MDPI

Article

Review Paper on Algorithms and Advancements in Detecting Deepfakes

Ángel Salazar [1,‡,*], Johel Cuasapaz [1,‡] and Iván Fierro [1,‡]

[1] School of Physical Sciences and Nanotechnology, Yachay Tech University, Hacienda San José s/n, San Miguel de Urcuquí 100119, Ecuador; angel.salazar@yachaytech.edu.ec (A.S.); edison.cuasapaz@yachaytech.edu.ec (J.C.); ivan.fierro@yachaytech.edu.ec (I.F.)

[*] Correspondence: angel.salazar@yachaytech.edu.ec

[‡] These authors contributed equally to this work.

Abstract

Over the past decade, misinformation on social media has increased due to the proliferation of synthetic media, including video, audio, and images generated or modified by artificial intelligence (AI), commonly referred to as deepfakes. The prevalence of deepfakes has grown in parallel with advancements in AI technology. This review evaluates recent automated deep learning models for deepfake detection. It analyzes how various architectures, such as Convolutional Neural Networks (CNNs), Autoencoders, Generative Adversarial Networks (GANs), and Diffusion Models, distinguish deepfakes from authentic content in real time across large datasets. The review emphasizes the comparative accuracy of these models in scenarios involving manipulated images and fabricated audiovisual content.

Keywords: deepfake detection; machine learning; transfer learning; CNNs; GANs; synthetic media; artificial intelligence

1. Introduction

Currently, the exponential growth of generative technologies has radically transformed the digital media landscape. Deep learning-based tools, such as Generative Adversarial Networks (GANs) and large language models (ChatGPT, DeepSeek, Copilot, etc.), have enabled the creation of hyperrealistic synthetic content, widely known as deepfakes. Deepfakes are modified images and videos that are falsified with the malicious objective of creating misinformation and fraud. To generate such counterfeit videos, two neural networks are used: a generative network and a discriminative network with a FaceSwap technique [1,2]. The generative network creates fake images using an encoder and a decoder, while the discriminative network determines the authenticity of the newly generated images. The combination of these two networks is called Generative Adversarial Networks (GANs), proposed by Ian Goodfellow [3]. According to a report by DeepStrike (2025), the number of deepfake files increased from 500,000 in 2023 to more than 8 million in 2025, representing growth of over 1,500% [4]. Additionally, deepfake-related fraud attempts rose by 3,000% in 2023, with a 1,740% increase in North America alone. These statistics reflect not only a serious problem regarding the increase of deepfakes, but also ethical and legal challenges when attempting to remove this type of information. For example, in the American context: "The prospect of a government entity attempting to distinguish real news from fake news—and suppressing the latter—raises the First Amendment concerns described above in relation to election-lies laws" [5].

Received:
Revised:
Accepted:
Published:

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

A study by Jumio (2024) revealed that 60% of consumers were exposed to at least one deepfake video in the past year, while the human detection rate for high-quality manipulated videos remains critically low at approximately 24.5% [6]. These phenomena pose serious risks to public trust, journalistic integrity, and democratic processes, especially as this content becomes increasingly indistinguishable from authentic material. The OECD (2023) warns that deepfakes pose a serious threat as powerful tools for spreading disinformation that can undermine trust in democratic processes and increase social polarization [7]. The detection and mitigation of these digital threats have therefore become a critical challenge for researchers, journalists, and civil society organizations. For instance, UC Berkeley researchers developed an approach to transfer body movements from one person to another in video [8], while NVIDIA introduced a style-based generator architecture for GANs for synthetic image generation [9].

In Ecuador, the cybercrime bulletin from the Ministry of the Interior indicates that in 2023, news outlets recorded incidents in schools where students utilized AI to generate sexualized images of their peers, actions that constitute the distribution of child sexual content and cyberbullying. This underscores the necessity to enhance prevention measures and responses to digital risks exacerbated by AI. In 2024, the Attorney General recorded 8,724 ICT-related 'news of crime', including electronic fraud (1,030), computer falsification (105), child pornography involving minors (118), 105 cases of criminals exploiting deepfakes for cybercrime (identity theft and fraudulent calls), offers of sexual services with minors via electronic means (9), and online contact for sexual purposes with minors (116). Police statistics also note detentions for distribution of pornographic material to minors (1) and for offering sexual services with minors via electronic means (2). The bulletin further warns that criminal groups already deploy deepfakes to impersonate victims, extort, and even simulate kidnappings, deceiving families of missing migrants [10].

Since the creation of deepfakes, users no longer require expert knowledge or training; many applications are democratizing deepfake creation [11]. Therefore, deepfakes have shifted from being a novelty to causing a trust crisis because humans alone cannot accurately detect fake digital media. This is dangerous because deepfakes can reach an unprecedented audience through social media [12]. Although it has been shown that humans are still good at detecting fake-manipulated images, they fail more often when tested on recognizing deepfake videos. In fact, an online experiment carried out by Groh et al. (2021) [13], in which over 15,000 participants guessed which image was fake and which was real from a large dataset, showed that the accuracy of people recognizing fake images is around 88%. However, this is not the case for deepfake videos. According to Nils et al. (2021) [14], in their experiment involving 16 videos, participants' overall accuracy at distinguishing deepfakes from real videos was only 58%, i.e., barely above chance. Performance was asymmetric: while viewers were above chance on most authentic clips, they were at or below chance on the majority of deepfake clips, revealing a systematic vulnerability to synthetic media. Another problem that emerges in this kind of experiment is the bias due to the confidence of participants. Nils et al. (2021) [14] revealed that, despite being told the set contained a 50/50 mix, participants showed a strong bias toward "authentic" judgments (67% of responses), consistent with a conservative tendency to believe videos are real. Together, these results show that unaided human detection is unreliable, especially for identifying fakes, underscoring the need for robust algorithmic detectors.

The failure of humans to detect deepfakes provides additional motivation to move toward machine learning and deep learning models to automate deepfake detection [15]. The difference between machine learning and deep learning lies in the neural network architecture. Deep learning models are more powerful since they use hundreds to thousands of computational layers to train the model, while machine learning models typically use only one or two layers [13]. Over the last few years, many attempts have been made to detect deepfake videos, but the challenges are increasing due to the rise of new AI models [16]. In fact, in the study reported by Khan et al. (2021) [17], the accuracy of models for deepfake detection achieved only 65%. Although diverse models are used for deepfake detection, their evolution is slow, and due to their architecture, it has been difficult to continue improving their precision [18].

Although deepfakes cover various types of manipulated content, deepfake detection can be classified into three categories: Feature-Based Methods (Handcrafted Features), which are algorithms that identify artificially constructed features [19]; Deep Learning-Based Methods (Model-Based), which are models that focus on learning patterns from large datasets of original and fake videos to differentiate true from manipulated content [20]; and Hybrid Methods (Combining Feature-Based and Deep Learning Approaches), which combine both feature-based and model-based approaches [21]. Currently, deep learning methods, which represent mostly the second and third types of deepfake detection, are being used to effectively detect fake videos [22].

Furthermore, since there are large datasets and volumes of deepfake videos available, deep learning methods are becoming increasingly effective [23]. Over recent years, new applications for creating deepfakes have emerged; this software can combine video and audio at high resolution [24]. Recent studies suggest a new approach to confronting these technologies: mapping multiple modalities or features into a shared embedding space so that items expressing similar affective cues are located near one another. Affective cues are specific features that convey rich emotional and behavioral information to human observers, helping them distinguish between different emotions. This review focuses on the second and third types of fake detection, with the priority of analyzing how these models behave when they are tested with real-world data and challenges.

This literature review aims to provide the community with concise information in the area of deepfake detection. In light of this scenario, this research establishes the following objectives: to analyze the effectiveness of current machine learning approaches in detecting deepfakes and synthetic accounts in order to identify the models and algorithms with the highest precision and robustness; to examine and compare the reference datasets used to train and evaluate these models in order to determine their suitability and identify potential gaps; and to investigate the ethical implications arising from the use of automated systems for verifying and detecting manipulated content, considering aspects such as privacy and algorithmic biases. These objectives seek to provide the basis for the development of practical and ethically responsible verification tools.

This leads to the following research questions, which are aligned with the objectives outlined above:

RQ1: How effective are current machine learning approaches in detecting deepfakes and synthetic accounts?

RQ2: Which models or algorithms offer the highest accuracy and robustness in identifying manipulated content?

2. State of the Art

This section provides a comprehensive overview of how deepfake detection operates and its theoretical development. Research focused on deepfake detection emerged between 2018 and 2020. During these years, research focused on detecting visual inconsistencies and artifacts generated by GANs. Thus, the first efforts to detect deepfakes based on inconsistencies in the facial textures of images used CNNs. For analyzing videos, researchers explored Recurrent Neural Networks (RNNs) models [25]. During 2021 and 2022, detection techniques evolved by incorporating new architectures such as Transformers, Autoencoders, and Hybrid Models. In these years, researchers also explored audio, text, and multimodal detection methods, which helped create new models based on audio-visual detection capabilities. Despite having a variety of models for deepfake detection, the lack of accuracy against real data led recent studies to develop enhanced models such as DAG-FDD (Demographic-Agnostic Fair Deepfake Detection) and DAW-FDD (Demographic-Aware Fair Deepfake Detection) [25]. The trend of detection techniques indicates a rise since 2020. In this review, we analyze the most widely used detection techniques in the literature, which over the past five years have been machine learning and deep learning.

2.1. Machine Learning Methods

These methods are suitable because they allow for a tree-based machine learning approach, such as decision trees or random forests.

Support Vector Machine (SVM) A Support Vector Machine (SVM) is a computer algorithm that learns by example to assign labels to objects [26]. In the field of deepfakes, data such as images are input and characteristics are extracted to classify the data.

Logistic Regression (LR) With logistic regression, we apply the logit transformation to the probabilities, meaning that we have a linear model for the log-odds of success instead of the probability of success [27].

Multilayer Perceptron (MLP) As Almeida states [28], MLP is a neural network widely used in the detection of deepfakes. It is composed of inputs whose sum becomes a nonlinearity called an activation function. These are connected in a feedforward form, meaning they do not form loops. They are used because, as non-linear models, they can model complex relationships between different characteristics, although they require a large amount of data to be effective.

2.2. Deep Learning Methods

Zhang et al. [29] propose a GAN simulator, AutoGAN, which can simulate the artifacts produced by the common pipeline shared by several popular GAN models.

Convolutional Neural Network (CNN) CNNs are neural networks designed to process spatial data such as images. They are inspired by the biological nervous system and consist of a large number of computational nodes that are interconnected to learn collectively. They are used to detect pattern behavior within images, which allows specific image characteristics to be encoded in the architecture [30].

Recurrent Neural Networks (RNN) RNNs are a type of neural network used to detect patterns in data sequences, which are applicable to images if they are broken down into a series of parts and treated as a sequence [31].

Figure 1 summarizes the types of networks used to detect deepfakes.

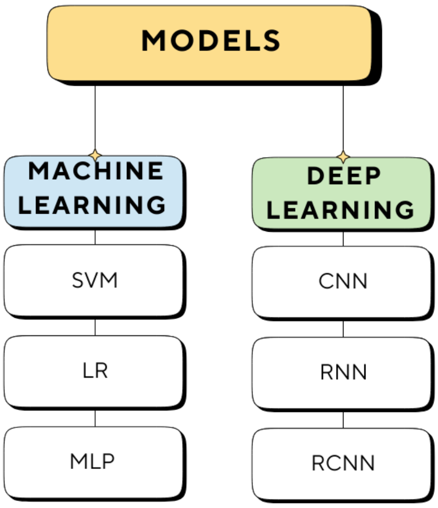

Figure 1. List of deepfake detection models.

Table 1. Overview of Machine Learning (ML) and Deep Learning (DL) approaches for Deepfake detection: type, main models, and datasets.

Type	Model	Dataset(s)	Reference
DL	CNN + Swin Transformer	FF++	[32]
DL	Conv-LSTM (CNN-based)	HOHA	[33]
ML	S-MIL, S-MIL-T	FF++, DFDC, Celeb	[34]
ML	Dynamic texture (F, B)	FF++	[35]
DL	Triplet loss network	FF++	[36]
DL	Metric learning models	DFDC	[37]
DL	DeepRhythm	FF++, DFDC	[38]
DL	DeepFakesON-Phys	Celeb-DF v2, DFDC	[39]
DL	Xception, Eff-B3, RNN, R3D, L3D	FF++	[40]
DL	Xception, ResNet 3D, Res2Net-101	Celeb-DF, FF++	[41]
DL	ID-Miner	Celeb-DF, VoxCeleb	[42]
DL	EfficientNet, XceptionNet, InceptionV3	DFDC	[43]
ML	SVM, LDA, KNN	FF++	[44]
ML	Xception with triplet loss	FF++, Celeb-DF	[45]
DL	Face X-ray	FF++	[46]
DL	AutoGAN detector	FF++	[29]

2.3. Datasets and Benchmarks

In recent years, the large-scale development and public use of artificial intelligence technologies have led to a significant increase in deepfakes. This, in turn, has created a pressing need to develop tools aimed at deepfake detection. However, information falsification methods evolve at a pace equal to or greater than the tools designed to detect them.

The primary tools currently available to combat deepfakes are large-scale deepfake datasets. These datasets store deepfake data to help detection tools identify them in real-world environments. Notable examples include the UADFV dataset, DeepFake-TIMIT, FaceForensics++, the Google DeepFake Detection Dataset (DFD), and the Facebook DeepFake Detection Challenge (DFDC). According to Li et al. (2020), several of these datasets exhibit variations in image quality, which causes problems when applying them in sophisticated environments. [46]

Building upon these datasets, other technologies can be employed for deepfake detection. According to Yan et al. (2023) [47], current deepfake detection can be divided into three categories: the naive detector, the spatial detector, and the frequency detector.

Naive Detector: This approach utilizes simple models that work with unprocessed video data, such as Convolutional Neural Networks (CNNs). These are considered simple models because they attempt to differentiate real from fake content based solely on direct visual differences present in the datasets.

Spatial Detector: These models explore differences and inconsistencies at the pixel level. They can detect incompatible edges, facial distortion, texture discontinuity, among other artifacts. However, they remain unreliable when post-processing is applied to the video or image.

Frequency Detector: This is a more sophisticated approach as it does not work with what is apparent to the human eye, but rather in the frequency domain. It uses tools like the Fourier Transform to detect abrupt changes in frequencies that are usually related to image or video manipulations.

Although databases may present challenges due to image resolution, it is not impossible to analyze them. For example, a 2020 study by Ciftci, Demir, and Yin demonstrated that their FakeCatcher detector, based on biological signals such as PPG maps and convolutional neural networks (CNN), successfully detected fake videos with high accuracy. The authors found that human physiological signals, particularly the spatial coherence and temporal consistency of these signals, are not adequately preserved in AI-generated content. Their method achieved up to 99.39% pairwise separation accuracy, maintaining robustness regardless of the resolution and image quality of the analyzed videos. This approach is innovative as it analyzes biological signals in synthetic portraits, establishing solid foundations for future research in deepfake detection. [48]

The training process through datasets is crucial, as it determines the reliability and accuracy with which a tool can detect a deepfake. For example, in 2020, a study conducted by Dolhansky et al. used the DFDC Dataset (a database created by Meta) as a reference for evaluating deepfake detectors. The authors found that models trained with this dataset achieved an average precision of 0.753 and a ROC-AUC score of 0.734, a metric that measures the model's ability to distinguish between real and fake content, even under varying lighting or video quality conditions. These results demonstrated that the diversity and quality of the dataset are determining factors for developing detectors with good generalization capability and ethical performance in real-world scenarios. [49]

2.4. Measurement Metrics

In this review, we consider two metrics to evaluate the performance of detection techniques reported by different authors. The first metric is Accuracy (AC), which is defined as the fraction of correctly predicted samples among the total number of samples that have been classified. The second metric is the Area Under the ROC Curve (AUC), where ROC stands for Receiver Operating Characteristic curve, which is used to measure the performance of binary classifiers. The ROC curve ranges from (0,0) to (1,1) on a 2D plane [50].

3. Materials and Methods

This research employed a structured methodology consisting of four primary phases: identification of pertinent literature, systematic screening, application of predefined eligibility criteria, and final selection of articles for inclusion. To conduct this systematic literature review, the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) were followed. This process is illustrated in Figure 2.

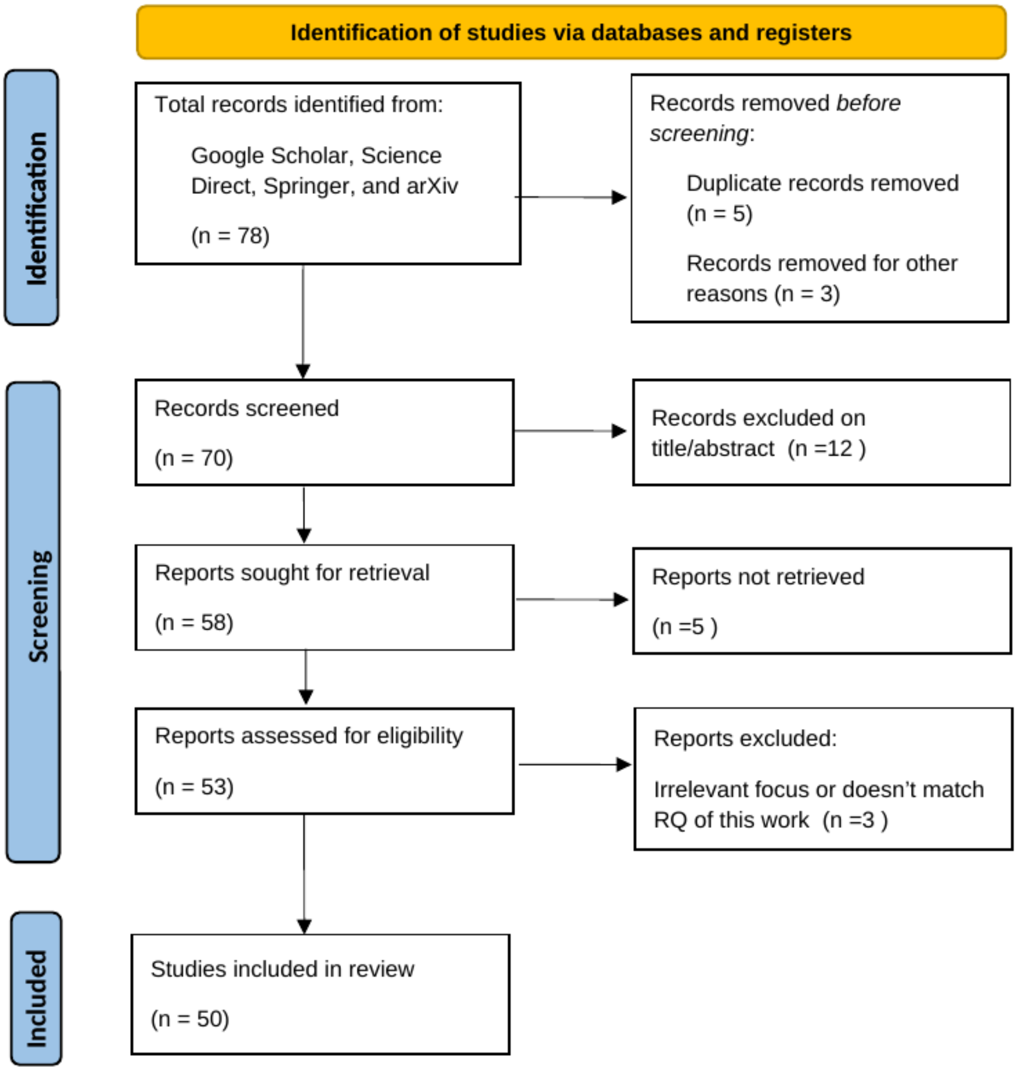
Identification of studies via databases and registers
Identification
Screening
Included
Total records identified from:
Google Scholar, Science Direct, Springer, and arXiv
(n = 78)
Records removed *before screening*:
Duplicate records removed (n = 5)
Records removed for other reasons (n = 3)
Records screened
(n = 70)
Records excluded on title/abstract (n =12)
Reports sought for retrieval
(n = 58)
Reports not retrieved
(n =5)
Reports assessed for eligibility
(n = 53)
Reports excluded:
Irrelevant focus or doesn't match RQ of this work (n =3)
Studies included in review
(n = 50)

Figure 2. PRISMA flow diagram for systematic review

3.1. Identification of pertinent literature

Approximately 78 articles were initially identified and compiled for the preliminary review list. The majority of these publications were sourced through a systematic search on Google Scholar, supplemented by targeted queries in databases such as arXiv, Springer and ScienceDirect. The search used certain keywords, such as "deepfake videos," "statistics of misinformation," "deepfake detection," "deepfake methods and challenges," and "deepfake future." Only articles published within the last five years were considered to ensure the currency and relevance of the reviewed literature. Eight articles did not meet this requirement and were eliminated, leaving a total of 70 articles.

3.2. Systematic screening

From the initial list, all articles were explored based on the topic of this review, and 12 of them were classified as irrelevant to this study. After this procedure, 58 research papers remained for the next filter.

3.3. Eligibility criteria

Research papers that were primarily focused on human detection of deepfakes or on the advancements of AI for content creation were excluded. Additionally, papers that test models and methods for detecting deepfakes but do not clearly show the accuracy of the method when tested with large datasets were not considered. Eight reports were identified in these categories and were excluded from the main list.

3.4. Selected Articles

The examination and filtering of the research studies resulted in 50 articles in the final list. It was ensured that the research papers are from the last five years so that this review analyzes the latest models of deepfake detection. All of these selected studies contain relevant details and results of different detection methods.

3.5. Data Extraction and Synthesis

In this phase, articles that passed the screening and eligibility criteria were studied. We identified the most relevant detection methods based on machine learning and deep learning. We also focused on the following components: datasets used by the study authors, methodologies applied, and the metrics used to evaluate detection techniques (in this review, AC and AUC). Finally, we synthesized the data by comparing findings from the data extraction process. We visualized the collected data through histograms, tables, and plots. The entire process is summarized in Figure 3.

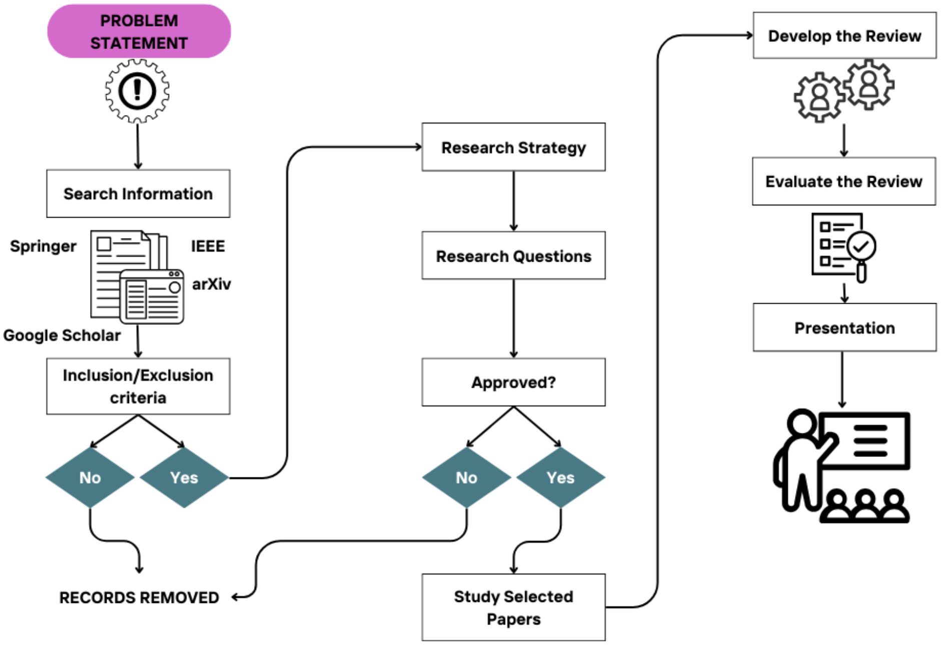

Figure 3. Flow diagram of the methodology followed in this review

4. Results

This section presents a comprehensive analysis of the performance metrics reported in the selected studies, examining both accuracy (AC) and area under the ROC curve (AUC) across different model architectures and benchmark datasets. The results are structured to address our research questions regarding the effectiveness and robustness of current detection approaches.

4.1. Overall Performance Comparison: Deep Learning vs. Machine Learning

Based on the reported metrics in Tables 2 and 3, deep learning approaches clearly dominate the recent deepfake detection literature, both in terms of the number of models and performance superiority. As illustrated in Figure 4, deep learning models achieve an average accuracy of 94.2% and an average AUC of 95.8%, substantially outperforming machine learning methods, which obtain 73.0% and 78.3% respectively.

The performance gap between these two paradigms is statistically significant and practically meaningful. Many deep architectures—including CNN+Swin Transformer [32], S-MIL variants [34], DeepRhythm [38], DeepFakesON-Phys [39], and Res2Net-101 [41]—reach accuracies above 90% on at least one dataset, with several works reporting values close to or above 97%. In stark contrast, classical machine learning methods (SVM, LDA, KNN, Xception+Tri) consistently obtain lower scores on FaceForensics++ (FF++), with accuracies ranging from 60% to 83%. This disparity can be attributed to deep learning's superior capacity for automatic feature extraction and hierarchical representation learning, which proves crucial for detecting subtle manipulation artifacts that traditional handcrafted features may miss.

Table 2. Accuracy (AC) reported for each model and dataset.

Paper	Model	Dataset	Accuracy (reported)
[32]	CNN + Swin (small)	FF++	95.60%
[32]	CNN + Swin (base)	FF++	95.62%
[33]	Conv-LSTM	HOHA	97.1%
[34]	S-MIL-T	FF++	97.14%
[34]	S-MIL	FF++	96.79%
[34]	S-MIL	DFDC	83.78%
[34]	S-MIL-T	DFDC	85.11%
[34]	S-MIL	Celeb	99.23%
[34]	S-MIL-T	Celeb	98.84%
[35]	F	FF++	69.76%
[35]	B	FF++	71.67%
[36]	Triplets	FF++	86.74%
[38]	DeepRhythm	FF++	98.0%
[38]	DeepRhythm	DFDC	64.1%
[39]	DeepFakesON-Phys	Celeb-DF v2	98.7%
[39]	DeepFakesON-Phys	DFDC	94.4%
[40]	Xcept	FF++	24.28%
[40]	Eff-B3	FF++	16.43%
[40]	RNN	FF++	18.57%
[40]	R3D	FF++	56.43%
[40]	L3D	FF++	78.57%
[41]	Xception	Celeb-DF	97.0%
[41]	ResNet 3D	Celeb-DF	97.0%
[41]	Res2Net-101	Celeb-DF	98.95%
[41]	Xception	FF++	91.05%
[41]	ResNet 3D	FF++	90.36%
[41]	Res2Net-101	FF++	93.48%
[42]	ID-Miner	Celeb-DF	77.4%
[42]	ID-Miner	VoxCeleb	86.5%
[43]	EfficientNet-B1	DFDC	97.63%
[43]	EfficientNet-B0	DFDC	96.24%
[43]	XceptionNet	DFDC	86.62%
[43]	InceptionV3	DFDC	92.07%
[44]	SVM	FF++	79.07%
[44]	LDA	FF++	83.33%
[44]	KNN	FF++	81.70%
[45]	Xception+Tri	FF++	61.3%
[45]	Xception+Tri	Celeb-DF	60.0%

Table 3. AUC reported for each model and dataset.

Paper	Model	Dataset	AUC (reported)
[32]	Hybrid CNN + Swin	FF++	95.62%
[34]	S-MIL	FF++	99.84%
[35]	F	FF++	76.68%
[35]	B	FF++	78.34%
[36]	Triplets	FF++	99.2%
[37]	MDS-based FD	DFDC	90.55%
[37]	Capsule	DFDC	53.30%
[37]	Meso4	DFDC	75.30%
[39]	DeepFakesON-Phys	Celeb-DF v2	99.9%
[39]	DeepFakesON-Phys	DFDC	98.2%
[41]	Xception	Celeb-DF	99%
[41]	ResNet 3D	Celeb-DF	99%
[41]	Res2Net-101	Celeb-DF	100%
[41]	Xception	FF++	96%
[41]	ResNet 3D	FF++	96%
[41]	Res2Net-101	FF++	97%
[43]	EfficientNet-B1	DFDC	97.63%
[43]	EfficientNet-B0	DFDC	96.24%
[43]	XceptionNet	DFDC	86.62%
[43]	InceptionV3	DFDC	92.07%

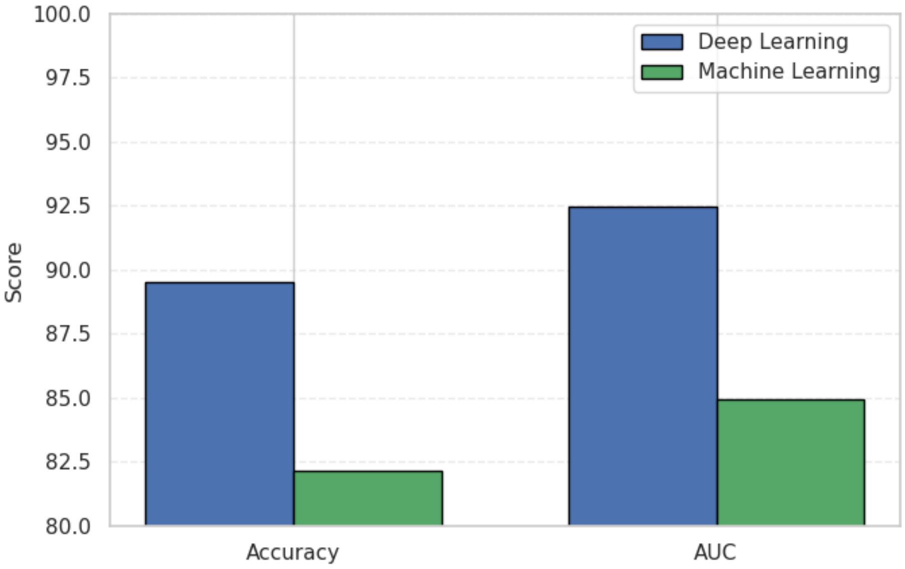

Figure 4. Average accuracy (AC) and area under the curve (AUC) for each model and dataset.

4.2. Model Architecture Analysis

The dispersion plot in Figure 5 provides critical insights into the distribution and consistency of detection performance across different architectures. For deep learning approaches, most data points cluster in the upper region of both graphs (AC $>$ 90%, AUC $>$ 95%), confirming that high performance is not an isolated phenomenon driven by a single architecture but rather represents a consistent pattern across diverse network families including CNNs, Transformers, and hybrid architectures. This clustering suggests that the field has reached a certain level of maturity in terms of baseline performance on standard benchmarks.

However, important variations exist even within deep learning methods. Notable outliers include some early Xception and RNN variants on FF++, which achieve only 16–24% accuracy [40], highlighting that not all deep architectures are equally suited for deepfake detection. These low-performing models typically rely on spatial features alone without considering temporal consistency, suggesting that effective deepfake detection requires both spatial and temporal analysis.

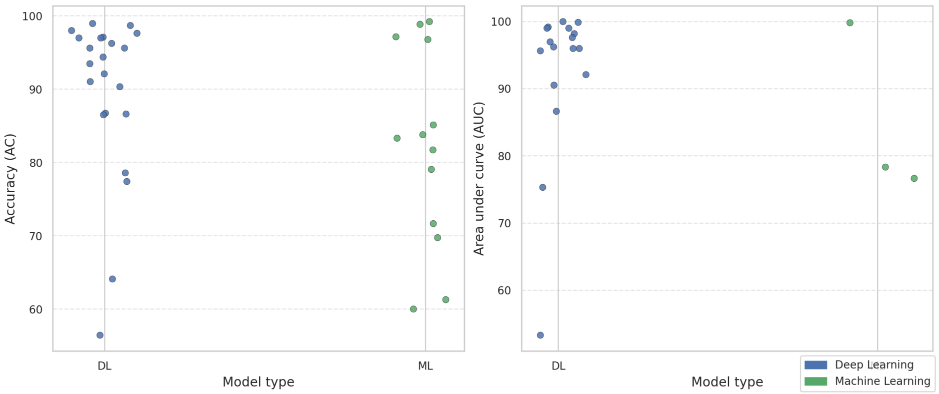

Figure 5. Dispersion of accuracy (AC) and area under the curve (AUC) for each model and dataset.

On the machine learning side, performance scores are concentrated in a noticeably lower band (60–83% accuracy), with no traditional ML model reaching the performance levels of the best deep networks in either accuracy or AUC. This limitation stems from their reliance on manually engineered features, which may not capture the complex, high-dimensional patterns that characterize synthetic media manipulations. The best-performing ML method (LDA) achieves 83.33% accuracy on FF++ [44], which, while respectable, still falls short of the 95–98% range achieved by state-of-the-art deep learning models. 298 299 300 301 302 303 304

4.3. Dataset-Specific Performance and Generalization Challenges 305

Figure 6 reveals a significant imbalance in the benchmark datasets used across the reviewed works. FF++ dominates the evaluation landscape, accounting for nearly half of all reported experiments (25/55, 45.5%), followed by DFDC (16/55, 29.1%) and Celeb-DF (10/55, 18.2%). This concentration on FF++ has important implications for interpreting reported performance metrics and understanding generalization capabilities. 306 307 308 309 310

The dataset distribution directly correlates with performance disparities observed across benchmarks. Models consistently exhibit higher accuracies and AUC values on FF++ and Celeb-DF compared to DFDC, suggesting that DFDC presents a more challenging and realistic detection scenario. This pattern manifests across multiple architectures: 311 312 313 314

- **DeepRhythm** [38]: 98.0% accuracy on FF++ vs. 64.1% on DFDC (33.9% drop) 315
- **S-MIL** [34]: 96.79% on FF++ vs. 83.78% on DFDC (13.0% drop) 316
- **DeepFakesON-Phys** [39]: 98.7% on Celeb-DF v2 vs. 94.4% on DFDC (4.3% drop) 317

These performance degradations, particularly the 33.9% drop in DeepRhythm, raise critical questions about overfitting to specific dataset characteristics and the real-world applicability of models trained and evaluated primarily on FF++. The DFDC dataset, created by Facebook with greater diversity in compression artifacts, lighting conditions, and demographic representation, appears to serve as a more rigorous test of model robustness [51]. 318 319 320 321 322 323

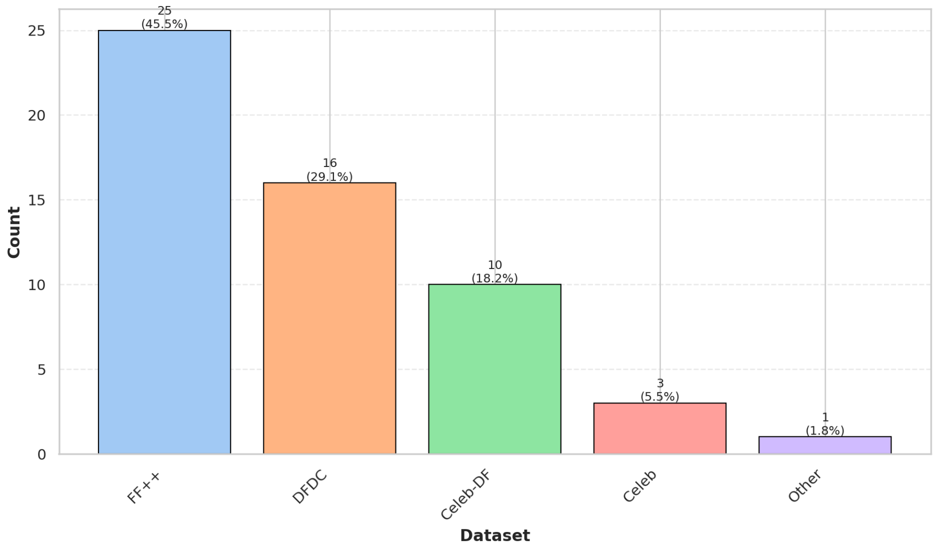

Figure 6. Distribution of datasets used in the reviewed works.

5. Discussion

5.1. Answering the Research Questions

Through systematic analysis of 50 studies (2020-2025), we obtained definitive answers to both research questions guiding this review.

RQ1: How effective are current machine learning approaches in detecting deepfakes?

Our analysis reveals a clear performance gap between traditional machine learning and deep learning approaches. Traditional ML methods (SVM, LDA, KNN) achieve limited effectiveness with accuracies between 60–83% on FaceForensics++, falling short of practical deployment requirements. In contrast, deep learning approaches demonstrate substantially higher effectiveness, with average accuracies of 94.2% and AUC scores of 95.8%. State-of-the-art models (Res2Net-101, DeepFakesON-Phys, S-MIL, EfficientNet-B1) consistently exceed 95% on multiple benchmarks. However, performance degrades significantly on diverse datasets like DFDC (drops of 4–34%), indicating that generalization to real-world scenarios with varied manipulation techniques and demographic diversity remains challenging.

RQ2: Which models offer the highest accuracy and robustness?

Four models emerge as top performers: Res2Net-101 [41] achieves the highest raw accuracy (98.95% on Celeb-DF, 100% AUC) through multi-scale feature extraction. DeepFakesON-Phys [39] demonstrates superior robustness with only 4.3% performance drop across datasets by incorporating physiological signals via rPPG. EfficientNet-B1 [43] offers optimal accuracy-efficiency balance (97.63% on DFDC), while S-MIL variants [34] achieve exceptional AUC scores (99.84%) through multiple instance learning.

Common success factors include: (i) spatio-temporal feature integration, (ii) attention mechanisms for facial regions, (iii) multi-scale extraction, and (iv) multimodal inputs. Purely spatial models achieve only 16–56% accuracy, confirming the necessity of comprehensive spatio-temporal modeling.

5.2. Performance Trends and Dataset Challenges

Deep learning models demonstrate clear superiority (90–98% success) over traditional ML approaches (60–83%). However, significant cross-dataset variation reveals generalization challenges. Performance on FF++ (95–98%) drops substantially on DFDC (64–94%), reflecting DFDC's greater scale (124,000 vs 5,000 videos) and manipulation diversity. Models struggle to generalize beyond training conditions, highlighting the gap between benchmark and real-world performance.

5.3. Implications for Practice and Future Research

While deep learning establishes clear superiority, performance degradation across datasets (e.g., 33.9% drop for DeepRhythm) indicates critical gaps in real-world applicability. Future research should prioritize cross-dataset evaluation, domain adaptation, meta-learning, and adversarial training on diverse manipulation techniques. The success of multimodal approaches (DeepFakesON-Phys) suggests that incorporating signals beyond visual appearance, such as physiological patterns, offers a promising path toward more robust, evolution-resistant detection systems.

6. Conclusions

This systematic literature review of 50 studies (2020–2025) provides definitive answers to both research questions guiding this analysis of automated deepfake detection methods.

Deep learning demonstrates clear superiority over traditional machine learning approaches in detecting deepfakes. However, significant performance degradation occurs when models trained on one dataset are evaluated on others, revealing critical generalization challenges for real-world deployment where manipulation techniques and content diversity vary substantially.

Four models emerge as most effective: Res2Net-101 achieves the highest raw accuracy, DeepFakesON-Phys exhibits superior cross-dataset robustness, EfficientNet-B1 offers optimal accuracy-efficiency balance, and S-MIL variants demonstrate exceptional detection capabilities. Common success factors include spatio-temporal feature integration, attention mechanisms, multi-scale extraction, and multimodal inputs such as physiological signals.

While deepfake detection has advanced significantly through deep learning architectures incorporating temporal analysis and multi-scale features, cross-dataset performance variation highlights persistent limitations in real-world applicability. Future research should prioritize domain adaptation, meta-learning, and adversarial training to enhance generalization capabilities. The success of multimodal approaches suggests that incorporating signals beyond visual appearance offers a promising path toward robust detection systems capable of addressing evolving manipulation techniques.

Bibliographic References

1. Hui, J. How deep learning fakes videos (deepfake) and how to detect it. https://jonathan-hui.medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9, 2018. Accessed 2025-10-16.
2. Oberoi, G. Exploring deepfakes. https://goberoi.com/exploring-deepfakes-20c9947c22d9, 2018. Accessed 2025-10-16.
3. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 2672–2680.
4. DeepStrike. Deepfake Statistics 2025. *DeepStrike* **2025**. Accessed: 2024-11-20.
5. Chesney, R.; Citron, D. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal* **2019**. https://doi.org/10.2139/ssrn.3213954.
6. Jumio. 2024 Online Identity Study: Global Consumer Research. https://www.jumio.com/2024-identity-study/, 2024. Accessed 2025-10-16.
7. Organisation for Economic Co-operation and Development (OECD). Disinformation and misinformation. https://www.oecd.org/en/topics/disinformation-and-misinformation.html, 2023. Accessed 2025-10-16.
8. Chan, C.; Ginosar, S.; Zhou, T.; Efros, A.A. Everybody Dance Now. *arXiv preprint arXiv:1808.07371* **2018**.

9. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405. https://doi.org/10.1109/CVPR.2019.00453.
10. Ministerio del Interior del Ecuador, Dirección de Ciberdelitos. Boletín de Análisis de la Ciberdelincuencia: «La nueva era de la ciberdelincuencia, el lado oscuro de la Inteligencia Artificial». Boletín, Subsecretaría de Combate al Delito, Dirección de Ciberdelitos, Ministerio del Interior del Ecuador, Quito, Ecuador, 2025. Equipo responsable: Jorge Fernando Illescas Peña (Director) y colaboradores técnicos listados en el documento.
11. Nirkin, Y.; Keller, Y.; Hassner, T. Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7184–7193.
12. Diakopoulos, N.; Johnson, D. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New media & society* **2021**, *23*, 2072–2098.
13. Groh, M.; Epstein, Z.; Obradovich, N.; Cebrian, M.; Rahwan, I. Human detection of machine-manipulated media. *Communications of the ACM* **2021**, *64*, 40–47.
14. Köbis, N.C.; Doležalová, B.; Soraperra, I. Fooled twice: People cannot detect deepfakes but think they can. *Iscience* **2021**, *24*.
15. Diel, A.; Lalgi, T.; Schröter, I.C.; MacDorman, K.F.; Teufel, M.; Bäuerle, A. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports* **2024**, *16*, 100538.
16. She, H.; Hu, Y.; Liu, B.; Li, J.; Li, C.T. Using graph neural networks to improve generalization capability of the models for deepfake detection. *IEEE Transactions on Information Forensics and Security* **2024**.
17. Zhao, L. Event prediction in big data era: A systematic survey. arxiv preprint. *ArXivorg* **2020**.
18. Alrashoud, M. Deepfake video detection methods, approaches, and challenges. *Alexandria Engineering Journal* **2025**, *125*, 265–277.
19. Kaddar, B.; Fezza, S.A.; Hamidouche, W.; Akhtar, Z.; Hadid, A. On the effectiveness of handcrafted features for deepfake video detection. *Journal of Electronic Imaging* **2023**, *32*, 053033–053033.
20. Shah, Y.; Shah, P.; Patel, M.; Khamkar, C.; Kanani, P. Deep Learning model-based Multimedia forgery detection. In Proceedings of the 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 564–572. https://doi.org/10.1109/I-SMAC49090.2020.9243530.
21. Rajesh, N.; Prajwala, M.S.; Kumari, N.; Rayyan, M.; Ramachandra, A.C. Hybrid Model for Deepfake Detection. In Proceedings of the Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication; Tomar, A.; Malik, H.; Kumar, P.; Iqbal, A., Eds., Singapore, 2022; pp. 639–649.
22. Rafique, R.; Gantassi, R.; Amin, R.; Frnda, J.; Mustapha, A.; Alshehri, A.H. Deep fake detection and classification using error-level analysis and deep learning. *Scientific reports* **2023**, *13*, 7422.
23. Soudy, A.H.; Sayed, O.; Tag-Elser, H.; Ragab, R.; Mohsen, S.; Mostafa, T.; Abohany, A.A.; Slim, S.O. Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications* **2024**, *36*, 19759–19775.
24. Nguyen, T.T.; Nguyen, C.M.; Nguyen, D.T.; Nguyen, D.T.; Nahavandi, S. Deep learning for deepfkes creation and detection. *arXiv preprint arXiv:1909.11573* **2019**, *1*, 2.
25. Rana, M.S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake detection: A systematic literature review. *IEEE access* **2022**, *10*, 25494–25513.
26. Noble, W.S. What is a support vector machine? *Nature biotechnology* **2006**, *24*, 1565–1567.
27. Buis, M.L. Logistic regression: When can we do what we think we can do. *Unpublished note* **2017**, *2*.
28. Almeida, L.B. Multilayer perceptrons. In *Handbook of neural computation*; CRC Press, 2020; pp. C1–2.
29. Zhang, X.; Karaman, S.; Chang, S.F. Detecting and Simulating Artifacts in GAN Fake Images (Extended Version). *arXiv preprint arXiv:1907.06515* **2019**. Extended version.
30. O'shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* **2015**.

31. Schmidt, R.M. Recurrent neural networks (rnns): A gentle introduction and overview. *arXiv preprint arXiv:1912.05911* **2019**.
32. Wang, S.; Du, C.; Chen, Y. A New Deepfake Detection Method Based on Compound Scaling Dual-Stream Attention Network. *EAI Endorsed Transactions on Pervasive Health & Technology* **2024**, *10*.
33. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2018, pp. 1–6.
34. Guarnera, L.; Giudice, O.; Battiato, S. Fighting deepfake by exposing the convolutional traces on images. *IEEE access* **2020**, *8*, 165085–165098.
35. Bonomi, M.; Pasquini, C.; Boato, G. Dynamic texture analysis for detecting fake faces in video sequences. *Journal of visual communication and image representation* **2021**, *79*, 103239.
36. Durall, R.; Keuper, M.; Pfreundt, F.J.; Keuper, J. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686* **2019**.
37. Kumar, A.; Bhavsar, A. Detecting deepfakes with metric learning. *arXiv preprint arXiv:2003.08645* **2020**.
38. Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Liu, Y.; Zhao, J. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In Proceedings of the Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 4318–4327.
39. Hernandez-Ortega, J.; Tolosana, R.; Fierrez, J.; Morales, A. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400* **2020**.
40. Ganiyusufoglu, I.; Ngô, L.M.; Savov, N.; Karaoglu, S.; Gevers, T. Spatio-temporal features for generalized detection of deepfake videos. *arXiv preprint arXiv:2010.11844* **2020**.
41. Mishra, S.R.; Mohapatra, H.; Edalatpanah, S.A.; Gourisaria, M.K. Advanced deepfake detection leveraging swin transformer technology. *Engineering Review: Međunarodni časopis namijenjen publiciranju originalnih istraživanja s aspekta analize konstrukcija, materijala i novih tehnologija u području strojarstva, brodogradnje, temeljnih tehničkih znanosti, elektrotehnike, računarstva i građevinarstva* **2024**, *44*, 45–56.
42. Wang, W.H.; Yeh, C.Y.; Chen, H.W.; Yang, D.N.; Chen, M.S. In anticipation of perfect deepfake: Identity-anchored artifact-agnostic detection under rebalanced deepfake detection protocol. *arXiv preprint arXiv:2405.00483* **2024**.
43. Singh, A.; Saimbhi, A.; Singh, N.; Mittal, M. DeepFake video detection: a time-distributed approach. SN Comput Sci 1: 212, 2020.
44. Li, X.; Lang, Y.; Chen, Y.; Mao, X.; He, Y.; Wang, S.; Xue, H.; Lu, Q. Sharp multiple instance learning for deepfake video detection. In Proceedings of the Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1864–1872.
45. Feng, D.; Lu, X.; Lin, X. Deep detection for face manipulation. In Proceedings of the International conference on neural information processing. Springer, 2020, pp. 316–323.
46. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics **2020**. pp. 3207–3216. Accessed on [Fecha en la que accediste al documento, e.g., 22 de octubre de 2023].
47. Inoue, N.; Sakaguchi, K.; Dibia, C.; Clark, J.; Inui, K.; Tandon, N.; Fried, D.; Brahman, F.; Pyatkin, V.; Kim, J.; et al. WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs **2023**. *36*, 64339–64365. Accessed on [Fecha en la que accediste al documento, e.g., 23 de mayo de 2024].
48. Ciftci, U.A.; Demir, I.; Yin, L. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 8101–8115. https://doi.org/https://doi.org/10.48550/arXiv.1901.02212.
49. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of the IEEE International Conference on Computer Vision* **2019**, pp. 1–11. Preprint available at arXiv:2006.07397.
50. Altuncu, E.; Franqueira, V.N.; Li, S. Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data* **2024**, *7*, 1400024.
51. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* **2020**.