Multimedia Tools and Applications (2025) 84:22187–22229
https://doi.org/10.1007/s11042-024-19906-1

A systematic literature review on deepfake detection techniques

Vishal Kumar Sharma[1] · **Rakesh Garg**[2] · **Quentin Caudron**[3]

Received: 20 January 2024 / Revised: 12 June 2024 / Accepted: 17 July 2024 /
Published online: 2 August 2024

Abstract

Big data analytics, computer vision, and human-level governance are key areas where deep learning has been impactful. However, its advancements have also led to concerns over privacy, democracy, and national security, particularly with the advent of deepfake technology. Deepfakes, a term coined in 2017, primarily involve face-swapping in videos. Initially easy to detect, rapid advancements in machine learning have made deepfakes increasingly realistic and challenging to distinguish from reality. Generative Adversarial Networks (GANs) and other deep learning methods are instrumental in creating deepfakes, leading to the development of applications like Faceapp and Fake App. These technological advancements, while impressive, pose significant risks to individual integrity and societal trust. Recognizing this, the necessity to develop systems capable of instantaneously identifying and assessing the authenticity of digital visual media has become paramount. This study aims to evaluate deepfake detection methods by discussing manipulations, optimizations, and enhancements of existing algorithms. It explores various datasets for image, video, and audio deepfake detection, including performance metrics to gauge detection algorithm effectiveness. Through a comprehensive review, this paper identifies gaps in current research, proposes future research directions, and provides a detailed quantitative and qualitative analysis of existing deepfake detection techniques. By consolidating existing literature and presenting new insights, this study serves as a valuable resource for researchers and practitioners aiming to advance the field of deepfake detection.

Keywords Deepfake detection · Generative adversarial network · Neural network · Face manipulation · Systemetic literature review · Deepfake datasets · Perfomance metrics

✉ Rakesh Garg
rkgarg06@gmail.com

Vishal Kumar Sharma
vishalsharma3003@gmail.com

Quentin Caudron
quentincaudron@gmail.com

[1] Amity School of Engineering & Technology, Amity University, Amity Rd, Sector 125, Noida, Uttar Pradesh 201301, India

[2] Department of Computer Science & Engineering, Gurugram University, Gurugram, Haryana, India

[3] Sound Agriculture, 6401 Hollis St STE 100, Emeryville, CA 94608, United States

1 Introduction

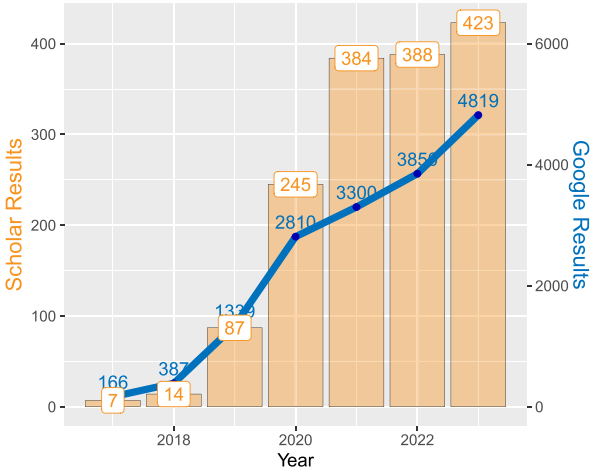
With the rapid advancement of technologies, every field—including healthcare, media, computer science, and mechanical engineering—has experienced significant changes. These changes bring both opportunities and challenges, and nowhere is this dichotomy more apparent than in the realm of digital media manipulation. Deepfake technology, in particular, has evolved swiftly, making it increasingly difficult to distinguish between genuine and manipulated media.

The concept of deepfakes, which emerged in 2017, involves the use of artificial intelligence, specifically Generative Adversarial Networks (GANs) [31], to create hyper-realistic fake videos and images. Initially, detecting these manipulated media was relatively straightforward. However, as machine learning techniques have advanced, deepfakes have become more sophisticated and harder to detect, posing serious threats to privacy, democracy, and national security.

The implications of deepfake technology are profound. On one hand, it enables impressive advancements in entertainment, such as realistic visual effects in films and innovative virtual reality experiences. On the other hand, it presents significant risks, including the potential for misinformation, identity theft, and the erosion of trust in digital content.

Given these challenges, there is an urgent need to develop robust systems capable of detecting and assessing the authenticity of digital media in real time. This review paper aims to evaluate current deepfake detection methods by examining manipulations, optimizations, and enhancements of existing algorithms. It also explores various datasets used for image, video, and audio deepfake detection, assessing their effectiveness through a range of performance metrics.

Figure 1 demonstrates the rise in the blogs and scholarly articles on Google. The figure represents a yearly comparison of scholar articles on Google and blogs on Google. The data is captured with the mandatory keyword “Deepfake Detection”.



Scholar here is referred to as Google Scholar and Google results refers to the advanced search on Google search engine.

Fig. 1 Scholar results vs Google results

This study aims to provide unique insights by identifying gaps in the current research, proposing future research directions, and offering a comprehensive comparative analysis of the effectiveness of various deepfake detection methods. Structured into five sections, the study begins with this Section 1, followed by a Section 2 detailing the methodology used. The Section 3 presents answers to the research questions identified, while the Section 4 explores future research areas in deepfake detection. The study concludes by summarizing the findings and implications in the context of deep learning's dual-edged impact on society.

2 Systematic literature review

A systematic literature review (SLR) is essential for evaluating and interpreting all available research relevant to a specific research question or topic area. This review follows the methodology outlined by Brereton et al. [11], which divides the SLR process into three broad phases: Planning the Review, Conducting the Review, and Documenting the Review. These phases are further subdivided into ten steps as shown in Fig. 2 to ensure a thorough and unbiased review.

A systematic literature review is "a means of evaluating and interpreting all available research relevant to a particular research question or topic area or phenomenon of interest" as mentioned by Kitchenham [52]. The primary studies for a SLR are summarized in the review, while the secondary study is the review itself. The SLR highlights and accumulates the shreds of evidence to offer new insights or highlight the shortcomings of the primary studies. Based on the definition above, the evaluation and interpretation of available (primary) researches are required to perform and SLR. Hence, the SLR should present the evidence to recognize the outcomes of the studies in a consistent and fair manner. One of the essential elements in

Fig. 2 Process of SLR with 3 broad phases and 10 sub-phases

conducting the SLR is the establishment of a protocol for the study. The aim of this protocol is to minimize biasing among primary studies. This protocol is imperative and should be validated separately. Any changes in the protocol should be reflected against all the studies in the secondary review.

Phase 1: Plan review

This is the initial phase of SLR, where the planning is performed in terms of research questions and protocols.

A. Specify research questions

A critical step in any SLR is the formulation of clear and precise research questions. These questions form the core of the review and guide the search for primary studies. For this review, the research questions shown in Table 1 are designed to cover all relevant aspects of deepfake detection.

Table 1 Research questions

RQID	Supplementary research question	Objective
RQ 1	Overview of deepfakes	Describe deepfake and the technologies related to the generation of deepfakes.
SRQ 1.1	What are deepfakes?	Explain the concept of deepfake.
SRQ 1.2	What are Generative Adversarial Networks and how do they work?	Explain the concept and working of GANs.
SRQ 1.3	How are deepfakes trained to generate a new object?	Identify how the deepfakes are gnerated using the learning mechanism of GANs.
SRQ 1.4	What are the pros and cons of deepfakes and what are the popular implementations of deepfake?	Discover the apps where deepfakes are implemented
SRQ 1.5	How to detect deepfake?	Explain the process of detecting deepfake.
RQ 2	What are the major events highlighted due to deepfakes in past?	Find all the major events in the past which had significant impact on society.
RQ 3	How can empirical tests be used to detect deepfakes?	Determine the process required to detect a deepfake.
SRQ 3.1	What are the available datasets to train and test deepfake?	Find the datasets that can be used for the training and testing of deepfakes.
SRQ 3.2	What are the existing deepfake detection techniques?	Discover as many techniques that have been implemented in the past to detect deepfakes.
SRQ 3.3	How to measure the performance of a deepfake detection method?	Idenitfy the measurement metrics that are majorly used for evaluation of deepfake detection models.
RQ 4	What are the best methods for detecting deepfakes based on existing literature?	Identify the best methods out of the existing literature available.
SRQ 4.1	What features are mostly considered while detecting deepfakes?	Analyse the most important features used in deepfake detection.

B. Develop review protocol

The review protocol provides a detailed plan for the review, specifying the conditions for selecting primary studies and the qualitative metrics to be applied. The protocol for this review includes:

- Search Strategy: We used databases such as Google Scholar, IEEE Xplore, ScienceDirect, Web of Science, Cornell University, Arxiv, Wiley Inter Science Journal Finder, and ACM Digital Library. Search terms included combinations of 'Deepfake', 'video manipulation', 'video forgery', 'digital media forensics', 'facial manipulation', 'detection', and 'identification' using Boolean operators (AND, OR).
- Inclusion Criteria:
  - Studies published between January 2018 and August 2023.
  - Articles including empirical evidence on deepfake detection techniques.
  - Papers written in English.
- Exclusion Criteria:
  - Non-English studies.
  - Papers without relevant keywords in the title or abstract.
  - Presentations or summaries of workshops.

C. Validate review protocol

The review protocol was validated by the author and co-authors to minimize bias. Validation involved checking the protocol against established guidelines to ensure comprehensive coverage and fairness. Any changes made to the protocol during the review were documented and reflected in all stages of the review process.

Phase 2: Conduct review

This is the second phase of SLR, where the execution of the steps planned in the first phase is performed. For example, when defining the review protocol, we identified the search strategy and resources to be searched. We initiated the review process based on the validated protocols in this phase. The sub-phases included in this phase are represented in Fig. 3

D. Identify relevant research

Using the search strategy defined in the review protocol (part B of Section 2), we identified a total of 81 relevant research studies. The search terms were included in the title, abstract, or keywords of the articles. Both empirical studies and reviews were considered if they included relevant keywords and were written in English. Non-English studies, PowerPoint presentations, and workshop summaries were excluded.

E. Select primary studies

Primary studies were selected through a two-stage process:

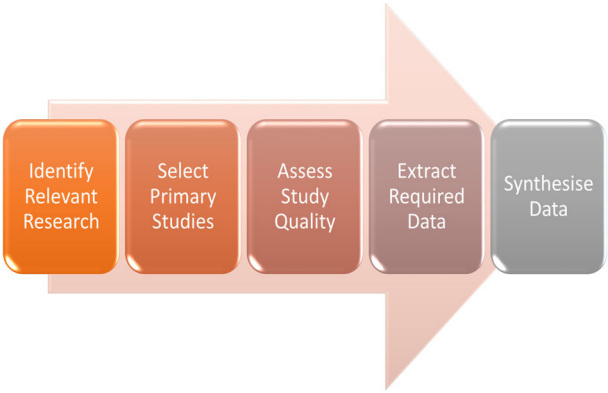


Fig. 3 Steps included in second phase of SLR - conducting a review

1. Initial Screening: Irrelevant papers were rejected based on the title and abstract. This step ensured that only potentially relevant studies were considered for further review.
2. Detailed Review: Full copies of the remaining papers were obtained and reviewed against the inclusion and exclusion criteria. This iterative process ensured a comprehensive selection of primary studies relevant to the research questions.

F. Assess study quality

The quality of the selected studies was assessed based on predefined criteria, such as methodological rigor, relevance to the research questions, and the robustness of the data presented. To aid this process, we created a word cloud (Fig. 4) of the study titles to visualize the focus of the selected studies and ensure alignment with the search strategy.

Fig. 4 Word cloud of all the titles included in this systematic literature review

G. Extract required data

Data extraction forms were used to systematically record information from the primary studies. This included:

- General Details:
  - Author(s)
  - Publication year
  - Venue
- Technical Details:
  - Detection techniques and algorithms used
  - Features considered
  - Datasets used
  - Evaluation metrics

This step involved both an extractor and a checker to ensure accuracy and relevance. The extractor gathered the data from the studies, and the checker verified its relevance to the research questions.

H. Synthesise data

The extracted data were synthesized to answer the research questions. Qualitative data were aggregated using tabular formats, facilitating a structured comparison of different deepfake detection techniques. This synthesis helped identify patterns, trends, and gaps in the existing research.

Phase 3: Document review

This is the final phase of SLR, where the review performed must be documented in such a way that it is a good read. The report should include precise answers to all the research questions and the potential questions of the reader or researcher. The size of the report, the formatting of the report, and the readability of the report are the main concerns that need to be addressed in this phase.

I. Write review report

The review report was written to provide clear and precise answers to the research questions. The report was structured according to the guidelines of the Multimedia Tools and Applications of Springer publications, ensuring clarity, coherence, and readability. Each section of the report was crafted to address specific research questions, present the findings of the SLR, and discuss their implications.

J. Validate report

The completed review report was validated by the authors and external experts to ensure it met the required standards for publication. This validation process included a thorough review of the report's content, structure, and adherence to the review protocol. Feedback from the validation process was incorporated to improve the final document.

3 Discussion

This section provides a detailed discussion based on the research questions identified in Section 2 A. The discussion integrates both quantitative and qualitative analyses to comprehensively address the state of deepfake detection research. By examining the existing literature, we highlight significant findings, identify gaps, and propose directions for future research. This section aims to enhance understanding of deepfake technologies, their implications, and the effectiveness of various detection methods.

RQ 1: Overview of deepfakes

Manipulated media is not a new phenomenon, but deepfakes represent a significant advancement in the creation of synthetic images and videos. Historically, the manipulation of photos and videos has been used to deceive or entertain audiences. With the advent of the internet, the proliferation and sophistication of media manipulation have increased dramatically. What began with simple photoshopped images has evolved into highly realistic deepfake videos, making it increasingly difficult to discern between real and fake content.

Deepfakes, a portmanteau of "deep learning" and "fake," involve the artificial production, modification, and manipulation of images, audio, and video using advanced automation techniques. This technology leverages deep learning algorithms, particularly Generative Adversarial Networks (GANs), to create highly convincing digital forgeries. Initially popularized through the swapping of celebrity faces onto pornographic video actors, deepfakes have now extended their reach to various sectors, including politics, entertainment, and social media.

The rise of deepfake technology poses significant risks to society, including the potential for spreading misinformation, compromising individual privacy, and undermining public trust in digital media. Recognizing these threats, there has been an urgent call for effective detection methods to identify and mitigate the impact of deepfakes. This research question aims to provide an overview of the development and use of deepfake technology, exploring its underlying mechanisms, applications, and the challenges it presents with the help of four Supplementary Research Questions (SRQ).

SRQ 1.1: What are deepfakes?

Deepfakes are a form of synthetic media created by leveraging deep learning, a subfield of machine learning that uses deep neural networks to produce realistic fake content. The term "deepfake" was coined in 2017 by a Reddit user who utilized face-swapping technology to post adult celebrity videos. The technology has since evolved, finding applications in various domains such as entertainment, politics, and social media.

At its core, a deepfake involves the artificial generation or alteration of images, audio, and video to create misleading content that appears authentic. The underlying technology relies heavily on Generative Adversarial Networks (GANs), which are pivotal in producing high-quality deepfakes. Originally developed by Goodfellow et al. [31], GANs have revolutionized the field of synthetic media by making it possible to generate content that is indistinguishable from real media.

Deepfakes have gained notoriety for their potential misuse, such as creating fake news, spreading misinformation, and violating individuals' privacy by producing explicit content without consent. These risks underscore the urgent need for effective detection methods to

preserve the integrity and trustworthiness of digital media. The following sections will delve deeper into the mechanisms behind deepfake technology and the methods used to detect such fabrications. Figure 5 represents a comparison of deepfake image with the real image.

SRQ 1.2: What are Generative Adversarial Networks and how do they work?

Generative Adversarial Networks (GANs) are a class of deep learning models designed for generative modeling, which is an unsupervised learning task aimed at discovering and learning the patterns in input data to generate new data samples that resemble the original dataset. Introduced by Ian Goodfellow and his colleagues in 2014, GANs have become a cornerstone of deepfake technology due to their ability to create highly realistic synthetic media.

A GAN consists of two neural networks: the generator and the discriminator. Figure 6 (Nishad [82]) represents the generator and discriminator sub-models in a GAN. These two networks are engaged in a constant adversarial process:

- **Generator:** The generator's role is to produce fake data samples that mimic the real data. It takes random noise as input and transforms it into plausible data instances.
- **Discriminator:** The discriminator's task is to evaluate the data it receives, distinguishing between real data (from the original dataset) and fake data (generated by the generator). It outputs a probability indicating the likelihood that a given sample is real or fake.

During training, the generator and discriminator play a minimax game. The generator aims to produce increasingly convincing fake samples to deceive the discriminator, while the discriminator strives to become better at detecting fake samples. This adversarial training continues until the generator produces data that is virtually indistinguishable from the real data. Figure 7 (Zucconi [115]) represents the autoencoder used to replicate the input image.

The power of GANs lies in this dynamic, adversarial training process, which enables the generator to learn and improve continually, resulting in highly realistic synthetic data. GANs

Fig. 5 Comparison of a deepfake image with the original image

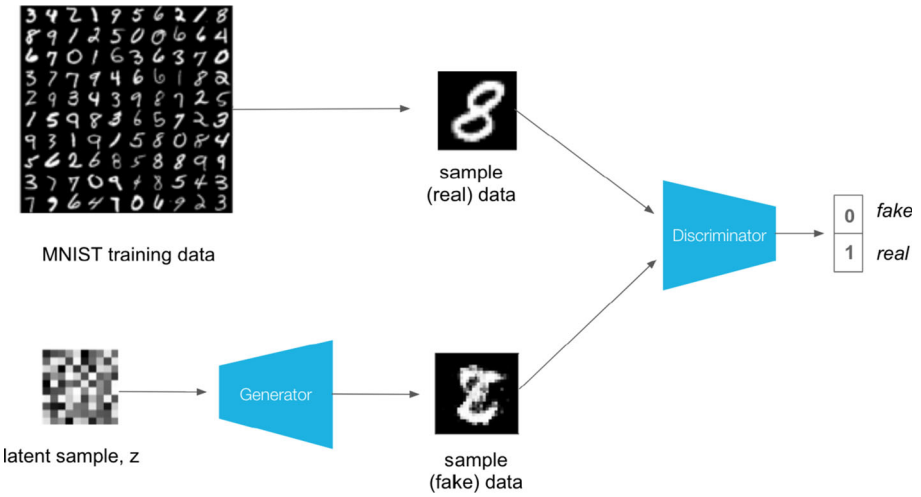


Fig. 6 Architecture of generative adversarial networks

have been instrumental in the development of deepfake technology, allowing for the creation of videos and images that are difficult to differentiate from authentic media.

Understanding how GANs operate is crucial for developing effective deepfake detection techniques, as it provides insight into the generative processes that need to be countered. The following sections will discuss various deepfake detection methods that leverage this understanding to identify and mitigate the impact of deepfake content.

SRQ 1.3: How are deepfakes trained to generate a new object?

Deepfakes primarily involve the manipulation and generation of realistic synthetic images or videos, often through face-swapping techniques. The process of creating a deepfake typically involves training deep learning models, such as autoencoders and GANs, to generate new, highly realistic objects or scenes.

Autoencoders in Deepfake Creation

Autoencoders play a crucial role in deepfake generation. An autoencoder consists of two main parts: an encoder that compresses the input into a latent space representation, and a decoder that reconstructs the input from this latent representation. When creating deepfakes, particularly for face-swapping, two autoencoders are often trained to encode and decode faces.

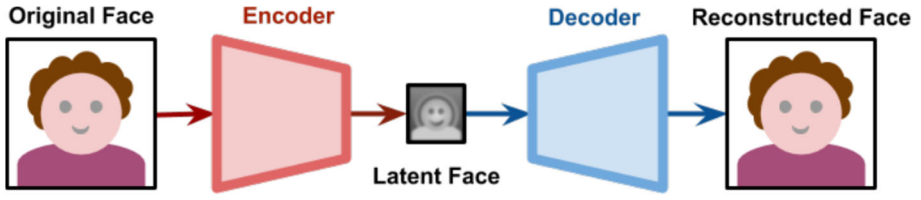


Fig. 7 An autoencoder with encoder and decoder replicating the source image

Training Process:

1. **Shared Encoder:** A shared encoder is used to create a common latent space for both source and target faces. This shared encoder learns to extract facial features that are common across different faces, ensuring compatibility in the latent space.
2. **Separate Decoders:** Two separate decoders are trained independently for the source and target faces. Each decoder learns to reconstruct faces from the shared latent space but retains the unique characteristics of its respective face. For instance, decoder A is trained with faces of subject A, and decoder B with faces of subject B. For example, in Fig. 8, the faces of only A object are used to train the decoder A; the faces of only B object are used to train the decoder B.
3. **Face Swapping:** During the face-swapping process, the face of subject A can be encoded into the latent space and then decoded using decoder B, resulting in the face of subject A appearing on the body of subject B. As seen in Fig. 9, the Decoder B will try to rejuvenate Subject B from the features relative to Subject A.

Generative Adversarial Networks (GANs) GANs further enhance the realism of deepfakes by refining the generated outputs. The generator creates fake images or videos, while the discriminator evaluates their authenticity. Through adversarial training, the generator improves its capability to produce highly convincing fakes that can deceive the discriminator. The combined use of autoencoders and GANs enables the creation of deepfakes that are not only visually convincing but also difficult to detect, highlighting the importance of developing robust detection methods.

SRQ 1.4: What are the pros and cons of deepfakes and what are the popular implementations of deepfake?

Deepfake technology offers a range of applications with both positive and negative implications. Understanding these pros and cons is essential for leveraging the technology's benefits while mitigating its risks.

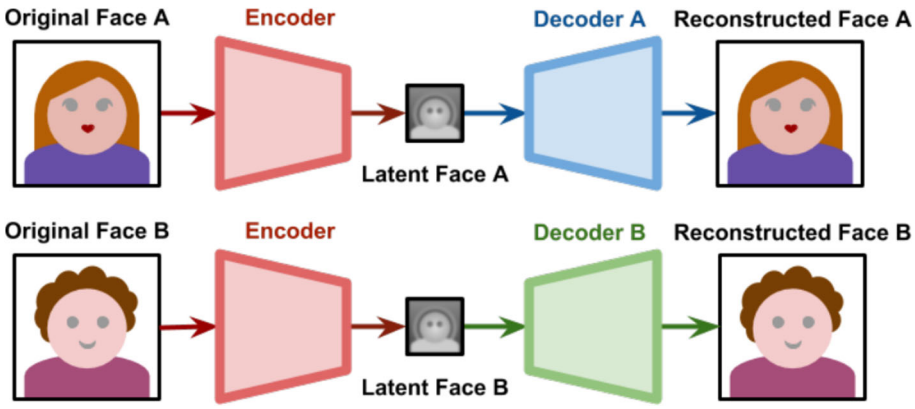


Fig. 8 Same encoder with different decoders to train deepfake

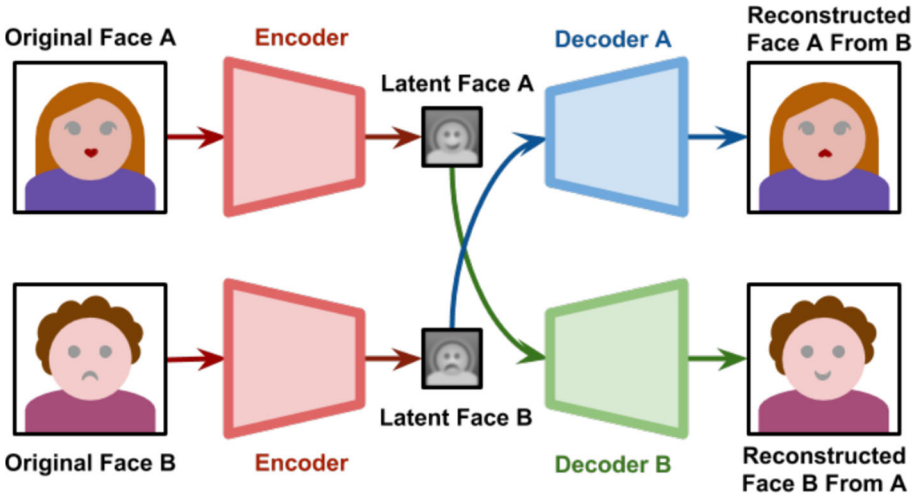


Fig. 9 Reversal of latent space to generate deepfake

Pros of deepfake technology

1. **Entertainment and Media:**
   - **Visual Effects and Film Industry:** Deepfakes can create realistic special effects, allowing filmmakers to generate scenes that would be impossible or too expensive to produce traditionally. They can also resurrect deceased actors for sequels, as seen with Paul Walker in "Fast and Furious 7."
   - **Enhanced Gaming Experience:** Deepfakes can significantly improve the realism in video games, providing more immersive and interactive experiences [101].
   - **Saving Finances:** With the potential to create a video for everything, deepfake technology can potentially save costs in advertisements, games, and movies without actually filming the scenes. This can be cheaper as compared to conventional methods. One instance of the same is visible in the de-aging of actor Harrison Ford in the new Indiana Jones movie [76].
2. **Education and Training:**
   - **Historical Reconstructions:** Deepfakes can bring historical figures to life, providing engaging educational content and immersive learning experiences. For example, students could watch realistic reenactments of historical events or lectures from virtual representations of famous scientists [16].
   - **Simulation Training:** Medical and military training can benefit from deepfakes by creating realistic scenarios for simulation-based learning.
3. **Personalization and Assistive Technologies:**
   - **Voice Assistance:** Deepfake technology can personalize voice assistants (e.g., Apple's Siri, Google Assistant) by mimicking voices of historical figures or celebrities, enhancing user engagement [12]. For example, "LumiereNet" by Kim and Ganapathi [51] is a system to enhance the approach to academic content generation

on learning platforms. Just imagine studying physics from the voice and video lecture from Albert Einstein.
- **Rehabilitation and Therapy:** Virtual avatars created using deepfakes can assist in therapeutic settings, helping individuals with speech disorders or providing companionship to those in need.

Cons of deepfakes technology

1. **Misinformation and Fake News:**
   - **Political Manipulation:** Deepfakes can be used to create fake videos of politicians making controversial statements, potentially influencing public opinion and election outcomes. A notable example is the manipulated video of U.S. House Speaker Nancy Pelosi appearing to slur her speech [12].
   - **Media Integrity:** The spread of fake news through deepfakes undermines trust in media and digital content, making it difficult for the public to discern the truth.
2. **Privacy Violations:**
   - **Non-consensual Explicit Content:** One of the earliest and most notorious uses of deepfakes has been the creation of explicit videos featuring celebrities without their consent, violating their privacy and damaging their reputations.
   - **Identity Theft:** Deepfakes can be used to impersonate individuals, leading to identity theft and fraud.
3. **Legal and Ethical Concerns:**
   - **Evidence Tampering:** Deepfakes pose a significant risk to the judicial system by enabling the creation of fake evidence, potentially affecting court decisions and undermining justice [84].
   - **Ethical Dilemmas:** The ability to create highly realistic fake content raises ethical questions about consent, authenticity, and the potential harm caused by such fabrications.

3.0.1 Popular implementations of deepfake technology

Table 2 presents a summary of popular deepfake tools and their typical features.

SRQ 1.5: How to detect deepfake?

From the previous section, we know that learning of a GAN is of utmost importance in the generation of deepfake. Hence, the larger the input data leveraged to train the model, the better the generated deepfake. As a matter of fact, there is sufficient data available over the web for celebrities and renowned personalities, which makes it possible to create believable deepfakes and hence spread fake news and rumors over social networking sites and adult sites that can have an underlying and severe impact on their lives and our society. Therefore, the detection of deepfake images and videos has become imperative and increasingly critical. Considering this encouragement and need for deepfake detection, many organizations such as Facebook, Google, United States Defense Advanced Research Projects Agency (DARPA) has launched various initiative towards attempting the prevention and detection of deepfake.

Table 2 Popular implementations of deepfake technology along with their features

Tool	Link	Features	References
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN	An auto-encoder architecture is expanded with a perceptual loss (VGGface) and adversarial loss.	
DeepFaceLab	https://github.com/iperov/DeepFaceLab	- Adds new models to the Faceswap technique, such as the H64, H128, LIAEF128, and SAE. - Supports several face extraction techniques, including manual, S3FD, MTCNN, and dlib.	Perov et al. [83]
DFaker	https://github.com/dfaker/df	Face reconstruction is done using the DSSIM loss function, which was implemented using the Keras library.	
DeepFake tf	https://github.com/StromWine/DeepFake_tf	Tensorflow was used to develop something comparable to Dfaker.	
AvatarMe	https://github.com/lattas/AvatarMe	- From arbitrary "in-the-wild" images, generates 3D faces. - Is capable of reconstructing real 3D faces with a 4K by 6K resolution from a single low-quality image.	Lattas et al. [58]
DiscoFaceGAN	https://github.com/microsoft/DiscoFaceGAN	– Produces virtual person face images using independent latent variables for identification, emotion, position, and illumination. - Adversarial learning incorporated with 3D priors.	Deng et al. [23]
StyleRig	https://gvv.mpi-inf.mpg.de/projects/StyleRig	- Self-supervised without manual annotations. - Uses a fixed, pre-trained Style-GAN to manage rig-like portrait pictures of faces.	Tewari et al. [94]
FaceShifter	https://lingzhili.com/FaceShifterPage	Face swapping in high-fidelity without requiring subject-specific training by utilizing and combining the target properties.	Li et al. [59]
FSGAN	https://github.com/YuvalNirkin/fsgan	- A face swapping and reenactment model that doesn't require training on the target faces and may be applied to pairs of faces. - Responds to changes in both position and expression.	Nirkin et al. [81]

Table 2 continued

Tool	Link	Features	References
StyleGAN	https://github.com/NVlabs/stylegan	- On the basis of research in the field of style transfer, a novel generator architecture for GANs is suggested. - The new architecture offers automatic, unsupervised segmentation of high-level characteristics along with easy, scale-specific management of image synthesis.	Karras et al. [45]
Face2Face	https://justusthies.github.io/posts/face2face/	- Re-renders the changed resulting video in a photo-realistic mode after animating the target video's facial expressions using a source actor. - Facial replication of a monocular targeted video sequence in real-time, such as one from a YouTube video.	Thies et al. [95]
Neural Textures	https://github.com/SSRSGJYD/NeuralTexture	- Feature maps learned throughout the scene capture process and recorded as maps on top of 3D mesh proxies, allowing coherent real-time re-rendering or editing of ongoing video content for both static and dynamic circumstances.	Thies et al. [96]
DeepNude	https://app.deepnude.cc/upload	- Based on image-to-image translation technique - After entering an image, it will instantly produce the bare version of the image	dee [1]

Many approaches involving deep learning techniques, such as recurrent neural networks (RNN), convolutional neural networks (CNN), region-based convolutional neural networks (RCNN), and even hybrid approaches, have been presented to detect deepfake images and videos.

The primary artifacts to look out for in a video or image can be:

- Paying attention to the face because manipulated videos are mostly facial transformations
- The concurrency between the age and the skin or hair
- Attention to the color and size of lips and if they synchronize with the rest of the face
- Unexpected glare on glasses and difference in natural lightning
- How real are the facial hair, addition or removal of mustache or beard
- The shadows look natural or not
- Uneven blinking in eyes (blinking too much or not blinking enough)
- The synchronization in the lips movement and the speech

RQ 2: What are the major events highlighted due to deepfakes in past?

Deepfake technology has been at the center of several significant events, creating both panic and intrigue as its applications have grown more sophisticated. It has emerged as a means to create chaos and prank unaware individuals with strange manipulations. Since its beginning, it has been used and misused and some of the deepfakes made significant headlines. Figure 10 represents some of the major events in the past (2017-2021) that included the use of deepfake technology.

RQ 3: How can empirical tests be used to detect deepfakes?

Empirical tests play a crucial role in evaluating the effectiveness of various deepfake detection methodologies. This section presents a comprehensive analysis of how empirical testing can be leveraged to assess and improve the performance of deepfake detection techniques, ranging from traditional computer vision methods to advanced deep learning approaches.

Designing Empirical Tests for Deepfake Detection:

1. **Dataset Selection:**
   - **Diverse and Representative Datasets:** To ensure the robustness of detection methods, empirical tests must be conducted on diverse datasets that include a wide range of deepfake types, qualities, and sources. Popular datasets such as FaceForensics++, DFDC (Deepfake Detection Challenge), and Celeb-DF provide a comprehensive set of real and fake media for evaluation.
2. **Evaluation Metrics:**
   - **Accuracy, Precision, Recall, and F1-Score:** These metrics are essential for measuring the performance of detection algorithms. Accuracy indicates the overall

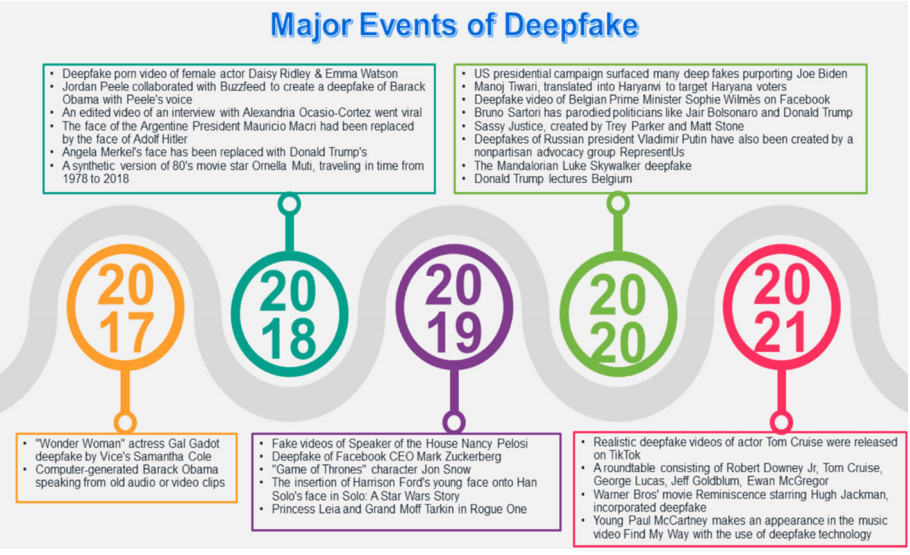


Fig. 10 The timeline of major events caused due to deepfakes from 2017 to 2021

correctness, precision measures the rate of true positives among detected positives, recall assesses the ability to detect all actual positives, and the F1-score provides a harmonic mean of precision and recall.

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** This metric evaluates the trade-off between true positive rates and false positive rates, providing a comprehensive measure of detection performance.

3. **Baseline Comparisons:**

   - **Traditional vs. Deep Learning Methods:** Empirical tests should include baseline comparisons between traditional computer vision methods (e.g., SVM, HOG, LBP) and deep learning techniques (e.g., CNNs, RNNs, GANs). This comparison helps in understanding the advancements and limitations of modern approaches.

4. **Cross-Validation:**

   - **K-Fold Cross-Validation:** Implementing cross-validation techniques ensures that the detection methods generalize well to unseen data. K-fold cross-validation, in particular, splits the dataset into k subsets, training on k-1 subsets and testing on the remaining one. This process is repeated k times, providing a robust assessment of model performance.

Empirical Testing of Deep Learning Techniques:

1. **Convolutional Neural Networks (CNNs):**

   - **Feature Extraction and Classification:** Empirical tests involving CNNs focus on their ability to extract fine-grained features from images and classify them as real or fake. Performance is evaluated based on their accuracy in detecting subtle artifacts unique to deepfakes.

2. **Recurrent Neural Networks (RNNs):**

   - **Temporal Consistency Analysis:** RNNs, especially LSTM networks, are tested on their capability to analyze temporal consistency in videos. Empirical tests measure their effectiveness in identifying frame-by-frame inconsistencies and unnatural movements indicative of deepfakes.

3. **Autoencoders and GANs:**

   - **Reconstruction Error and Adversarial Training:** Autoencoders are tested for their reconstruction error rates, which can indicate anomalies in deepfake images. GANs, particularly their discriminator models, are evaluated based on their adversarial training performance, distinguishing between real and fake media.

Hybrid and Ensemble Methods:

- **Combining Models for Enhanced Detection:** Empirical tests often involve hybrid approaches that combine multiple models to leverage their strengths. For instance, a combination of CNNs for feature extraction and RNNs for temporal analysis can improve detection accuracy. Ensemble methods, which aggregate predictions from several models, are tested for their ability to enhance robustness and reduce false positives.

Empirical testing is indispensable for the development and refinement of deepfake detection methods. By systematically evaluating various techniques on diverse datasets and using robust evaluation metrics, researchers can identify the strengths and weaknesses of different approaches. This iterative process of testing and improvement is vital for advancing the state of deepfake detection technology and ensuring the integrity of digital media in the face of evolving threats.

The following sections, with SRQ 3.1 to SRQ 3.3, will provide the details about the popular datasets to be used in the training and testing of deepfake, various deepfake detection techniques, and metrics to measure the performance of deepfake detection models.

SRQ 3.1: What are the available datasets to train and test deepfakes?

In the process of detecting deepfake, the first step is data collection. Data collection can be primary as well as secondary. In the case of primary data collection, one can generate a dataset of original and manipulated videos. However, in secondary data collection, we can use the already existing datasets for deepfake. These datasets include a combination of original and manipulated videos, which can be used to train and test the detection model. Some of the datasets which exist and are usually used in the research are the following:

- **FaceForensics++** [86]: This dataset consists of 1,000 real videos with manipulations done using Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The source of these videos were 977 YouTube videos with trackable frontal face to generate realistic forgeries. It also includes 1,000 deepfake models for the augmentation of new data.
- **Deepfake Detection Challenge (DFDC)** [24]: This is one of the largest publically available datasets consisting of over 100,000 clips from 3,426 paid actors produced with multiple deepfake, GAN-based, and non-learned models. It comprises both fake and real videos for training and testing purposes.
- **FaceForensics** [85]: This dataset consists of more than 500,000 frames with faces captured from 1,004 videos to study forgeries in images and videos. These videos have been created using the Face2Face approach. The videos have been taken from YouTube and trimmed to short clips containing frontal faces.
- **WildDeepfake** [113]: This dataset consists of 7,314 face sequences taken from 707 deepfake videos from the web. This dataset can be used for testing the model as it includes videos that are difficult to be detected as deepfakes.
- **DeeperForensics-1.0** [42]: This large-scale dataset includes 48,475 real videos and 11,000 synthesized videos. The high-quality videos were sourced from 100 paid actors from 26 countries with their consent. The model used for video manipulation is the many-to-many end-to-end face-swapping method, DF-VAE. The total size of this dataset combines 284 GBs.
- **Korean DeepFake Detection Dataset (KoDF)** [57]: This is a large-scale dataset that constitutes real and synthesized videos on Korean subjects.
- **DFDM** [41]: This dataset includes 6,450 deepfake videos generated by GAN models. The five autoencoder models used to create these deepfakes were used with variations in the encoder, decoder, intermediate layer, and input resolution.
- **DeepFake MNIST+** [39]: This dataset contains 10,000 real videos from VoxCeleb1 dataset and 10,000 animation videos generated from real videos.
- **ForgeryNet** [36]: This is one of the largest publicly available datasets for deep face forgery detection, with 2.9 million images and 221,247 videos. This dataset can be used for multiple purposes, and deepfake detection is one of them.

- **Celeb DF v1.0** [64]: This dataset contains 408 real videos from YouTube covering different ethnic groups, genders, and ages, along with 795 synthesized videos from the actual videos.
- **Celeb DF v2.0** [64]: This dataset includes 590 real videos from YouTube covering different ethnic groups, genders, and ages, along with 5,639 synthesized videos from the real videos.
- **Deepfake detection dataset** [26]: This dataset was developed in collaboration with Google & JigSaw, and it contains over 3,000 synthesized videos of 28 actors in different scenes.
- **Deepfake TIMIT** [55]: This database includes pairs of 16 similar people available from the public VidTIMIT database contributing to 32 subjects which were trained on a lower quality model (64 x 64 input/output size) and higher quality model (128 x 128 input/output size). A total of 620 videos with swapped faces are constructed in this database without manipulating the audio track. The faces were swapped using an open-source GAN-based approach.
- **Vid TIMIT** [88]: This dataset comprises videos with audio recordings of 43 people saying short sentences. The videos are stored as a sequence of JPEG images (with a resolution of 512 x 384), and the audio is mono, 16-bit, 32 kHz WAV files.
- **Deep Fakes Dataset** [17]: This dataset includes 142 videos for 32 minutes with a size of approximately 17 GBs. The data was gathered from several sources like forums, news articles, research presentations, and applications. This dataset is also publicly released and available for educational use.
- **WaveFake** [28]: This dataset can be used for audio deepfake detection. It consists of a large-scale dataset of over 100,000 generated audio clips.
- **iFakeFaceDB** [79]: This dataset comprises approximately 87,000 Style-GAN Model generated synthetic face images transformed with the GANprintR method and are used in the study of artificial face manipulation detection. The standard size of all images in this dataset is 224 x 224.
- **CelebA-Spoof** [110]: This is a large-scale face anti-spoofing dataset with 625,537 images with 10,177 subjects, including 43 rich attributes on the environment, face, spoof types, and illumination. Amongst 43 rich attributes, 40 attributes include facial components and features like eyes, nose, skin, lip, hat, hair, eyebrows, and eyeglasses, and the other 3 attributes include environments, illumination, and spoof types. This dataset can be used for training and evaluation of face anti-spoofing algorithms.
- **CASIA-SURF HiFiMask** [65]: This is yet another large-scale High-Fidelity Mask dataset containing 54,600 videos from 75 subjects with 225 realistic masks by 7 kinds of sensors.
- **VideoForensicsHQ** [27]: This dataset contains forged videos of unmatched quality, including the synthesis of videos which is challenging to identify by the human eye. It introduces a new family of detectors to examine spatial and temporal features.
- **FVI (Free-form Video Inpainting)** [14]: This dataset can be used for training as well as testing video inpainting models. It includes 12,600 videos from YouTube-BoundingBoxes and 1,940 videos from the YouTube VOS dataset.
- **UADFV dataset** [63]: This dataset includes 50 videos for one individual, each with a length of approximately 30 seconds and at least one blink of an eye.
- **FFHQ (Flickr-Faces-HQ)** [45]: This is a high-quality image dataset developed as a convention for GAN. It contains 70,000 PNG images of high-quality (1024 x 1024 resolution) with variations in terms of the image background, ethnicity, and age.

- **VoxCeleb1** [77]: This dataset contains 100,000 utterances from 1,251 celebrities which were extracted from YouTube videos.

The distribution in the Table 3 is done based on modalities so as to assist researchers in identifying the dataset required for their needs. If a dataset includes Real & Fake objects, then it can be utilized for both training and testing purposes. The modality of research can be matched with the research purposes. For example, the research on deepfake detection on images can utilize the datasets with Image modalities and the research on video deepfake detection can utilize the datasets with Video modalities.

These datasets are fundamental for advancing deepfake detection research. They provide the necessary diversity and complexity to train robust models capable of identifying deepfakes across different contexts and manipulation methods. By utilizing these datasets, researchers can benchmark their algorithms, identify strengths and weaknesses, and drive improvements in detection technology.

Table 3 List of deepfake detection datasets used in various studies

Dataset name	Papers	Year	Modalities	Includes
iFakeFaceDB	2	2019	Image	Real
FFHQ (Flickr-Faces-HQ)	589	2019	Image	Real
CelebA-Spoof	11	2020	Image	Real
VoxCeleb1	NA	2017	Video	Real
FaceForensics	57	2018	Video	Real
Deepfake TIMIT	55	2018	Video	Fake
UADFV dataset	9	2018	Video	Real
DFDC	69	2019	Video	Real & Fake
Deepfake detection	NA	2019	Video	Fake
FVI (Free-form Video Inpainting)	6	2019	Video	Real
Celeb DF v1.0	NA	2020	Video	Real & Fake
Celeb DF v2.0	76	2020	Video	Real & Fake
Deep Fakes Dataset	3	2020	Video	Real
KoDF	6	2021	Video	Real & Fake
DeepFake MNIST+	1	2021	Video	Real & Fake
VideoForensicsHQ	2	2021	Video	Fake
DFDM	1	2022	Video	Fake
WaveFake	4	2021	Audio	Real
Vid TIMIT	NA	2002	Audio, Video	Real
FaceForensics++	194	2018	Image, Video	Real & Fake
DeeperForensics-1.0	15	2020	Image, Video	Real & Fake
WildDeepfake	17	2021	Image, Video	Fake
ForgeryNet	15	2021	Image, Video	Fake
HiFiMask	5	2021	Image, Video	Real

Note: The distribution of table is based on modalities and sorted on year of publishing. Every section divided by horizontal rule contains the datasets for same modalities sorted by publishing year in increasing order

SRQ 3.2: What are the existing deepfake detection techniques?

Deepfake detection can be achieved by various different techniques. The ongoing researches are exploring the latest technologies to detect deepfakes. Based on the existing literature, we have categorized the techniques of deepfake detection as Deep Learning based techniques, Machine Learning based techniques, and Statistical techniques. The existing literature in each category has been presented in a chronological order.

1. Deep Learning based techniques

This is the most implemented technique for deepfake detection. It works well for both deepfake image and deepfake video detection. Tao et al. [93] has focused on the fact that motion compensation needs to be done along with a proper frame alignment to achieve better results in detecting deepfakes. An SPMC layer for sub-pixel motion compensation is introduced using the CNN framework. FlowNet-S CNN is used with the motion transformer module (MCT) to accomplish the frame alignment and motion compensation. The authors provided a way to organize multiple frames input to yield better results. Rezende et al. [21] used a pre-trained ResNet-50 algorithm to detect computer graphic-generated images. The topmost layer was changed to a fully-connected layer, and the classification was performed using support vector machines. The softmax layer reported an accuracy of 92.28%; however, the authors of this paper reported an accuracy of 94.05%. Şengür et al. [89] used VGG16 and AlexNet to extract facial features. This paper also used the concept of transfer learning by importing the trained weights and replacing the dense layer with SVM for real or fake classification. The prediction effectiveness was improved by combining features from both nets and providing additional information simultaneously. They achieved an accuracy of 94.01% on the CASIA dataset.

Mo et al. [75] detected deepfakes using a simple CNN model with three groups of max-pooling and convolutional layers. The real or fake classification was performed on images using spatial high-pass filters to highlight delicate details and amplify the noise. The authors successfully achieved an accuracy of 99.4% on real images of the CELEBA-HQ dataset, which was extended using the GAN-based approach [44] to generate synthetic faces. Khodabakhsh et al. [50] also used deep learning techniques to cope with fraudulent images from YouTube videos. The idea behind this paper was to generalize the proposed model for non-public datasets. The study was performed on 53,000 images from 150 YouTube videos with manipulated faces generated using deepfake methods like FakeApp, face replacement, and CGI (Computer-Generated Imagery). The paper faced difficulty predicting the new artifacts despite a high accuracy from ImageNet test images. It also implemented several popular CNN architectures like AlexNet, ResNet, Inception, Xception, and VGG19.

Li and Lyu [62] developed an approach to detect deepfake with the help of face-warping artifacts. Since it uses these artifacts as the identifying characteristic to distinguish between real and fake photos, this method's key advantage is that it does not use deepfake-generated images as negative training samples. The authors tested the approach on UADFV and Deepfake-TIMIT. Afchar et al. [2] proposed an effective network to identify deepfake and videos tampered with from Face2Face. The proposed network focuses on the mesoscopic properties of images, and it is composed of a few layers. The experiments were performed on a private dataset with an accuracy of 98%. Marra et al. [70] implemented image-to-image translation with comparison on a set of deep learning algorithms on pairs of original and the corresponding fake images generated using GAN architecture. The primary purpose of this study was to assess image-to-image translation recognition using steganalysis features.

It showed good accuracy of 89.55% by XceptionNet in identifying fake images considering the compressed scenario.

Güera and Delp [33] proposed a system that was aware of temporal features to detect deepfake videos using CNN for extracting frame-level features. Further classification was performed using a feed-forward recurrent neural network (RNN). The maximum accuracy acquired in this approach was 97.1% on high-resolution images. This paper illustrates the generation of deepfake videos, which can also be used to understand the concept. In the same way, Li et al. [63] described a method to detect the blinking of eyes in videos which was usually ignored while creating deepfake videos earlier. They used a combination of CNN and LRCN to differentiate between the states of eyes, i.e., open state or closed state.

Luo et al. [66] has worked on the problem faced by CNN algorithms with the fixed-size conversion of images given as input. A new spatial pyramid pooling layer was introduced between the convolution layer and the fully connected layer restraining the information loss. Two prediction frames with 11x11 grid images modified from CelebA, FDDB, and AFW datasets were used with respective frame coordinates, probability, and confidence. The authors explained how the accuracy of the CNN algorithm is dropped due to certain factors. Amerini et al. [6] detected deepfake videos using optical flow vectors, which were calculated using methods based on CNN with two consecutive frames. The manipulated videos exhibit disorders in such vectors, which made it possible to detect deepfake. These vectors were converted to a 3-channel color image to extract spatial features and classify the video as fake or real using VGG16 and ResNet-50 models. VGG16 resulted in a detection accuracy of 81.61% on Face2Face videos, whereas, ResNet50 detected videos with an accuracy of 75.46%. This is a unique paper because of the consideration of inter-frame dissimilarities; however, the usual papers relied on intra-frame inconsistencies.

Hsu et al. [38] used CNNs based on contrastive loss functions to detect deepfake. The authors suggested using a fully connected layer connected to the feature extraction network to combine features collected from real and fake images for the following prediction. When comparing the feature representations of the input pair of photos, the contrastive loss was used to assess whether the images were comparable. The contrastive loss may provide essential assistance in learning features linked to any image alteration in the context of fake image identification by contrasting the actual properties of natural images. After being trained, the suggested deep learning model can deal with the counterfeit spots in the feature representation of the images, resulting in impressive performance even in fake images created by five GANs architectures. According to the authors, the average precision and recall were 88% and 87%, respectively. Furthermore, Korshunov and Marcel [56] showed that sophisticated facial recognition systems based on deep learning are susceptible to deepfake films. The VidTIMIT dataset, which the authors provided, was used to test several baseline approaches. The authors found that methods based on visual quality criteria, which are frequently employed in presentation attack detection, perform the best. They also demonstrate how difficult it is for conventional and face recognition systems to recognize deep fake videos produced by GANs. Additionally, they claim that the situation worsens when new deepfake technologies arise.

The research conducted by Malolan et al. [68] focuses on developing explicable models to identify deepfake videos. The work combines two explainable AI techniques to comprehend the image's protruding regions, including Layer-Wise Relevance Propagation (LRP) and Local Interpretable Model-Agnostic Explanations. It trains a CNN on a face database (LIME). The authors further demonstrated the model's rotational invariance and robustness in detecting deepfake images by the authors' presentation of a set of understandable results for the model's predictions for image slices, heat maps, and input perturbation. Frank et al. [29]

proposed a study that examines the frequency domain of GAN-generated images. The findings of the experiments revealed severe artifacts brought on by the upsampling processes present in the current GAN architectures, pointing to a structural and underlying issue with the GAN's image production process.

Similarly, Ciftci et al. [18] distinguished deep forgeries and authentic flicks and identified the distinct generator model behind deepfakes. The research contends that the generator's residuals hold information that can be used to distinguish between manufactured artifacts and biological signals. The work uses 32 unprocessed photoplethysmogram (PPG) data from various facial locations stored in a Spatio-temporal block, or PPG cell, along with their spectral density. The contradictions that deepfake creation introduces into videos and provides a Spatio-temporal hybrid model of capsule networks combined with long short-term memory networks was discussed by Mehra [73]. The authors started with a Capsule Network to identify spatial discrepancies in a single frame. Then LSTM is integrated with the Capsule Network to identify Spatio-temporal discrepancies across many frames with an accuracy of 83.42%. As a baseline for uni-frame detection and in combination with LSTM for multi-frame detection, the state-of-the-art deepfake detection model XceptionNet is used. Finally, the model is evaluated on two more data sets using various deepfake creation techniques, augmentations, and facial filters like the flower crown and dog filters. To address the zero and few-shot transfer issue, Aneja and Nießner [7] presented a transfer learning-based strategy dubbed Deep Distribution Transfer (DDT). This method's key concept was a novel distribution-based loss formulation that may effectively bridge the gap between the domains of various facial forgery techniques or cryptic datasets. The suggested approach significantly surpasses the baselines in both zero-shot and few-shot learning. A ResNet-18 neural network model developed for the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC) was used for this activity. For zero shots and very few shot cases, greater detection efficiencies of 4.88% and 8.38%, respectively, are attained. The proposed method uses zero and a few shot transfers to generalize forgery detection tools.

To recognize videos with indications of face alteration, Wodajo and Atnafu [103] merged a CNN model with a Vision Transformer (ViT) architecture. The authors employ the VGG-16 CNN model to extract features from the video frames, and then they apply the ViT model to these feature maps to determine if the video is authentic or counterfeit. Two components comprise a convolutional vision transformer: a CNN and a ViT. ViT classifies the gathered data using the attention technique, whereas CNN pulls the learnable attributes. The model's accuracy was 91.5%, its loss value was 0, and its AUC was 0.91 after training on the DFDC dataset. The capability of this technique can swiftly ascertain whether or not the images are authentic and detect deepfake.

More recently, Arshed et al. [8] also explored the application of Vision Transformer (ViT) in deepfake detection. Their approach, focusing on extracting global features, has yielded remarkable results in terms of detection accuracy, precision, recall, and F1 rate, especially in datasets with real-world applications like Snapchat-filtered images. This study underscores the potential of transformer networks in the realm of deepfake detection.

In the study of Jung et al. [43], GAN-generated deepfakes were detected using a novel method to examine a major change in the blinking pattern, which is a voluntary and spontaneous activity that does not require conscious effort. The authors identified that if eye blinks are continually repeated in a concise amount of time, the suggested approach called DeepVision was used as a measure to validate an anomaly based on the expired eye blink time, repeated number, and period. In seven out of eight different videos, DeepVision correctly identified Deepfakes (87.5% accuracy rate), indicating the constraints of integrity verification algorithms that are just based on pixels.

Agarwal et al. [4] recently developed a neural model for identifying fraudulent images created using GANs, and it relies on both the frequency spectrum of the images and Capsule Networks. To perform the task of deepfake detection, Wang and Dantcheva [99] offered a comparison of three different 3D-CNN models, namely I3D, ResNet 3D, and ResNeXt 3D. Four video modification approaches were taken into account by the authors, who provided consistent outcomes for various training and assessment settings.

To identify deepfake videos, Kolagati et al. [53] built a deep hybrid neural network model. Facial landmarks detection is used to extract data from the videos on a variety of facial traits. This data is put together into a multilayer perceptron in order to comprehend the distinction between actual and fake videos (i.e., MLP). The hybrid system's high speed and minimal processing resource requirements make it perfect for deepfake video screening. Given that most algorithms are getting incredibly accurate in creating genuine human faces, Coccomini et al. [19] concentrated on video deepfake detection on faces in their work. The authors mixed different kinds of Vision Transformers with an EfficientNet B0 convolutional feature extractor to achieve results equivalent to some relatively recent approaches, including Vision Transformers. They didn't employ either distillation or ensemble methods, in contrast to cutting-edge methods. Additionally, they also provided a simple voting-based inference approach for managing many faces in a single video frame, with their top model attaining an AUC of 0.951 and an F1 score of 88% on the DFDC dataset.

On the other hand, Suganthi et al. [92] implemented a deep learning-based approach of fisher face utilizing Local Binary Pattern Histogram (FF-LBPH) to detect deepfake face images. By employing LBPH to reduce the dimension in the face space, the Fisherface method is used to recognize faces. Further, DBN and RBM were used to create a classifier for deepfake detection. This research was performed on public data sets FFHQ, 100K-Faces DFFD, and CASIA-WebFace. A simple combination of ResNext (a CNN algorithm) and LSTM was also used by Vamsi et al. [97] to detect deepfake videos on the Celeb-Df dataset with an accuracy of 91%. Similarly, Deng et al. [22] added a variation of EfficientNet-V2 to determine the veracity of images and videos but post the discovery of authentic images and videos, a spectacular visualization on strengthening the detection system's precision in differentiating between real and synthetic faces was also presented. Given the research evidences that the voices and faces in deepfake videos typically have the wrong people behind them, a voice-face similarity detection model is created to gauge how closely these two match on a general audio-visual dataset. Cheng et al. [15] recommended executing the deepfake detection from this uncharted voice-face matching angle. However, the current defense against deepfake is mainly focused on identifying authenticity rather than attribution.

Zhao et al. [111] introduced the ISTVT model, which uniquely combines spatial-temporal self-attention with a self-subtract mechanism. This approach not only captures spatial artifacts and temporal inconsistencies effectively but also enhances model interpretability through relevance propagation algorithms. Their extensive experiments across multiple datasets like FaceForensics++ and Celeb-DF demonstrate its robustness in both intra and cross-dataset scenarios. Similarly, Ke and Wang [47] proposes a novel method to counter degraded deepfakes. By focusing on feature restoration in the feature space rather than the image domain, DF-UDetector uses an image feature extractor, a feature transforming module, and a discriminator to uplift feature quality. This method has shown promising results, particularly in detecting deepfakes in uncontrolled, real-world conditions. This is another step forward towards addressing the generalizability concern in deepfake video detection.

Yu et al. [108] contributed to the field with "PVASS-MDD: Predictive Visual-audio Alignment Self-supervision for Multimodal Deepfake Detection." This study tackles the challenge of audio-visual inconsistencies in deepfake videos. By employing a three-stream network and

a cross-modal predictive align module, PVASS-MDD efficiently bridges the gap between modalities, leading to a significant performance improvement in capturing subtle inconsistencies across various multimodal deepfake datasets. However, Yin et al. [107] address the shortcomings of frame-based detection methods by focusing on long and short-range inter-frame motions. Their dynamic difference learning method, incorporating a fine-grained difference capture module and a multi-scale spatio-temporal aggregation module, differentiates between manipulations and natural facial motions, thereby enhancing the accuracy of spatio-temporal inconsistency modeling. This approach has demonstrated superiority over existing methods in several benchmark datasets.

2. Machine Learning based techniques

Traditional machine learning (ML) algorithms play a crucial role in understanding the reasoning behind any choice that can be described in human words. These techniques work well in deepfake since the processes and data are well understood. Additionally, it is considerably easier to adjust hyper-parameters and modify model designs. The decision process is displayed as a tree in tree-based ML techniques, such as Decision Trees, Random Forest, Extremely Randomized Trees, and many others. Consequently, a tree-based approach has no explainability problems, but once there are a thousand trees it gets complicated. For example, SHAP values can be used as a way to understand ensemble tree models and it's noticeable that these are "generally explainable to some extent" but still there are limitations as proved by Marcílio and Eler [69].

The model that McCloskey and Albright [72] presented leverages saturation cues to identify deep fakes. Images can be differentiated as being created by a camera or a GAN using saturation cues. The power of this technique can expose two sorts of GAN-generated imagery. HDR camera photographs typically have areas that are underexposed in saturation. The assumption made for this approach is that the normalizing processes taken by the generator will reduce the frequency of saturated and underexposed pixels. They recommended using a GAN image detector, which can easily measure the frequency of saturated and underexposed pixels in each image. A linear Support Vector Machine (SVM) classifies these features after being trained with Matlab's fitcsvm function. They applied this methodology on two separate datasets, GAN Crop and GAN Full. With a 0.7 AUC, their technique demonstrates a clear performance improvement when spotting entirely GAN-generated images. In order to detect Deepfake movies, this study offers an alternative to the attributes that can be taken into account.

Similarly, a technique to identify Deepfake videos employing Support Vector Machine (SVM) regression has been put forth by Kharbat et al. [48]. Their process trains Artificial Intelligence (AI) classifiers to recognize fake videos using feature points taken from the video. They have developed the feature point extraction algorithms HOG, ORB, BRISK, KAZE, SURF, and FAST. In order to take advantage of this discrepancy, this research suggests a method that extracts feature points using conventional edge feature detectors. There are 98 videos in the collection, half fraudulent and the other half actual. Each video is roughly 30 seconds long and is in mp4 format. The use of the HOG feature point extraction approach attained an accuracy of 95%. This study also argues that the above-mentioned conventional technique can also be employed for the process of feature detection, which is the most crucial component of this activity.

Ismail et al. [40] also used the best tree model, the XGBoost, as a machine-learning technique for identifying deepfakes. The algorithm takes advantage of more accurate approximations, and they have used it flexibly. In their paper, they have focused on the flaws in the deepfake generation pipeline, i.e., visual artifacts or discrepancies.

In a more recent study, Altaei and others [5] address the challenge of detecting fake face images by proposing a machine learning model utilizing Support Vector Machine (SVM) as a classifier. The process involved preprocessing steps like converting images from RGB to YCbCr and applying gamma correction, followed by edge detection using a Canny filter. Their approach includes two detection methods: SVM with Principal Component Analysis (PCA) and SVM alone, achieving a high accuracy of 96.8% with SVM-PCA and 72.2% with SVM alone.

Regarding the performance issue with machine learning-based Deepfake algorithms, it has been found that these methods can detect Deepfakes with up to 98% accuracy. The training and testing sets' alignment, the features employed, and the type of dataset all affect performance, though. When the experimentation employs a matching dataset, the study can get a better result by dividing it into a specific level of ratio, for instance, 80% for a training sample and 20% for a testing sample. It is assumed that the performance declines by nearly 50% for the unrelated dataset.

3. Statistical Techniques

The Expectation-Maximization (EM) technique was used by Guarnera et al. [32] to extract a group of regional features and simulate a fundamental generating convolutional structure. Following the extraction, they use preliminary tests and naive classifiers to apply ad-hoc validation to various designs, including GDWCT, STARGAN, ATTGAN, STYLEGAN, and STYLEGAN2. Through the development of a statistical framework by Maurer [71] for identifying the Deepfakes, Agarwal and Varshney [3] conducted a test of the hypothesis. The shortest route between distributions of the original and GAN-generated images is first defined by this method. This distance represents the ability to detect based on the findings of this hypothesis. For instance, when this distance is raised, Deepfakes are easily identified. On the other hand, the distance typically grows if the GAN offers a lower level of accuracy. Additionally, a very exact GAN is required to produce high-resolution altered images that are more difficult to identify. Recently, Hou et al. [37] proposed a novel approach, StatAttack, to evade DeepFake detection. This method minimizes statistical differences between real and fake images in spatial and frequency domains. It involves adding statistical-sensitive natural degradations (like exposure, blur, noise) in an adversarial manner and using a distribution-aware loss to align the feature distributions of fake images closer to real images. Additionally, they enhance this method with MStatAttack, which applies multi-layer degradations. Extensive testing on various detectors and datasets validates the effectiveness of their approach in both white-box and black-box settings.

The originality of the data can be understood by calculating several statistical measures, such as the average normalized cross-correlation scores between the original and suspicious data. Koopman et al. [54] investigated photo response nonuniformity (PRNU) to identify deepfakes in video frames. The distinctive noise pattern known as PRNU appears in digital photographs as a result of flaws in the light-sensitive detectors of the camera. It is sometimes referred to as the fingerprint of digital photos due to its uniqueness. The study creates a series of frames from the input videos and keeps them in directories that are organized chronologically. To maintain and make the PRNU sequence more clear, each video frame is cropped to the identical pixel range. Then eight equal groups are formed from these frames. Then, using the second-order FSTV approach, it creates the typical PRNU pattern for each frame as implemented by Baar et al. [9]. Then, it correlates them by computing the variances between the correlation values and the average correlation value for every frame, normalized cross-correlation value, and normalized correlation values for every frame.

Finally, the authors run a t-test similar to Welch [100] on the data to see whether Deepfakes and authentic videos have statistically significant differences.

While each deepfake detection technique has its strengths and limitations, their combined application can significantly enhance detection accuracy and robustness. We have listed the strengths and limitations for detection techniques in Table 4 Ongoing research should focus on addressing the computational challenges of deep learning methods, improving the generalizability of machine learning models, and enhancing the robustness of statistical techniques to keep pace with the evolving sophistication of deepfake generation technologies.

SRQ 3.3: How to measure the performance of a deepfake detection method?

The process of deepfake detection is a binary classification problem with the classes being *'fake'* & *'real'*. When a model is trained from the training data, it needs to be validated using the testing data. This validation can be measured using performance metrics. We tried to gather common performance metrics as well as how they can be utilized in the deepfake detection research.

1. The Confusion Matrix

The Confusion Matrix is a metric to measure performance used in machine learning classification problems. The outcome of these problems can be either binary classification or multiclass classification. The confusion matrix is a tabular representation of the actuals and predictions. It represents the counts of 4 occurrences while evaluating the model. The output *'TN'* expands to True Negative, which denotes the count of negative values correctly classified. Likewise, *'TP'* expands to True Positive, which denotes the count of positive values correctly classified. The output *'FP'* expands to False Positive, which denotes the count of positive values incorrectly classified. Similarly, *'FN'* expands to False Negative, which denotes incorrectly classified negative values.

In the case of deepfake detection, the positive values can be considered the 'fake' class, and the negative values can be considered the 'real' class. Based on this, please find the confusion matrix in Table 5, which can be used in case of deepfakes. The values in green are the favorable outcomes that are good for a model, and the ones in red are the incorrect classifications.

2. Precision & Recall

Precision is defined as the proportion of positives correctly identified, whereas Recall is defined as the proportion of the actual positives which are correctly identified, i.e., true positive rate.

Mathematically, precision is

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

and recall is

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

In terms of deepfake detection, precision can be described as

$$Precision = \frac{\text{Fake predicted fake}}{\text{Fake predicted fake} + \text{Fake predicted real}} \tag{3}$$

Table 4 List of features, strengths and limitations along with the papers in which they are implemented

Reference	Features	Strengths	Limitations
Tao et al. [93]	SPMC, Motion Transformer	Improves frame alignment and motion compensation	High computational requirements
De Rezende et al. [21]	ResNet-50, SVM	High accuracy with pre-trained model	Requires large labeled datasets
Şengür et al. [89]	VGG16, AlexNet, SVM	Combines features for improved prediction	May not capture complex patterns as deep learning
Mo et al. [75]	CNN, Spatial High-Pass Filters	Highlights subtle details, high accuracy	Requires high-quality input data
Khodabakhsh et al. [50]	CNN, Various Architectures	Generalizes well for non-public datasets	Struggles with new artifacts
Li and Lyu [62]	Face-Warping Artifacts	Effective without needing deepfake images for training	Limited to face-warping artifacts
Afchar et al. [2]	Mesoscopic Properties	Focuses on mesoscopic properties	Relies on private dataset
Marra et al. [70]	Image-to-Image Translation, Steganalysis	Good accuracy in compressed scenarios	Sensitive to initial parameter selection
Güera and Delp [33]	CNN, RNN	Captures temporal features	Computationally intensive
Li et al. (2018)	Eye-Blinking Patterns	Analyzes eye-blinking patterns	Overlooks other facial movements
Luo et al. [66]	Spatial Pyramid Pooling	Improves fixed-size image conversion	Complex implementation
Amerini et al. [6]	Optical Flow Vectors	Uses optical flow vectors for detection	Requires high-quality video
Hsu et al. [38]	CNN, Contrastive Loss Functions	Combines features effectively	Sensitive to training data quality
Korshunov and Marcel [56]	Facial Recognition Systems	Demonstrates vulnerability of facial recognition systems	Limited to specific datasets
Malolan et al. [68]	LRP, LIME	Provides explainable results	Requires complex explanations
Frank et al. [29]	Frequency Domain Analysis	Analyzes frequency domain artifacts	Sensitive to data quality
Ciftci et al. [18]	PPG Data	Uses physiological signals for detection	Needs high-resolution data
Mehra [73]	Capsule Networks, LSTM	Combines spatial and temporal analysis	Complex model training
Aneja and Nießner [7]	ResNet-18, DDT	Effective zero-shot and few-shot learning	Requires domain adaptation
Wodajo and Atnafu [103]	CNN, ViT	Combines CNN and transformer models	High computational cost
Arshed et al. [8]	Vision Transformers	High accuracy in real-world datasets	Resource-intensive
Jung et al. [43]	Eye-Blinking Patterns	Uses natural eye-blinking for detection	Overlooks other features

Table 4 continued

Reference	Features	Strengths	Limitations
Agarwal et al. [4]	Frequency Spectrum, Capsule Networks	Uses both frequency and spatial information	High computational requirements
Wang and Dantcheva [99]	3D-CNN Models	Provides consistent results	Dataset-specific performance
Kolagati et al. [53]	Facial Landmarks	Focuses on facial landmarks	Requires accurate landmark detection
Coccomini et al. [19]	Vision Transformers, EfficientNet B0	Combines transformers with efficient feature extraction	High computational requirements
Suganthi et al. [92]	FF-LBPH, Fisherface	Effective face recognition and detection	Needs extensive preprocessing
Vamsi et al. [97]	ResNext, LSTM	Combines CNN and RNN for video analysis	High computational cost
Deng et al. [22]	EfficientNet-V2	Enhances precision in detection	Requires real-time adaptation
Cheng et al. [15]	Voice-Face Similarity Detection	Addresses voice-face mismatches	Needs large multimodal datasets
Zhao et al. [111]	ISTVT Model	Robust across multiple datasets	Sensitive to hyperparameters
Ke and Wang [47]	DF-UDetector	Focuses on feature restoration	Complex model structure
Yu et al. [108]	PVASS-MDD	Bridges audio-visual inconsistencies	Complex cross-modal fusion
Yin et al. [107]	Dynamic Difference Learning	Models inter-frame motions	Sensitive to frame inconsistencies

Note: Not all algorithms are included in this table

Table 5 Confusion matrix for deepfake detection

		Predicted	
		Fake	Real
Actual	Fake	True Positive (TP)	False Positive (FP)
	Real	False Negative (FN)	True Negative (TN)

Similarly, recall is

$$Recall = \frac{\text{Fake predicted fake}}{\text{Fake predicted fake} + \text{Real predicted fake}} \tag{4}$$

3. Accuracy

Accuracy is defined as the ratio of the correctly predicted values with the total values, i.e., it explains to us how accurately the classifier is able to predict the labels provided in the problem statement.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

Or, if we talk in terms of deepfake detection, then

$$Accuracy = \frac{\text{Fake predicted fake} + \text{Real predicted real}}{\text{Total values}} \tag{6}$$

Accuracy metric is best used in case of balanced dataset. A balanced datset is where the proportion of positives and negatives (fake and real in case of deepfake detection) are equal.

4. TPR & FPR

TPR expands to a True Positive Rate, whereas FPR expands to a False Positive Rate. TPR and FPR can be referred to as metrics for positive samples. Alternatively, TPR is also known as Correct Decision Rate (or CDR), and FPR is also known as False Alarm Rate (or FAR). TPR is the ratio of positive values predicted among the positive samples. FPR is the fraction of positive values predicted among the samples which are actually negative.

$$TPR = \frac{TP}{TP + FN} = 1 - FNR \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

Please note that TPR is an alternative name for **recall**.

5. Sensitivity & Specificity

The previously mentioned positive rates (TPR & FPR) and negative rates (TNR & FNR) are also known as sensitivity and specificity, respectively. Specificity is another name for **recall**. Hence, recall, sensitivity, and TPR are similar metrics. However, when it comes to

deepfake detection, then sensitivity and specificity are only sometimes used because these metrics' sole purpose is to classify biological and medicinal data.

6. ROC Curve & AUC

ROC expands to Receiver Operating Characteristics and ROC curves are commonly used to compute the performance in case of a binary classifier which provides score or probability of prediction as output. To understand the concept of ROC curve, let S be the set of all test samples and let $f(s)$ be the output scores where all $s \in S$ and $f(s)$ lie in interval $[a_1, a_2]$. Let t be a prediction threshold where $t \in [a_1, a_2]$ and the classification occurs as follows:

$$\text{class}(s) = \begin{cases} \text{positive}, & \text{if } f(s) \geq t, \text{ and} \\ \text{negative}, & \text{otherwise.} \end{cases} \tag{9}$$

It may be noted that, if $t = a_1$, the samples will be predicted as positive which leads to TN = FN = 0 and hence FPR = TPR = 1, while $t = a_2$, the samples will be predicted as negative which leads to TP = FP = 0 and hence FPR = TPR = 0.

We can statistically determine the ROC curve for a random classifier to be the TPR = FPR line, with an area under the ROC curve (AUC) of 0.5. It is based on the assumption that $f(s)$ is distributed uniformly on $[a_1, a_2]$ for the testing sample. We can also quantitatively demonstrate that given a binary classifier that outperforms a random predictor, its AUC is always greater than 0.5, and 1 being the optimal solution. It should be noted that no binary classifier can have an AUC below 0.5 because one only needs to reverse the prediction outcome to have a classifier with an area under the curve (AUC) of 1. Figure 11 shows graphically how the ROC and AUC relate to one another.

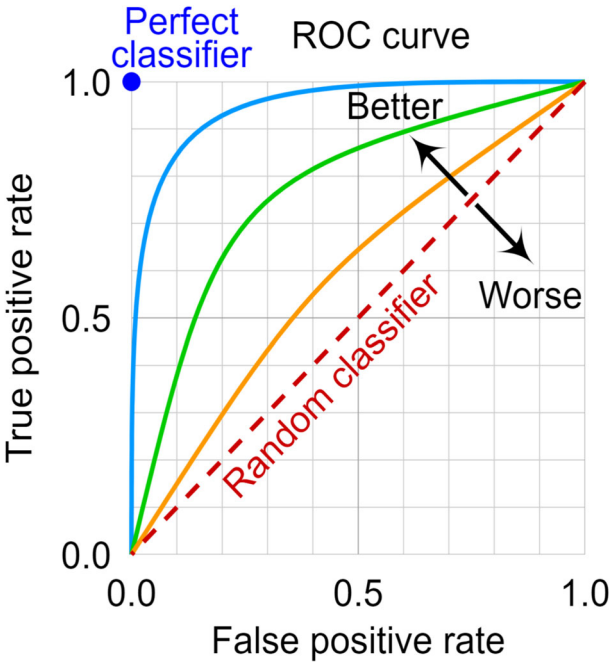


Fig. 11 The ROC space for a better and worse classifier [Wikipedia [102]]

7. F-score

While performing binary classification in statistical analysis, F-measure or F-score is a metric for accuracy. It is derived from the precision and recall calculated from the trial. The F-score is a family of F metrics where F_1 score (balanced F-score) is the harmonic mean of precision and recall.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = \frac{2.TP}{2.TP + FP + FN} \tag{10}$$

Whereas a more generic form of F-score is represented as F_β which applies extra weights, to value either precision or recall more than the other.

F-score can have a maximum value of 1.0, which indicates perfect precision and recall; however, a minimum value of 0 is possible if either the precision or recall is 0.

$$F_\beta = (1 + \beta^2).\frac{\text{precision.recall}}{\beta^2.\text{precision} + \text{recall}} \tag{11}$$

In the above equation, β is chosen such that recall is evaluated β times as significant as precision. If we consider equal importance to precision and recall, then it is the case of F_1 score because we say that recall is *1* times as important as precision. Alternatively, if we represent the same equation in terms of Type I and type II errors, then it becomes

$$F_\beta = \frac{(1 + \beta^2).TP}{(1 + \beta^2).TP + \beta^2.FP + FP} \tag{12}$$

8. Log Loss

This is another performance metric which is widely used for binary classifiers. It can be used to return a score of probability for the predicted label. It is defined as negative log-likelihood of a binary classifier returning `y_pred` probabilities for its training data `y_true`. For single sample with true label $y \in \{0, 1\}$ and an estimated probability $p = Pr(y = 1)$, the log loss is defined as

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{13}$$

RQ 4: What are the best methods for detecting deepfakes based on existing literature?

To determine the best methods for detecting deepfakes, we created two categories of methods: Deep Learning Methods and Non-Deep Learning Methods. The Non-Deep Learning methods comprise Machine Learning based methods and Statistical methods. The performance was measured based on Accuracy, AUC, Precision, and Recall.

We found that deep learning-based models surpassed non-deep learning-based models after evaluating these models using performance metrics. As shown in Fig. 12, the outcomes show that deep learning models perform much better in terms of accuracy and precision as compared to non-deep learning models. However, performance in AUC is comparable, although non-deep learning techniques performed well for Recall. The overall outcomes show the advantage of models based on deep learning over models not based on deep learning.

To quantitatively analyze the performance metrics of various deepfake detection algorithms we considered accuracy, precision, recall, and F1-score, metrics and evaluated them across multiple datasets. We present the performance metrics of six prominent deepfake detection algorithms: XceptionNet, ResNet-50, VGG16, Capsule Networks, 3D-CNN, and

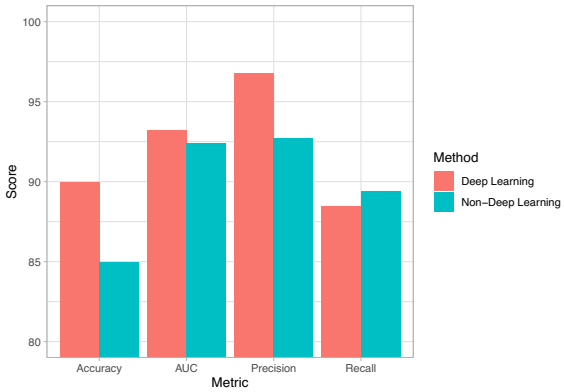


Fig. 12 Comparison of Deepfake detection methods classified in 2 categories - deep learning and non-deep learning

EfficientNet-V2. The metrics are evaluated on six different datasets: FaceForensics++, UADFV, CASIA, Deepfake-TIMIT, DFDC, and Celeb-Df. This analysis shown in Fig. 13 helps in understanding the strengths and limitations of each algorithm in different scenarios.

SRQ 4.1: What features are mostly considered while detecting deepfake?

While classifying images or videos as deepfake, there are different features used in the detection models. Out of these features the following features have been used frequently while detecting deepfakes:

Table 6 represents the papers with respect to the features.

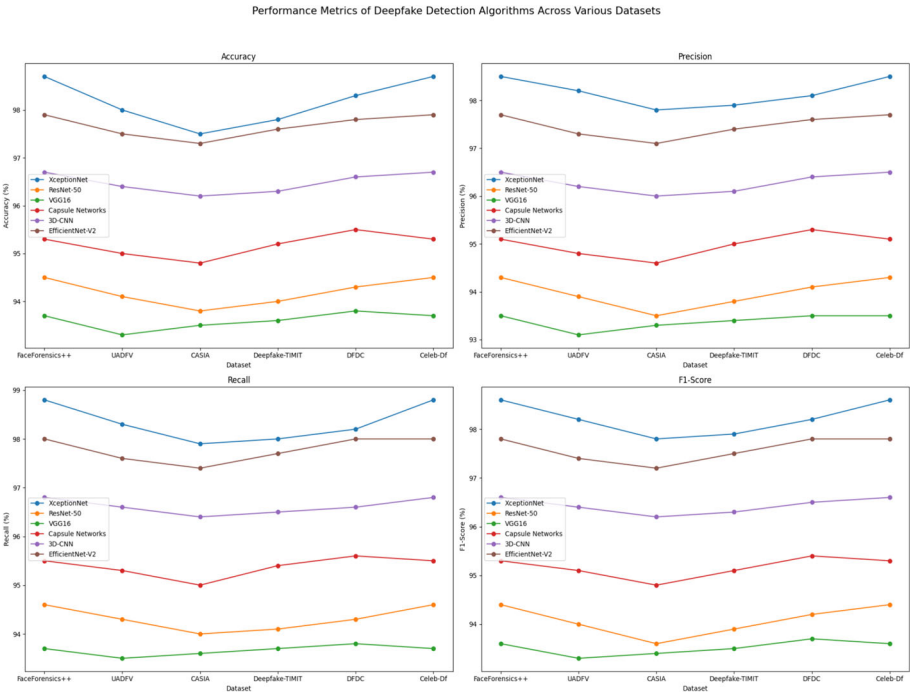


Fig. 13 Performance matrix of deepfake detection algorithms across various datasets

Table 6 List of features and datasets along with the papers in which they are implemented

Features	Reference	Method	Dataset
	Khodabakhsh and Busch [49]	CNN	FaceForensics
Special artifacts	Han and Gevers [35]	CNN	UADFV, Celeb-DF, DF-TIMIT, FaceForensics++
	Cozzolino et al. [20]	CNN	Celeb-DF, DFD, FaceForensics++
Visual artifacts	Zhang et al. [109]	CNN	Private Dataset
	Haliassos et al. [34]	CNN, MSTCN	DFDC, Celeb-DF, FaceForensics++, Deepfake 1.0
Biological artifacts	Chang et al. [13]	CNN	Celeb-DF, FaceForensics++
	Sabir et al. [87]	CNN	FaceForensics++
Face landmarks	Mittal et al. [74]	CNN	DFDC, DF-TIMIT
	Li et al. [60]	Multiple Instance Learning	Celeb-DF, FaceForensics, DFDC, FaceForensics++
Spatio-temporal consistency	Ganiyusufoglu et al. [30]	CNN	DFDC, FaceForensics++, Deepfake 1.0
	Bonomi et al. [10]	SVM	FaceForensics++
Texture	Zhu et al. [112]	CNN	DFDC, DFD, FaceForensics++
	Du et al. [25]	CNN	Celeb-A, FaceForensics++
Latent feature	Nguyen et al. [80]	CNN	FaceForensics++
	Afchar et al. [2]	CNN	Deepfake, FaceForensics
Mesoscopic features	Kawa and Syga [46]	CNN	FaceForensics++
	Sohrawardi et al. [91]	RNN	FaceForensics++
Intra-frame inconsistency	Li et al. [61]	CNN	FaceForensics, DF-TIMIT, Private dataset

4 Future research directions

4.1 Limitations in existing literature

Despite significant advancements in deepfake detection, several limitations persist in the current literature. One primary limitation is the generalizability of detection models. Many deepfake detection methods demonstrate high accuracy on specific datasets but struggle to maintain performance across diverse datasets with varying types of manipulations. This is crucial since potential deepfake types are typically unknown in real-world circumstances, and a model trained on one particular media must be able to compete against other unknown media. Additionally, the robustness of these models against adversarial attacks remains inadequate, posing a critical vulnerability. Existing detection techniques mainly concentrate on the flaws of the deepfake generating pipelines, i.e., finding weaknesses in the rivals to attack. In adversarial contexts, where attackers frequently try to conceal such deepfake production methods, this kind of information and knowledge is not always accessible. The high computational cost associated with deep learning-based detection methods also limits their applicability in real-time scenarios and on resource-constrained devices. Furthermore, there is a lack of comprehensive approaches that integrate multi-modal data, which can leverage audio-visual cues to improve detection accuracy. These are significant obstacles to the evolution of detection techniques, and future studies should concentrate on developing more reliable, scalable, and, most importantly, generalizable techniques.

4.2 Future directions

To address these limitations and enhance the efficacy of deepfake detection, several future research directions are proposed:

1. Robust and Generalizable Detection Algorithms:
Developing robust detection algorithms that can generalize well across different types of deepfakes and diverse datasets is crucial. Future research should focus on creating models that can adapt to new and unseen manipulation techniques. Techniques such as domain adaptation, transfer learning, and meta-learning can be explored to enhance model generalizability. As noted by Verdoliva [98], blending various techniques, data sources, and modalities can significantly improve deepfake detection accuracy.

2. Adversarial Robustness:
Improving the robustness of deepfake detection models against adversarial attacks is a critical area of research. Adversarial training, which involves training models with adversarial examples, can be employed to enhance model resilience. Additionally, exploring robust optimization techniques and incorporating adversarial defense mechanisms into existing detection frameworks can mitigate the impact of adversarial attacks.

3. Real-Time Detection:
As deepfakes become more prevalent, the need for real-time detection systems increases. Future research should aim to optimize the computational efficiency of detection models to enable real-time processing. Techniques such as model pruning, quantization, and the development of lightweight neural network architectures can be pursued to achieve this goal. Implementing detection algorithms on edge devices and distributed systems can also facilitate real-time applications.

4. Multi-Modal Deepfake Detection:

Integrating multi-modal data, such as audio-visual information, can significantly improve deepfake detection accuracy. Future research should explore the development of multi-modal deepfake detection frameworks that analyze and fuse data from different modalities. This approach can help detect inconsistencies between audio and visual cues, such as lip-sync errors and mismatched audio-visual signals [67, 90].

5. Explainability and Transparency:

Enhancing the interpretability of deepfake detection models is essential to gain user trust and facilitate their adoption. Developing explainable AI techniques that provide clear justifications for detection decisions can help achieve this. Methods such as Layer-Wise Relevance Propagation (LRP), Local Interpretable Model-Agnostic Explanations (LIME), and SHAP values can be employed to create interpretable models.

6. Ethical and Legal Frameworks:

As deepfake technology evolves, addressing its ethical and legal implications is imperative. Collaborative efforts involving policymakers, legal experts, and technology companies are needed to establish guidelines and regulations for the use and dissemination of deepfake content. Ensuring ethical AI practices in the development and deployment of detection technologies is crucial to maintain fairness, accountability, and transparency.

7. Enhanced Dataset Creation:

The availability of diverse and high-quality datasets is critical for training and evaluating deepfake detection models. Future research should focus on creating comprehensive benchmark datasets that include a wide variety of deepfakes generated using different techniques. Data augmentation and synthesis techniques can be employed to create more robust training datasets, improving model training and evaluation.

8. Brain-Inspired Learning for Deepfake Detection:

Brain-inspired learning, particularly spike-based machine intelligence, offers promising avenues for improving deepfake detection. Techniques such as robust and energy-efficient learning frameworks (e.g., SIBOLS by Yang et al. [106]), effective surrogate gradient learning with high-order information bottleneck by Yang and Chen [104], and spike-driven multi-scale learning with hybrid mechanisms of spiking dendrites by Yang et al. [105] can be explored. These approaches leverage the efficiency and robustness of spiking neural networks, providing potential improvements in detection accuracy and energy efficiency.

9. Continuous Learning and Adaptation:

Developing deepfake detection models capable of continuous learning and adaptation is essential to keep pace with evolving manipulation techniques. Continual learning frameworks that can incrementally learn from new data without forgetting previously learned information can enhance model adaptability. Techniques such as robust spike-based continual meta-learning and neuromorphic architectures for spike-driven online learning can be pursued.

10. Collaborative Research and Open Initiatives:

Advancing deepfake detection requires collaboration across different sectors and active engagement with the research community. Promoting interdisciplinary collaboration between computer scientists, ethicists, legal experts, and industry practitioners can address the multi-faceted challenges posed by deepfakes. Supporting open research initiatives and data sharing can accelerate progress in deepfake detection, fostering innovation and improving detection capabilities.

4.3 Integrating detection into distribution channels

Another research approach involves integrating detection techniques into distribution channels like social media. On these sites, a filtering or screening system can be developed to make it easier to identify deepfakes [114]. Furthermore, to reduce the consequences of deepfakes, legal constraints might be put on the tech corporations that control these platforms. Additionally, watermarking tools can be incorporated into the devices used to create fake digital content, utilizing blockchain technology to store original information about multimedia contents, ensuring integrity and authenticity.

4.4 Accents and detection accuracy

The bulk of detection techniques focuses on identifying the kind of deepfake without considering other elements that could influence detection accuracy. Accents, which are described as a specific set of people's characteristic speech patterns, notably those of the inhabitants or natives of a given nation, are one such element. Therefore, accents may have an impact on detection accuracy. However, research on this topic needs to be done in the deepfake literature. Accents impact the effectiveness of the suggested approaches in other audio disciplines, such as speaker recognition [78]. Further research is required on languages with a wide variety of accents, such as South Asian English, to solve this issue. One nation will frequently have speakers with various accents, and the Asian English language will be no exception, with speakers who have accents from India, Sri Lanka, Bhutan, the Maldives, and many other places. The likelihood that the classifier will acquire a more generic model for the detection job rises as the variety of accents grows, prompting more exploration.

5 Conclusion

Deepfakes have weakened people's faith in media material since seeing them no longer equates to believing in them. Moreover, they disturb the people who are being targeted, have an adverse influence on them, intensify hate speech and disinformation, and possibly intensify political unrest, enflame the community, or even start a war. It is incredibly significant because deepfake technologies are becoming more accessible, and social media platforms can quickly disseminate those fake information.

This study offers an up-to-date systematic literature review of 81 studies within five years from 2017 to 2023 for creating and detecting deepfakes and discusses problems, prospective trends, and future possibilities in this field. The contributions mentioned above were made with the help of a defined methodology of performing an excellent systematic literature review and specifying appropriate research questions to assist in future research for identifying deepfake media. Therefore, this work will be helpful for the artificial intelligence research community to create efficient techniques for combating deepfakes.

Data Availability We do not analyze or generate any datasets, because our work proceeds within a theoretical and mathematical approach towards reviewing the techniques of Deepfake Detection. One can obtain the relevant materials from the references below or from SRQ 3.1.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

References

1. (2019) Deepnude. https://deepnude.ca/
2. Afchar D, Nozick V, Yamagishi J et al (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International workshop on information forensics and security (WIFS), IEEE, pp 1–7, https://doi.org/10.1109/wifs.2018.8630761
3. Agarwal S, Varshney LR (2019) Limits of deepfake detection: a robust estimation viewpoint. arXiv:1905034 https://doi.org/10.48550/arXiv.1905.03493
4. Agarwal S, Girdhar N, Raghav H (2021) A novel neural model based framework for detection of gan generated fake images. In: 2021 11th International conference on cloud computing, data science & engineering (Confluence), IEEE, pp 46–5 https://doi.org/10.1109/confluence51648.2021.9377150
5. Altaei MSM, others (2023) Detection of deep fake in face images based machine learning. Al-Salam J Engr Tech 2(2):1–1 https://doi.org/10.55145/ajest.2023.02.02.001
6. Amerini I, Galteri L, Caldelli R et al (2019) Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International conference on computer vision workshops, pp 0–0 https://doi.org/10.1109/iccvw.2019.00152
7. Aneja S, Nießner M (2020) Generalized zero and few-shot transfer for facial forgery detection. https://doi.org/10.48550/arXiv.2006.11863
8. Arshed MA, Alwadain A, Faizan Ali R et al (2023) Unmasking deception: empowering deepfake detection with vision transformer network. Mathematics 11(17):3710. https://doi.org/10.3390/math11173710
9. Baar T, van Houten W, Geradts Z (2012) Camera identification by grouping images from database, based on shared noise patterns. https://doi.org/10.48550/arXiv.1207.2641
10. Bonomi M, Pasquini C, Boato G (2021) Dynamic texture analysis for detecting fake faces in video sequences. J Visual Commu Image Represent 79:103239. https://doi.org/10.1016/j.jvcir.2021.103239
11. Brereton P, Kitchenham BA, Budgen D et al (2007) Lessons from applying the systematic literature review process within the software engineering domain. J Syst Software 80(4):571–583. https://doi.org/10.1016/j.jss.2006.07.009
12. Buo SA (2020) The emerging threats of deepfake attacks and countermeasures. https://doi.org/10.48550/arXiv.2012.07989
13. Chang X, Wu J, Yang T et al (2020) Deepfake face image detection based on improved vgg convolutional neural network. In: 2020 39th Chinese Control Conference (CCC), IEEE, pp 7252–725https://doi.org/10.23919/ccc50068.2020.9189596
14. Chang YL, Liu ZY, Lee KY et al (2019) Learnable gated temporal shift module for deep video inpainting. https://doi.org/10.48550/arXiv.1907.01131
15. Cheng H, Guo Y, Wang T et al (2022) Voice-face homogeneity tells deepfake. https://doi.org/10.1145/3625231
16. Chesney R, Citron D (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. Calif L Rev 107:1753. https://doi.org/10.2139/ssrn.3213954
17. Ciftci UA, Demir I, Yin L (2020a) Fakecatcher: detection of synthetic portrait videos using biological signals. IEEE Trans pattern anal machine intell https://doi.org/10.1109/tpami.2020.3009287
18. Ciftci UA, Demir I, Yin L (2020b) How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, pp 1–1 https://doi.org/10.1109/ijcb48548.2020.9304909
19. Coccomini DA, Messina N, Gennaro C et al (2022) Combining efficientnet and vision transformers for video deepfake detection. In: International conference on image analysis and processing, Springer, pp 219–22 https://doi.org/10.1007/978-3-031-06433-3_19
20. Cozzolino D, Rössler A, Thies J et al (2021) Id-reveal: identity-aware deepfake video detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15108–151 https://doi.org/10.1109/iccv48922.2021.01483
21. De Rezende ER, Ruppert GC, Carvalho T (2017) Detecting computer generated images with deep convolutional neural networks. In: 2017 30th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), IEEE, pp 71–7 https://doi.org/10.1109/sibgrapi.2017.16

22. Deng L, Suo H, Li D (2022) Deepfake video detection based on efficientnet-v2 network. Comput Intell Neurosci 2022. https://doi.org/10.1155/2022/3441549
23. Deng Y, Yang J, Chen D et al (2020) Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5154–516 https://doi.org/10.1109/cvpr42600.2020.00520
24. Dolhansky B, Howes R, Pflaum B et al (2019) The deepfake detection challenge (dfdc) preview dataset. https://doi.org/10.48550/arXiv.1910.08854
25. Du M, Pentyala S, Li Y et al (2020) Towards generalizable deepfake detection with locality-aware autoencoder. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 325–33 https://doi.org/10.1145/3340531.3411892
26. Dufour N, Gully A (2019) Contributing data to deepfake detection research. Google AI Blog 1(3) https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html
27. Fox G, Liu W, Kim H, et al (2021) Videoforensicshq: detecting high-quality manipulated face videos. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6 https://doi.org/10.1109/icme51207.2021.9428101
28. Frank J, Schönherr L (2021) Wavefake: a data set to facilitate audio deepfake detection. https://doi.org/10.48550/arXiv.2111.02813
29. Frank J, Eisenhofer T, Schönherr L et al (2020) Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning, PMLR, pp 3247–325 https://doi.org/10.48550/arXiv.2003.08685
30. Ganiyusufoglu I, Ngô LM, Savov N et al (2020) Spatio-temporal features for generalized detection of deepfake videos. https://doi.org/10.48550/arXiv.2010.11844
31. Goodfellow I, Pouget-Abadie J, Mirza M et al (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C et al (eds) Advances in Neural Information Processing Systems, vol 27 Curran Associates, Inc., https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
32. Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 666–66 https://doi.org/10.1109/cvprw50498.2020.00341
33. Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on Advanced Video and Signal based Surveillance (AVSS), IEEE, pp 1–6 https://doi.org/10.1109/avss.2018.8639163
34. Haliassos A, Vougioukas K, Petridis S et al (2021) Lips don't lie: a generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5039–504 https://doi.org/10.1109/cvpr46437.2021.00500
35. Han J, Gevers T (2020) Mmd based discriminative learning for face forgery detection. In: Proceedings of the Asian conference on computer vision, https://doi.org/10.1007/978-3-030-69541-5_8
36. He Y, Gan B, Chen S et al (2021) Forgerynet: a versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4360–43 https://doi.org/10.1109/cvpr46437.2021.00434
37. Hou Y, Guo Q, Huang Y et al (2023) Evading deepfake detectors via adversarial statistical consistency. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12271–1228 https://doi.org/10.1109/cvpr52729.2023.01181
38. Hsu CC, Lee CY, Zhuang YX (2018) Learning to detect fake face images in the wild. In: 2018 International Symposium on Computer, Consumer and Control (IS3C), IEEE, pp 388–39 https://doi.org/10.1109/is3c.2018.00104
39. Huang J, Wang X, Du B et al (2021) Deepfake mnist+: a deepfake facial animation dataset. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1973–198 https://doi.org/10.1109/iccvw54120.2021.00224
40. Ismail A, Elpeltagy M, S. Zaki M et al (2021) A new deep learning-based methodology for video deepfake detection using xgboost. Sensors 21(16):5413. https://doi.org/10.3390/s21165413
41. Jia S, Li X, Lyu S (2022) Model attribution of face-swap deepfake videos. https://doi.org/10.1109/icip46576.2022.9897972
42. Jiang L, Li R, Wu W et al (2020) Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2889–289 https://doi.org/10.1109/cvpr42600.2020.00296
43. Jung T, Kim S, Kim K (2020) Deepvision: deepfakes detection using human eye blinking pattern. IEEE Access 8:83144–83154. https://doi.org/10.1109/access.2020.2988660
44. Karras T, Aila T, Laine S et al (2017) Progressive growing of gans for improved quality, stability, and variation. https://doi.org/10.48550/arXiv.1710.10196

45. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4401–441 https://doi.org/10.1109/cvpr.2019.00453
46. Kawa P, Syga P (2020) A note on deepfake detection with low-resources. https://doi.org/10.48550/arXiv.2006.05183
47. Ke J, Wang L (2023) Df-udetector: an effective method towards robust deepfake detection via feature restoration. Neural Networks 160:216–226. https://doi.org/10.1016/j.neunet.2023.01.001
48. Kharbat FF, Elamsy T, Mahmoud A et al (2019) Image feature detectors for deepfake video detection. In: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), IEEE, pp 1–4 https://doi.org/10.1109/aiccsa47632.2019.9035360
49. Khodabakhsh A, Busch C (2020) A generalizable deepfake detector based on neural conditional distribution modelling. In: 2020 international conference of the Biometrics Special Interest Group (BIOSIG), IEEE, pp 1–5
50. Khodabakhsh A, Ramachandra R, Raja K et al (2018) Fake face detection methods: Can they be generalized? In: 2018 international conference of the biometrics special interest group (BIOSIG), IEEE, pp 1–6 https://doi.org/10.23919/biosig.2018.8553251
51. Kim BH, Ganapathi V (2019) Lumièrenet: lecture video synthesis from audio. arXiv:1907.02253 https://doi.org/10.48550/arXiv.1907.02253
52. Kitchenham B (2004) Procedures for performing systematic reviews. Keele, UK, Keele University 33(2004):1–26
53. Kolagati S, Priyadharshini T, Rajam VMA (2022) Exposing deepfakes using a deep multilayer perceptron-convolutional neural network model. Int J Inf Manage Data Insights 2(1):100054. https://doi.org/10.1016/j.jjimei.2021.100054
54. Koopman M, Rodriguez AM, Geradts Z (2018) Detection of deepfake video manipulation. In: The 20th Irish machine vision and image processing conference (IMVIP), pp 133–136, https://shorturl.at/bmLRY
55. Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? assessment and detection. https://doi.org/10.48550/arXiv.1812.08685
56. Korshunov P, Marcel S (2019) Vulnerability assessment and detection of deepfake videos. In: 2019 International Conference on Biometrics (ICB), IEEE, pp 1–6 https://doi.org/10.1109/icb45273.2019.8987375
57. Kwon P, You J, Nam G, et al (2021) Kodf: a large-scale korean deepfake detection dataset. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10744–1075 https://doi.org/10.1109/iccv48922.2021.01057
58. Lattas A, Moschoglou S, Gecer B et al (2020) Avatarme: realistically renderable 3d facial reconstruction" in-the-wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 760–76 https://doi.org/10.1109/cvpr42600.2020.00084
59. Li L, Bao J, Yang H et al (2019) Faceshifter: towards high fidelity and occlusion aware face swapping. https://doi.org/10.48550/arXiv.1912.13457
60. Li X, Lang Y, Chen Y et al (2020a) Sharp multiple instance learning for deepfake video detection. In: Proceedings of the 28th ACM international conference on multimedia, pp 1864–187 https://doi.org/10.1145/3394171.3414034
61. Li X, Yu K, Ji S et al (2020b) Fighting against deepfake: patch&pair convolutional neural networks (ppcnn). In: Companion proceedings of the web conference 2020, pp 88–89 https://doi.org/10.1145/3366424.3382711
62. Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. https://doi.org/10.48550/arXiv.1811.00656
63. Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing ai created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, pp 1–7 https://doi.org/10.1109/wifs.2018.8630787
64. Li Y, Yang X, Sun P et al (2020c) Celeb-df: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3207–321https://doi.org/10.1109/cvpr42600.2020.00327
65. Liu A, Zhao C, Yu Z et al (2022) Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. IEEE Trans Inf Forensics Secur 17:2497–2507. https://doi.org/10.1109/TIFS.2022.3188149
66. Luo M, Xiao Y, Zhou Y (2018) Multi-scale face detection based on convolutional neural network. In: 2018 Chinese Automation Congress (CAC), IEEE, pp 1752–175 https://doi.org/10.1109/cac.2018.8623411

67. Lyu S (2020) Deepfake detection: current challenges and next steps. In: 2020 IEEE international conference on multimedia & expo workshops (ICMEW), IEEE, pp 1–6 https://doi.org/10.1109/icmew46912.2020.9105991
68. Malolan B, Parekh A, Kazi F (2020) Explainable deep-fake detection using visual interpretability methods. In: 2020 3rd International Conference on Information and Computer Technologies (ICICT), IEEE, pp 289–293 https://doi.org/10.1109/icict50521.2020.00051
69. Marcílio WE, Eler DM (2020) From explanations to feature selection: assessing shap values as feature selection mechanism. In: 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI), Ieee, pp 340–347 https://doi.org/10.1109/sibgrapi51738.2020.00053
70. Marra F, Gragnaniello D, Cozzolino D et al (2018) Detection of gan-generated fake images over social networks. In: 2018 IEEE conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, pp 384–389 https://doi.org/10.1109/mipr.2018.00084
71. Maurer UM (2000) Authentication theory and hypothesis testing. IEEE Trans Inf Theory 46(4):1350–1356. https://doi.org/10.1109/18.850674
72. McCloskey S, Albright M (2019) Detecting gan-generated imagery using saturation cues. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp 4584–4588 https://doi.org/10.1109/icip.2019.8803661
73. Mehra A (2020) Deepfake detection using capsule networks with long short-term memory networks. Master's thesis, University of Twent. https://doi.org/10.5220/0010289004070414
74. Mittal T, Bhattacharya U, Chandra R et al (2020) Emotions don't lie: an audio-visual deepfake detection method using affective cues. In: Proceedings of the 28th ACM international conference on multimedia, pp 2823–2832 https://doi.org/10.1145/3394171.3413570
75. Mo H, Chen B, Luo W (2018) Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM workshop on information hiding and multimedia security, pp 43–47 https://doi.org/10.1145/3206004.3206009
76. Mok A (2022) Take a look at the digitally de-aged harrison ford in the trailer for the new indiana jones movie. Accessed 24 Dec 2022 https://shorturl.at/bhIU5
77. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. In: Proc. interspeech 2017, pp 2616–2620 https://doi.org/10.21437/Interspeech.2017-950
78. Najafian M (2013) Modeling accents for automatic speech recognition. In: University of Birmingham Graduate School Research Poster Conference 2013, Prizewinners from the Graduate School Research Poster Conference 2013, research Supervisor: Prof Martin Russell http://epapers.bham.ac.uk/1736/
79. Neves JC, Tolosana R, Vera-Rodriguez R et al (2020) Ganprintr: improved fakes and evaluation of the state of the art in face manipulation detection. IEEE J Select Topics Signal Process 14(5):1038–1048. https://doi.org/10.1109/jstsp.2020.3007250
80. Nguyen HH, Yamagishi J, Echizen I (2019) Capsule-forensics: using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 2307–2311 https://doi.org/10.1109/icassp.2019.8682602
81. Nirkin Y, Keller Y, Hassner T (2019) Fsgan: subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7184–7193 https://doi.org/10.1109/iccv.2019.00728
82. Nishad G (2019) Mnist-gan: detailed step by step explanation & implementation in code. accessed: 2022-08-10, https://shorturl.at/jkyFK
83. Perov I, Gao D, Chervoniy N et al (2020) Deepfacelab: Integrated, flexible and extensible face-swapping framework. https://doi.org/10.1016/j.patcog.2023.109628
84. Pfefferkorn R (2019) "deepfakes" in the courtroom. BU Pub Int LJ 29:245
85. Rössler A, Cozzolino D, Verdoliva L et al (2018) Faceforensics: a large-scale video dataset for forgery detection in human faces. https://doi.org/10.48550/arXiv.1803.09179
86. Rössler A, Cozzolino D, Verdoliva L et al (2019) FaceForensics++: learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV), https://doi.org/10.1109/iccv.2019.00009
87. Sabir E, Cheng J, Jaiswal A et al (2019) Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3(1):80–87 https://doi.org/10.48550/arXiv.1905.00582
88. Sanderson C, Lovell BC (2009) Multi-region probabilistic histograms for robust and scalable identity inference. In: International conference on biometrics, Springer, pp 199–208 https://doi.org/10.1007/978-3-642-01793-3_21
89. Şengür A, Akhtar Z, Akbulut Y et al (2018) Deep feature extraction for face liveness detection. In: 2018 International conference on artificial Intelligence and Data Processing (IDAP), Ieee, pp 1–4 https://doi.org/10.1109/idap.2018.8620804

90. Sharma VK, Garg R, Caudron Q (2023) Spatio-temporal convolutional neural networks for deepfake detection: an empirical study. In: 2023 Second International Conference on Informatics (ICI), IEEE, pp 1–7 https://doi.org/10.1109/ICI60088.2023.10420892
91. Sohrawardi SJ, Chintha A, Thai B et al (2019) Poster: towards robust open-world detection of deepfakes. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pp 2613–2615 https://doi.org/10.1145/3319535.3363269
92. Suganthi S, Ayoobkhan MUA, Bacanin N et al (2022) Deep learning model for deep fake face recognition and detection. Peer J Computer Science 8:e881. https://doi.org/10.7717/peerj-cs.881
93. Tao X, Gao H, Liao R et al (2017) Detail-revealing deep video super-resolution. In: Proceedings of the IEEE international conference on computer vision, pp 4472–4480 https://doi.org/10.1109/iccv.2017.479
94. Tewari A, Elgharib M, Bharaj G et al (2020) Stylerig: rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6142–6151 https://doi.org/10.1109/cvpr42600.2020.00618
95. Thies J, Zollhofer M, Stamminger M et al (2016) Face2face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2387–2395 https://doi.org/10.1109/cvpr.2016.262
96. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. ACM Trans Graphics (TOG) 38(4):1–12 https://doi.org/10.1145/3306346.3323035
97. Vamsi VVVNS, Shet SS, Reddy SSM et al (2022) Deepfake detection in digital media forensics. Global Trans Proceed. https://doi.org/10.1016/j.gltp.2022.04.017
98. Verdoliva L (2020) Media forensics and deepfakes: an overview. IEEE J Select Topics Signal Process 14(5):910–93 https://doi.org/10.1109/jstsp.2020.3002101
99. Wang Y, Dantcheva A (2020) A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In: 2020 15Th IEEE international conference on automatic face and gesture recognition (FG 2020), IEEE, pp 515–519 https://doi.org/10.1109/fg47880.2020.00089
100. Welch BL (1947) The generalization of 'student's' problem when several different population varlances are involved. Biometrika 34(1–2):28–35 https://doi.org/10.1093/biomet/34.1-2.28
101. Westerlund M (2019) The emergence of deepfake technology: a review. Technology Innovation Management Review 9(11) https://doi.org/10.22215/timreview/1282
102. Wikipedia (2022) Receiver operating characteristic. https://en.wikipedia.org/wiki/Receiver_operating_characteristic, accessed: 2022-10-26
103. Wodajo D, Atnafu S (2021) Deepfake video detection using convolutional vision transformer. https://doi.org/10.48550/arXiv.2102.11126
104. Yang S, Chen B (2023) Effective surrogate gradient learning with high-order information bottleneck for spike-based machine intelligence. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2023.3329525
105. Yang S, Pang Y, Wang H et al (2023a) Spike-driven multi-scale learning with hybrid mechanisms of spiking dendrites. Neurocomputing 542:126240 https://doi.org/10.1016/j.neucom.2023.126240
106. Yang S, Wang H, Chen B (2023b) Sibols: robust and energy-efficient learning for spike-based machine intelligence in information bottleneck framework. IEEE Trans Cognitive Develop Syst https://doi.org/10.1109/TCDS.2023.3329532
107. Yin Q, Lu W, Li B et al (2023) Dynamic difference learning with spatio-temporal correlation for deepfake video detection. IEEE Trans Inf Forensics Secur. https://doi.org/10.1109/tifs.2023.3290752
108. Yu Y, Liu X, Ni R et al (2023) Pvass-mdd: predictive visual-audio alignment self-supervision for multimodal deepfake detection. IEEE Trans Circuits Syst Video Tech. https://doi.org/10.1109/tcsvt.2023.3309899
109. Zhang W, Zhao C, Li Y (2020a) A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. Entropy 22(2):249 https://doi.org/10.3390/e22020249
110. Zhang Y, Yin Z, Li Y et al (2020b) Celeba-spoof: large-scale face anti-spoofing dataset with rich annotations. In: European conference on computer vision, Springer, pp 70–85 https://doi.org/10.1007/978-3-030-58610-2_5
111. Zhao C, Wang C, Hu G et al (2023) Istvt: interpretable spatial-temporal video transformer for deepfake detection. IEEE Trans Inf Forensics Secur 18:1335–1348. https://doi.org/10.1109/tifs.2023.3239223
112. Zhu X, Wang H, Fei H et al (2021) Face forgery detection by 3d decomposition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2929–2939 https://doi.org/10.1109/cvpr46437.2021.00295

113. Zi B, Chang M, Chen J, et al (2020) Wilddeepfake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM international conference on multimedia, pp 2382–2390 https://doi.org/10.1145/3394171.3413769
114. Zubiaga A, Aker A, Bontcheva K et al (2018) Detection and resolution of rumours in social media: a survey. ACM Comput Surveys (CSUR) 51(2):1–36 https://doi.org/10.1145/3161603
115. Zucconi A (2018) Understanding the technology behind deepfakes. https://shorturl.at/ctGO1, Accessed 08 Aug 2022