

[24] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, ''Wild Deepfake: A challenging real-world dataset for Deepfake detection,'' in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 2382–2390.

[25] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu; "Celeb-DF: A Large-scale Challenging Dataset for Deepfake Forensics"; arXiv:1909.12962v4 [cs.CR] 16 Mar 2020.

[26] Liu, Ziwei and Luo, Ping and Wang, Xiaogang, and tang, Xiaoou. "Deep learning face attributes in the wild", in Dec. 2015.

[27] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016) Deep Learning (No. 2). MIT Press, Cambridge.

[28] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. Neural Computation, 9, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[29] Schuster, M. and Paliwal, K.K. (1997) Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing, 45, 2673-2681. https://doi.org/10.1109/78.650093.

[30] P.L, Chithra and P, Bhavani, A Study on Various Image Processing Techniques (May 7, 2019). International Journal of Emerging Technology and Innovative Engineering Volume 5, Issue 5, May 2019, Available at SSRN: https://ssrn.com/abstract=3388008.

[31] H. Farid, ''Image forgery detection,'' IEEE Signal Process. Mag., vol. 26, no. 2, pp. 16–25, Mar. 2009.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ''Generative adversarial nets,'' in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–9.

[33] P. Baldi, ''Autoencoders, unsupervised learning, and deep architectures,'' in Proc. ICML Workshop Unsupervised Transf. Learn., 2012, pp. 37–49.

[34] Bañuelos, J. (2022). Evolution of Deepfake: semantic fields and discursive genres (2017-2021). Revista ICONO 14. Revista Científica de Comunicación y Tecnologías Emergentes, 20(1). https://doi.org/10.7195/ri14.v20i1.1 773.

[35] Scharre, Paul, et al. "The Artificial Intelligence Revolution." ARTIFICIAL INTELLIGENCE, 2018, pp. 3–4,19 Oct. 2022

[36] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016) Deep Learning (No. 2). MIT Press, Cambridge.

[37] Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning Long-Term Dependencies with Gradient Descent Is Difficult. IEEE Transactions on Neural Networks, 5, 157-166. https://doi.org/10.1109/72.279181

[38] Li, Y., Chang, M. and Lyu S. (2018): Exposing AI generated fake face videos by detecting Eye blinking. 2018 IEEE workshop and Security, Hong kong, 11-13 December 2018, 1-7.

[39] Tariq, S., Lee, S., Kim, H., Shin, Y. and Woo, S.S. (2018) Detecting Both Machine and Human Created Fake Face Images in the Wild. Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, Toronto, 15 October 2018, 81-87.

[40] Li, H., Li, B., Tan, S. and Huang, J. (2018) Detection of Deep Network Generated Images Using Disparities in Colour Components.

[41] Lu Y, Chai J, Cao X (2021) Live speech portraits: real-time photorealistic talking-head animation. ACM Trans Graph 40:1–17

[42] Lomnitz, Michael & Hampel-Arias, Zigfried & Sandesara, Vishal & Hu, Simon. (2020). Multimodal Approach for Deepfake Detection. 1-9. 10.1109/AIPR50011.2020.9425192.

[43] Prajwal K, Mukhopadhyay R, Namboodiri VP, Jawahar C (2020) A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia, pp 484–492.

[44] Kim H, Garrido P, Tewari A, Xu W, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. ACM Trans Graph 37:163–177.

[45] Li, Yuezun, and Siwei Lyu. "Exposing Deepfake videos by detecting face warping artifacts." arXiv preprint arXiv:1811.00656 (2018).

[46] Wang, Junke, et al. "M2tr: Multi-modal multi-scale transformers for Deepfake detection." Proceedings of the 2022 International Conference on Multimedia Retrieval. 2022.

[47] Hatmaker Taylor, "DARPA is funding new tech that can identify manipulated videos and Deepfake", web blog post. May 01, 2018.

[48] Y. Mirsky and W. Lee, ''The creation and detection of deepfakes: A survey,'' ACM Comput. Surv., vol. 54, no. 1, pp. 1–41, Jan. 2022.

[49] M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, ''Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,'' 2021, arXiv:2103.00484.

[50] B. Fan, L. Wang, F. K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional LSTM," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4884-4888: IEEE.

[51] J. Charles, D. Magee, and D. Hogg, "Virtual immortality: Reanimating characters from tv shows," in European Conference on Computer Vision, 2016, pp. 879-886: Springer.

[52] K. M. Malik, H. Malik, and R. Baumann, "Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks," in 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 523-528: IEEE.

[53] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A lightweight replay detection framework for voice controlled iot devices," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 982-996, 2020.

[54] Y. Chen et al., "Sample efficient adaptive text-to-speech," arXiv preprint arXiv:1809.10460, 2018.

[55] H. Lu et al., "One-Shot Voice Conversion with Global Speaker Embeddings," in INTERSPEECH, 2019, pp. 669-673.

[56] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance," in Interspeech, 2018, pp. 496-500.

[57] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," arXiv preprint arXiv:.05742, 2019.

[58] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," arXiv preprint arXiv:.02812, 2018.

[59] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 266-273: IEEE.

[60] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 666-667.

[61] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 417-432.

[62] Z. He, M. Kan, J. Zhang, and S. Shan, "PA-GAN: Progressive Attention Generative Adversarial Network for Facial Attribute Editing," arXiv preprint arXiv:2007.05892, 2020.

[63] R. Wang et al., "Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces," arXiv preprint arXiv:.06122, 2019.

[64] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[65] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU School of Computer Science, vol. 6, no. 2, 2016.

[66] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815-823.

[67] Chen, W.; Huang, H.; Peng, S.; Zhou, C.; Zhang, C. YOLO-face: A real-time face detector. Vis. Computer. 2021, 37, 805–813.

[68] Jin, B.; Cruz, L.; Gonçalves, N. Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis. IEEE Access 2020, 8, 123649–123661.

[69] Narayan, Kartik, et al. "DF-Platter: Multi-Face

Heterogeneous Deepfake Dataset." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[70] Ilyas, Hafsa, Ali Javed, and Khalid Mahmood Malik. "AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection." Applied Soft Computing 136 (2023): 110124.

[71] Mohiuddin, S., Sheikh, K.H., Malakar, S. et al. A hierarchical feature selection strategy for deepfake video detection. Neural Comput & Applic 35, 9363–9380 (2023). https://doi.org/10.1007/s00521-023-08201-z.

[72] Salvi, D.; Liu, H.; Mandelli, S.; Bestagini, P.; Zhou, W.; Zhang, W.; Tubaro, S. A Robust Approach to Multimodal Deepfake Detection. J. Imaging 2023, 9, 122. https://doi.org/10.3390/jimaging9060122.