Deepfake Image Detection: A Comprehensive Review of Techniques, Challenges, And Future Directions

Neha Verma[1], Prof. Dr. D.N. Goswami[2], Prof. Dr. Anshu Chaturvedi[3]
[1,2]*School of Studies in Computer Science & Application, Jiwaji University, Gwalior*
[3]*Department of Computer Science and Engineering, MITS, Gwalior*

***Abstract*—In this research, the authors offer an extensive overview of the methods currently developed to detect deepfake images—from traditional machine learning methods to deep learning, hybrid models, and transformer-based approaches. It provides a depth analysis of benchmark datasets, evaluation metrics, and key challenges like generalization, real-time requirements, and adversarial attacks. Moreover, it is focusing on new research trends such as federated learning, explainable AI, lightweight models, and blockchain integration. In the current context of fast-evolving generative AI technologies, the review highlights the importance of systems capable of detecting AI-generated content that are robust, scalable, and privacy-protection, to maintain authenticity and trust in digital media.**

***Index Terms*—Deepfake detection, Artificial intelligence, Deep learning, GANs, CNNs, Transformer models, Image forgery, Synthetic media, Federated learning, Explainable AI.**

I. INTRODUCTION

With the advancement of AI, it has become much easier to produce lifelike fake images, commonly referred to as deepfakes. These altered images can cause misinformation, harm reputations and pose security threats. Thus, the creation of reliable deepfake image detection methods is crucial to foster digital trust, safeguard privacy, and preserve the authenticity of online content. With the development of AI and deep learning, deep fake technology has progressed at a fast pace. There are many image editing techniques that can be easily recognized, but recent ones that are based on GANs and diffusion models can create realistic fake images [1].

This change has brought not only innovative uses of the technology but also significant risks of misuse, privacy issues, and cybersecurity concerns.

Figure 1 Deepfake detection

The use of deep fake images can have a profound impact on society, as they can be used to spread false information and mislead people. They can destroy personal reputations, cause political confusion and lessen trust in digital media. The rise of deepfakes has also posed new challenges for cyber security, fraud and internet harassment.

As it is being used more frequently, it is clear that there is an urgent need for reliable detection systems and digital awareness. Identifying manipulated images is essential to block the spread of fake images online, and this is where deepfake image detection plays a role. Effective detection tools can safeguard people, organizations and society against misinformation, identity misuse and cyber threats. Precise detection systems also enhance trust on online platforms and help ensure the authenticity and trustworthiness of visuals. The primary goal of this review is to examine and analyse the current deepfake image detection techniques, datasets and evaluation measures. It also intends to pinpoint existing challenges, make comparisons, and point out future research avenues. This review will enable researchers to gain insight into recent developments and to build more efficient and effective detection systems. The paper summarizes the

research on the salient aspects of the detection of deepfake images, the datasets, challenges, and recent developments. It also discusses applications and future research opportunities. The paper is organized into sections on deepfakes generation methods, deepfakes detection approaches, performance analysis, existing limitations and future directions on how to improve the accuracy and reliability of the detection methods [2].

II. FUNDAMENTALS OF DEEPFAKE IMAGE GENERATION

A. Definition of Deepfake Images

Deepfake images are images that have been digitally altered or generated that look real but have false or manipulated information in the image. Usually, they are created with deep learning approaches, including Generative Adversarial Networks (GANs). These images can be used to alter the expressions or faces, or create fake faces, which are hard to distinguish from real images [3].

B. Types of Deepfake Manipulations

There are various types of deepfake manipulations, all of which involve the creation of fake or altered digital images and videos. There are basically two types of poker:

- Face Swapping – Replacing one person's face with another in images or videos.
- Facial Expression Manipulation – Changing facial emotions or movements artificially.
- Lip-Sync Manipulation – Modifying lip movements to match fake audio.
- Identity Morphing – Combining features of two individuals into one image.
- Attribute Editing – Altering features such as age, hairstyle, skin tone, or gender.
- Synthetic Image Generation – Creating completely fake human faces using AI models.
- Body Manipulation – Changing body posture or movements digitally.
- Background Manipulation – Editing or replacing the background to create misleading scenes[4].

C. Generative Models Used in Deepfake Creation

Generative models are Artificial Intelligence techniques to generate realistic fake images and videos. Typical examples of generative models are Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models. The models are able to learn patterns from the real data and creating the synthetic content with high visual quality that makes the creation of deepfakes more advanced and difficult to detect [5].

Generative Adversarial Networks (GANs) are deep learning techniques for producing realistic fake images, videos and audio. The generator and the discriminator are the two neural networks used in a GAN. The generator generates fake images, and the discriminator determines if they're real or fake. Both networks are in competition with each other and over time evolve the quality of images created.

GANs can generate highly realistic synthetic content, which makes them a popular choice for deepfake creation. Variational Autoencoders (VAEs) are a type of deep learning that can be employed for generating plausible synthetic images and data. They operate by encoding input data into a reduced latent representation, and subsequently decoding it.

VAEs are effective for learning data patterns and are used in various applications such as image generation, deepfake creation, data compression, and learning features. Diffusion models are the powerful AI models that produce highly realistic images by adding and removing noise from data over a period of time. They are taught to recognize image patterns by multiple processing steps and can generate fine-grained synthetic image content. The models are popular for deepfake creation in present days since they generate high-quality and visually convincing images.

D. Popular Deepfake Generation Tools and Frameworks

Some popular deepfake generation frameworks and tools include DeepFaceLab, FaceSwap, StyleGAN, and Stable Diffusion. Using deep learning and artificial intelligence, these programmes generate realistic-looking synthetic photos and movies. While they are often used for research, entertainment and content creation, they can also be misused in the context of fraud, disinformation and privacy violations [6].

E. Publicly Available Deepfake Datasets

Deepfake datasets that are accessible to the public are sets of authentic and fraudulent photos or videos that

are used to test and train detection algorithms. FaceForensics++, Celeb-DF, DeepFake Detection Challenge (DFDC), and FFHQ are examples of popular datasets. Such datasets can help researchers test the accuracy of detection and compare algorithms, and improve the accuracy of a deepfake identification system [7].

III. TAXONOMY OF DEEPFAKE IMAGE DETECTION TECHNIQUES

The methods used for deepfake detection encompass traditional machine learning, deep learning, hybrid models, Explainable AI (XAI), frequency analysis, biometric analysis, transformer-based models and multimodal deepfake detection. These techniques exploit the visual patterns, hidden features and artefacts, human physiology and cross-modal inconsistencies. These two go together to increase the accuracy, robustness, transparency and reliability of detection of increasingly realistic and complex deepfake content [8].

A. Traditional Machine Learning-Based Methods

Traditional machine learning based approaches use handcrafted features like texture, color inconsistencies and image artifacts to detect deep fake images. Some popular algorithms used for classification are Support Vector Machine (SVM), Random Forest and Decision Tree. They need manual feature extraction and are much less successful in complex deepfakes than current deep learning based methods generally would be in this case [9].

B. Deep Learning-Based Detection Approaches

Deep learning-based detection methods involve creating a neural network that can automatically detect patterns and hidden artifacts in deepfake images. These models, like Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and autoencoders, can extract features directly from data without having to be extracted manually. These methods can achieve higher accuracy and good performance in the detection of complex and realistic deepfakes.

C. Hybrid Detection Techniques

Combining traditional machine-learning techniques with deep-learning techniques to boost deepfake detection accuracy is the hybrid detection techniques. Both handcrafted features and automatic features learnt from neural networks are used in these techniques. Hybrid models combine several techniques to enhance the robustness, detection accuracy, and performance in handling complex and highly realistic deepfake images [10].

D. Explainable AI (XAI)-Based Detection

Explainable AI (XAI) detection techniques help users understand how AI-based deepfake detection models make decisions. These methods highlight large regions of the image, features or patterns of images that affect the classification results. The increased transparency, trust, and reliability of deepfake detection models provided by XAI make it easier for researchers and users to assess and validate model predictions [11].

E. Frequency Domain and Noise Pattern Analysis

Frequency domain analysis and noise pattern analysis can detect a deepfake by examining the patterns of the image, which is not visible to the human eye. These methods investigate abnormalities in the noise of the sensors, their frequency components and their resulting compression artefacts in the picture processing. They precisely identify the abnormal distribution of pixels, distortion and unusual texture, which helps in the identification of phoney images [12].

F. Biometric and Physiological Feature Analysis

Biometric and physiological feature analysis is used to detect deepfakes by examining human characteristics such as facial expressions, blinking, skin tone, heart rate, and more. These natural patterns are often not accurately reproduced in deepfake images. These techniques help enhance the likelihood of detection by detecting mismatches between actual human behavior and the visual content created by humans [13].

G. Transformer-Based Detection Models

Transformer-based detection algorithms use attention mechanisms to analyze relationships between image regions to help detect deepfakes. For example, models such as Vision Transformers (ViTs) excel at processing the features across the entire image, which is a capability that is not performed by traditional CNNs. These techniques are highly accurate in identifying subtle changes, inconsistencies, or

imperfections in very realistic deepfake images and synthetic visual information [14].

H. Multimodal Detection Approaches

The deepfakes are discovered by multimodal detection methods, which are built on the analysis of multiple data modalities such as text, audio, video and photos. These methods combine information from different sources to boost resilience and to enhance detection accuracy. Multimodal models have been able to better detect more complex, incredibly realistic deepfake information on digital platforms by detecting discrepancies among their respective modes [15].

IV. DEEP LEARNING ARCHITECTURES FOR DEEPFAKE DETECTION

CNNs, RNNs, Vision Transformers, Capsule Networks, auto encoders, attention-based models, and transfer learning techniques are examples of deep learning architectures for deepfake detection. These methods include the analysis of artifacts in the hidden content, temporal information, spatial correlations, and visual patterns. Together, they create new capabilities of efficiency, robustness, feature extraction and detection accuracy over complex deepfake techniques [16].

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are commonly used deep learning models for image analysis and deepfake detection. Textures, borders and patterns are detected in photos, and added to the image automatically. CNNs are often used for the recognition of fake and fabricated digital images due to their ability of high accuracy to recognize concealed alteration artefacts [17].

B. Recurrent Neural Networks (RNNs)

Deep learning models called Recurrent Neural Networks (RNNs) are made to handle sequential and time-dependent input. RNNs analyze the temporal features and frame relations between videos and identify discrepancies to aid in deepfake detection. They are helpful for precisely identifying modified video-based deepfake content since they can record motion and behavioural changes over time [18].

C. Vision Transformers (ViTs)

Images are analysed with deep learning models, known as Vision Transformers (ViTs), which are based on the transformer architecture. Unlike CNNs, ViTs use attention processes to capture the overall connections between regions of a picture and view the image as a series of patches. They perform very well in deepfake identification as they are able to precisely detect minute changes and complicated visual discrepancies [19].

D. Capsule Networks

Advanced neural network models called capsule networks are made to maintain the spatial relationships between features in a picture. They can more accurately describe the position, orientation, and structure of objects compared to conventional CNNs. Capsule Networks help to identify minor facial distortions and manipulation artefacts in deepfakes, making the identification of fake and synthetic facial images more accurate [20].

E. Auto Encoder-Based Models

Auto encoder-based models are examples of deep learning approach to feature learning and image reconstruction. After reducing the images into smaller representations, they recreate the images. These algorithms help identify reconstructed images more accurately and efficiently, by recognizing reconstruction errors and unobserved discrepancies between real and fake images in deepfake detection [21].

F. Attention Mechanism-Based Models

Attention mechanism models focus on important parts and features of an image when detecting deepfakes. The altered areas are also factored into the equation more than the rest, which makes it easier for the neural networks to recognize the slight differences. Attention methods have been shown to improve the detection accuracy, resilience and performance of realistic and complex deepfake images through the improvement of feature selection and information processing [22].

G. Transfer Learning Approaches

When detecting deepfakes, attention mechanism-based models concentrate on key areas and characteristics in a picture. This approach assigns more weight to altered areas, enabling neural networks

to detect subtle differences more effectively. Attention mechanisms improve feature selection and information processing, which increases the detection accuracy, resilience and performance of realistic and complex deepfake images [23].

V. DATASETS AND BENCHMARKING

When conducting deepfake detection research, it is essential to rely on benchmark datasets, preprocessing techniques, assessment metrics, and comparison analysis. These help the efficient training, testing and evaluation of detection models. Furthermore, they enable the comparison of performance, help identify limitations in the datasets, and improve the accuracy, robustness, and reliability of models in real-world applications for detecting artificial and modified photos [24].

A. Overview of Benchmark Datasets

Benchmark datasets are standardised sets of real and fake photos or videos used for training and testing deepfake detection algorithms. Popular datasets include FaceForensics++ and Celeb-DF, DFDC and DeepFakeTIMIT. Researchers can compare detection methods, assess model performance, and enhance the precision and dependability of deepfake detection systems with the use of these datasets, which offer a variety of altered samples [25].

B. Dataset Characteristics and Limitations

The size, realism, diversity, image quality, and the way the images are manipulated are all different between deepfake datasets. The features have an effect on the training and testing of detection models. There are however, a number of limitations on the quality of a data set that can reduce generalisation and detection performance, including little diversity, poor quality samples, disproportion between true and false data and no real-world scenarios [26].

C. Data Preprocessing Techniques

Data pretreatment methods are used to prepare pictures and videos before they are used to train deepfakes detection models. Some of the typical techniques include image scaling, image normalisation, noise reduction, face extraction, data augmentation, and feature improvement. These techniques help to improve the quality of data, reduce unwanted variations, and help deep learning models learn meaningful patterns more efficiently, leading to better detection performance and accuracy [27].

D. Evaluation Metrics

The performance of deepfake detection models is evaluated using evaluation metrics. Common Metrics include accuracy, precision, recall, F1-score, and ROC-AUC. These measures help researchers to compare different detection systems, evaluate the performance and reliability of a model, and identify the pros and cons of successfully differentiating between real and fake photos, as well as the benefits of the model's reliability and performance [28].

Accuracy is one of the many metrics that can be employed to assess the overall performance of a deepfake detection model. It displays the percentage of successful detection of genuine and imposter photographs out of all the analysed ones. High accuracy values are a measure of better model performance; however, when the data set is highly imbalanced, it does not necessarily mean high reliability.

- Precision and recall are two metrics that are used to measure the efficacy of deepfake detection. Recall gives the percentage of the correctly recognized images that are false, while precision indicates the percentage of the images that are falsely detected that are in fact false. A high recall ensures that a low number of altered photos are not detected and a high precision reduces the number of false alarms.
- The F1-Score is a metric for evaluating the performance of deepfake detection models, providing a comprehensive measure of precision and recall. As it balances false positives and false negatives, it is especially useful in datasets that are unbalanced. A higher F1-Score means that the model is more reliable and there is a better overall accuracy in the detection.
- An evaluation indicator called ROC-AUC is employed to gauge the effectiveness of a deepfake detection algorithm at distinguishing between real and fake photographs. AUC – is the area under the curve and ROC – is the ROC curve. Better classification performance, accuracy, and dependability of the detection model are indicated by a higher ROC-AUC value.

E. Comparative Performance Analysis

Comparative performance analysis evaluates and compares different deepfake detection techniques based on criteria such as accuracy, precision, recall, F1-score and ROC-AUC. It helps researchers assess how robust and efficient a model is, and its strengths and weaknesses. The use of appropriate detection techniques for enhancing performance against sophisticated deepfake manipulations is supported by this analysis [29].

VI. CHALLENGES IN DEEPFAKE IMAGE DETECTION

Some challenges faced by deepfake detection include:The following are some of the challenges that deepfake detection needs to overcome: The impact of these problems is on deployment, precision and reliability of detection. These challenges need to be overcome to create reliable, effective and trustworthy deepfake detection systems for use in practice [30].

A. High Visual Realism of Deepfakes

The capacity of AI-generated images to closely mimic actual human faces and expressions is referred to as high visual realism of deepfakes. The sophisticated deep learning algorithms create highly detailed and realistic-looking information, making it difficult to tell a phoney image from real. The realism makes it increasingly complicated to use detection systems and poses issues regarding false information and misuse [31].

B. Generalization Issues Across Datasets

When a deepfake detection model works well on one dataset but not on another, there are problems with generalisation across datasets. Changes in image quality, diversity of data, lighting and processing all affect model performance. These issues make deepfake detection less reliable, and show the importance of developing more resilient and flexible deepfake detection model [32].

C. Adversarial Attacks and Robustness

Adversarial attacks are ways of subtly manipulating deepfake images to make the detection models misclassify them and to reduce their accuracy. These changes are often hard for humans to see and could fool AI systems. Robustness is the ability to detect models to resist such attacks and still perform robustly when faced with hostile or altered deepfake content [33].

D. Computational Complexity

The amount of processing power, memory, and time needed to train and operate deepfake detection algorithms is referred to as computational complexity. High-performance computing power and large data sets are often needed for more complex deep learning algorithms. Costs and limited real-time detection were a problem because of the high computational complexity of the widespread deployment of effective and scalable detection systems [34].

E. Limited Availability of Realistic Datasets

One of the main challenges in detection research is the lack of realistic deepfake datasets. Many existing data sets lack diversity, variation and quality interventions. This means that there is less training done on the model, and less generalisation performance. This is because detection systems are hard to detect sophisticated deepfake images and recently created deepfake images because of the lack of realistic data [35].

F. Real-Time Detection Constraints

The challenge of immediately recognising deepfake photos or videos during live processing is referred to as real-time detection restrictions. Often, high memory, CPU and computing power is required for the more sophisticated detection models. These limitations hinder the practical application of accurate deepfake detection methods to real-time systems and online platforms, as they can decrease detection speed and effectiveness [36].

G. Ethical and Privacy Concerns

Deepfake technology can be abused to disseminate false information, sway public opinion, harm reputations, and infringe upon personal privacy, raising ethical and privacy concerns. Indigenous identity theft and harassment can occur upon the unauthorized creation of fake photographs/films. The problems highlight the need for safe detection systems, legal regulations, and responsible development of AI technologies [37].

VII. APPLICATIONS OF DEEPFAKE DETECTION

Social media verification, information security, digital forensics, law enforcement, journalism, authentication using biometrics, healthcare, and identity protection are all areas where deepfake detection is crucial. These systems aid in the detection of fraudulent content, the prevention of false information, the enhancement of digital security, the protection of individual identities, and the assurance of reliable communication and authentication across a variety of real-world digital platforms and services[38].

A. Social Media Content Verification

Social Media Content Verification involves verifying the authenticity of content, including images, videos, and information posted on social media platforms. Deepfake Identification Technologies help identify altered or false information before it's amplified. This ensures that media contents are accurate and reliable, reducing inaccuracies and protecting users from misleading information, thereby boosting trust in digital communications [39].

B. Cybersecurity and Digital Forensics

Deepfake detection methods are used by cybersecurity and digital forensics to detect altered photos, videos, and digital evidence. These techniques aid in the prevention of phishing assaults, identity theft, cyber fraud, and false information. In digital investigations, deepfake detection plays a crucial role in verifying data authenticity, supporting investigative efforts, and contributing to the identification of criminal activities, ensuring digital security and integrity [40].

C. Law Enforcement and Criminal Investigation

Deepfake detection technologies are used by law enforcement and criminal investigation organisations to detect phoney digital evidence, stop cybercrime, and look into online fraud. These methods aid in confirming the legitimacy of photos and videos that are utilised in investigations. Furthermore, the deepfake detection improves the authenticity of digital forensic evidence, helps criminal investigation and protects individuals from ID theft [41].

D. Biometric Authentication Systems

Biometric Identification Systems use unique human characteristics such as voice recognition, fingerprints, iris patterns and facial features for identity verification. By spotting phoney or altered biometric data, deepfake detection enhances these systems. This enhances the security and reliability of digital authentication solutions, thwarting identity fraud, spoofing attacks, and unauthorized access [42].

VIII. COMPARATIVE ANALYSIS OF EXISTING STUDIES

TABLE 1

Authors/year	Methodology	Research gap	Performance matrix	Problem statement
Tanvir/2025[43]	Evaluate deepfake detection using CIFAKE, FakeGPT, PFake datasets.	Limited generalization and robustness in deepfake detection models.	Accuracy 94%, PFake accuracy 96.61%, ConvNextLarge accuracy 92.54% achieved.	Detecting AI-generated images accurately across diverse datasets remains challenging.
Murali/2024[44]	Two-step GAN and CNN based fake image detection method proposed.	Limited robustness in detecting GAN-generated fake images accurately.	CNN accuracy 87%, GAN+CNN accuracy 94.4% for fake image detection.	Detecting GAN-generated fake images accurately remains a major challenge.
Preeti/2024[45]	GAN–CNN generative replay method reduces catastrophic forgetting in detection.	CNN models suffer catastrophic forgetting in deepfake detection tasks.	Training accuracy 98.67%, testing accuracy 70.08%, loss 0.0337 achieved.	Improving deepfake detection while preventing catastrophic forgetting in CNNs.

Arash/2024[46]	Blockchain-based federated learning improves deepfake detection using CNN and SegCaps.	Limited privacy-preserving, scalable deepfake detection with heterogeneous data sources.	Accuracy, precision, recall, F1-score, AUC, loss, and convergence rate. Accuracy improved 6.6%, AUC enhanced 5.1% over benchmark models.	Ensuring accurate, privacy-preserving deepfake detection across distributed heterogeneous data.
Arash/2024[47]	Review of deepfake detection methods using CNN and multimedia classification.	Limited multimodal analysis and overemphasis on accuracy optimization only.	CNN most-used method; accuracy parameter emphasized in majority of reviewed studies.	Limited multimodal deepfake detection with insufficient robust evaluation methods.
Ahmad M/2024[48]	ELA-based CNN detects image forgery using CASIA 2 dataset.	Limited robustness in detecting diverse image forgeries accurately.	Training accuracy 99.05%, testing accuracy 94.14%, precision 94.1%, recall 94.07%.	Detecting image forgeries accurately across varied manipulation techniques.
Yogesh/2023[49]	Deepfake detection using improved deep-CNN trained on multi-GAN datasets.	Limited inter-frame analysis and generalization across diverse GAN models.	Accuracy 98.33% AttGAN, 99.33% GDWCT, 99.17% StarGAN achieved.	Detecting GAN-based deepfakes with high generalization and accuracy.
Wasin/2023[50]	CNN-based fake face detection using multi-color space features.	Limited multi-color feature fusion for fake face detection.	RGB, HSV, YCbCr features improve CNN-based fake face detection performance.	Detecting fake faces using robust multi-color CNN models.
Fadi/2023[51]	Synthetic data-based face recognition with privacy-preserving deep learning approach.	Limited real-synthetic balance and evaluation in face recognition systems.	High-resolution synthetic data improves face recognition accuracy while preserving privacy.	Effective face recognition using privacy-preserving synthetic data methods.
Tripty/2023[52]	Face recognition using FaceMesh and LWF dataset deep learning model.	Limited robustness under varying conditions in face recognition systems.	Face recognition accuracy achieved 94.23% using FaceMesh and LWF dataset.	Robust face recognition under varying illumination and conditions needed.

A. Strengths and Weaknesses of Existing Techniques

Existing methods generally demonstrate excellent accuracy and robustness in controlled settings with existing CNN, GAN, federated learning, and synthetic data, excelling in deepfake and face detection. However, they come with its limitations including poor multimodal integration, narrow generalisation, high computing costs, catastrophic forgetting and poor performance in real-world and diverse environments [53].

IX. FUTURE RESEARCH DIRECTIONS

A. Robust and Generalizable Detection Models

Accurately identifying deepfake images in a variety of datasets and real-world scenarios is the goal of robust

and generalizable detection methods. These models work well with varying lighting, positioning, and manipulation methods, reduce overfitting, and enhance adaptability. They employ ensemble methods, deep learning and transfer learning techniques to enhance their performance, consistency and reliability in detecting tasks [54].

B. Lightweight Models for Real-Time Applications

The lightweight models designed to detect deepfakes quickly and cost-effectively are developed through deep learning algorithms for real-time applications. The memory usage and processing time can be reduced using optimised architectures such as MobileNet or compressed CNNs. These models are suitable for real-time situations that require rapid and effective detection, like surveillance systems, mobile devices, and social media websites [55].

C. Federated and Privacy-Preserving Learning

Multiple devices or organisations can cooperatively train deepfake detection models without sharing raw data thanks to federated and privacy-preserving learning. Rather, data security and privacy are ensured by sharing only model changes. Using this approach, model methods are more accurate and able to effectively ensure user privacy in critical applications like healthcare, banking, and surveillance [56].

D. Explainable and Trustworthy AI Models

The issue with explainable and reliable AI models is that they aim to make the decisions made by deepfake detection models clear and intelligible. These models provide succinct explanations of their predictions using feature importance and visual explanations. This increases the accountability, model reliability and trust of the users. They are crucial in crucial applications where comprehension of AI judgements is just as crucial as precision [57].

X. CONCLUSION

In this review paper, the main ideas and challenges of deepfake picture detection methods, datasets, and possible future research areas are summarized. It highlights the evolution of deep learning, transformer models, and hybrid models from traditional machine learning. Important discoveries reveal that although great accuracy is attained, problems with generalisation, real-time detection, and dataset constraints still exist. The study puts forth the importance of creating strong, lightweight, explainable, and privacy-preserving AI models. Some of the future directions are federated learning, multimodal analysis, blockchain integration, and synthetic media detection. With that said, this report provides a comprehensive understanding of the progress made in deepfake detection and provides recommendations to develop more reliable and secure AI-driven digital forensics solutions.

REFERENCES

[1] B. A. Kumar, N. Kumar, N. Pathak, and S. Ahmadpour, “Hybrid CMNV2: DeepFake faces classification and recognition using deep learning methods,” *Results in Engineering*, vol. 28, p. 107513, 2025, doi: 10.1016/j.rineng.2025.107513.

[2] K. Somoray, D. J. Miller, and M. Holmes, “Human performance in deepfake detection: A systematic review,” 2025, doi: 10.1155/hbe2/1833228.

[3] S. Wang, D. Zhu, J. Chen, J. Bi, and W. Wang, “Deepfake face discrimination based on self-attention mechanism,” Pattern Recognition Letters, vol. 183, pp. 92–97, 2024, doi: 10.1016/j.patrec.2024.02.019.

[4] Z. Sha, “DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models,” pp. 3418–3432, 2022, doi: 10.1145/3576915.3616588.

[5] A. Badale, L. Castelino, and J. Gomes, “Deep fake detection using neural networks,” vol. 9, no. 3, pp. 349–354, 2021.

[6] J. Frank, T. Eisenhofer, and L. Sch, “Leveraging frequency analysis for deep fake image recognition,” 2020.

[7] V. Roy, G. Mishra, R. Mannadiar, and S. Patil, “Fake currency detection using image processing,” vol. 8, no. 4, pp. 88–93, 2019.

[8] N. M. Alshariah, A. Khader, and J. Saudagar, “Detecting fake images on social media using machine learning,” vol. 10, no. 12, pp. 170–176, 2019.

[9] S. Tariq, S. Lee, H. Kim, and S. S. Woo, “Detecting both machine and human created fake face images in the wild,” pp. 81–87, 2018.

[10] S. Wei, G. Gu, Y. Wei, C. Tan, and Y. Zhao, "Learning on gradients: Generalized artifacts representation for GAN-generated images detection," pp. 12105–12114, 2023.
[11] J. Chen, J. Yao, and L. Niu, "A single simple patch is all you need for AI-generated image detection," 2024.
[12] S. A. Fezza, M. Y. Ouis, B. Kaddar, W. Hamidouche, and A. Hadid, "Evaluation of pre-trained CNN models for geographic fake image detection," 2022.
[13] L. Nataraj et al., "Detecting GAN generated fake images using co-occurrence matrices," 2019.
[14] P. S. Yu, "TI-CNN: Convolutional neural networks for fake news detection," 2019.
[15] W. Alkishri, S. Widyarto, and J. H. Yousif, "Evaluating the effectiveness of a GAN fingerprint removal approach in fooling deepfake face detection," vol. 1, pp. 85–103, 2024, doi: 10.58346/JISIS. 2024.I1.006.
[16] H. Li, J. Zhou, Y. Li, B. Wu, B. Li, and J. Dong, "FreqBlender: Enhancing DeepFake detection by blending frequency knowledge," 2024.
[17] A. Hatem, S. Omnia, S. Hala, T. R. Ragab, and S. Mohsen, "Deepfake detection using convolutional vision transformers and convolutional neural networks," Neural Computing and Applications, vol. 36, no. 31, pp. 19759–19775, 2024, doi: 10.1007/s00521-024-10181-7.
[18] I. Balafrej and M. Dahmane, "Enhancing practicality and efficiency of deepfake detection," pp. 1–11, 2024.
[19] A. Heidari, N. J. Navimipour, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," pp. 1–45, 2024, doi: 10.1002/widm.1520.
[20] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang, "Depth map guided triplet network for deepfake face detection," Neural Networks, vol. 159, pp. 34–42, 2023, doi: 10.1016/j.neunet.2022.11.031.
[21] B. Chen, X. Liu, Z. Xia, and G. Zhao, "Privacy-preserving DeepFake face image detection," Digital Signal Processing, vol. 143, p. 104233, 2023, doi: 10.1016/j.dsp.2023.104233.
[22] B. Le, K. Moore, and S. S. Woo, Why Do Facial Deepfake Detectors Fail? New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3595353.3595882.
[23] N. M. Alnaim, Z. M. Almutairi, and M. S. Alsuwat, "DFFMD: A Deepfake face mask dataset for infectious disease era with deepfake detection algorithms," IEEE Access, vol. 11, pp. 16711–16722, 2023, doi: 10.1109/ACCESS.2023.3246661.
[24] S. D. Johnson and B. Kleinberg, "Testing human ability to detect 'deepfake' images of human faces," Journal of Cybersecurity, vol. 9, pp. 1–18, 2023, doi: 10.1093/cybsec/tyad011.
[25] P. Trojovský and N. Bacanin, "Evaluation of deepfake detection using YOLO with local binary pattern histogram," 2022, doi: 10.7717/peerj-cs.1086.
[26] N. Guhagarkar, S. Desai, S. Vaishyampayan, and A. Save, "Deepfake detection techniques: A review," vol. 1, no. 4, pp. 1–10, 2021.
[27] S. Lee and S. S. Woo, "One detector to rule them all," pp. 3625–3637, 2019.
[28] S. Pashine, S. Mandiya, and P. Gupta, "Deep fake detection: Survey of facial manipulation detection solutions," 2017.
[29] P. Neekhara and B. Dolhansky, "Adversarial threats to DeepFake detection: A practical perspective," 2020.
[30] H. Hao et al., "Deepfakes detection with automatic face weighting," 2020.
[31] P. Kwon, J. You, G. Nam, and S. Park, "KoDF: A large-scale Korean DeepFake detection dataset," pp. 10744–10753, 2021.
[32] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu, "Improving fairness in deepfake detection," pp. 4655–4665, 2023.
[33] A. Ciamarra, R. Caldelli, F. Becattini, L. Seidenari, A. Del Bimbo, and U. Mercatorum, "Deepfake detection by exploiting surface anomalies: The SurFake approach," pp. 1024–1033, 2023.
[34] Y. Zhang, B. Colman, X. Guo, A. Shahriyari, and G. Bharaj, "Common sense reasoning for deepfake detection," 2024.
[35] A. V. Nadimpalli and A. Rattani, "GBDF: Gender balanced DeepFake dataset towards fair DeepFake detection," 2022.
[36] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," 2021.

[37] X. Cao and N. Z. Gong, "Understanding the security of deepfake detection," 2021.

[38] H. Cheng and T. Wang, "Fair deepfake detectors can generalize," 2025.

[39] P. Salehpour, H. S. Aghdasi, and C. Author, "Robust DeepFake face detection leveraging Xception model and novel snake optimization technique," vol. 5, no. 5, pp. 1444–1456, 2024, doi: 10.18196/jrc. v5i5.22473.

[40] F. Becattini, "Head pose estimation patterns as deepfake detectors," vol. 20, no. 11, 2024.

[41] Y. Xu, M. Pedersen, and K. Raja, "Analyzing fairness in deepfake detection with massively annotated databases," vol. 5, no. 1, pp. 93–106, 2024, doi: 10.1109/TTS.2024.3365421.

[42] P. Peer, "Cross-dataset deepfake detection: Evaluating the generalization capabilities of modern DeepFake detectors," 2024.

[43] T. Islam and I. Hyun, "MEXFIC: A meta ensemble explainable approach for AI-synthesized fake image classification," Alexandria Engineering Journal, vol. 116, pp. 351–363, 2025, doi: 10.1016/j.aej.2024.12.031.

[44] K. Murali and K. Murali, "Enhancing fake image detection: A novel two-step approach combining GANs and CNNs," Procedia Computer Science, vol. 235, pp. 810–819, 2024, doi: 10.1016/j.procs.2024.04.077.

[45] P. Sharma, M. Kumar, and H. K. Sharma, "GAN-CNN ensemble: A robust deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique," Procedia Computer Science, vol. 235, pp. 948–960, 2024, doi: 10.1016/j.procs.2024.04.090.

[46] A. Heidari, N. Jafari, H. Dag, S. Talebi, and M. Unal, "A novel blockchain-based deepfake detection method using federated and deep learning models," Cognitive Computation, pp. 1073–1091, 2024, doi: 10.1007/s12559-024-10255-7.

[47] A. M. Nagm et al., "Detecting image manipulation with ELA-CNN integration: A powerful framework for authenticity verification," pp. 1–18, 2024, doi: 10.7717/peerj-cs.2205.

[48] Y. Patel, S. Tanwar, and S. Member, "An improved dense CNN architecture for deepfake image detection," IEEE Access, vol. 11, pp. 22081–22095, 2023, doi: 10.1109/ACCESS.2023.3251417.

[49] W. Alkishri, S. Widyarto, J. H. Yousif, and M. Al-bahri, "Fake face detection based on colour textual analysis using deep convolutional neural network," vol. 3, pp. 143–155, 2023, doi: 10.58346/JISIS. 2023.I3.009.

[50] F. Boutros, V. Struc, J. Fierrez, and N. Damer, "Synthetic data for face recognition: Current state and future prospects," Image and Vision Computing, vol. 135, p. 104688, 2023, doi: 10.1016/j.imavis.2023.104688.

[51] S. Hangaragi, T. Singh, and N. Neelima, "Face detection and recognition using face mesh and deep neural network," Procedia Computer Science, vol. 218, pp. 741–749, 2023, doi: 10.1016/j.procs.2023.01.054.

[52] M. Kumar and H. K. Sharma, "A GAN-based model of deepfake detection in social media," Procedia Computer Science, vol. 218, pp. 2153–2162, 2023, doi: 10.1016/j.procs.2023.01.191.

[53] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," 2023.

[54] H. Sabah, M. Sahib, and M. Altaei, "Detection of deep fake in face images using deep learning," pp. 60–71, 2022.

[55] S. A. Khan, "Video transformer for deepfake detection with incremental learning," pp. 1821–1828, 2021, doi: 10.1145/3474085.3475332.

[56] İ. Demir, "Where do deep fakes look? Synthetic face detection via gaze tracking," 2014.