International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)

IJCI

المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

August 2024

A Systematic Review on Deep Fake Image Generation, Detection Techniques, Ethical Implications, and Overcoming Challenges

Weeam Khimi
Effat University, Kingdom of Saudi Arabia
wkhimi0003@stu.kau.edu.sa

Kholood Albarqi and Kendah Saif
King Abdulaziz University, Kingdom of Saudi Arabia

Salma Elhag
Associate Professor, King Abdulaziz University, Kingdom of Saudi Arabia

Abstract

Deepfake technology, rooted in sophisticated machine learning techniques, utilizes deep neural networks to create highly realistic fake content such as videos, audio recordings, and images. This technology has rapidly evolved due to advancements in deep learning models, computational power, and data availability. The ethical implications, social impact, misuse, and legal frameworks surrounding Deepfake technology have been extensively studied. Detection techniques using deep learning approaches have been developed to combat the challenges posed by Deepfake content. Recommendations for future research include enhancing detection techniques, integrating explainable AI, and exploring real-time detection systems. Industry and policy implications emphasize the need for robust detection technologies, comprehensive legal frameworks, and collaborative efforts to address ethical concerns and regulate Deepfake content.

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)
August 2024

This systematic review explores the landscape of deep fake image generation, detection techniques and challenges, in addition to ethical considerations. By synthesizing existing research, we aim to provide insights into deep-fake technology's advancements, limitations, and societal implications. This review underscores the urgent need for interdisciplinary collaboration and robust frameworks to address the multifaceted issues surrounding deep fakes in the digital age.

Keywords: Images Manipulation, Deepfake detection, Generative models, Ethical implications, Misuse, Legal frameworks, Industry implications.

1. Introduction

Deep fake technology represents a significant advancement in the field of artificial intelligence, specifically in the realm of synthetic media. This technology involves the use of deep learning algorithms to create highly realistic and often deceptive fake content, such as videos, audio recordings, or images. The term "deep fake" is derived from "deep learning" and "fake," highlighting its roots in sophisticated machine learning techniques.

- **Definition and Overview of Deep Fake Technology**

Deep fake technology relies on deep neural networks, a subset of machine learning algorithms inspired by the human brain's neural architecture. These networks are trained on vast amounts of data to understand patterns and nuances, allowing them to generate content that mimics the appearance and behavior of real media. The term is commonly associated with the manipulation of video and audio content, where individuals or objects can be convincingly superimposed or synthesized within existing footage. This technology has evolved rapidly, driven by advancements in deep learning models, increased computational power, and the availability of large

datasets. Generative models, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), play a crucial role in creating realistic deep fake content by generating new data that is indistinguishable from authentic media.

- **Significance and Pervasiveness in Various Domains**

The significance of deep fake technology extends across various domains, raising both opportunities and concerns. In the entertainment industry, deep fake applications have been used for digital doubles of actors, enabling filmmakers to recreate scenes seamlessly or rejuvenate actors for roles. However, the widespread use of deep fakes also poses challenges to the integrity of visual media, as the lines between reality and manipulation become increasingly blurred.

Beyond entertainment, deep fakes have implications in politics, journalism, and cybersecurity. Politicians and public figures may find themselves targets of manipulated content, leading to misinformation and potential damage to reputations. Journalistic integrity is threatened as deep fake technology can be exploited to create fabricated news stories or alter the context of real events. Additionally, the use of deep fakes in cybersecurity poses risks, as attackers can manipulate audio or video to impersonate individuals for malicious purposes. As deep fake technology continues to advance, society must grapple with the ethical, legal, and societal implications of its use. Striking a balance between leveraging its positive applications and mitigating the potential for misuse is crucial for navigating the evolving landscape of synthetic media.

- **Research questions**

1- What are the ethical implications of Deepfake?

2- What are used and proposed techniques to detect Deepfake?

3- What are the challenges of Deepfake detection development?

In order to answer previous questions, we conducted a systematic review methodology in which we collected sixty papers related to our subject based on specific criteria, then we synthesized every paper by extracting useful information and categorizing according to our main questions, finally, we analyzed our findings to give final answers followed by visualization for easier understanding.

The structure of this paper is as follows: Section one presents the abstract for an overview about the study. Section two presents an introduction to our topic. Section three presents the methodology including steps and summary of scientific research papers. Section four presents. Section five presents analysis and results of our synthesis. Section six presents conclusions and discussion including gaps and future research. Section seven presents other information related to the conduction of this study. Finally, section seven presents all the used references.

2. Methods

In our systematic review, we reviewed many papers, then narrowed it down to sixty papers. During our synthesis, we ensured that the selected papers met our standards. We filtered the date to get only recent papers between 2019 and 2024, as the technical field is rapidly changing; therefore, we need to cope with the most recent changes, so we excluded papers that were conducted before 2019, as well as papers that did not clearly mention the date of publication. Moreover, selected papers have been chosen according to their connectivity to our subject. As we have three main objectives, we ensured that every single paper could benefit us in one or more objectives, and we excluded all papers that showed weak connections. Papers were selected from the Google Scholar engine due to the wide variety provided and the filtration features that ease the process. Finally, we only selected papers that are written in English. As English is considered the dominant and standardized language

in the academic world, we excluded papers that were originally written in any other language, even if they were translated. Table 1 summarizes our inclusion and exclusion criteria, as shown:

Table 1. List of Inclusion and Exclusion Criteria

PICOS	Included	Excluded
Relativity	Related	Unrelated, Weak relation
Date	>= 2019	< 2019, Unmentioned
Source	Reliable	Unreliable
Language	English	Other languages

For our systematic review we depended on academic journals collected through search engines, which represent a combination of primary and secondary sources. Primary sources helped us highlight latest discoveries, making strong evidence for the purpose of these papers, and providing authoritative information. While secondary sources helped us to gain a more comprehensive perception about the subject and related aspects.

In a review of multiple papers examining ethical implications and effect on media trust that come along with the creation of Deepfake, tools used to generate Deepfake, methods investigated to detect Deepfake, and challenges associated with the development of Deepfake detection we used several terms to extract most related papers. We illustrated these terms using Boolean operators for clarification, in which OR connects similar terms while AND connects the flow of different terms, detailed as follows:

(Images* OR Pictures* OR frames* OR Videos) AND (Detection* OR Reveal* OR Discover* OR Find) AND (Fake* OR Manipulated* OR Forgery) AND (Technique*

OR Approach* OR Method) AND (Concerns* OR Issues* OR Implications) AND (Improve* OR Enhance* OR Mitigate) AND (Prevent* OR Protect* OR Preserve) AND (Attack* OR Violate* OR Abuse) AND (Ethics* OR Morals) AND (Challenges* OR Limitations* OR Obstacles) AND (Authenticity* OR Originality* OR Genuineness) In our pursuit to address the core research inquiries, we meticulously sifted through fifty papers, each dedicated to unraveling Deepfake ethical implications, detection techniques and challenges. These selected studies provide in-depth understanding of the Deepfake field, including implications, detection techniques and challenges. To comprehensively structure our findings, we organized the amassed information into distinct categories, namely objective, methodology, findings, and limitations or future areas of exploration. This strategic categorization facilitated a thorough investigation into our primary research queries, ensuring a comprehensive and detailed understanding of the landscape concerning Deepfake technology.

- **Comprehensive Overview: Ethical Implications, Detection Techniques and Challenges**

(1) Ethical Implications and Social Impact of Deep Fake Image Generation

A. Ethical Considerations in Deep Fake Technology

The collection of studies and research on Deepfake technology reveals a multitude of ethical concerns regarding its potential misuse and harm. Themes such as deception, manipulation, and the erosion of authenticity in media content recur throughout the literature [1], [2], [3], [5], [6], [8], [12], [13]. Ethical considerations, including consent, privacy, and the impact on public trust, emerge as central issues across various contexts [1], [2], [3], [6], [7], [8], [12], [13], [14], [15], [16], [17]. Notably, non-consensual use, particularly in areas like pornography, raises significant concerns about privacy invasion and reputational harm [1], [3], [4], [6],

[9], [10], [15]. Furthermore, scholars stress the need to address broader ethical implications, including those related to cybercrimes, terrorism, and political manipulation [3], [6], [9], [15]. The literature highlights the importance of ethical guidelines, regulatory frameworks, and interdisciplinary collaboration to mitigate risks, protect individuals, and ensure responsible use of Deepfake technology [3], [6], [11], [12], [54]. Challenges such as attribution techniques and content detection underscore the complexity of addressing these ethical concerns [11], [12]. In essence, the discourse surrounding Deepfakes emphasizes the urgent need for comprehensive strategies to navigate their ethical dimensions and safeguard societal well-being [52], [53]. Deepfake technology raises ethical concerns regarding the manipulation of audio and visual content. The creation and dissemination of Deepfakes without consent can lead to deception and harm. Ethical guidelines and regulations are essential to address the potential misuse of Deepfake technology and protect individuals and society from its negative impacts.

B. Social Impact on Individuals and Society

The array of studies and research provided delve into the profound social implications of Deepfake technology, highlighting its potential to disrupt trust, manipulate perceptions, and spread misinformation [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17][52],[53],[54]. Across these studies, a recurring theme emerges regarding the erosion of trust in media, institutions, and public figures due to the proliferation of Deepfakes. Moreover, the impact extends beyond traditional media as Deepfakes infiltrate social interactions and interpersonal relationships, affecting self-perception and memories. The misuse of Deepfake technology, including its role in spreading misinformation, inciting violence, and perpetuating non-consensual pornography, poses significant risks to individuals and society. Victims of Deepfake pornography, in particular, experience profound psychological and professional consequences, exacerbating gendered disparities in

online spaces. Additionally, the continuous exposure to Deepfakes can instill skepticism and confusion, challenging societal trust and mental well-being. It is evident that addressing the social impact of Deepfake technology requires interdisciplinary efforts, awareness campaigns, and educational initiatives to foster critical thinking and mitigate harm. Deepfake technology's social impact includes spreading fake news, manipulating public opinion, and undermining trust in media, necessitating a thorough understanding to address its negative consequences [52], [53], [Social Impact on Individuals and Society]. Deepfake technology can have a significant social impact by influencing public perception, spreading misinformation, and undermining trust. Individuals and society may face challenges in distinguishing between real and fake content, leading to confusion and potential harm. The proliferation of deep fakes can disrupt social dynamics, impact relationships, and have broader implications on media consumption and communication. Deepfake technology has the potential to manipulate information, influencing public opinion and eroding trust in media and online content. This manipulation can lead to misinformation and deception, impacting individuals and society at large. Researchers and policymakers are actively working to develop strategies and tools to protect against the harmful effects of Deepfakes, aiming to safeguard the integrity of information and public discourse. Being targeted by Deepfakes can lead to loss of trust, credibility, and potential harm to personal and professional lives. The social impact of Deepfakes underscores the need for awareness and ethical considerations in their creation and use [54].

C. Misuse and Malicious Applications of Deepfakes

The compilation of studies and research underscores the wide-ranging potential for malicious misuse of Deepfake technology across various domains, from politics to personal relationships. Malicious actors exploit Deepfakes for purposes such as political manipulation, fraud, harassment, and spreading misinformation, posing

significant threats to national security, public safety, and individual well-being [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17],[52],[53],[54]. Victims of malicious Deepfakes experience reputational damage, emotional distress, and personal harm, with particular vulnerabilities highlighted in cases such as non-consensual sexual imagery. The ease of access to Deepfake creation tools exacerbates these risks, amplifying concerns about the widespread dissemination and impact of malicious content. Privacy infringements, defamation, and psychological harm are among the grave consequences associated with the misuse of Deepfake technology. To combat these negative ramifications, a multi-pronged approach encompassing awareness campaigns, detection tools, regulatory frameworks, and technological solutions is imperative to mitigate the harmful effects of malicious Deepfakes and safeguard societal well-being. Deepfakes can be maliciously used to create false and damaging content, leading to defamation and misinformation. Addressing this misuse is crucial to prevent harm and protect the integrity of information and media [52], [53], [54]. Deep fakes can be misused for spreading disinformation, manipulating public opinion, and creating fake news. Individuals and organizations may use deep fake technology to deceive, defame, or manipulate others for personal or political gain. The potential harm, discord, and erosion of trust in media and information are significant concerns. Deepfakes can be misused for spreading fake news, creating revenge pornography, and manipulating images or videos for malicious purposes. The ability to deceive through fake content poses a significant threat, highlighting the importance of addressing the negative implications of Deepfake technologies. Establishing comprehensive legal frameworks and ethical guidelines is crucial to prevent misuse and protect individuals and society from the harmful effects of Deepfakes [52], [53], Misuse and Malicious Applications of Deepfakes: Deepfakes can be used for spreading misinformation, defamation, and manipulation. Malicious actors may use Deepfakes

to deceive, intimidate, or blackmail individuals. The misuse of Deepfakes poses significant threats to privacy, security, and the integrity of information in various contexts. The misuse of Deepfake technology poses significant risks, including identity theft, the spread of misinformation, and reputational damage. Malicious actors can exploit Deepfakes to create deceptive content that can be used for fraudulent purposes or to manipulate public perception. Detecting and countering Deepfakes is essential to prevent their misuse and protect individuals and organizations from potential harm [3], [6].

D. Legal and Regulatory Frameworks Addressing Deepfake Concerns

The collection of studies and research underscores the pressing need for legal and regulatory frameworks to address the multifaceted challenges posed by Deepfake technology [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17],[52],[53],[54]. These frameworks are essential in combating the malicious creation and dissemination of Deepfakes, holding accountable those responsible for their production, and safeguarding individuals and society from the detrimental effects they can inflict. Efforts toward establishing comprehensive legal measures encompass criminalizing the creation and distribution of Deepfakes without consent, imposing penalties on offenders, and developing guidelines for their detection and mitigation [1], [3], [4], [6], [7], [8], [11], [12], [13], [15], [16], [17]. Collaboration among governments, regulatory bodies, technology experts, and law enforcement agencies is imperative to develop effective strategies and ensure international cooperation in combating the multifarious threats posed by Deepfakes [1], [2], [7], [16]. Furthermore, the scope of these efforts extends beyond mere prevention, encompassing the protection of individuals' fundamental rights to privacy, consent, and digital self-representation [2], [8], [10], [15]. Legislative endeavors also emphasize the importance of transparency, ethical use of generative modeling, and the implementation of robust detection tools to mitigate the deleterious impacts of

Deepfake technology [10], [11], [13], [14], [15], [16]. As the legal and regulatory landscape surrounding Deepfake technology continues to evolve, the development of comprehensive and adaptive regulations remains paramount to effectively distinguish between authentic and manipulated content, uphold individuals' rights, and ensure responsible use across diverse sectors [17]. Legal frameworks are crucial in addressing Deepfake concerns by establishing guidelines for their creation, distribution, and detection, thereby mitigating negative impacts and safeguarding individuals and society [52], [53], [54]. These frameworks aim to provide clear regulations and penalties to deter the misuse of Deepfakes while protecting individuals' rights. However, the legal framework for Deepfakes currently lags behind the rapid advancement of the technology, creating challenges in addressing concerns related to manipulation and misuse. Developing a comprehensive legal framework is crucial to enable Deepfake recognition software to effectively distinguish between fake and authentic content. Alongside legal regulations, ethical considerations are also essential to ensure the responsible use of Deepfake technologies and mitigate potential risks associated with their misuse [54]. As Deepfake technology advances, legal and regulatory frameworks are evolving to address the challenges posed by its misuse. Issues related to authentication, verification, and the admissibility of Deepfake evidence in legal proceedings are being actively considered. Initiatives and collaborations between organizations, technology companies, and policymakers aim to develop effective detection methods and establish guidelines within legal boundaries to combat the negative impacts of Deepfakes [3]. Developing legal and regulatory frameworks is essential to address the challenges posed by Deepfake technology. Regulations can help deter malicious use, protect individuals from harm, and hold accountable those responsible for creating and disseminating Deepfakes. Establishing clear guidelines and

consequences for Deepfake misuse can contribute to mitigating the negative impact on individuals and society [6].

(2) Deep Learning Approaches for Deep Fake Detection

A. Current and proposed techniques

Study [22] proposed the FST-Matching Deepfake detection model which aims to identify highly-compressed altered videos by separating irrelevant features and disentangling source/target-irrelevant representations from visual concepts. The model achieved success in detecting manipulated compressed videos with a high accuracy and AUC values. Similarly, study [30] has also introduced the FST-Matching Deepfake Detection Model to enhance detection performance by analyzing artifact-relevant features and image matching. The model improved forgery detection on compressed videos, showcasing an average AUC of 97.0% across various scenarios, which is considered a high rate. Another study [57] introduced two complementary face recognition networks to obtain identity cues for the face and its context by leveraging deep neural networks for face identification. These networks are designed to focus on specific facial regions, specifically the segmented face and its surrounding context. Results show that the proposed method achieves the best AUC scores on all benchmarks and exhibits improved generalization abilities compared to baseline methods with an AUC of 0.7555 for FSGAN manipulation method and 0.9262 for 3DMM-based swap. One study [23] conducted an experiment on preserving picture authenticity through the use of watermarking. The process involved two steps: first, embedding watermarks into facial features using a neural network with an encoder and decoder to make images face-swap sensitive, and second, verifying the watermark's presence to determine authenticity of the image. This method showed effectiveness with an average detection accuracy exceeding 80%. Another study [29] introduced the FaceGuard framework, utilizing

deep-learning-based watermarking to embed semi-fragile watermarks into real face images. FaceGuard was trained then tested on datasets containing real and Deepfake images, demonstrating high effectiveness in detecting Deepfakes with an accuracy rate of 99.5% overall by embedding watermarks into real face images proactively.

Study [25] introduced a detection method called Disruptive Technique (DDPM) to address data poisoning in training data by implementing purification steps using a denoising diffusion probabilistic model. Experimental validation showed an enhanced detection accuracy and robustness in identifying Deepfakes even when the dataset is poisoned, achieving accuracy rates ranging from 11.24% to 45.72% compared to traditional methods, even in scenarios that have 100% poisoned data. Moreover, study [40] presented NoiseScope, a blind detection method for Deepfakes that does not require prior knowledge of generative models or access to fake images. NoiseScope utilizes deep neural networks to analyze unique patterns from generative models when creating fake images. It achieved over 90% F1 score in detecting fake images across diverse datasets and generative models. NoiseScope demonstrated effectiveness in various scenarios, including compressed images and different post-processing techniques, showcasing resilience against countermeasures like fingerprint spoofing attacks and attempts to evade detection by adapting GAN models.

Study [34] introduced Pair-wise Self-Consistency Learning (PCL) to extract source features from Deepfake images by assessing consistency within image patches. The study also introduced the Inconsistency Image Generator (I2G) for generating forged images with annotated manipulated regions, achieving high performance with AUC scores of 99.11% to 99.98%. Also, study [36] presented the Common Fake Feature Network (CFFN), a framework for detecting fake faces and general images generated by GANs. CFFN utilized pairwise learning as well, cross-layer feature representations, and contrastive loss to capture discriminative features across

different GANs. Experimental evaluations on CelebA and ILSVRC12 datasets demonstrated CFFN's superior performance in fake face and general image detection, with precision and recall rates of 0.936 to 0.930.

Study [41] introduces a CNN-based image forgery detection system that focuses on variations in image compression to detect different types of image forgeries. Trained on the CASIA 2.0 database, the model achieves high accuracy in distinguishing genuine from tampered images, outperforming existing methods with an accuracy of 92.23% on test images. Another study [45] presents a method for detecting Deepfake images using DL techniques, combining Error Level Analysis and CNNs, by utilizing pre-trained CNN models like Alex Net and Shuffle Net, the study achieves accuracies of 86.1% to 88.2% with SVM and KNN classifiers, demonstrating the efficacy of DL in efficiently detecting Deepfakes. Study [48] introduces an evolutionary learning algorithm for automatic creation of CNN architectures for Deepfake detection. By utilizing genetic algorithms to generate diverse CNN structures and optimizing parameters, the model achieves high accuracies of 98.45% to 99.75% on different datasets, surpassing existing architectures and demonstrating effectiveness in detecting manipulations without extensive preprocessing. Also, study [50] presents a method for detecting Deepfake-forged content using a Separable CNN and image segmentation techniques. Trained on FaceForensics++ and tested on DeepFaceLab and StyleGAN images, the model achieves high accuracy and AUC values, outperforming state-of-the-art methods, and demonstrating effectiveness in detecting Deepfake content.

Study [24] proposed two methods, DAG-FDD and DAW-FDD, for Deepfake detection to enhance fairness. DAG-FDD targets Deepfakes without demographic annotations, while DAW-FDD focuses on considering annotations, using conditional Value-at-Risk (CVaR) and demographic factors like ethnicity and age to reduce bias and disparity between groups. Another study [27] introduced the Deepfake Disrupter

algorithm to detect and disrupt fake images by adding imperceptible perturbations, showing improvements in disrupting Deepfake methods compared to baseline techniques, increasing F1-score by 10% to 20% and achieving high success rates in detecting real and perturbed inputs. Study [37] suggested a deep learning-based approach for Deepfake detection using transfer learning techniques like Xception, NAS-Net, Mobile Net, and VGG16, achieving a high accuracy rate of 94% by analyzing facial attributes to identify anomalies indicative of Deepfake manipulation. And study [38] investigated the use of Discrete Cosine Transform (DCT) in Deepfake image recognition using GANs, showing that DCT-transformed images were linearly separable, enabling a simple linear classifier to achieve 100% accuracy and outperform classifiers trained on raw pixels, with greater resistance to image perturbations except for noise

Study [42] introduces the DeepfakeStack model for detecting manipulated videos and images using deep ensemble learning techniques. By combining multiple base-learners pretrained on ImageNet weights, the model achieves significant performance improvement, with precision, recall, and F1-score values close to 1.0. The DeepfakeStackClassifier (DFC) model attains high accuracy and AUROC scores, surpassing individual deep learning models in detecting Deepfakes. This approach demonstrates effectiveness in accurately identifying manipulated multimedia content, providing a robust basis for real-time Deepfake detection systems. Another study [43] presents a novel multi-attentional Deepfake detection framework that efficiently captures local features crucial for distinguishing between real and fake faces. The model integrates an Attention Module, texture enhancement block, and Bilinear Attention Pooling to address subtle discrepancies in Deepfake videos. Evaluation results show high accuracy on FaceForensics++ and Celeb-DF datasets, highlighting the framework's robustness and accuracy in detecting Deepfake content across diverse datasets. Study [44] introduces a method for

detecting Deepfake images generated by various GANs using the Expectation Maximization (EM) algorithm. By extracting features from images produced by GAN architectures and employing classification with K-NN, SVM, and LDA classifiers, the model achieves impressive accuracy rates in differentiating between authentic and generated images. It effectively distinguishes Deepfakes from different GAN architectures based on their convolutional traces, showcasing its success in detecting manipulated content.

The study [46] introduces a Deepfake detection method that utilizes the Vision Transformer (ViT) model, fine-tuned on a balanced dataset of real and synthetic images. The ViT model, with patch embedding and self-attention mechanisms, achieves exceptional accuracy rates, reaching 100% in identifying synthetic images from StyleGAN and Snapchat filters. When tested on a combined dataset of real and synthetic images, the model consistently achieves high accuracy rates ranging from 99.66% to 100%, showcasing its robustness in Deepfake detection. In contrast, the study [47] presents ADD, an attention-based digital video authentication system for detecting Deepfakes. ADD focuses on facial regions in videos, extracting discriminative features using attention maps and data augmentation techniques. Evaluated on challenging datasets like Celeb-DF (V2) and WildDeepfake, ADD significantly enhances detection accuracy rates compared to baseline models. For instance, on Celeb-DF (V2), the ResNet architecture achieves a 98.37% detection accuracy rate with ADD, representing a substantial improvement. Similarly, on WildDeepfake, the Exception architecture attains an 80.13% accuracy rate with ADD, surpassing previous state-of-the-art methods. While in study [56] ResNet-Swish-Dense54 model was designed to effectively capture and analyze visual features in images, results proved improved detection accuracy and robustness against adversarial attacks across different datasets and manipulation types with an accuracy of 99.26%.

Furthermore, the Enhanced Model for Fake Image Detection (EMFID) proposed in the study [49] offers a comprehensive approach for identifying forged digital images. EMFID integrates image pre-processing, histogram-based and Discrete Wavelet Transform (DWT)-based feature extraction, and Convolutional Neural Network (CNN) classification. The model achieves impressive performance metrics on benchmark datasets like CASIA v1.0 and CelebA, demonstrating high sensitivity, specificity, precision, accuracy, and F-Measure. These results highlighted the robustness and efficacy of EMFID in distinguishing manipulated images from authentic ones. Study [51] introduces CD-Net, a novel framework for face forgery detection, utilizing deep learning techniques. CD-Net comprises two main components: DICM and IDM. DICM captures communal features across multiple frames, enhancing stability in detecting forgery patterns, while IDM adaptively adjusts discriminative centers based on individual instance features, improving detection accuracy. Extensive experiments on datasets like FF++, DFDC, and Celeb-DF v2 to show the effectiveness of CD-Net, achieving an AUC of 0.952 on FF++ and significant improvements in detection performance and stability metrics with DICM and IDM. While study [55] investigated and compared the effectiveness of handcrafted features (SIFT, HoG) and deep features (Xception, CNN+RNN) in detecting Deepfake videos. It was found that handcrafted features performance is poor and may not be suitable for detecting Deepfakes due to their limitations in capturing facial details. On the other side, deep learning methods showed high performance in Deepfake detection, especially Xception which achieved nearly perfect results with an accuracy of 98.77% in original Deepfake test. As shown in table 2, we gathered all detection techniques and algorithms among the synthesized studies.

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)

IJCI

المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

August 2024

Table 2. Deepfake Detection Techniques for Each Study

Technique / Study	DC-GAN	Convolutional Trees and Expectation Maximization	RESNET-18	EfficientNet Architecture	FST-Matching	watermarking	DAG-FDD	DAW-FDD	DDPM	XceptionNet	MesoNet	EfficientNet B7	EfficientNet B3	3D CNN	CNN + RNN	Deepfake Disrupter,	EN-B7 Selim	XN WM Team WM	EN-B3 WM Team WM	Sequence-Based Model	PCL	DenseNet	VGGNet	ResNet	CNN	CFFN	Mobile Net	NAS-Net	VGG16	FF-LBPH-DBN	NoiseScope	DeepfakeStack	Expectation Maximization	Error Level Analysis	Vision Transformer	ADD framework	EMFID	Discrete Wavelet Transform	CD-Net framework	CVM classification	Data Augmentation	PCA	unsupervised contrastive learning	ResNet-Swish-Dense54	Face Recognition Model
[18]	X	X	X	X																																									
[19]																																													
[20]																																													
[21]																																													
[22]					X																																								
[23]						X																																							
[24]							X	X																																					
[25]									X																																				
[26]										X	X	X	X	X	X																															
[27]																X																														
[28]																																													
[29]						X																																							
[30]					X																																								
[31]																																													
[32]																	X	X	X																										
[33]										X	X									X																									
[34]																					X																								
[35]																						X	X	X	X																				
[36]																					X					X																			
[37]										X																	X	X	X																
[38]																														X															
[39]																														X															
[40]																															X														
[41]																									X																				
[42]																																X													
[43]																																													
[44]																																	X												
[45]																																		X											
[46]																																			X										
[47]																																				X									
[48]																						X		X	X																				
[49]																																					X	X							
[50]																									X																				
[51]																																							X						
[55]										X					X																														
[56]																																												X	
[57]																																													X
[58]																									X																X	X	X			
[59]																																												X		
[60]																									X																				
Total	1	1	1	1	2	2	1	1	1	4	2	1	1	1	2	1	1	1	1	1	2	2	1	2	6	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Percentage	2.50%	2.50%	2.50%	2.50%	5.00%	5.00%	2.50%	2.50%	2.50%	10.00%	5.00%	2.50%	2.50%	2.50%	5.00%	2.50%	2.50%	2.50%	2.50%	2.50%	5.00%	5.00%	2.50%	5.00%	15.00%	2.50%	2.50%	2.50%	2.50%	5.00%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%

B. Evaluation Metrics for Deepfake Detection

The evaluation of various studies on Deepfake detection techniques involves a comprehensive range of metrics to assess performance and reliability of detection

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)
IJCI
August 2024

models. As shown in table 3, we gathered all evaluation metrics used among the studies.

Table 3. Evaluation Metrics for Each Study

Tool / Study	Precision	Recall	F1- score	AUC	FNR	FPR	TPR	ASR	DR	PSNR	SSIM	CONFUSION	ROC	Cross-Validation	PIPS	CSIM	Shapely Value	Stability Analysis	ACC	Metric Q	Transferability Rate	L*	H.264 Quantization Factor	Average Percision	EER	specificity	geometric mean	DCT	sensitivity	MAE	RMSE	K-NN	LDA	T-SNE	PUP	CR	AUC-ROC	F2 - score	Total Score	Threshold-based Trade-off	MAP
[18]	X	X	X								X																														
[19]																																									
[20]							X				X				X	X																									
[21]				X	X	X		X																																	
[22]																	X	X																							
[23]			X	X					X	X	X								X																					
[24]				X		X	X												X																						
[25]	X	X	X	X								X	X	X					X																						
[26]					X	X													X																						
[27]	X	X	X																																						
[28]										X	X								X																						
[29]					X	X											X		X																						
[30]				X													X		X	X																					
[31]																																									
[32]								X				X									X																				
[33]								X											X			X	X																		
[34]				X																				X	X																
[35]	X	X	X	X															X																						
[36]	X	X																																							
[37]	X	X	X																X						X	X															
[38]																												X												
[39]				X					X	X									X						X	X		X		X	X										
[40]	X	X	X																																						
[41]	X	X	X																X																						
[42]		X	X										X						X																						
[43]				X															X																						
[44]																																	X	X							
[45]	X	X	X																X																						
[46]	X	X	X																X																						
[47]													X																												
[48]																			X						X																
[49]	X		X																X										X												
[50]				X									X																						X						
[51]				X															X																	X	X				
[55]	X	X	X																X																		X				
[56]	X	X	X																X																						
[57]	X	X	X																																		X		X	X	X
[58]	X	X	X																X					X													X				
[59]				X																																					
[60]	X	X	X																X																						
Total	16	16	17	12	3	4	2	3	2	3	4	2	4	1	1	1	3	1	23	1	1	1	1	2	3	2	1	2	1	1	1	1	1	1	1	2	1	1	1	1
Percentage	40.00%	40.00%	42.50%	30.00%	7.50%	10.00%	5.00%	7.50%	5.00%	7.50%	10.00%	5.00%	10.00%	2.50%	2.50%	2.50%	7.50%	2.50%	57.50%	2.50%	2.50%	2.50%	2.50%	5.00%	7.50%	5.00%	2.50%	5.00%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	5.00%	2.50%	2.50%	2.50%	2.50%

International Journal of Computers and Informatics, London Vol (3), No (8), 2024
https://doi.org/10.59992/IJCI.2024.v3n8p3 E-ISSN 2976-9361

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)

IJCI

المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

August 2024

C. Benchmark Datasets for Deep Fake Detection

Several studies have contributed to the advancement of detecting fake images and videos using machine learning tools and innovative techniques. As shown in table 4, we gathered all used datasets among studies.

Table 4. Datasets for Each Study

Study \ Dataset	CelebA	FFHQ	DFDC	DFD	DF	WDF	DeeperForensics-1.0	FaceForensics++/FF++/c23	Celeb-DF	VoxCeleb	VoxCeleb2	ForgeryNet	C40	CelebA-HQ	LUSN	WildDeepFake	All-in-One-DF	Trump Cage	Nicolas Cage	CD2	DFDC-P	StyleGAN	Flickr	ILSVRC12	100K-Faces	DFFD	CASIA-WebFace	StyleGAN-Face1	StyleGAN-Face2	PGGAN-Face	CycleGAN-Zebra	CASIA 2.0	STYLEGAN2	GDWCT	ATTGAN	RFFD	HHF	CASIA v1.0	Celeb-DF v2	UADF and DFTMIT	Real and Fake Face-Detection	140K	Deeper Forensics	UADFV	STARGAN
[18]	x	x																																											
[19]			x		x	x																																							
[20]								x																																					
[21]								x	x	x		x																																	
[22]			x					x	x				x																																
[23]	x													x																																
[24]			x	x				x	x																																				
[25]	x			x			x				x				x	x																													
[26]			x					x																																					
[27]	x									x																																			
[28]																	x																												
[29]			x					x										x	x																										
[30]								x																																					
[31]																																													
[32]			x																																										
[33]			x					x																																					
[34]								x												x	x																								
[35]																						x	x																						
[36]	x																							x																					
[37]									x																																				
[38]	x	x													x																														
[39]		x																							x	x	x																		
[40]																												x	x	x	x														
[41]																																x													
[42]								x																																					
[43]			x					x	x																																				
[44]	x																					x											x	x	x									
[45]																																				x									
[46]																						x																						
[47]									x							x																													
[48]																																					x						x			
[49]	x																																				x								
[50]								x																																					
[51]			x				x	x																																x					
[55]								x																																					
[56]			x					x	x																																				
[57]			x					x																																x					
[58]			x						x																																	x	x	x	x		
[59]								x	x																																				x		
[60]	x	x																				x																							x
Total	9	4	12	2	1	1	2	17	9	2	1	1	1	1	2	2	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	2	1	1	2	1	1	1
Percentage	22.50%	10.00%	30.00%	5.00%	2.50%	2.50%	5.00%	42.50%	22.50%	5.00%	2.50%	2.50%	2.50%	2.50%	5.00%	5.00%	2.50%	2.50%	2.50%	2.50%	2.50%	10.00%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	5.00%	5.00%	5.00%	2.50%	2.50%	2.50%	5.00%	2.50%	2.50%	5.00%	2.50%	2.50%	2.50%

(3) Challenges and Limitations of Deep Fake Detection Techniques

A. Machine Learning Generated Threats

A study [18] highlighted machine learning-generated threats, including Adversarial Perturbation Attacks, in which imperceptible perturbations are added to manipulate input data. Frequency Domain Manipulation involves eliminating manipulation traces in the frequency domain of Deepfake content to bypass detection methods. Image Filtering Techniques apply filters or transformations to fake images. Additionally, concerns about the robustness and security of machine learning systems have been raised, including vulnerabilities to adversarial attacks or manipulation, as mentioned in Study [60]. These challenges underscore the importance of addressing ethical, technical, and societal considerations in the development and deployment of machine learning technologies to mitigate potential risks and ensure responsible use.

Another study introduced Fakepolisher [19], aiming to enhance Deepfake evasiveness through shallow reconstruction techniques, altering pixel values and color gradients to confuse detection mechanisms. Black-box attacks by Denoising Diffusion Models (DDMs) [20] deceive detection systems through conditional image synthesis and guided post-processing. Adversarial attacks [21] manipulate Deepfake content to evade detection by exploiting vulnerabilities in detection models. Various attack scenarios [26] include White-Box Attacks, where adversaries with complete model knowledge craft imperceptible modifications to bypass detectors. Transferable Attacks allow adversarial examples to deceive different detection models. Universal Adversarial Perturbations are highly transferable perturbations added to any input data. A trace removal attack method, TR-Net [28], enhances Deepfake spuriousness and evades detection by forensic detectors.

The impact of adversarial examples on deceiving machine learning models and gradient-based attacks on CNN-based Deepfake detection systems are discussed [32]. Adversarial perturbations designed to fool one model can deceive other unseen CNN-based detection methods. Adversarial attacks [33] demonstrate the ability to bypassing Deepfake detectors by adversarially modifying fake videos. The emerging threat landscape of machine learning-generated content [36] focuses on fake images produced by advanced Generative Adversarial Networks (GANs), posing challenges in image forgery detection.

Concerns from machine learning-generated threats, particularly Deepfake images, have escalated [38]. The ability to create convincing fake images raises apprehensions about their misuse in spreading misinformation, identity theft, and privacy violations. Machine Learning Generated Threats (MLGTs) [40] encompass risks associated with generative models like GANs, facilitating various malicious activities. Addressing these challenges is crucial to mitigate the detrimental effects of machine learning-generated threats.

One notable challenge mentioned in Study [58] is the potential for biased or unfair outcomes produced by machine learning algorithms. These biases can arise from the data used to train the algorithms, leading to discriminatory or inaccurate results, particularly in sensitive domains such as healthcare or criminal justice. Another challenge highlighted is the issue of interpretability and explainability of machine learning models, as discussed in Study [59]. Complex algorithms, such as deep learning neural networks, may produce highly accurate predictions but lack transparency in how they arrive at those decisions, making it difficult for humans to understand or trust the outcomes.

B. Other Practical Challenges

Study [18] illustrates that current detection techniques face limitations in requiring extensive datasets for training, leading to resource-intensive demands and a lack of universal applicability [19]. The sophistication of Deepfake generating models continuously enhances the realism of fake images, making them visually convincing and harder to detect, thereby exposing vulnerabilities in existing detection systems [20]. Detection methods struggle with poor generalization abilities across different Deepfake types, impacting detection performance and susceptibility to transfer attacks [21]. While effective on familiar datasets, detection models encounter difficulties when faced with new data containing novel Deepfakes, emphasizing the challenge of generalizing detection systems to unseen datasets [22]. Additionally, computational overhead and the need for further development to adapt to various Deepfake types beyond facial features present challenges in Deepfake detection [23].

Limited datasets for training pose a challenge in ensuring the accuracy and effectiveness of detection models, especially with the rapid evolution of Deepfake techniques and adversarial attacks [24]. Data poisoning, affecting the integrity of training datasets, leading to decreased accuracy in detection, emphasizing the need to enhance generalization capabilities for real-world applicability [25]. Assessing the robustness of adversarial examples in compressed videos and evaluating transferability among different Deepfake detectors are key challenges in the field [26]. The presence of subtle visual artifacts and synthesis signals in Deepfake content complicates the differentiation between authentic and manipulated media, posing challenges for detection processes [27]. The complexity of trace removal mechanisms, aimed at enhancing the authenticity of manipulated content while evading detection, presents a significant challenge in Deepfake detection [28]. Robustness to post-processing operations, such as compression and resizing, is crucial for maintaining detection accuracy amidst various image distortions [29].

Deepfake detection models face performance drops when tested on compressed images or videos, highlighting the challenge of maintaining accuracy under compression and understanding artifact features learned by these models [30].

In [31], the rapid advancements in generative AI pose a challenge in keeping pace with evolving Deepfake techniques, potentially eroding public trust in media authenticity. Accessibility to Deepfake creation tools by a wide user base further complicates detection efforts, while biases in training data can propagate harmful stereotypes through generated content. Adversarial attacks, transferability across models, neural network complexity, and temporal consistency are key challenges outlined in [32], emphasizing the need for robust detection mechanisms capable of adapting to diverse datasets and sophisticated manipulation techniques.

Adversarial examples overriding detection mechanisms, challenges in classifying compressed videos, and exploiting temporal dependencies in manipulated content are highlighted in [33], underscoring the complexity introduced by adversarial attacks and compression artifacts. Training Pair-wise Self-Consistency Learning (PCL) for Deepfake detection, as discussed in [34], requires detailed annotations and diverse training data, posing computational and annotation challenges. Issues related to dataset diversity, model robustness, generalization, and computational complexity are addressed in [35], emphasizing the need for fine-tuning models to minimize false positives and negatives while ensuring scalability for real-time applications.

The diverse characteristics of fake images generated by various GANs, lack of transparency in technical details, and challenges in dataset curation are discussed in [36], highlighting the importance of robust learning strategies for effective detection. The complexity of Deepfake generation processes, sophistication of techniques, and demand for accurate detection methods in digital forensics and cybersecurity are emphasized in [37], indicating the necessity of addressing computational complexity

and dataset handling challenges. Vulnerability to adversarial examples, limitations in mitigating artifacts, and the need for ongoing research and development to enhance resilience are key challenges outlined in [38], emphasizing the continuous effort that is required to improve Deepfake detection models' effectiveness and reliability.

Study [39] faces challenges with substantial computational resources needed for training on large datasets and adapting to various Deepfake manipulations and image quality variations. Another study [40] focuses on enhancing NoiseScope's resilience against countermeasures like JPEG compression and denoising attacks. Practical challenges in Deepfake detection models [41] include optimizing performance across diverse computational resources and addressing hardware limitations. Challenges in a different model [42] include sensitivity to training data quality and diversity, robustness against emerging manipulation techniques, and deployment in real-world scenarios. Another study [43] highlights challenges in distinguishing subtle differences between real and fake faces and variability in textural patterns. Challenges in a different model [44] involve computational resources, interpretability of features, generalization across Deepfake generation techniques, and integration into forensic workflows. Common challenges discussed in a study [45] include adapting to evolving Deepfake techniques, robustness against adversarial attacks, computational complexities, and dataset diversity.

Practical challenges in Deepfake detection using the Vision Transformer (ViT) architecture [46] include high computational demands and the need for expertise in machine learning. Another study [48] faces issues with computational complexity, training time, and generalization to unseen data. The Enhanced Model for Fake Image Detection (EMFID) [49] grapples with computational complexity, dataset biases, and generalization across diverse image types. Challenges in a different study [50] involve data quality, model interpretability, and scalability in real-world

deployment. Addressing inconsistencies in forgery patterns is a challenge for a model [51], requiring understanding variations in fake images and refining discriminative centers.

in [55] accuracy of Deepfake detection models dropped dramatically when the training sets and test sets did not match during the experiment. Furthermore, a study [56] faced lack of robustness when introducing the model of unseen cases and adversarial attacks, in addition to lack of the explainability feature, which is essential for forensic analysts. Moreover, study [57] faces difficulties in acquiring high-quality and diverse datasets for training, in addition to difficulties in ensuring that detection techniques are used responsibly and do not infringe on privacy or civil liberties.

3. Analysis & Results

According to table 5 in which we have concluded fifty-one studies, which answer each research question, we specified all papers that were interested in ethical implications of deep fake, used and proposed detection techniques, and challenges related to development of detection techniques.

Table 5. Numbers of Relevant Studies for Each Research Question

Research question	Related studies	Number of studies	Total percentage
(1) What are the ethical implications of Deepfake?	[1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17], [52], [53], [54]	Twenty studies	33.33%
(2) What are used and proposed techniques to detect Deepfake?	[18], [22], [23], [24], [25], [26], [27], [29], [30], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [55], [56], [57], [58], [59], [60]	Thirty-five studies	58.33%
(3) What are the challenges of Deepfake detection development?	[18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [48], [49], [50], [55], [56], [57], [58], [59]	Thirty-seven studies	61.66%

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
August 2024
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

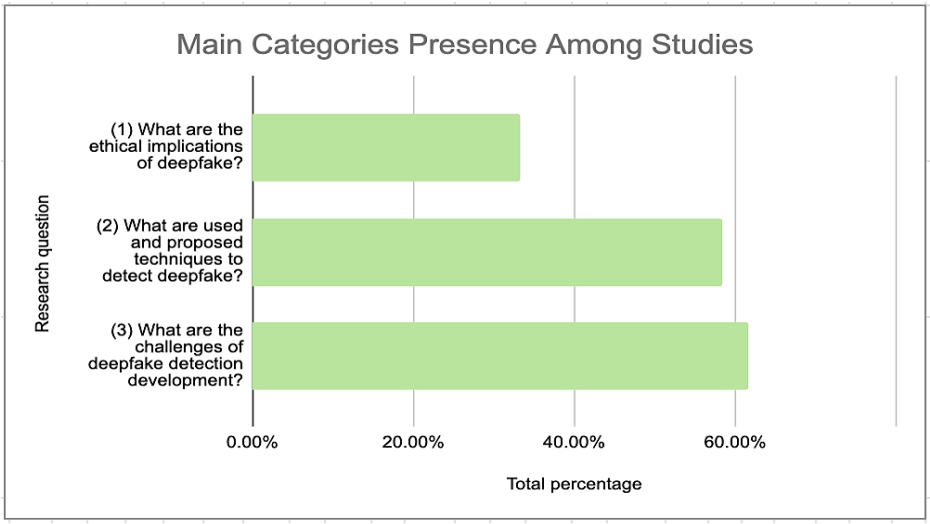

Figure 1. Main Categories Presence Among Synthesized Studies

As shown in figure 1, papers have been shown according to their relevance to the main aspects: (1) Ethical implications of Deepfake (2) used and proposed techniques to detect Deepfake (3) Challenges of Deepfake detection development. The figure illustrates that Twenty studies discussed ethical implications (33.33%), while twenty-nine studies discussed detection techniques (56.86%), and thirty-two studies discussed detection techniques development challenges (62.74%), noting that most papers which explored detection techniques discussed challenges and limitations as well, indicating transparency and providing a strong foundation for those who are willing to continue the study. While some other papers with a literature review nature discussed challenges as well, without proposing a detection technique or conducting any experimental validation, justifying why the last category was covered the most within synthesized papers.

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
August 2024
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

(1) Summary of Ethical Implications

According to table 6, we derived subcategories from the main category (1) The ethical implications of Deepfake, which included Ethical Considerations in Deepfake Technology, Social Impact on Individuals and Society, Misuse and Malicious Applications Deepfake, and Legal and Regulatory Frameworks Addressing Deepfake Concerns.

Table 6. Numbers of Relevant Studies for Each Sub-Category in Ethical Implications and Social Impact of Deepfake Image Generation

Ethical Implications and Social Impact of Deepfake Image Generation	Related Studies	Number of Studies	Total Percentage
Ethical Considerations in Deepfake Technology	[1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17], [52], [53], [54]	Twenty studies	25%
Social Impact on Individuals and Society	[1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17], [52], [53], [54]	Twenty studies	25%
Misuse and Malicious Applications Deepfake	[1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17], [52], [53], [54]	Twenty studies	25%
Legal and Regulatory Frameworks Addressing Deepfake Concerns	[1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17], [52], [53], [54]	Twenty studies	25%

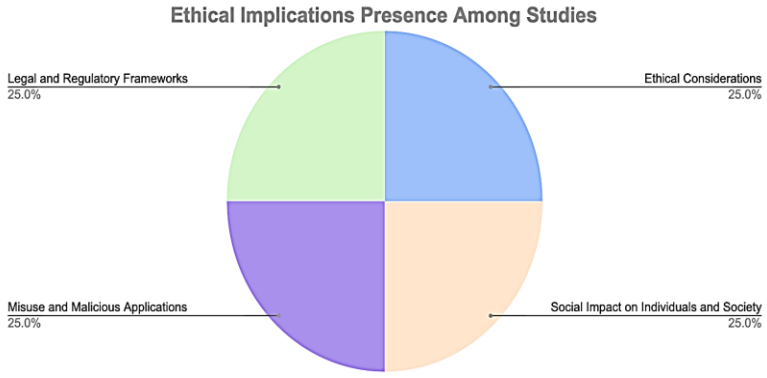

Figure 2. Ethical Implications Sub-Categories Presence Among Studies

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)
August 2024

As shown in figure 2, the twenty synthesized studies related to the first category (1) Ethical implications of Deepfake, had four sub-categories which were covered in all the twenty studies, which are: Legal and regulatory framework, ethical considerations, misuse and malicious applications, and social impact on individuals and society. Indicating that any study discussing ethical implications necessarily covered all sub-subjects, and providing us insights into the interconnection of these sub-subjects.

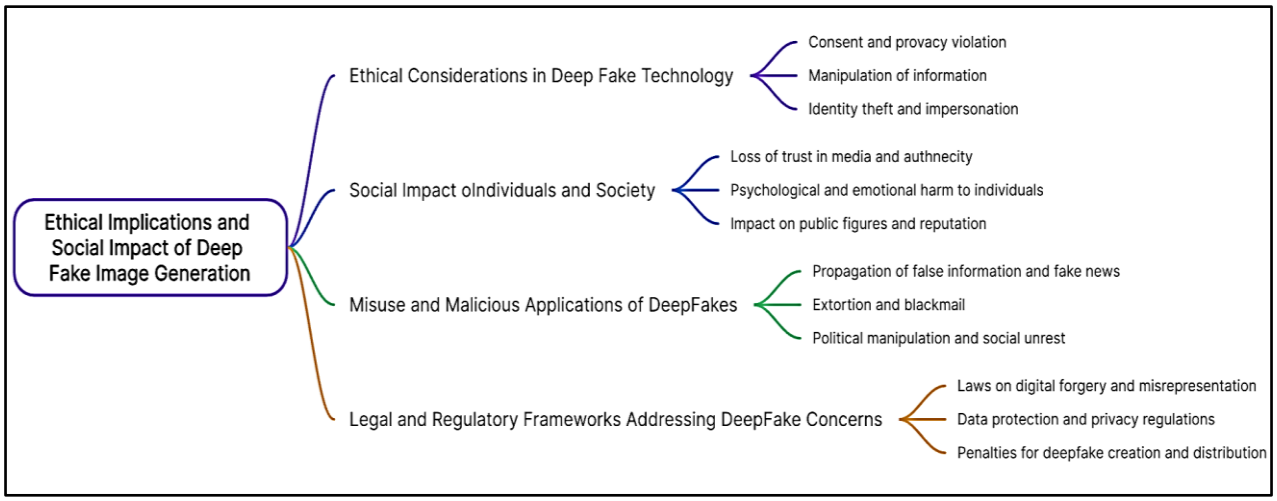

Figure 3. Summary of Ethical Implications and Social Impact of Deepfake

As shown in figure 3 below which summarizes all ethical implications, Deepfake technology presents ethical challenges, including concerns about consent, privacy, and the potential harm to individuals and society. It raises issues such as identity theft, non-consensual use for explicit content, and the erosion of public trust. The social impact involves the spread of misinformation, potential harm to personal and professional lives, and challenges in distinguishing real from fake content. Deepfakes can be misused for malicious purposes, posing threats to national security and individual well-being. Addressing these concerns requires legal and regulatory

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
August 2024
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

frameworks to criminalize malicious creation, hold creators accountable, and encourage international cooperation.

(2) Summary of detection approaches

A. Current and suggested detection techniques

Synthesis of the studies [34]-[60] illustrated a variety of detection techniques, some of them are currently applicable while the others are newly proposed as enhancements in this field, including forty-five different detection techniques and algorithms as shown in figure 4.

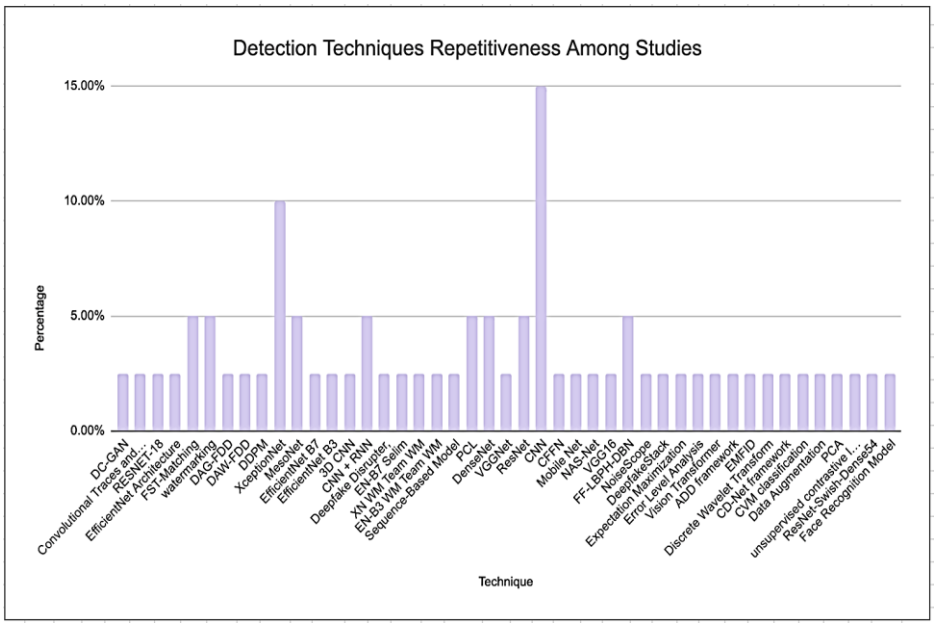

Figure 4. Detection Techniques Repetitiveness Among Studies

As shown in figure 4, results of our analysis demonstrate the percentage of repetitiveness of Detection Techniques Among Studies, it shows that the most used technique within deepfake detection field is Convolutional Neural Networks (CNN), while the least used techniques are DC-GAN, Convolutional Traces and Expectation-

Maximization, RESNET-18, EfficientNet Architecture, DAG-FDD, DAW-FDD, DDPM, EfficientNet B7, EfficientNet B3, 3D CNN, Deepfake Disrupter, EN-B7 Selim, XN WM Team WM, EN-B3 WM Team WM, Sequence-Based Model, VGGNet, CFFN, Mobile Net, NAS-Net, VGG16, NoiseScope, DeepfakeStack, Expectation Maximization, Error Level Analysis, Vision Transformer, ADD framework, EMFID, Discrete Wavelet Transform, CD-Net framework, CVM classification, Data Augmentation, PCA, unsupervised contrastive learning, ResNet-Swish-Dense54, and Face Recognition Model.

This analysis provides valuable insights into the prevailing detection techniques employed within the Deepfake research domain. It becomes evident that certain techniques are favored over others, largely due to their demonstrated effectiveness and applicability in identifying and mitigating the proliferation of Deepfake content. This preference may stem from various factors, including the robustness of the technique, its scalability to different types of Deepfake media, and its ability to adapt to evolving manipulation methods. By understanding the predominant use of specific detection techniques, researchers and practitioners can gain a clearer understanding of the current landscape and focus their efforts on refining and advancing these methods to address emerging challenges in Deepfake detection.

B. Summary of Evaluation Metrics

Synthesis of the studies [34]-[60] illustrated a significant use of accuracy measures to evaluate results, including forty-one different measurement metrics as shown in figure 5.

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
August 2024
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

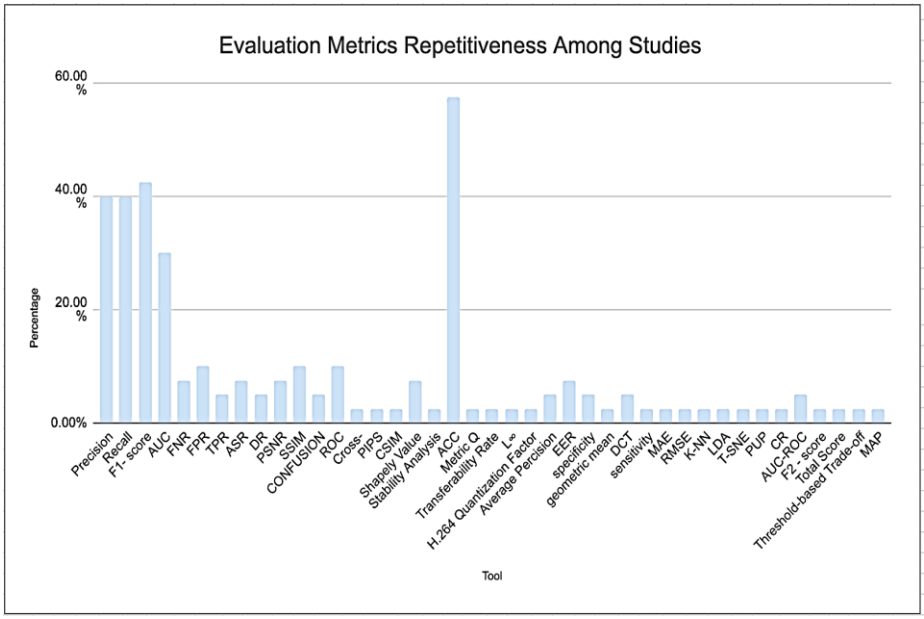

Figure 5. Evaluation Metrics Repetitiveness Among the Studies

As shown in figure 5, results of our analysis demonstrate the percentage of repetitiveness of evaluation metrics, it shows that the most used metrics within Deepfake detection field is Accuracy (ACC), while the least used metrics are Cross-Validation, Perceptual Image Patch Similarity (PIPS), Cosine Similarity Index Measure (CSIM), Stability Analysis, Metric Q, Transferability Rate, L∞, H.264 Quantization Factor, Average Precision, Geometric Mean, Sensitivity, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), K-NN, Support Vector Machine (SVM), LDA, T-SNE, Proportion of Unstable Predictions (PUP), and Correction Rate (CR).

Moreover, it can be noticed also that some studies used evaluation metrics more than the others, as the highest number of metrics used in one study is eight evaluation metrics, while the lowest number of used metrics is one only, some studies did not use any metrics due to its literature review nature. Consequently, the usage of

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
August 2024
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

evaluation metrics among a study aligns proportionally to its certainty, reliability and credibility, the more metrics used the better the results.

C. Summary of Datasets

Synthesis of the studies [34]-[60] illustrated the use of many different datasets to train detection models using deep learning techniques, including forty-four different datasets as shown in figure 6.

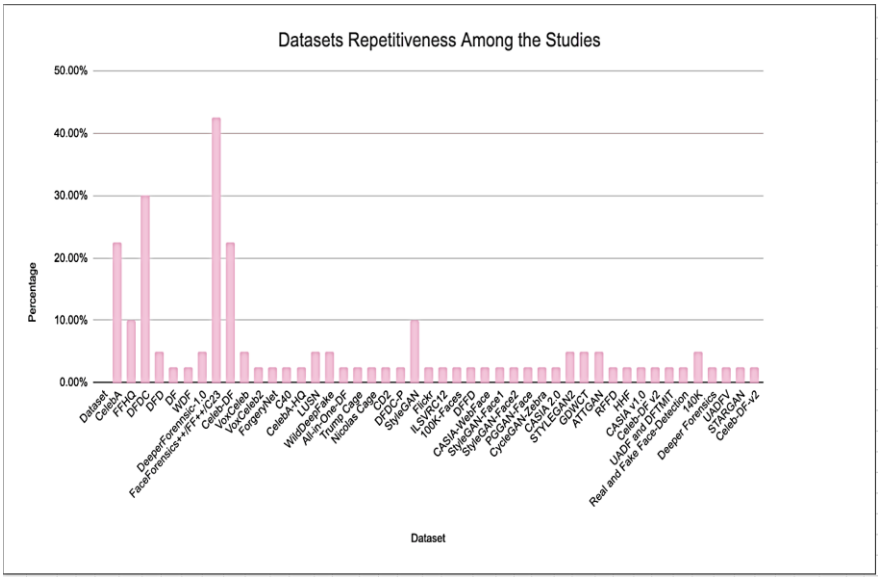

Figure 6. Datasets Repetitiveness Among Studies

As shown in figure 6, results of our analysis illustrate the percentage of repetitiveness of many different datasets among the synthesized studies, moreover, it illustrates that the most widely used dataset for models training is FaceForensics++ (FF++), while the least used datasets are DF, WDF, VoxCeleb2, ForgeryNet, C40, CelebA-HQ, All-in-One-DF, Trump Cage, Nicolas Cage, CD2, DFDC-P, Flickr, ILSVRC12, 100K-Faces, DFFD, CASIA-WebFace, StyleGAN-Face1, StyleGAN-Face2, PGGAN-Face, CycleGAN-Zebra, CASIA 2.0, RFFD, HHF, CASIA v1.0, UADF,

DFTMIT, Real and Fake Face-Detection, Deeper Forensics, UADFV, and STARGANK.

This analysis gives us insights about the mostly used datasets in researches within Deepfake field, as some datasets are preferable more than others mostly due to its large-scale as it contains a numerous number of images and videos, as well as the variety of characteristics it provide, making them a good option to train models and to enhance its capabilities effectively. Moreover, it can be noted that some studies used a number of datasets more than others, as the highest number of datasets used in one study is six, while the lowest number of used datasets is one only, some studies did not use any datasets due to its literature review nature. Furthermore, it can be said that the more used datasets the better the detection model abilities, performance and accuracy, in addition to that the number of datasets among a study aligns directly to its certainty, and inversely with the level of bias.

(3) Summary of challenges of Deepfake detection development

As shown in figure 7 below which summarizes all challenges and limitations of Deepfake detection techniques among synthesized studies, challenges were divided into machine learning generated threats, which refer to the advanced techniques built by machine learning specifically to complicate the detection process, including adversarial attacks, transferable attacks, trace removal attacks, FakePolisher and others. While other challenges focused on other obstacles related to the detection process development itself, including generalization abilities, intensive computational resources, the need for continuous evolution and adapting, the need for numerous datasets for training and others.

International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (8)
IJCI
August 2024
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (3)، العدد (8)

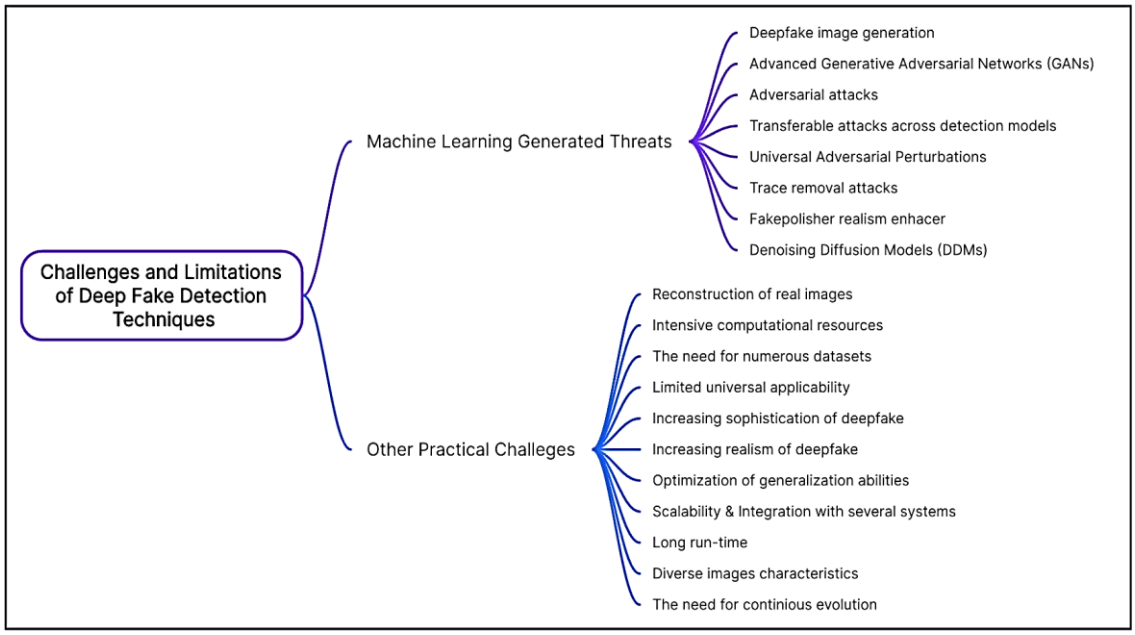

Figure 7. Summary of Challenges and Limitations of Deepfake Detection Techniques

4. Conclusion & Discussion

- **Summary of Key Findings**

In conclusion, our systematic review provided useful insights into Deepfake technology ethical implications and impact on society and media, detection techniques development and related challenges.

1- Ethical Implications of Deepfake: our review has highlighted many several ethical concerns that surround Deepfake technology, which include issues related to deception, manipulation and the decrease of authenticity in media content. Main themes like consent, privacy, as well as the impact of public trust have emerged as critical considerations along with the emergence of Deepfake. Moreover, non-consensual use especially in some sensitive areas raises significant concerns about individuals reputation harm.

2- Detection techniques for Deepfake: this review has explored and identified a wide range of techniques that are used or proposed for Deepfake content detection, leveraging different deep learning methods such as watermarking, FST-Matching and DAG-FDD. Moreover, evaluation metrics played an important role for verifying reliability of the results and assessing effectiveness of detection models, which included Accuracy, AUC, FPR, TPR, precision and many others. Furthermore, several datasets were utilized for the training and testing of detection models, which involves FF++, CelebA, FFHQ, DFDC and others.

3- Challenges Associated with Deepfake Detection Development: This review covered practical constraints faced during the exploration of Deepfake detection techniques, such as the need for extensive high-quality training and testing datasets, the need for intensive computational resources, and ensuring responsible use of detection techniques without violating privacy or civil liberties. These challenges reflect the nature of developing Deepfake detection strategies and emphasize the need to overcome difficulties in order to advance the efficacy and reliability of detection techniques, to mitigate the potential harm posed by this rapidly evolving technology.

- **Conclusion**

Deepfake technology presents significant ethical implications, social impact, and challenges that necessitate robust detection techniques and regulatory frameworks. The synthesis of relevant studies highlights the importance of addressing ethical concerns, enhancing detection methods, and overcoming related challenges to mitigate the risk of Deepfake content. Efforts to address issues related to the emergence of Deepfake are crucial to ensure the reliability, transparency, and ethical use of synthetic media in the digital age.

- **Recommendations for Future Research**

While our review provided valuable insights into this field, several gaps and areas for further research have been identified, which involves enhancing the robustness and scalability of Deepfake detection techniques to keep pace with evolving Deepfake generation methods. Moreover, it is important to investigate the integration of explainable AI techniques to improve the interpretability and transparency of Deepfake detection models. Furthermore, researchers must explore the development of real-time Deepfake detection systems to address the challenges of detecting rapidly evolving Deepfake content. We also suggest conducting studies on the impact of Deepfake technology on various sectors such as politics, journalism, and entertainment to understand its broader implications. Finally, we recommend collaboration across disciplines to develop comprehensive frameworks.

- **Implications for Industry and Policy**

Industry stakeholders need to invest in robust Deepfake detection technologies to safeguard their platforms and users from malicious activities. In addition to that, policymakers should prioritize the development of comprehensive legal frameworks to regulate the creation, distribution, and malicious use of Deepfake content. Furthermore, collaboration between industry, policymakers, and technology experts is crucial to establish guidelines and policies that address the ethical, legal, and societal implications of Deepfake technology. Finally, it is important to implement measures to enhance transparency, accountability, and trust in digital media platforms.

5. Other Information

In regard to additional information, this study was not officially published, therefore, it does not have a registration number. The study was not financially supported. The

authors declare no lack of conflicts of interest related to this study. Furthermore, no applicable Institutional Review Board (IRB) statement is provided, and the same for informed consent statements. Finally, there is no data availability statement as it is not applicable to this study.

References

[1] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1520.

[2] Hancock, J. T., & Bailenson, J. N. (2021). The social impact of Deepfakes. Cyberpsychology, behavior, and social networking, 24(3), 149-152.

[3] Arshed, M. A., Mumtaz, S., Ibrahim, M., Dewi, C., Tanveer, M., & Ahmed, S. (2024). Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model. Computers, 13(1), 31.

[4] Nightingale, S. J., & Wade, K. A. (2022). Identifying and minimising the impact of fake visual media: Current and future directions. Memory, Mind & Media, 1, e15.

[5] Hoque, M. A., Ferdous, M. S., Khan, M., & Tarkoma, S. (2021). Real, forged or deep fake? enabling the ground truth on the internet. IEEE Access, 9, 160471-160484.

[6] Ramachandran, V., Hardebolle, C., Kotluk, N., Ebrahimi, T., Riedl, R., & Jermann, P. (2023). A multimodal measurement of the impact of Deepfakes on the ethical reasoning and affective reactions of students.

[7] Al-Khazraji, S. H., Saleh, H. H., KHALID, A. I., & MISHKHAL, I. A. (2023). Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 23, 429-441.

[8] De Ruiter, A. (2021). The distinct wrong of Deepfakes. Philosophy & Technology, 34(4), 1311-1332.

[9] Jabiyev, B., Onaolapo, J., Stringhini, G., & Kirda, E. (2021, October). e-Game of FAME: Automatic Detection of FAke MEmes. In TTO (pp. 1-11).

[10] Li, M., & Wan, Y. (2023). Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of Deepfake information. Internet Research.

[11] De Ruiter, A. (2021). The distinct wrong of Deepfakes. Philosophy & Technology, 34(4), 1311-1332.

[12] Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1), e2110013119.

[13] Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. Computers and Education: Artificial Intelligence, 2, 100040.

[14] Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with Deepfakes–an interdisciplinary examination of the state of research and implications for communication studies. SCM Studies in Communication and Media, 10(1), 72-96.

[15] Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022, June). Limits and possibilities for "Ethical AI" in open source: A study of Deepfakes. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2035-2046).

[16] Wazid, M., Mishra, A. K., Mohd, N., & Das, A. K. (2024). A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact. Cyber Security and Applications, 100040.

[17] Neethirajan, S. (2021). Is seeing still believing? Leveraging Deepfake technology for livestock farming. Frontiers in Veterinary Science, 8, 740253.

[18] Guarnera, L., Giudice, O., Guarnera, F., Ortis, A., Puglisi, G., Paratore, A., ... & Battiato, S. (2022). The face Deepfake detection challenge. Journal of Imaging, 8(10), 263.

[19] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence, 53(4), 3974-4026.

[20] Ivanovska, M., & Struc, V. (2024). On the Vulnerability of Deepfake Detectors to Attacks Generated by Denoising Diffusion Models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1051-1060).

[21] Wang, L., Meng, X., Li, D., Zhang, X., Ji, S., & Guo, S. (2023). DeepfakeR: A Unified Evaluation Platform for Facial Deepfake and Detection Models. ACM Transactions on Privacy and Security.

[22] Dong, S., Wang, J., Liang, J., Fan, H., & Ji, R. (2022, October). Explaining Deepfake detection by analysing image matching. In European Conference on Computer Vision (pp. 18-35). Cham: Springer Nature Switzerland.

[23] Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive Deepfake defence via identity watermarking. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 4602-4611)

[24] Ju, Y., Hu, S., Jia, S., Chen, G. H., & Lyu, S. (2024). Improving Fairness in Deepfake Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4655-4665).

[25] Park, J., Park, L. H., Ahn, H. E., & Kwon, T. (2024). Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge. IEEE Access.

[26] Hussain, S., Neekhara, P., Dolhansky, B., Bitton, J., Ferrer, C. C., McAuley, J., & Koushanfar, F. (2022). Exposing vulnerabilities of Deepfake detection systems with robust attacks. Digital Threats: Research and Practice (DTRAP), 3(3), 1-23.

[27] Wang, X., Huang, J., Ma, S., Nepal, S., & Xu, C. (2022). Deepfake disrupter: The detector of Deepfake is my friend. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14920-14929).

[28] Liu, C., Chen, H., Zhu, T., Zhang, J., & Zhou, W. (2023). Making Deepfakes more spurious: evading deep face forgery detection via trace removal attack. IEEE Transactions on Dependable and Secure Computing.

[29] Yang, Y., Liang, C., He, H., Cao, X., & Gong, N. Z. (2021). Faceguard: Proactive Deepfake detection. arXiv preprint arXiv:2109.05673.

[30] Dong, S., Wang, J., Liang, J., Fan, H., & Ji, R. (2022, October). Explaining Deepfake detection by analysing image matching. In European Conference on Computer Vision (pp. 18-35). Cham: Springer Nature Switzerland.xa

[31] George, A. S., & George, A. H. (2023). Deepfakes: The Evolution of Hyper realistic Media Manipulation. Partners Universal Innovative Research Publication, 1(2), 58-74.

[32] Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, C. C. (2021). Adversarial threats to Deepfake detection: A practical perspective. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 923-932).

[33] Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021). Adversarial Deepfakes: Evaluating vulnerability of Deepfake detectors to adversarial examples. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 3348-3357).

[34] Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for Deepfake detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 15023-15033).

[35] Shad, H. S., Rizvee, M. M., Roza, N. T., Hoq, S. M., Monirujjaman Khan, M., Singh, A., ... & Bourouis, S. (2021). Comparative analysis of Deepfake image detection method using convolutional neural network. Computational Intelligence and Neuroscience, 2021.

[36] Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10(1), 370.

[37] Raza, A., Munir, K., & Almutairi, M. (2022). A novel deep learning approach for Deepfake image detection. Applied Sciences, 12(19), 9820.

[38] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020, November). Leveraging frequency analysis for deep fake image recognition. In International conference on machine learning (pp. 3247-3258). PMLR.

[39] Suganthi, S. T., et al. "Deep learning model for deep fake face recognition and detection." PeerJ Computer Science 8 (2022): e881.

[40] Pu, J., Mangaokar, N., Wang, B., Reddy, C. K., & Viswanath, B. (2020, December). Noisescope: Detecting Deepfake images in a blind setting. In Annual computer security applications conference (pp. 913-927).

[41] Ali, S. S., Ganapathi, I. I., Vu, N. S., Ali, S. D., Saxena, N., & Werghi, N. (2022). Image forgery detection using deep learning by recompressing images. Electronics, 11(3), 403.

[42] Rana, M. S., & Sung, A. H. (2020, August). Deepfakestack: A deep ensemble-based learning technique for Deepfake detection. In 2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom) (pp. 70-75). IEEE.

[43] Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional Deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194).

[44] 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. 666-667).

[45] Rafique, R., Nawaz, M., Kibriya, H., & Masood, M. (2021, November). Deepfake detection using error level analysis and deep learning. In 2021 4th International Conference on Computing & Information Sciences (ICCIS) (pp. 1-4). IEEE.

[46] Arshed, M. A., Alwadain, A., Faizan Ali, R., Mumtaz, S., Ibrahim, M., & Muneer, A. (2023). Unmasking Deception: Empowering Deepfake Detection with Vision Transformer Network. Mathematics, 11(17), 3710.

[47] khormali, A., & Yuan, J. S. (2021). Add: Attention-based Deepfake detection approach. Big Data and Cognitive Computing, 5(4), 49.

[48] Jasim, R. M., & Atia, T. S. (2023). An evolutionary-convolutional neural network for fake image detection. Indonesian Journal of Electrical Engineering and Computer Science, 29(3), 1657-1667.

[49] Sabitha, R., Aruna, A., Karthik, S., & Shanthini, J. (2021). Enhanced model for fake image detection (EMFID) using convolutional neural networks with histogram and wavelet based feature extractions. Pattern Recognition Letters, 152, 195-201.

[50] Yu, C. M., Chang, C. T., & Ti, Y. W. (2019). Detecting Deepfake-forged contents with separable convolutional neural network and image segmentation. arXiv preprint arXiv:1912.12184.

[51] Song, L., Fang, Z., Li, X., Dong, X., Jin, Z., Chen, Y., & Lyu, S. (2022, October). Adaptive face forgery detection in cross domain. In European Conference on Computer Vision (pp. 467-484). Cham: Springer Nature Switzerland.

[52] Ali, S., DiPaola, D., Lee, I., Hong, J., & Breazeal, C. (2021, May). Exploring generative models with middle school students. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-13).

[53] Allein, L., Moens, M. F., & Perrotta, D. (2023). Preventing profiling for ethical fake news detection. Information Processing & Management, 60(2), 103206.

[54] De Ruiter, A. (2021). The distinct wrong of Deepfakes. Philosophy & Technology, 34(4), 1311-1332.

[55] Xu, Y., & Yayilgan, S. Y. (2022). When Handcrafted Features and Deep Features Meet Mismatched Training and Test Sets for Deepfake Detection. arXiv preprint arXiv:2209.13289.

[56] Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. The Visual Computer, 39(12), 6323-6344.

[57] Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). Deepfake detection based on the discrepancy between the face and its context. arXiv preprint arXiv:2008.12262.

[58] Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy, 2(1), 89-106.

[59] Fung, S., Lu, X., Zhang, C., & Li, C. T. (2021, July). Deepfakeucl: Deepfake detection via unsupervised contrastive learning. In 2021 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

[60] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An improved dense cnn architecture for deepfake image detection. IEEE Access, 11, 22081-22095.