IJCI V12-2(2025*) 60-80*

International Journal of Computers and Information (IJCI)

Available online at *https://ijci.journals.ekb.eg/*

A Comprehensive Overview of Deep Learning for Deepfakes: Generation, Detection, Dataset: A Survey

Eman AbdElfattah[a], **Nader Mahmoud**[b], **Hamdy M. Mousa**[a], ***Ashraf Elsisi***[a]

a Computer Science Dept., Faculty of Computers and Information, Menoufia University, Egypt.
b Cyber security Dept. , Engineering and Information Technology College, Buraydah Private colleges, the Kingdom of Saudi Arabia.

eman4cs@gmail.com, eng.nader.mahmoud@gmail.com, hamdimmm@hotmail.com, ashraf.elsisi@ci.menofia.edu.eg.

Abstract

The rapid evolution of deep learning techniques, particularly through generative adversarial networks (GANs), has enabled the creation of hyper-realistic synthetic media, heightening concerns in domains such as politics, entertainment, and security. Consequently, there has been a heightened interest in developing robust deepfake creation and detection systems. This paper provides a comprehensive survey of state-of-the-art deepfake detection methodologies, more specifically we focus on video-based, and image based approaches and their applications. It seeks to enhance the reader's understanding of recent developments, vulnerabilities in existing security measures, and areas for improvements. This research is among the few studies that extensively review datasets, highlighting their strengths and weaknesses. Which are crucial for training and validation purposes of deepfake models, such as: FaceForensics++, DeepFake Detection Challenge (DFDC), and Celeb-DF, emphasizing the need for diverse and realistic datasets to ensure robust model generalization. It thoroughly addresses every aspect of identifying and creating deepfakes, providing the reader with a comprehensive understanding of these topics in a single study. Recent studies primarily use Convolutional Neural Networks (CNNs) for deepfake detection, with the main goal of optimizing detection accuracy, also focuses on comparing different approaches. From previous studies we conduct that XceptionNet, EfficientNet, and I3D are top models for deepfake detection, each excelling in different areas: XceptionNet identifies subtle manipulations, EfficientNet offers balanced and efficient video frame analysis, and I3D specializes in real-time video sequence processing.

Keywords: ***Deepfake, Detection, Generation, CNN, RNN, GAN;***

1. Introduction

In recent years, advancements in multimedia manipulation techniques have achieved unprecedented levels of realism, making it difficult to differentiate between authentic and synthetic media [1]. This development has opened up new possibilities for innovative applications in various fields, including creative arts, advertising, film production, video games and character assassination, and deepfake pornography [2]. Examples include a manipulated video of U.S. Speaker of the House Nancy Pelosi appearing inebriated [3], President Donald Trump supporting his political opponent [4] and a deepfake video on tiktok [5]. However, it has also posed significant security risks. These deepfakes, generated using advanced DL techniques such as GANs and Auto-encoders [6].

T. Zhang, et al [1] provides an overview of the latest deepfake generation techniques, detection methods and available datasets. The authors focus on only two categories of deepfake generation: face swapping and facial reenactment. Patil, Kundan et al [7] focuses on the generation and detection of deepfakes, examining existing detection techniques and their challenges. They focus on only biological indicators, such as eyebrow recognition, eye blink detection, and pulse monitoring, and compares their effectiveness using various classifiers and methods. Also discusses the difficulty in verifying integrity in GAN-generated images, highlighting the role of the discriminator in this process. Heidari, Arash et all [8], provide a comprehensive analysis of deep learning (DL) in deepfake detection, including an overview of the field, a detailed review of existing methods, key DL techniques used for detection, and an evaluation of critical features in various strategies. It also identifies areas where these techniques can be improved. But the authors focus on deepfake detection without delving into any details related to deepfake creation.

This article explores the application of DL techniques in deepfake detection and addresses various related concerns. Additionally, it provides a detailed discussion of future research directions, emphasizing key issues that need to be resolved. In summary, the key **contributions** of this article are as follows:

- Offering a comprehensive overview of DL approaches in the field of deepfake detection.
- Explore the underlying algorithms and architectures used for deepfake creation, highlighting their strengths, limitations, and advancements
- Critically analyzing each DL-based strategy for deepfake detection and generation, focusing on their distinctive features.
- Examines the most widely used datasets in deepfake research, detailing their scope, diversity.

The remainder of our survey is organized as follows. In section 2 approaches of deepfake creation is introduced. In section 3, a discussion of deepfake detection approaches is provided. Available datasets used for deepfake creation and detection are explored in section 4. Section 5, provides a comparison of the state-of-the-art detection approaches. A conclusion and future work directions are discussed in section 6.

2. Deepfake Creation Techniques for Images and Videos

The term "deepfake" merges "deep learning" and "fake," emphasizing the utilization of deep neural networks to generate or alter audio, images, text, and videos in a very realistic way[9]. However, alongside the advancements in deepfake generation, there has been substantial progress in detection methodologies. Researchers are developing sophisticated algorithms to identify subtle inconsistencies and artifacts that escape human eyes but can be detected through DL techniques [10]. In this section, we will explore the roles of DL Methods and Traditional Tools in Media Creation with a particular emphasis on images and videos.

A. Conventional Media Editing Tools

Conventional media editing tool also called a traditional method that is frequent operations includes adding, duplicating, or removing objects. Fig. 1 illustrates the process of image splicing, where multiple image segments or pieces are meticulously combined to form a cohesive whole. As demonstrated in the splicing section, parts a and b have been seamlessly integrated to generate a new composite image, represented as c. This technique is commonly used in photography, digital art, and graphic design to create larger compositions or to merge different elements seamlessly. In the copy-move [11] section, a new object can be introduced by duplicating it either from another image or even from the same image itself, a process known as copy-move.

In contrast, an existing object can be removed by expanding the background to seamlessly conceal it, a technique known as inpainting, similar to the widely recognized exemplar-based inpainting method [12]. These tasks can be efficiently executed using widely accessible image editing software. Furthermore, post-processing steps, such as resizing, rotation, or color adjustments, may be essential to ensure the object is flawlessly integrated into the scene, enhancing its visual coherence and maintaining consistent perspective and scale

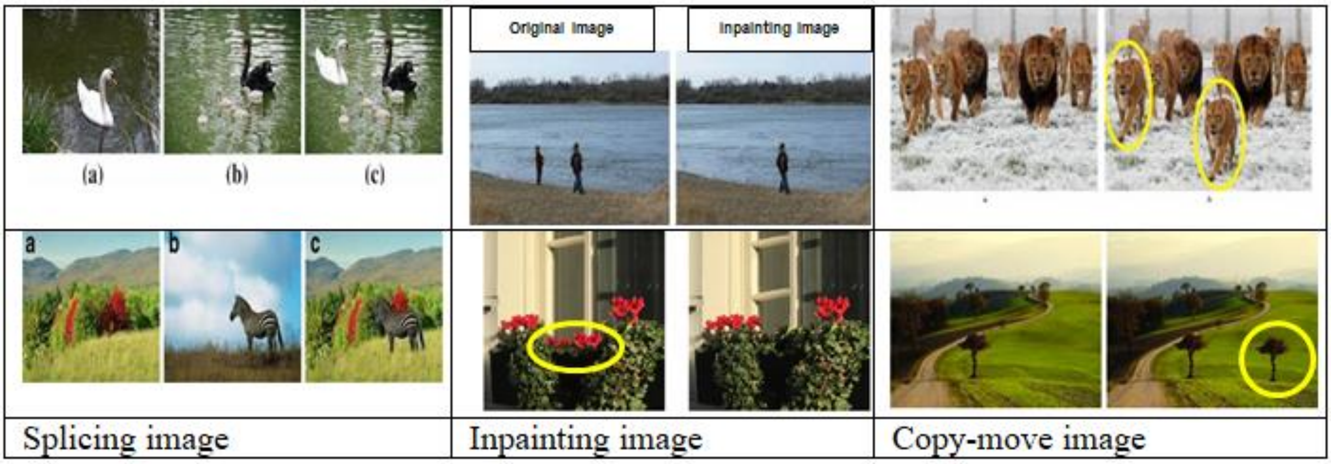


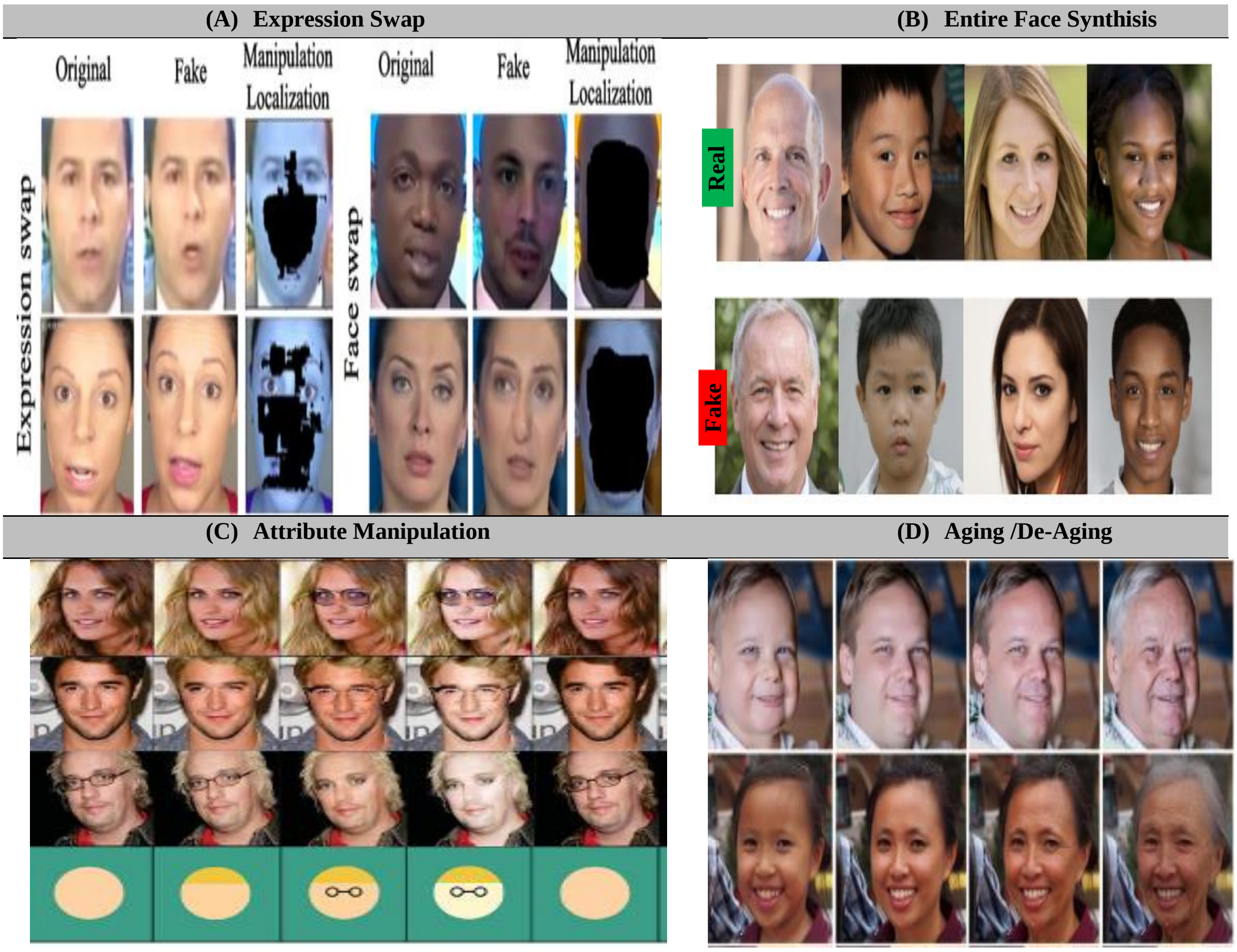
Fig 1. *Examples for manipulated image by traditional method..*

Weiwei et al [12], introduced PiiGAN, an innovative generative adversarial network for pluralistic image inpainting that achieves not only higher quality results but also a diverse range of realistic outputs. Additionally, developed a new extractor to enhance the GAN by capturing the style vectors of training samples during each iteration and incorporating a consistency loss to enable the generator to learn various styles corresponding to the semantics of the input image. This model has been validated to inpaint the same missing regions with multiple plausible results that align with the high-level semantics of the image, and its effectiveness have been assessed on several datasets.

Khalid M. Hosny et al [13], introduced a CNN-based method for detecting copy-move forgery in images. The model distinguishes between forged and original images by generating feature vectors and using a fully connected layer to identify feature correspondences. Once trained, it accurately detects tampered images. The method was evaluated on different benchmark datasets, such as MICC-F2000, MICC-F600, and MICC-F220—where it outperformed other techniques, yielding 100% accuracy within 35 epochs. Additionally, the model achieved fast processing time within 47.48 seconds.

B. Deep Learning Methods

Several advanced techniques are employed to manipulate images and videos, harnessing the capabilities of DL [9, 14] such as: entire face synthesis, identify swap, attribute manipulation, expression swap, [15] as shown in Fig.2.


(A) Expression Swap

(B) Entire Face Synthisis

(C) Attribute Manipulation

(D) Aging /De-Aging

Fig 2. *Examples for manipulated image by deep learning method [35].*

I. Expression Swap

Expression swapping, also known as face reenactment, entails transferring a facial expression from a source image or video to another person's face in a target image or video, as illustrated in Fig.2(A). This technique alters the target's expression to match that of the source, creating a reenactment effect. Current approaches employ image-level manipulation using well-known GAN architectures such as Face2Face [16]. Stehouwer et al [17] implement an attention mechanism to enhance feature maps specifically for classification, rather than using multi-task learning to simultaneously detect manipulated images and identify altered regions. The attention maps focus on the most pertinent areas, improving the binary classification of real versus fake faces while also visualizing the modified regions. This research also developed a large-scale database of various facial forgeries to aid in detecting and localizing manipulated faces and performed a comprehensive analysis of data-driven fake face detection. Their results indicate that the attention mechanism significantly enhances both the detection and localization of forgeries.

II. Identity Swap

The identity swap technique, also known as face-swapping, is widely used to replace one individual's face in an image or video with that of another person, In this process, the source image contributes the identity while the target image provides the features, resulting in a face that has been swapped.as illustrated in Fig. 2(A). This manipulation is usually done with GANs such as Face Swap-GAN [20]. Ruben et al [21] provide an extensive analysis of deepfake advancements, emphasizing facial regions and detection accuracy. They compare early DeepFake datasets like UADFV and FaceForensics++ with newer datasets such as Celeb-DF [22] and DFDC [23]. Two detection approaches were evaluated: one using the entire face and the other focusing on specific regions like the eyes and nose. The results indicate that second-generation deepfakes are significantly harder to detect, showing equal error rates (EER) of 20-30% compared to 1-3% in first-generation datasets. Enhanced realism, especially in areas like the nose, mouth, and face edges, resulted in EERs of 24-44% for these regions. This study provides valuable insights into improving detection by analyzing various facial regions under different conditions such as lighting and camera distance.

III. Entire Face Synthesis

These set of approaches typically generate syhtnetic face images which are completely new, as shown in Fig.2(B), artificial face is created that does not actually exist, using advanced GANs like styleGan[34]. Nicholas and Hany [18] demonstrate the forensic classifiers that are susceptible to a variety of attacks that drastically lower their accuracy, in some cases to nearly 0%. They conducted five case studies on attacks against a state-of-the-art classifier that achieves an area under the ROC curve (AUC) of 0.95 for nearly all image generators, even when trained on just one generator. With complete access to the classifier, the authors discovered that flipping the lowest bit of each pixel in an image reduces the AUC to 0.0005; changing just 1% of the image area lowers the AUC to 0.08; and adding a single noise pattern in the synthesizer's latent space decreases the AUC to 0.17. Furthermore, they developed a black-box attack that does not require access to the classifier, which reduced the AUC to 0.22. These results reveal significant vulnerabilities in certain image-forensic classifiers.

IV. Attribute Manipulation

Face editing, or retouching, involves modifying facial features like skin color, gender and hair style, as illustrated in Fig.2(C). This process is usually done with GAN models, such as CycleGAN [31]. Christian et al. [20] studied the effects of moderate facial retouching on advanced face recognition systems. Their results indicate that while minor retouching has little impact on recognition accuracy if image quality is preserved, it becomes important for applications like social media and automated border control. For contexts like anti-photoshop legislation, detecting retouched images is crucial. To address this, they developed a detection system based on Photo Response Non-Uniformity features, using score-level fusion for final detection. This system, requiring minimal training, performed well but struggled with compressed images when matching the retouched images' average file size in JPEG format. Despite this, the PRNU-based approach successfully distinguished between original and edited images, attaining an average detection rate of 86.3% across different retouching techniques.

V. Aging/ De-Aging

These methods primarily rely on deep neural networks, including CNNs and GANs, to model the complex transformations that occur as a person ages. In aging, the models enhance certain facial attributes such as wrinkles, skin sagging, and graying hair, simulating the natural progression of time. Conversely, de-aging techniques work by reversing these changes, effectively rejuvenating the face to appear younger by reducing wrinkles, firming skin, and adding features commonly associated with youth. As shown in Fig.2 (D).

Martis, J. E et al [28] this study demonstrated the potential of combining de-aging networks with sketch generation to improve forensic facial recognition. The authors proposed system effectively addresses aging challenges, enhancing identification accuracy and reliability. Tested on the CUHK and AR Face Sketch Databases, it showed significant improvements in realism, with better FID scores, SSIM, and PSNR. The system uses a deepfake-based neural network for de-aging and a pix2pix-based GAN for sketch generation, producing highly realistic sketches across diverse facial features. This approach outperforms existing methods and emphasizes the need for ongoing advancements in de-aging technology for forensic applications.

3. Techniques for Detecting DeepFakes

This section delves into the various techniques available for detecting deepfake images and videos. Before exploring these methods, we will provide an overview of the deepfake detection process.

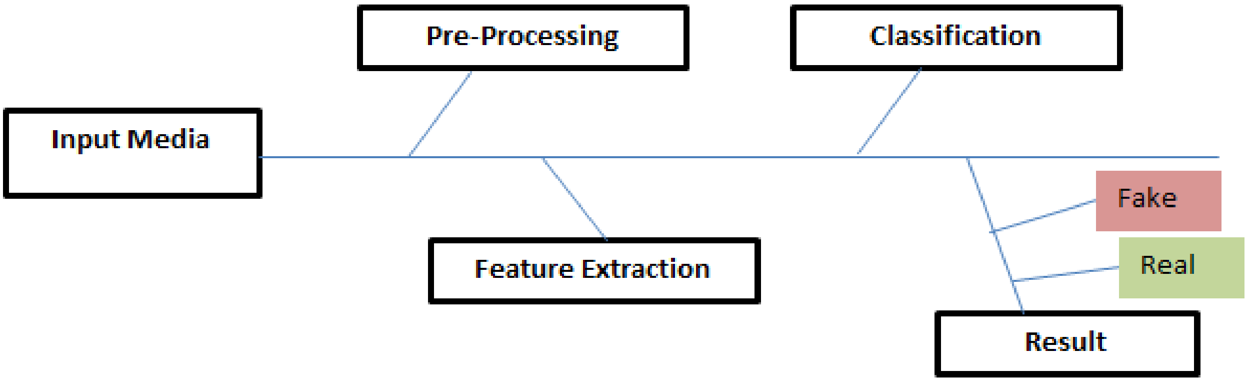


Fig 3. *Generalized Schematic of DeepFake Detection Process.*

3.1 DeepFake Detection Pipeline

Figure 3 illustrates the main phases involved in the deepfake detection process. The process begins with the input media, which undergoes pre-processing to prepare the data for analysis. Next, feature extraction is performed to identify and isolate key attributes indicative of deepfakes. These features are then passed to a classification phase, where the media is categorized as either real or fake. The final result indicates the outcome of this classification, distinguishing between authentic and manipulated content. We will present all the above process in the form of sequential steps to explain each part in detail, including the techniques and features used at each stage.

3.1.1 First Step: Pre-Processing Phase

This step covers important elements of handling feature and target variables for exploratory data analysis and model building. It emphasizes that feature variables can include person-specific data such as age, height, and weight, or image data that must be flattened and normalized. Data preprocessing is essential and varies based on the data type, whether it's time series, audio, image formats (PNG/JPEG), or structured and unstructured data. Table 1 provides commonly used pre-processing techniques in detection approaches [27].

Table 1. *Pre-Processing Techniques*

Pre-Processing Techniques	**Description**
Face Detection	Detecting and isolating faces from images or video frames allows the analysis to concentrate on the relevant areas of interest
Face Alignment	Standardizing the orientation and size of the extracted faces ensures uniformity across samples, minimizing variability from pose and angle differences.
Scaling	Normalizing pixel values to a standard range, like [0, 1] or [-1, 1], helps stabilize the training process.
Mean Subtraction	Subtracting the dataset's mean pixel value from each pixel centers the data, which can facilitate faster convergence during training.
Geometric Transformations	Implementing random rotations, translations, flips, and cropping enhances the variability of the training data and strengthens the model's robustness.
Color Jittering	Begin by adjusting the brightness, contrast, saturation, and hue of the images to ensure the model is resilient to varying lighting conditions.
Noise Addition	Incorporating random noise into the images helps simulate different real-world scenarios and enhances the model's robustness.

3.1.2 Second Step: Feature Extraction Phase

This phase focuses on identifying and analyzing subtle inconsistencies and artifacts that arise during the creation of DeepFakes, serving as the foundation for accurate detection. This phase employs a combination of advanced techniques to capture critical patterns: (1) Handcrafted features, including geometric features, shape descriptors, texture analysis, and color or lighting inconsistencies; (2) DL features, leveraging convolutional neural networks (CNNs), pre-trained models, and layer-wise representations for robust high-level embeddings; (3) Frequency domain features, (4) Temporal featuresand (5) Hybrid features.

A. Handcrafted Features

These features can be categorized into several groups such as: geometric features Concentrate on detecting and analyzing facial landmarks, including the eyes, nose, and mouth, by assessing their relative positions, distances, and movements to identify anomalies. Shruti at el [25], explained a forensic method for identifying deepfake videos by identifying mismatches between phonemes and visemes (mouth shapes). Specifically, sounds linked to the phonemes P, M, and B necessitate complete closure of the mouth., which is often not accurately reproduced in deepfakes. Two detection methods are proposed: one uses handcrafted features requiring no large datasets, and the other is a CNN model. While current results are person-specific (e.g., trained on videos of Barack Obama), broader datasets could improve performance. Expanding the analysis to include all visemes could enhance detection but poses challenges. Despite limitations, the technique effectively detects lip-sync deepfakes and could be refined as part of a broader forensic toolkit for combating deepfake advancements.

Shape Descriptors, Which capture the outlines and shapes of facial features, are also effective in detecting inconsistencies in altered media. McCrae et al [26] presented a novel classification model aimed at identifying semantic discrepancies between video content and text captions in social media news posts. Unlike existing systems focused on text and images, this approach handles the complexities of video, which includes multiple scenes and audio. The authors developed a multi-modal fusion framework that combines textual analysis, audio transcription, video semantics, object detection, named entity consistency, and facial verification to identify mismatches. They curated a dataset of 4,000 Facebook news posts to train and test the model, achieving 60.5% accuracy in detecting mismatches, outperforming single-modality models. Ablation studies showed that combining multiple modalities is crucial for success.

Texture Analysis Another significant category involves techniques such as Local Binary Patterns (LBP) [27], which analyze textures by comparing each pixel to its neighboring pixels, assisting in the detection of irregularities in skin texture and facial features. The formula for LBP is given by eq. (1):

$$\boldsymbol{LBP = \sum_{p=0}^{p-1} sign\left(I(x_p) - I(x_c)\right) . 2^p} \tag{1}$$

Where, $I(x_p)$ is the intensity value of a pixel x_p in the neighborhood,$I(x_c)$ is the intensity value of the center pixelx_c, P is the number of neighboring pixels, used 8 pixels , the output is a binary pattern used for texture classification.

Likewise, the Histogram of Oriented Gradients (HOG) [9] captures the gradient orientations in specific regions of an image to identify unnatural edges and textures that deepfake algorithms often introduce. It’s formula shown below at eq. (2).

$$\boldsymbol{HOGFeature = \sum_{i=1}^{N} histogram\ of\ gradients\ for\ cell\ i} \tag{2}$$

Where, N is the number of cells in the image, each histogram is formed by computing gradient directions and magnitudes in the local regions.

Color and Lighting Features color histograms analyze the color distribution within an image to identify inconsistencies that could indicate manipulation, whereas lighting consistency assesses the uniformity of lighting and shading across the face, as deepfake algorithms often struggle to accurately reproduce natural lighting. The formula shown below at eq. (3).

$$\boldsymbol{H(c) = \sum_{i=1}^{N} \delta(c - I(i))} \tag{3}$$

Where, H(C) is the histogram count for color c, I(i) is the color of the i-th pixel in the image, δ is the Kronecker delta function (1 if the condition is true,0 otherwise).

B. Deep Learning Features

There are several feature extraction techniques includes: CNNs are used to automatically learn and extract hierarchical features from raw images through layers of convolutions, pooling, and activations, capturing complex patterns and details indicative of deepfakes [10]. Pre-trained Models, such as VGG and ResNet, are employed in transfer learning, leveraging their pre-training on large datasets to extract features from new images and recognize complex and abstract patterns [10, 24]. Layer-Wise Feature Extraction enables the capture of multi-level features from different layers of a deep network, spanning from basic edges and textures to more complex semantic information.

C. Frequency Domain Features

It is essential for identifying manipulations in images includes Fourier transform is used for frequency analysis [10], the images are converted into the frequency domain to reveal patterns and artifacts that are not visible in the spatial domain, and these frequency anomalies often indicate manipulation. The 2D Discrete Fourier Transform (DFT) for an image I(x,y) is given by eq. (4)

$$\boldsymbol{F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y).exp(-j2\pi(\frac{ux}{M} + \frac{uy}{N}))} \tag{4}$$

Where, F(u,v) is the frequency representation of the image, I(x,y) is the spatial domain image at pixel coordinates (x,y), M and N are the width and height of the image, respectively, u and v are the frequency domain coordinates, J is the imaginary unit.

Wavelet Transform complements this by enabling multi-resolution analysis, decomposing images into different frequency components for a detailed examination of both fine and coarse features [29]. The Continuous Wavelet Transform (CWT) is given by eq. (5).

$$\boldsymbol{CWT(a,b) = \int_{-\infty}^{\infty} x(t).\psi_{a,b}^{*}(t)dt} \tag{5}$$

Where, $\int_{-\infty}^{\infty}$:Represents the integral over the entire signal duration. x(t): the input signal to be analyzed.(t) : The complex conjugate of the scaled and shifted version of the mother wavelet ψ(t), a: The scale parameter that controls the dilation or compression of the wavelet (scale inversely corresponds to frequency), b: The translation parameter that determines the shift in time, ψa,b(t): The scaled and shifted wavelet function is defined as eq. (6).

$$\boldsymbol{\psi_{a,b}^{*}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right)} \tag{6}$$

D. Temporal Features

Are crucial for analyzing videos to detect manipulations using Optical Flow Analysis focuses on motion patterns [30] by examining the movement of pixels between consecutive frames, which helps identify inconsistencies that may reveal deepfakes. Temporal Coherence involves ensuring consistency over time [9], checking that facial movements, expressions, and audio-visual synchronization remain natural across video frames. deepfakes often struggle to maintain this natural temporal coherence, making this an important indicator of manipulation.

E. Hybrid Features

Enhance detection performance by integrating various methods. This approach involves combining handcrafted features with DL features or merging spatial and temporal features. By leveraging the strengths of multiple techniques, hybrid methods offer a more comprehensive analysis and improve the overall effectiveness of detection systems.

3.1.3 Third Step: Classification Phase

As a last phase, classification techniques are applied for detection process, there is some classification techniques applied in this process such as: Support Vector Machines (SVMs) [10, 24] categorize data by determining the hyper plane that most effectively separates various classes in a high-dimensional feature space generated from extracted features. Random Forest [31, 29] classifiers utilize an ensemble of decision trees to make classifications based on media features, offering robustness against over fitting and effective handling of high-dimensional feature spaces. Logistic Regression [33] models predict the probability of a media sample being authentic or a deepfake through a linear combination of extracted features, which is then, transformed using a logistic function**.**

3.2 Fake Image/Video Detection

The current cutting-edge techniques for detecting deepfakes can be divided into feature-based and DL-based approaches, as illustrated in Fig. 4. These methods utilize specific traits found in deepfake media to distinguish them from genuine content. DL is vital for extracting these identifying features and accurately recognizing deepfake media [34, 35].

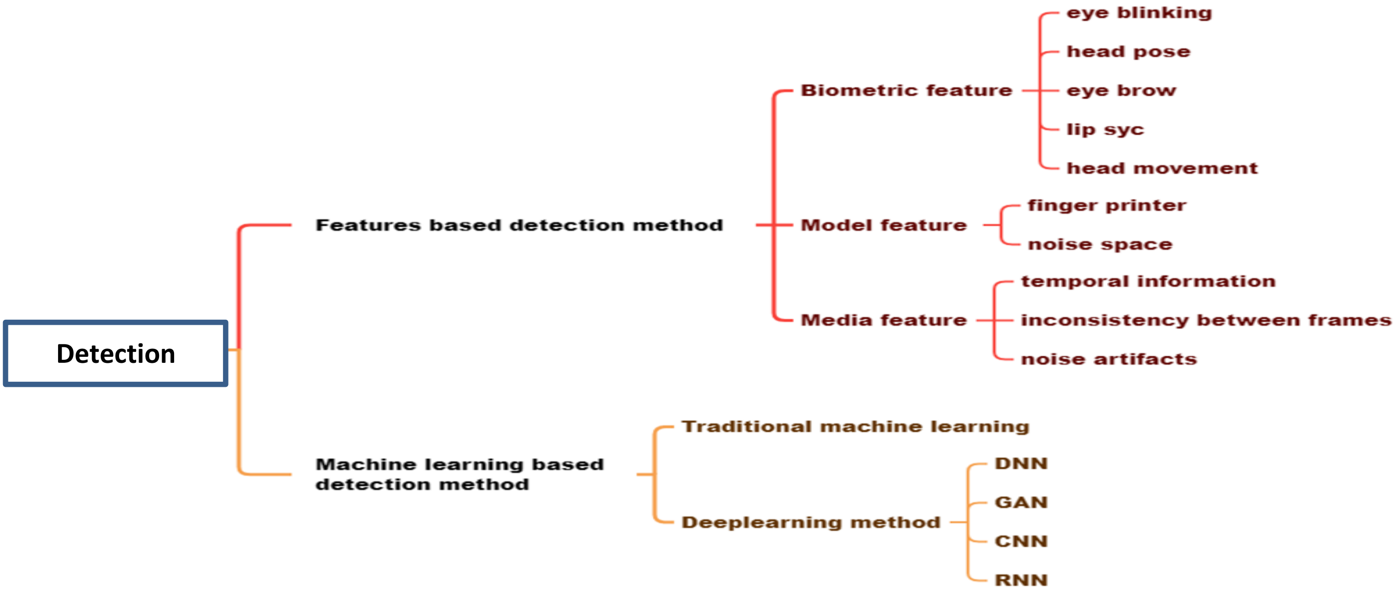


Fig 4. *Techniques for Detecting DeepFakes.*

3.2.1 Features-based Detection Approaches

Due to their method of creation, deepfake images and videos exhibit distinct characteristics that differentiate them from genuine media. These characteristics encompass biometric attributes, model-specific features, and media-specific traits [31].

A. Biometric Features

Offer unique traits for differentiating deepfake media from genuine content. These indicators encompass factors such as eye blinking, head movements, lip-sync accuracy and head pose, along with cues related to color, texture, and shape. Recent research has investigated various approaches to utilize these features for detection. Li et al. [35] observed variations in eye-blinking patterns between deepfake and authentic videos, proposing detection methods based on these patterns. DeepVision [36] similarly utilizes eye blinking behaviors to detect GAN-generated deepfake videos, achieving an accuracy rate of 0.875%. Other approaches include using eyebrow movements [37] and 3D head pose estimation [38] for detection. Techniques like Face-Aware Liquify (FAL) [39] analyze facial geometry parameters, while methods such as global consistency analysis and illumination estimation [40] aim to identify manipulated faces in images. Furthermore, studies have explored features like lip-syncing inconsistencies [41], frame-to-frame facial expression differences [42], low-level image patch inconsistencies [43], and affective cues related to perceived emotion [44] as indicators of deepfake media. These efforts highlight diverse approaches to effectively detect deepfake videos using biometric and model-specific features.

B. Model Features

Primarily stem from DL techniques, notably GANs, which are widely used for generating realistic deepfake images and videos, as outlined in Table 2. Current research works have identified specific model fingerprints embedded within GAN-generated media. Yu et al. [45] highlighted distinctive GAN fingerprints that can be employed to detect deepfake content created using GAN-based methods. Similarly, Pu et al. [46] utilized unique patterns within the noise space of GAN-generated images to identify deepfake images, achieving a high average F1 score of 0.9968 across various datasets. Additionally, convolutional traces left during the generative process of GANs have been explored for detection purposes by Guarnera et al. [47], who proposed an expectation-maximization algorithm specifically designed to extract these traces from GAN-generated images. However, the reliance on GAN fingerprints for detection has shown limitations in robustness. For instance, methods like GANprintR [48] have demonstrated effectiveness in removing GAN fingerprints from deepfake videos, rendering GAN fingerprint-based detection methods ineffective in some cases.

C. Media Features

In addition to biometric and model features, deepfake media exhibit distinctive characteristics that facilitate their detection. These include temporal inconsistencies, frame anomalies, and noise artifacts, all of which serve as critical indicators for identifying manipulated content [53].

3.2.2 DL Based Detection Approaches

DL based detection methods can be further categorized into traditional approaches which employ traditional pattern recognition techniques such as SVM and random forest models, and DL techniques such as GANs, CNNs, LSTM and RNNs. DL approaches offer automated feature extraction capabilities, leading to improved detection accuracy [24] as discussed below:

A. CNN

CNNs are special type of DL architecture that is used in computer vision and robotics [24]. CNNs were further evolved into models such as: LeNet-5 [31] is a pioneer in CNN architecture, designed for handwritten digit recognition and consists of 7 layers including convolutional, subsampling (pooling), and fully connected layers. It is simple and computationally efficient, suitable for low-resource environments. The weakness are limited applicability to more complex datasets due to its small and shallow architecture and lacks modern improvements like ReLU activation, batch normalization, and dropout.

AlexNet [10] transformed DL for image classification by introducing ReLU activation and employing dropout to mitigate over fitting and demonstrated the effectiveness of GPUs for DL. The weaknesses are computationally expensive and memory-intensive due to a large number of parameters and lacks modularity compared to more modern architectures.

VGGNet [50] is deep network with 16-19 layer, Known for its simplicity and use of small convolutional filters, achieves high accuracy for image classification and recognition tasks, designed for image classification and recognition. The limitation of VGGNet it is a very large model size, with millions of parameters, leading to high memory and computational requirements. Inefficient compared to newer architectures like ResNet or DenseNet.

GoogLeNet [51] introduced Inception modules to handle different filter sizes, used global average pooling instead of fully connected layers, designed for image classification and recognition. But it is more complex architecture, making it harder to implement and tune. Also initial versions lacked residual connections, which were later shown to improve training of deeper networks.

Inception-v3 [52] enhanced version of GoogLeNet with more efficient architectures and factorized convolutions and efficient grid size reduction. But it is still relatively complex compared to simpler models like VGGNet or ResNet. Furthermore, it may not be as parameter-efficient as models like DenseNet or MobileNet for specific applications, making it less ideal in certain contexts where computational efficiency is critical.

ResNet [10] used residual connections to enable training of very deep networks, alleviating vanishing gradients, very deep networks with up to 152 layers, designed for image classification and recognition. Despite all these features it increased depth can lead to higher computational cost and relatively large model size compared to more compact architectures like MobileNet.

DenseNet [9, 10] improved information flow between layers by densely connecting layers to each other's, resultant in improved gradient flow and parameters efficiency. Its computation is expensive in terms of memory usage due to dense connections and the implementation is more complex than simpler architectures like VGGNet.

MobileNet [52] optimized for Mobile and embedded vision applications, utilizing depth wise separable convolutions to reduce computational cost. Slightly lower accuracy compared to larger architectures like ResNet or EfficientNet and less effective for very complex datasets without additional tuning.

EfficientNet [53, 54] employs a balanced scaling approach across network width, depth, and resolution, enabling it to deliver high performance with significantly fewer parameters compared to traditional architectures. However, achieving its full potential requires precise tuning of compound scaling factors, which adds complexity to its configuration. Additionally, EfficientNet heavily depends on pre-training or transfer learning to attain optimal results, making it less straightforward to deploy without prior fine-tuned models.

CNNs have several weaknesses that limit their applicability in certain scenarios. One major challenge is their high computational and memory requirements, making them resource-intensive for training and deployment, especially on devices with limited hardware capabilities. CNNs also require large labeled datasets to perform well, as their numerous parameters make them prone to overfitting on small datasets. Additionally, CNNs struggle to model long-range dependencies in data effectively, as they primarily focus on local features through convolutional operations. They can also be sensitive to variations in input data, such as rotation, scaling, or translation, unless robust data augmentation is applied during training. Furthermore, the design and tuning of CNN architectures often involve significant manual effort and expertise, as finding the optimal configuration can be complex and time-consuming. Lastly, like many DL models, CNNs lack interpretability, making it difficult to understand the rationale behind their predictions, which can be a drawback in critical applications requiring transparency.

Kohli et al [55] used the leverages frequency domain features to detect facial forgeries. They utilizes a frequency-based CNN to distinguish between authentic and manipulated faces. The effectiveness of the fCNN is evaluated using the FaceForensics++ dataset, demonstrating that it reliably detects forgeries in both high-quality and low-quality video scenarios. Visualization of activation maps shows that their proposed architecture learns unique frequency features for various manipulation techniques like DeepFake, Face2Face, and FaceSwap. The fCNN achieves the highest recall for detecting DeepFakes, with scores of 0.9256, 0.8639, and 0.8399 for raw, c23, and c40 formats, respectively. It is also tested on the Celeb-DF (v2) dataset and the automatic FaceForensic benchmark, where it outperforms state-of-the-art methods, confirming its effectiveness in detecting facial manipulations.

The basic structure of a Convolutional Neural Network (CNN), illustrated in Fig. 5, consists of several essential components. The input layer represents the input image or data. The convolutional layers apply a set of filters to extract key features such as edges, corners, and curves; each filter is convolved with the input to generate a feature map that captures specific patterns in the data.

The pooling layers reduce the dimensionality of the extracted features by down-sampling, which decreases computational complexity and the number of model parameters while preserving critical information. The fully connected layers handle the final classification or regression tasks by processing the features extracted by earlier layers. Finally, the output layer provides the ultimate result, which can be a single value for regression tasks or a probability distribution over possible classes for classification tasks [24].

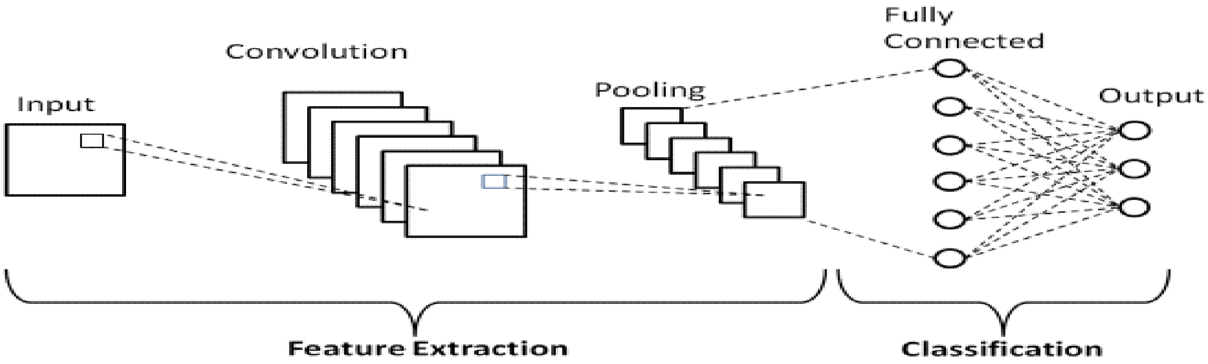


Fig 5. *CNN Architecture [13].*

B. GANs

GANs are first introduced in 2014 and consists of [24, 33, 56-58] a generative model and a discriminative model. GANs operate through an adversarial process where the Generator creates data to mimic real examples, while the Discriminator distinguishes between real and generated data. This dynamic interplay improves both models over time, with G aiming to capture data distribution and D assessing sample authenticity. Widely applied in image classification and generation, GANs have also used in medical research due to it is effective at distinguishing sophisticated forgeries. Despite challenges are training can be complex and time-consuming [31].

Kianoush et al [59] this study presents a diagnostic quality images generated using particular semantic information via SHI-GAN, which allows for the creation of an endless range of histologic images. These synthetic images can be used to expand datasets, improve model generalization by increasing image variety, and generate test cases for evaluating algorithm robustness. Tissue classifiers trained on synthetic data were found to have accuracy comparable to those trained on real data, demonstrating the quality and realism of the generated images. Expert pathologists were also able to assign Gleason grades to the synthetic images with consistency similar to real images, highlighting the potential of these images in testing human diagnostic ability and training pathologists.

GANs are powerful tools for data generation but come with several weaknesses. A significant challenge is their instability during training, as the adversarial process between the generator and discriminator often leads to issues such as vanishing gradients, mode collapse, or failure to converge. Mode collapse is particularly problematic, where the generator produces limited variations of outputs rather than the full diversity of the target distribution. GANs are also highly sensitive to hyper parameter tuning and require careful design and calibration of network architectures to achieve optimal performance, making them difficult to train effectively. Furthermore, GANs are computationally intensive, requiring significant resources for training, especially when applied to high-dimensional data like images or videos. Another drawback is their lack of interpretability; understanding how GANs generate data or why they fail in certain cases remains a challenge. Lastly, GANs may inadvertently reproduce biases present in the training data, raising ethical concerns, particularly in applications where fairness and diversity are crucial.

C. LSTM

LSTM [61, 62, and 50] network is an improvement of RNN. LSTMs excel at managing long-term dependencies, making them essential for tasks that require retaining information across extended sequences, such as language translation and time-series forecasting. The inclusion of gated mechanisms, such as input, output, and forget gates, provides precise control over information flow, improving learning efficiency and accuracy. Their capacity to process sequential data makes LSTMs ideal for applications like speech recognition and video analysis, offering flexibility and versatility in various fields. Moreover, the persistent cell state in LSTMs allows them to retain values over time, effectively capturing patterns in data that extend over long periods, while RNN can only handle short sequences.

Shobha Rani B R et al [63], present a hybrid architecture that combines ResNet50 and LSTM for deepfake video detection within a web framework using Python. By integrating ResNet50 and LSTM, the model takes advantage of the strengths of both architectures, enhancing the accuracy of deepfake video detection, particularly for videos that include both image-based and sequential data. A comparative analysis of different models was conducted using various datasets, including Celeb-DF and Face Forensic++. This developed model was evaluated for 20 and 40 epochs due to runtime constraints, yielding accuracy rates of 84.75% and 87.48%, respectively. The generated graphs indicate that both validation and testing accuracy improve as the number of epoch's increases. By utilizing the confusion matrix created during testing, the effectiveness of the system can be assessed.

Fig.6 presents the LSTM cell where is composed of several important elements that allow it to efficiently handle long-term dependencies in sequential data. The input gate regulates the transfer of information from both the current input and the previous hidden state into the memory cell, while the forget gate determines which information from the previous memory cell should be retained or discarded. The memory cell holds information that is modified according to the input and forgets gates, and the output gate governs the flow of information from the memory cell to the current hidden state and output. During the forward pass, the LSTM updates its memory cell and hidden state with each input in the sequence. The gates utilize sigmoid functions to control the information flow into and out of the memory cell, and the output gate combines sigmoid and tanh functions to generate the current hidden state and output. This selective process of remembering and forgetting information enables LSTMs to effectively capture long-term dependencies, making them well-suited for tasks such as language translation or sentiment analysis.

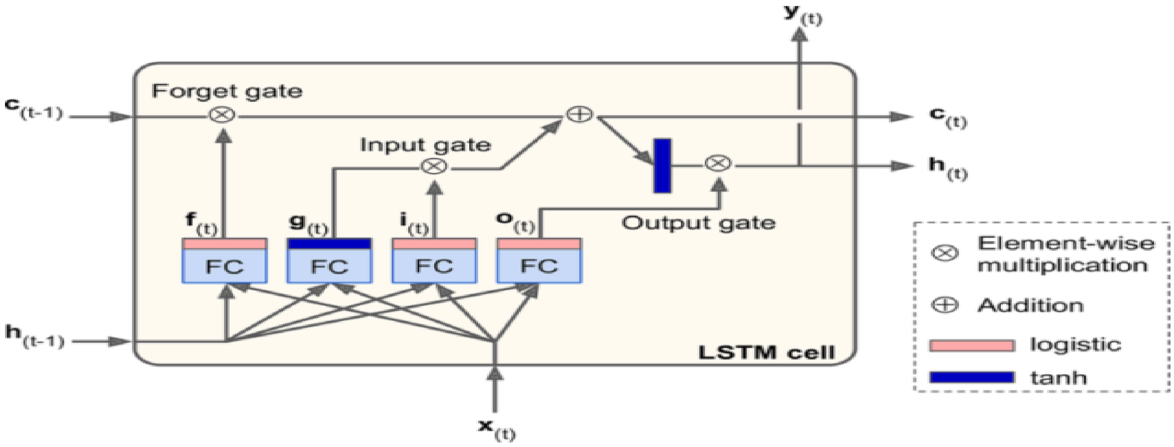


Fig 6. *Architecture of LSTM [65].*

LSTM networks address some of the weaknesses of traditional RNNs but still have their own limitations. One major drawback is their high computational cost, as the intricate gating mechanisms require more parameters and operations compared to simpler RNNs. This makes LSTMs slower to train and more resource-intensive, especially for large-scale datasets or long sequences. Despite their ability to handle longer dependencies, LSTMs can still struggle with extremely long sequences, as their memory cells have a finite capacity to store and manage information.

Additionally, like other DL models, LSTMs are prone to over fitting on small datasets unless regularization techniques, such as dropout or weight decay, are applied. They also require careful tuning of hyper parameters, such as the number of units, learning rate, and sequence length, which can be a complex and time-consuming process. Furthermore, LSTMs, like most neural networks, lack interpretability, making it difficult to understand their internal decision-making processes. These limitations have led to the adoption of alternative architectures like Transformers, which often outperform LSTMs in tasks involving very long sequences or large-scale datasets.

D. . RNN

RNN [24, 50, and 52] is a special type of neural network that are designed to process sequence data by maintaining a recurrent hidden state across multiple time steps. Unlike traditional feed forward networks, RNNs leverage sequential connections that allow them to capture temporal dependencies within data, making them particularly effective for tasks like speech recognition and natural language processing [24, 60]. RNNs operate in a loop, where the output from one time step serves as input to the next, enabling them to remember previous inputs and incorporate them into current predictions. This memory capability is crucial for handling sequential data and has been further enhanced by advancements like LSTM networks.

The unfolded RNN diagram in fig.7 processes an input vector X and produces an output vector y by scanning the data sequentially from left to right. At each time step, the hidden state is updated, resulting in an output. The network applies the same parameters at every time step, which means the parameters U, V, and W are uniformly used throughout. In this context, U is the weight matrix that controls the connection between the input layer X and the hidden layer h, W is the weight matrix that manages connections between hidden layers, and V denotes the weight linked to the connection from the hidden layer h to the output layer y. This parameter sharing allows the RNN to effectively capture temporal dependencies and handle sequential data by retaining information from prior inputs in the hidden state.

RNNs have several weaknesses that limit their effectiveness in certain tasks. One major drawback is their difficulty in capturing long-term dependencies due to the vanishing or exploding gradient problem, which often arises during back propagation through time (BPTT) in deep or lengthy sequences. This limitation makes RNNs less effective at handling tasks requiring the retention of information across long time spans. Additionally, RNNs are computationally expensive to train because their sequential nature prevents parallelization, leading to slower training times compared to architectures like Transformers. They are also prone to over fitting, especially on small datasets, and require careful regularization and tuning to perform well. Moreover, RNNs often struggle with handling highly complex or hierarchical sequences of data due to their relatively simple structure. Lastly, their training is sensitive to hyper parameter choices, and achieving convergence can require substantial expertise and computational resources. These limitations have led to the development of advanced RNN variants, such as LSTMs and GRUs, as well as the adoption of alternative architectures like Transformers.

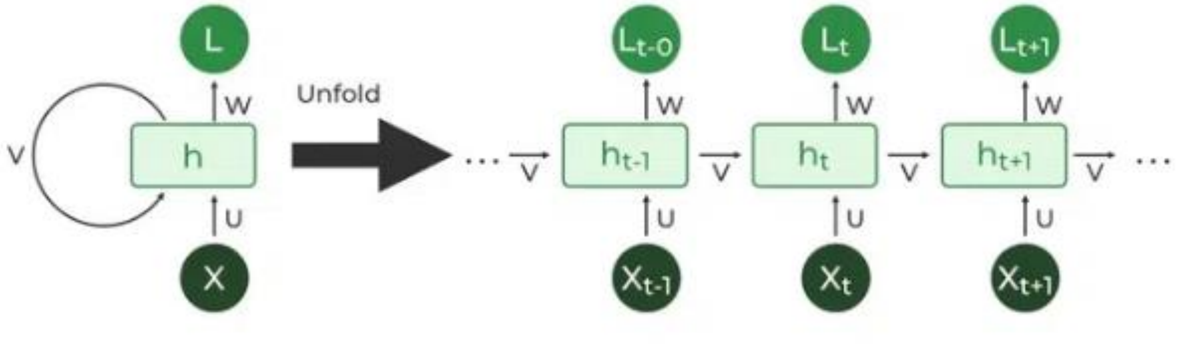


Fig 7. *RNN architecture [65].*

E. Error Level Analysis (ELA)

By utilizing ELA and DL-based forgery feature extraction, they effectively detect face-swap images. Rather than training the CNN model with original photos, they use ELA images, which enhance training efficiency because ELA images contain less information. ELA images emphasize regions where the error level surpasses a specific threshold, making differences more visible. This enhancement in contrast boosts the training effectiveness of the CNN model.

The model is specifically trained to detect counterfeit features in ELA images and assess image authenticity with precision. Leveraging just two convolutional layers, this method streamlines the process of identifying fake characteristics and evaluating the credibility of images. The study shows that features derived from ELA-processed images are effective for verifying authenticity and significantly improve the training efficiency of the CNN model. ELA is straightforward, quick, and suitable for initial screening; however, it is less effective on uniformly compressed images and is restricted to specific file formats [10].

Chakraborty et al, [64] proposed a hybrid approach that combines traditional handcrafted features with a DL model to distinguish between authentic and tampered images. They developed a dual-branch CNN that operates in conjunction with Error Level Analysis and noise residuals obtained from the Spatial Rich Model. The experiments utilized the publicly available CASIA dataset, and after training the dual-branch network for 16 epochs, it reached an accuracy of 98.55%. Furthermore, they performed a comparative analysis with earlier studies in the area of image forgery detection. The authors' results indicate that integrating DL models with traditional techniques can lead to enhanced outcomes in identifying image manipulation.

ELA has several weaknesses that limit its effectiveness. One of the primary challenges is its reliance on multiple compressions. This dependency can lead to false negatives in detecting edits on high-quality or uncompressed images. Additionally, ELA struggles with distinguishing between authentic image inconsistencies and those introduced by normal editing processes, such as resizing, color correction, or reformatting, which can result in false positives. It is also heavily influenced by the format and quality of the image; for example, highly compressed JPEG images may obscure traces of manipulation due to the dominance of compression artifacts. Furthermore, ELA requires manual interpretation of results, which can introduce subjectivity and depends on the expertise of the examiner. Its limited automation and lack of standardization make it less scalable or reliable in large-scale forensic analyses. Finally, ELA is not effective against advanced editing techniques that produce minimal compression differences, such as those generated by modern deepfake or AI-based tools.

4. Datasets

Table 2 provides an extensive overview of available datasets used for deepfake evaluation, used in deepfake research, highlighting their modalities, sizes, and unique characteristics. We emphasize the strengths and weaknesses of each dataset, which impact its usefulness for particular tasks.

FaceForensics++ [66] is a comprehensive dataset of controlled deepfake videos, primarily featuring face-swapping techniques. It stands out for its well-lit, meticulously controlled conditions and a focus on young celebrity faces. Widely used for benchmarking and evaluating deepfake detection systems, the dataset includes over 1,000 videos and showcases diverse manipulation techniques, The primary facial manipulation methods are representative and include DeepFakes, Face2Face, FaceSwap, FaceShifter, and Neural Textures. The data vary in compression levels and sizes. However, its demographic limitations (predominantly young celebrities) and relatively modest size (1,000+ videos) pose challenges to its applicability in real-world scenarios with greater diversity and complexity.

Celeb-DF [22] contains high-resolution deepfake videos that manipulated using FaceSwap and DFaker as the primary techniques. Although the dataset provides high-quality content, it is limited in demographic diversity, focusing mainly on celebrities. Publicly available with 5,000+ videos, Celeb-DF is primarily used for model training and evaluation, helping to test the robustness of detection systems in deepfake contexts. Yet, like FaceForensics++, it suffers from limited demographic diversity and focuses on celebrity faces, which limits its applicability for broader audiences.

FakeAVCeleb [69] offers high-quality deepfake videos that combine audio-visual manipulation, focusing on celebrity faces. The dataset includes over 10,000 videos available publicly and is useful for benchmarking deepfake detection systems. The dataset features a variety of manipulations, including face-swapping, lip-syncing, and voice cloning, created using state-of-the-art tools like GANs and voice synthesis technologies. Its strength lies in combining modalities, but the limited diversity in terms of age and ethnicity, along with its domain-specific nature, narrows its generalizability.

DFDC [23] dataset includes diverse subjects in deepfake videos, though the quality varies due to mixed lighting and conditions. With more than 119 k videos available publicly divided to 999 k fake and 19k real videos, DFDC is mainly used for large-scale deepfake detection challenges. The manipulation data are produced using deepfake methods, GAN-based techniques, and non-learned approaches, with resolutions varying from 320 × 240 to 3840 × 2160 and frame rates ranging from 15 fps to 30 fps. Despite limited ethnic representation, it remains a valuable resource for testing detection models on varied video quality and manipulation types. The dataset lacks sufficient representation of certain ethnic groups, which may bias detection models.

DeepFormantics-1.0 dataset [67] is a comprehensive resource designed for research in multimodal analysis, focusing on the relationship between visual forms and their semantic meanings. It includes a wide variety of annotated images, such as diagrams, info graphics, and stylized visual content, paired with corresponding textual descriptions and metadata. The dataset provides annotations for both low-level attributes, like shapes, colors, and spatial layouts, and high-level semantics, including thematic categories, relational structures, and inferred meanings. DeepFormantics-1.0 aims to advance research in areas such as computer vision, natural language processing, and cognitive computing by supporting tasks like visual-text alignment, diagrammatic reasoning, and content extraction driven by form. With its versatility, the dataset serves as a valuable tool for applications ranging from educational resources to automated content generation, fostering the development of more sophisticated multimodal AI systems.

DF-TIMIT dataset [68] is a widely used benchmark for research on deepfake detection and facial manipulation. It consists of deepfake videos created by applying face-swapping techniques using advanced deep learning models. The dataset is derived from the TIMIT dataset, which contains high-quality audio-visual recordings of individuals speaking various sentences. In DF-TIMIT, the faces in the videos are replaced with synthetic or altered versions while preserving the original audio. The dataset is available in two variations: one with high-resolution videos and another with lower-quality, compressed videos that mimic real-world conditions. This range of video quality makes DF-TIMIT an essential resource for testing the effectiveness and robustness of deepfake detection methods, playing a critical role in advancing research to combat the growing threat of deepfakes.

CIFAR-10 [70] is a well-known dataset containing high-quality images of various objects, without specific manipulation. It features 60,000 images of diverse objects, making it useful for general object classification tasks. While not focused on deepfake detection, it is widely used for image recognition models, offering a large and varied collection of non-manipulated data for testing and development. COCO: includes a vast collection of images with common objects in various settings. It is a general-purpose dataset with 330,000 images that feature broad diversity in terms of object categories and settings. While not related to deepfake detection, COCO is widely used in various image recognition and segmentation tasks, providing a rich resource for general computer vision projects.

Table 2. *Publicly accessible datasets*

Dataset Name	Content Overview	Quality of Media	Manipulation Types	Demographics & Diversity	Size and Availability	Modality	Metrix
FaceForensics++ [66]	Controlled deepfake videos (face swapping)	Well-lit, controlled conditions	Face swapping	Primarily young, celebrity faces	1,000+ videos, academic access	Video	F1, Accuracy
Celeb-DF [22]	Deepfake videos featuring celebrity faces	High-resolution, well-lit videos	Face swapping, expression manipulation	Limited diversity (ethnicity, gender, age)	5,000+ videos, publicly available	Video	F1, AUC
FakeAVCeleb [69]	Deepfake videos with audio-visual content	High-quality videos	Audio-visual manipulation	Limited diversity (age, ethnicity)	10,000+ videos, open access	Audio/ Video	Precision
DFDC [23]	Diverse subjects in deepfake videos	Mixed quality (varied lighting, conditions)	Face swapping, expression changes	Limited ethnic representation	3,000+ videos, open access	Video	F1, Recall, accuracy
DeepFormantics-1.0 [67]	Manipulated videos with comprehensive techniques	Mixed-quality media	GAN-based manipulation	Limited demographic diversity	50,000 videos, not fully detailed availability	Video	Precision
DF-TIMIT [68]	GAN-generated facial manipulations	Small-sized dataset	Expression manipulation	Limited subject diversity	620 videos	Video	Accuracy
CIFAR-10 [70]	Images of various objects	High-quality, small-size	No specific manipulation	Diverse (objects only)	60,000 images	Images	Precision
COCO [71]	Common objects in various settings	Mixed-quality images	No specific manipulation	Broad diversity	330,000 images	Images	F1, AUC

5. Benchmark

Table 3 presents an analysis of state-of-the-art deepfake detection methods due to their diversity in technology, datasets, and performance metrics. Shobha Rani B R et al. [72] utilized a ResNet50 + LSTM combination, achieving accuracies of 84% at 20 epochs and 87% at 40 epochs on the FaceForensics++ and Celeb-DF datasets. Ismail et al. [73] employed YOLO + CRNN, achieving a notable accuracy of 89.35% on Celeb-DF and FaceForensics++ (c23), emphasizing its potential for real-time applications. Priti Yadav et al. [87] used InceptionResNetV2 + LSTM, where increasing the epochs from 20 to 40 significantly improved accuracy from 84.75% to 91.48% across multiple datasets, including DFDC, FaceForensics++, and Celeb-DF.

Suratkar et al. [74] achieved high AUC scores (94% on DFDC and 98% on FaceForensics++) with CNN + RNN, while D. M. Montserrat et al. [75] attained 92.61% accuracy using CNN + GRU on DFDC. Fei et al. [76] leveraged CNN + LSTM to achieve 99.25% accuracy on FaceForensics++, demonstrating exceptional performance but raising potential overfitting concerns. Badrinarayan et al. [77] integrated Xception Net (CNN) with interpretability tools like LRP and LIME, achieving 90.17% accuracy on FaceForensics++. Finally, Usha Kosarkar [92] employed LSTM with GRU and VARMA on a comprehensive dataset of 200,000 samples, reaching an impressive accuracy of 99.8%, showcasing the potential of advanced hybrid architectures.

For future research, we recommend exploring advanced hybrid models to achieve high accuracy. If you are prepared to manage increased complexity, consider utilizing CNN combined with LSTM [76] or integrating LSTM with GRU and VARMA [78]. These approaches offer robust performance for complex tasks and data modeling. These approaches are yielding near-perfect results and offer cutting-edge potential. For real-time applications, YOLO + CRNN [73] is an excellent choice due to its ability to process data quickly while maintaining high accuracy. For interpretability in deepfake detection, Xception Net with interpretability tools [77] could be an interesting angle to explore. If you're looking for a simpler yet effective baseline, ResNet50 + LSTM or CNN + RNN could be good starting points. They strike a balance between performance and complexity.

Table 3. *Deepfake Detection Methods analysis.*

Research	Technologies	Dataset	Best performance
Shobha Rani B R et al [72]	*Resnet50 + LSTM*	*FaceForensics++ + Celeb-DF*	*Acc = 84% at epoch= 20* *Acc = 87% at epoch= 40*
Ismail et al [73]	*YOLO + CRNN*	*CelebDF + FaceForencics ++ (c23)*	*Acc = 89.35%*
Priti Yadav et al [52]	*InceptionResNetV2 + LSTM*	*The Dataset has been collected from DFDC dataset available on Kaggle, FaceForenscis and Celeb-deepfakeforensics*	*Acc = 84.75% at epoch= 20* *Acc = 91.48 % at epoch= 40*
Suratkar et al [74]	*CNN + RNN*	*DFDC* *Caleb-DF*	*AUC = 94% On DFDC* *AUC = 98% On FaceForencics++*
D. M. Montserrat et al [75]	*CNN + GRU*	*DFDC*	*Acc = 92.61%*
Fei et al [76]	*CNN + LSTM*	*Faceforensics++*	*Acc = 99.25%*
Badhrinarayan et al [77]	*Xception net (CNN) + LRP and LIME*	*Faceforensics++*	*Acc = 90.17%*
Usha Kosarkar [78]	*LSTM with GRU and VARMA*	*Collection of the DeepFake Datasets & Samples =200k*	*ACC=99.8%*

6. Conclusion and future work

This paper offers a comprehensive overview of deep learning (DL) approaches in the field of deepfake detection, examining the various strategies employed to identify and combat manipulated content. It delves into the underlying algorithms and architectures used in deepfake creation, emphasizing their strengths, limitations, and recent advancements. By critically analyzing each DL-based method for both deepfake detection and generation, the study highlights their distinctive features, providing insight into their performance and potential applications. Additionally, the paper explores the most widely used datasets in deepfake research, detailing their scope, diversity, and the implications these factors have on the development and evaluation of deepfake detection systems.

The technologies for both generating and detecting deepfakes are continually advancing, but the accuracy of current detection methods remains inadequate. As new techniques for creating deepfakes are developed, the effectiveness of existing detection methods diminishes, making it increasingly difficult to identify them. In contrast to previous studies [78] reaching an impressive accuracy of 99.8% by using LSTM with GRU and VARMA and 200000 images sample. Fei et al [76] leveraged CNN + LSTM to achieve 99.25% accuracy on FaceForensics++. Montserrat et al [75] attained 92.61% accuracy using CNN + GRU on DFDC.

Our results suggest that the combination of CNN and GRU provides more robust detection systems. Summary of the past three years, we conduct that the XceptionNet has emerged as one of the top-performing pre-trained models for deepfake detection, renowned for its use in the FaceForensics++ benchmark. Its efficient separable convolutions allow for accurate feature extraction, particularly in identifying subtle manipulations in facial images.

EfficientNet, with its balanced architecture that scales depth, width, and resolution, offers exceptional performance and efficiency, making it ideal for analyzing video frames while being computationally lightweight.

Additionally, I3D (Inflated 3D ConvNet) has proven highly effective in analyzing temporal and spatial features simultaneously, processing video sequences directly rather than individual frames, making it a strong contender for real-time deepfake video analysis.

Future work in deepfake detection must address the vulnerabilities of current systems to sophisticated attacks. One such challenge is distortion-minimizing attacks, which focus on reducing visible distortions in manipulated images or videos, making them appear more realistic and harder to detect. These attacks target the perceptible artifacts that detection algorithms typically rely on. Another significant threat is loss-maximizing attacks, designed to manipulate the loss function, thereby confusing detection algorithms and causing incorrect results, ultimately reducing system accuracy. Additionally, adversarial patch attacks involve introducing small, imperceptible changes, known as "patches," to the media. These subtle alterations effectively deceive detection systems, preventing them from identifying manipulations. Overcoming these advanced attack strategies is crucial for enhancing the robustness of future detection technologies.

Additionally, many deepfake detection methods lack robustness: the ability of a system to maintain good performance under diverse and unforeseen conditions. These systems often fail when exposed to adversarial perturbations: small, deliberate changes made to the data to mislead the model, and shallow reconstruction techniques: simple or non-deep techniques that do not capture complex manipulations, thus compromising performance. A significant challenge lies in the scarcity of high-quality, diverse datasets: datasets that contain accurate and diverse information crucial for training effective detection models. While large and varied datasets are essential for robust model training, existing datasets are often limited, focusing primarily on human faces, which restricts the model’s ability to generalize to other scenarios.

These datasets also lack real-world complexity, failing to represent a wide range of contexts. Privacy concerns: issues related to the protection of personal data, make it difficult to gather diverse and comprehensive datasets, especially when they involve real-world data containing identifiable individuals. Current datasets vary greatly in terms of resolution**,** length and diversity**:** the range of content in the dataset. Without standardized benchmarks, it is hard to evaluate and compare different detection techniques effectively. To address these challenges, future research must focus on establishing standardized benchmarks and developing automated testing methodologies: self-evaluating systems that can test detection models consistently and without human intervention, providing reliable results.

As deepfake generation techniques continue to evolve, the reliability of biometric features unique human biological characteristics such as eye blinking or facial movements diminishes, rendering these traditional detection methods increasingly ineffective. This calls for the development of more advanced, robust detection mechanisms. Moving forward, it is critical for research to focus on systematic approaches—structured, well-organized methods that integrate multi-modal cues from diverse sources, such as visual and audio data, to improve both detection performance and resilience. Recent advancements in deep learning (DL) algorithms have significantly elevated the quality of deepfake content, making conventional detection techniques less reliable and efficient, particularly for deepfake videos. To address these challenges, future research must prioritize the fusion of biometric, media, and model-based features, optimizing detection accuracy and effectiveness.

References

[1] T. Zhang, ‘‘Deepfake generation and detection, a survey,’’ *Multimedia Tools Appl*, vol. 81, no. 5, pp. 6259- 6276, Feb.2022.

[2] A. O. J. Kwok and S. G. M. Koh, ‘‘Deepfake: A social construction of technology perspective,’’ *Current Issues Tourism,* vol. 24, no. 13, pp. 1798 –1802, Jul. 2021.

[3] S. Awah Buo, ‘‘The emerging threats of deepfake attacks and countermeasures,’’ *arXiv:2012.07989,* 2020.

[4] J. Ice, ‘‘Defamatory political deepfakes and the first amendment,’’ Case W. Res. L. Rev, vol. 70, pp. 417- 455, 2019.

[5] Kiliç, Büşra and Mehmet Emin Kahraman. “Current Usage Areas of Deepfake Applications with Artificial Intelligence Technology,” *İletişim ve Toplum Araştırmaları Dergisi*, pp.301-332,2023.

[6] V.KarasavvaandA.Noorbhai,‘‘The real threat of deepfake pornography: A review of Canadian policy,’’ *Cyber psychology, Behavior, and Social Networking,* vol. 24, no. 3, pp. 203–209,Mar.2021.

[7] Patil, Kundan, Shrushti Kale, Jaivanti Dhokey and Abhishek A. Gulhane., “Deepfake Detection using Biological Features: A Survey*,” ArXiv abs/2301.05819*, 2023.

[8] Heidari, Arash, Nima Jafari Navimipour, Hasan Dag and Mehmet Unal.,“Deepfake detection using deep learning methods: A systematic and comprehensive review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14*, 2023.

[9] Patel, Yogesh, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent E. Davidson, Royi Nyameko, Srinivas Aluvala and Vrince Vimal., “ Deepfake Generation and Detection: Case Study and Challenges .” *IEEE Access 11*, pp. 143296-143323, 2023.

[10] Pei, Gan, Jiangning Zhang, Menghan Hu, Guangtao Zhai, Chengjie Wang, Zhenyu Zhang, Jian Yang, Chunhua Shen and Dacheng Tao., “Deepfake Generation and Detection: A Benchmark and Survey,” *ArXiv abs/2403.17881*, 2024.

[11] Tao, Zhang., “Deepfake generation and detection, a survey,” Multimedia Tools and Applications 81, pp.6259 – 6276, 2022.

[12] Cai, Weiwei and Zhanguo Wei., “PiiGAN: Generative Adversarial Networks for Pluralistic Image Inpainting,” *IEEE Access 8*, pp.48451-48463, 2019.

[13] Hosny, Khalid M., Akram M. Mortda, Mostafa M. Fouda and Nabil A. Lashin., “An Efficient CNN Model to Detect Copy-Move Image Forgery,” *in IEEE Access*, vol.10, pp. 48622-48632, 2022.

[14] Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Dung Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Viet Quoc Pham and Cu Nguyen. ,“Deep learning for deepfakes creation and detection: A survey.” Comput. Vis. Image Underst, pp.103525-103544, 2019.

[15] J. Thies, M. Zollh¨ofer, and M. Nießner., “Deferred neural rendering: image synthesis using neural textures,” *ACM Transactions on Graphics (TOG),* vol. 38, no. 4, 2019.

[16] Verdoliva, Luisa., “Media Forensics and DeepFakes: An Overview,” *IEEE Journal of Selected Topics in Signal Processing 14,* pp.910-932, 2020.

[17] Stehouwer, Joel, Hao Dang, Feng Liu, Xiaoming Liu and Anil K. Jain., “On the Detection of Digital Face Manipulation,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5780-5789, 2020.

[18] Carlini, Nicholas and Hany Farid., “Evading Deepfake-Image Detectors with White- and Black-Box Attacks*” IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp.2804-2813, 2020.

[19] Rathgeb, Christian, Angelika Botaljov, Fabian Stockhardt, Sergey Isadskiy, Luca Debiasi, Andreas Uhl and Christoph Busch., ” PRNU-based detection of facial retouching,” *IET Biom. 9,* pp. 154-164, 2020.

[20] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J,” Deepfakes and beyond: A Survey of face manipulation and fake detection,” *Information Fusion, 64,* pp.131–148, 2020.

[21] Tolosana, Rubén, Sergio Romero-Tapiador, Julian Fierrez and Rubén Vera-Rodríguez., “DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance,” *ArXiv abs/2004.07532,* 2020.

[22] Mao, M.; Lee, A.; Hong, M., “Deep Learning Innovations in Video Classification: A Survey on Techniques and Dataset Evaluations,” *Electronics*, 2024.

[23] Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang and Cristian Cantón Ferrer., “The DeepFake Detection Challenge (DFDC) Dataset.,” arXiv: Computer Vision and Pattern Recognition , 2020

[24] Kishan Vyas1, Preksha Pareek1, Ruchi Jayaswal1, Shruti Patil1., “Analysing the landscape of Deep Fake Detection: A Survey,” *International Journal of Intelligent Systems and Applications in Engineering IJISAE*, pp.40–55, 2024.

[25] Agarwal, Shruti, Hany Farid, Ohad Fried and Maneesh Agrawala. ,“Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp.2814-2822, 2020.

[26] McCrae, Scott, Kehan Wang and Avideh Zakhor., “Multi-Modal Semantic Inconsistency Detection in Social Media News Posts,” *Conference on Multimedia Modeling*, 2021.

[27] Bonomi M, Pasquini C, Boato G,” Dynamic texture analysis for detecting fake faces in video sequences,” *arXiv preprint arXiv:20071527*, 2020.

[28] Martis, J. E., Sannidhan, M. S., Hegde, N. P., & Sadananda, L., “Precision sketching with de aging networks in forensics,” *Frontiers in Signal Processing*, 4, 2024.

[29] Kaur, Achhardeep, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin and Feng Xia., “Deepfake video detection: challenges and opportunities,” *Artif. Intell. Rev. 57*, 2024.

[30] Ganiyusufoglu I, Ngô LM, Savov N, Karaoglu S, Gevers T., “Spatiotemporal features for generalized detection of deepfake videos,” *arXiv preprint arXiv:201011844*, 2020.

[31] Juefei-Xu, Felix, Run Wang, Yihao Huang, Qing Guo, Lei Ma and Yang Liu. “Countering Malicious DeepFakes: Survey, Battleground, and Horizon,” *International Journal of Computer Vision 130*, pp. 1678 – 1734, 2021.

[32] Malik, Asad, Minoru Kuribayashi, Sani M. Abdullahi and Ahmad Neyaz Khan., "DeepFake Detection for Human Face Images and Videos: A Survey," *IEEE Access 10*, pp.18757-18775, 2022.

[33] R. Singh, S. Shrivastava, A. Jatain, and S. B. Bajaj., ''Deepfake images, videos generation, and detection techniques using deep learning,'' *Machine Intelligence and Smart Systems: Proceedings of MISS 2021. Singapore: Springer Nature Singapore*, pp.501-514, 2022.

[34] FelixJuefei-Xu · RunWang · YihaoHuang · QingGuo · LeiMa · YangLiu., "Countering Malicious DeepFakes: Survey , Battleground, and Horizon," *arXiv:2103.00218v1 [cs.CV]*, 27 Feb 2021.

[35] Li Y, Chang M-C, Lyu S, In ictu oculi.," Exposing ai created fake videos by detecting eye blinking," *In: IEEE international workshop on information forensics and security (WIFS*). 2018.

[36] Jung T, Kim S, Kim K., DeepVision: "Deepfakes detection using human eye blinking pattern," *IEEE Access*,pp.83144–83154, 2020.

[37] Nguyen HM, Derakhshani R.,"Eyebrow recognition for identifying Deepfake videos," *In: international conference of the biometrics special interest group (BIOSIG)*, 2020.

[38] Yang X, Li Y, Lyu S., "Exposing deep fakes using inconsistent head poses," *In: IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019.

[39] J. Y. Zhu, T. Park, P. Isola, and A. Efros., "Unpaired image-toimage translation using cycle-consistent adversarial networks," *In: IEEE International Conference on Computer Vision*, 2017.

[40] Matern F, Riess C, Stamminger M., "Exploiting visual artifacts to expose Deepfakes and face manipulations," *In: IEEE winter applications of computer vision workshops (WACVW*), 2019.

[41] Korshunov P, Marcel S., "Speaker inconsistency detection in tampered video," *In: 26th european signal processing conference (EUSIPCO),IEEE*, 2018.

[42] Zhao Y et al., "Capturing the persistence of facial expression features for deepfake video detection," *Springer International Publishing, Cham*, 2020.

[43] Zhou P, et al., "Two-stream neural networks for tampered face detection." *In: IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, 2017.

[44] Mittal T, et al., "Emotions Don't Lie: An audio-visual deepfake detection method using affective cues," *In: Proceedings of the 28th ACM international conference on multimedia Association for Computing Machinery*, pp.2823–2832, 2020.

[45] Yu N, Davis L, Fritz M., "Attributing fake images to GANs: Analyzing fingerprints in generated images," *arXiv: 1811. 08180*, 2018.

[46] Pu J, et al., "NoiseScope: detecting deepfake images in a blind setting," *In: Annual computer security applications conference. Association for Computing Machinery, Austin*, pp.913–927, 2020.

[47] Guarnera L, Giudice O, Battiato S., "Deepfake detection by analyzing convolutional traces," *In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, 2020.

[48] Neves JC et al.," GANprintR: improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE Journal of Selected Topics in Signal Processing 14(5)*, pp.1038–1048, 2020.

[49] Sabir E et al., "Recurrent convolutional strategies for face manipulation detection in videos." *Interfaces (GUI)*, 2019.

[50] Tao Zhang1., "Deepfake generation and detection, a survey," *Multimedia Tools and Applications 81*, pp.6259–6276, 2022.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich., ''Going deeper with convolutions,'' *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1–9, Jun. 2015.

[52] Priti Yadav, Ishani Jaswal, Jaiprakash Maravi, Vibhash Choudhary and Gargi Khanna., "DeepFake Detection using InceptionResNetV2 and LSTM," *International Conference on Emerging Technologies: AI, IoT, and CPS for Science Technology Applications*, September, 2021.

[53] Pokroy, A. A., & Egorov, A. D., "EfficientNets for DeepFake detection: Comparison of pretrained models," *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 2021.

[54] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S., "Video face manipulation detection through ensemble of CNNS," *25th International Conference on Pattern Recognition (ICPR)*, 2021.

[55] Kohli, Aditi and Abhinav Gupta., "Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN," *Multimedia Tools and Applications 80*, pp.18461 – 18478, 2021.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A Efros.,"UnpairedImage-to-ImageTranslation usingCycle-ConsistentAdversarialNetworks," *arXiv: 1703.10593v7 [cs.CV]*, 24 Aug 2020.

[57] Tero Karras, Samuli Laine,Timo Aila., "AStyle Based Generator Architecture for Generative Adversarial Networks," *arXiv:1812.04948v3 [cs.NE]*, 29 Mar 2019.

[58] Luisa Verdoliva., "Media Forensics and DeepFakes: an overview," *arXiv:2001.06564v1 [cs.CV]*, 18 Jan 2020.

[59] Falah kheir khah, Kianoush, Saumya Tiwari, Kevin Yeh, Sounak Gupta, Loren Herrera-Hernandez, Michael R. McCarthy, Rafael E. Jimenez, John C. Cheville and Rohit Bhargava., “Deepfake histological images for enhancing digital pathology,” *ArXiv abs/2206.08308*, 2022.

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

[61] C. C. Ki Chan, V. Kumar, S. Delaney, and M. Gochoo., ‘‘Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media,’’ *in Proc. IEEE/ITU Int. Conf. Artif. Intell. for Good (AI4G)*, pp. 55–62, Sep. 2020.

[62] Houdt, Greg Van, Carlos Mosquera and Gonzalo Nápoles., “A review on the long short-term memory model,” *Artificial Intelligence Review 53*, pp. 5929 – 5955, 2020.

[63] R, Shobha Rani B, Piyush Kumar Pareek, Bharathi S and Geetha G., “Deepfake Video Detection System Using Deep Neural Networks,” *IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS),* pp.1-6, 2023.

[64] Chakraborty, Sunen, Kingshuk Chatterjee and Paramita Dey., “Detection of Image Tampering Using Deep Learning, Error Levels and Noise Residuals,” *Neural Process. Lett. 56*, 2024.

[65] Fang, Wei, Yupeng Chen and Qiongying Xue., “Survey on Research of RNN-Based Spatio-Temporal Sequence Prediction Algorithms,” *Journal on Big Data,* 2021.

[66] Mao, Maoyu and Jun Yang., “Exposing Deepfake with Pixel-wise AR and PPG Correlation from Faint Signals,” *ArXiv abs/2110.15561*, 2021.

[67] Liming Jiang, Wayne Wu, Ren Li, Chen Qian, and Chen Change Loy.,” DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection,” *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2020.

[68] Pavel Korshunov and Sebastien Marcel., “DeepFakes: a New Threat to Face Recognition? Assessment and Detection,” *arXiv preprint arXiv: 1812.08685,* 2018.

[69] Zhang, Yuanhan, Zhen-fei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao and Ziwei Liu., “CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations,” *European Conference on Computer Vision*, 2020.

[70] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer., “The Deepfake Detection Challenge (DFDC) Preview Dataset,” *arXiv preprint arXiv:1910.08854,* 2019.

[71] Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick., “Microsoft COCO: Common Objects in Context,” *European Conference on Computer Vision,* 2014.

[72] R, Shobha Rani B, Piyush Kumar Pareek, Bharathi S and Geetha G., “Deepfake Video Detection System Using Deep Neural Networks.” *IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS),* pp.1-6, 2023.

[73] Ismail, Aya, Marwa S. Elpeltagy, Mervat S. Zaki and Kamal A. Eldahshan. “Deepfake video detection: YOLO-Face convolution recurrent approach,” *PeerJ Computer Science 7,* 2021.

[74] Suratkar, S, Kazi, F., “Deep Fake Video Detection Using Transfer Learning Approach,” Arab J Sci Eng, 2022.

[75] D. M. Montserrat et al., “Deepfakes Detection with Automatic Face Weighting,” *In: IEEECVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 2020.

[76] Fei, J.; Xia, Z.; Peipeng, Yu.; Xiao, F., “Exposing AI-generated videos with motion magnification,” *Multimedia Tools Appl. 80(20),* pp.30789–30802, 2020.

[77] Badhrinarayan Malolan, Ankit Parekh, and Faruk Kazi., “Explainable deep-fake detection using visual inter pretability methods,” *In 3rd International Conference on Information and Computer Technologies (ICICT),* pp.289–293, 2020.

[78] Kosarkar, U., Sakarkar, G.,” Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis,” *Multimed Tools Appl*, 2024.

(السنة) المجلد (..) العدد (..)

المجلة الدولية للحاسبات والمعلومات

متاح على الإنترنت على الرابط : https://ijci.journals.ekb.eg/

Egyptian Knowledge Bank
بنك المعرفة المصري

دراسه استقصائيه شاملة لمنهجيات التعلم العميق لكشف وانشاء التزييف العميق ومجموعات البيانات

[أ] ايمان عبدالفتاح [ب] نادر محمود [أ] حمدي موسي [أ] اشرف السيسي

[أ] قسم علوم الحاسب ، كليه الحاسبات والمعلوامت ، جامعه المنوفيه، مصر.

[ب] قسم الامن السيبرانى ، كلية الهندسة وتكنولوجيا المعلومات ، بكليات بريدة الاهلية ، المملكة العربيه السعوديه.

, hamdimmm@hotmail.com, ashraf.elsisi@ci.menofia.edu.eg. , eng.nader.mahmoud@gmail.comeman4cs@gmail.com

ملخص البحث:

التطور السريع لتقنيات التعلم العميق ولا سيما من خلال الشبكات التوليديه التنافسيه أتاح إنشاء وسائط اصطناعية شديدة الواقعية، مما أثار مخاوف في مجالات مثل السياسة والترفيه والأمن. ونتيجة لذلك، زاد الاهتمام بتطوير أنظمة قوية لكشف التزييف العميق. تقدم هذه الورقة دراسة استقصائية شاملة لمنهجيات انشاء التزيف العميق والكشف عن التزييف العميق الحديثة، مع التركيز بشكل خاص على الأساليب المعتمدة على الفيديو والصور وتطبيقاتها. تهدف إلى تعزيز فهم القارئ للتطورات الحديثة والثغرات في التدابير الأمنية الحالية والمجالات التي تحتاج إلى تحسين. تُعد هذه الدراسة من بين الدراسات القليلة التي تستعرض مجموعات البيانات بشكل شامل، مثل FaceForensics++ وDFDC وCeleb-DF، مسلطةً الضوء على نقاط قوتها وضعفها، وهي أمر بالغ الأهمية لأغراض التدريب والتحقق من نماذج التزييف العميق، مع التأكيد على الحاجة إلى مجموعات بيانات متنوعة وواقعية لضمان تعميم النماذج بشكل قوي. تناقش الدراسة جميع الجوانب المتعلقة بتحديد وإنشاء التزييف العميق بشكل شامل، مما يوفر للقارئ فهمًا معمقًا لهذه الموضوعات في دراسة واحدة. تعتمد الدراسات الحديثة بشكل رئيسي على الشبكات العصبية الالتفافية للكشف عن التزييف العميق، مع الهدف الأساسي المتمثل في تحسين دقة الكشف، مع التركيز أيضًا على مقارنة الأساليب المختلفة. من الدراسات السابقة، توصلنا إلى أن XceptionNe , I3D , EfficientNet هي أفضل النماذج للكشف عن التزييف العميق، حيث يتميز كل منها في مجالات مختلفة: يتميز XceptionNetفي تحديد التعديلات الدقيقة، ويوفر EfficientNet تحليلًا متوازنًا وفعالًا لإطارات الفيديو، بينما يتخصص I3D في معالجة تسلسلات الفيديو في الوقت الفعلي.