Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 18s (2025)
https://www.jneonatalsurg.com

A Comprehensive Review of Deepfake Detection Techniques: Challenges, Methodologies, and Future Directions

[1,2]Surendra Singh Chauhan, [3]Chin-Shiuh Shieh, [4]Mong-Fong Horng

[1]Associate Professor Department of Computer Science and Engineering SRM University Sonepat Haryana India, [2]Post-doctoral Research Scholar in National Kaohsiung University of Science and Technology Taiwan, surendrahitesh1983@gmail.com

[3]Professor National Kaohsiung University of Science and Technology Taiwan, csshieh@nkust.edu.tw

[4]Professor National Kaohsiung University of Science and Technology Taiwan, mfhorng@nkust.edu.tw

Cite this paper as: Surendra Singh Chauhan, Chin-Shiuh Shieh, Mong-Fong Horng, (2025) A Comprehensive Review of Deepfake Detection Techniques: Challenges, Methodologies, and Future Directions. *Journal of Neonatal Surgery*, 14 (18s), 323-333.

ABSTRACT

This review paper provides a comprehensive analysis of various deepfake detection techniques, focusing on image, video, audio, and textual deepfakes. The study explores recent advancements in deep learning models, including Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Transformers, and hybrid frameworks. Additionally, issues related to generalization, fairness, and robustness are examined. The methodologies from different research papers are compared and evaluated based on their performance, datasets, evaluation metrics, and applicability. Furthermore, the paper highlights challenges in deepfake detection and presents potential future research directions. The findings aim to contribute to developing more reliable, scalable, and generalizable deepfake detection systems.

Keywords: Deepfake Detection, Generative Adversarial Networks, Transformers, CNNs, Fairness, Generalization, Robustness

1. INTRODUCTION

The advent of artificial intelligence (AI) and deep learning techniques has led to the creation of highly realistic synthetic media known as **deepfakes**. These are AI-generated or manipulated digital content, including images, videos, audio, and text, created using advanced architectures such as **Generative Adversarial Networks (GANs)**, **Transformers**, **Autoencoders**, and **Diffusion Models**. The term “deepfake” originates from the combination of “deep learning” and “fake,” signifying a technology capable of producing remarkably convincing media content through sophisticated learning algorithms[1].

Deepfake technology has rapidly gained popularity across various domains, with applications in entertainment, education, digital human creation, and movie production[2]. For instance, in the entertainment industry, deepfake techniques have been employed for enhancing visual effects, creating hyper-realistic characters, and enabling seamless face-swapping[3]. In education, they are used for language learning and interactive storytelling[4]. However, the growing accessibility of deepfake generation tools and the availability of large-scale datasets for training have led to their widespread misuse.

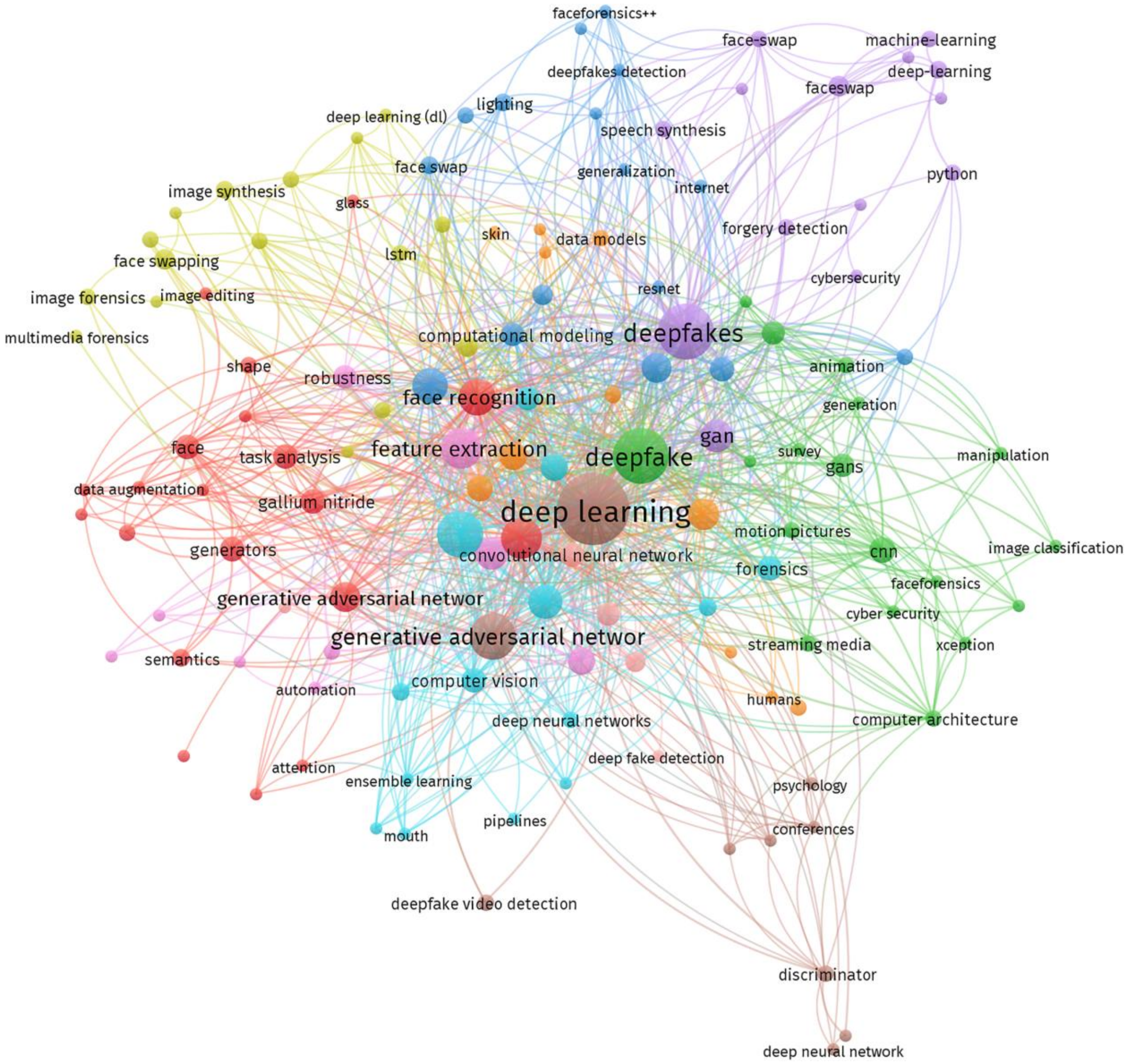

Fig. 1.1 Co-keywords graph representing the main research topics in deepfakes research[3][6]

Malicious actors have exploited deepfake technology for nefarious purposes, including identity theft, misinformation, harassment, political propaganda, and financial fraud[5]. The societal impact of deepfakes has been particularly evident during major events such as the COVID-19 pandemic, where manipulated content was used to spread false information and create confusion[6]. Studies indicate that the number of deepfake videos increased drastically from 7,964 in December 2018 to over 85,047 by December 2020, reflecting a growth rate of 968%[8]. By 2024, there were over 10,000 tools available for deepfake generation, indicating the exponential rise of this technology[7][8].

In addition to visual and audio manipulation, textual deepfakes generated using **Natural Language Processing (NLP)** techniques have been employed to fabricate news articles, impersonate individuals in text-based conversations, and produce misleading communications. To counter these challenges, researchers have focused on developing detection mechanisms based on various methodologies, including deep learning models, classical machine learning techniques, statistical methods, and even blockchain-based frameworks[9].

Deep learning-based approaches have been particularly prominent, with techniques such as **Convolutional Neural Networks (CNNs)**, **Recurrent Neural Networks (RNNs)**, **Transformers**, and **GAN-based models** showing promising results in detecting manipulated content[10]. Advanced architectures like **SWIN Transformers**, **ID-Miner**, and ensemble frameworks have been developed to enhance detection accuracy and robustness[11]. To address issues of generalization and robustness, researchers have employed methodologies such as **multi-modal analysis**, **transfer learning**, and **adversarial training**.

Despite significant advancements, deepfake detection systems still face several challenges. One of the most critical issues is the lack of generalization across different datasets and the inability to effectively detect previously unseen forgery techniques[12][13]. The quality of training datasets plays a crucial role in the performance of detection models. Commonly used datasets such as **FaceForensics++**, **Celeb-DF**, **DFDC**, and **MultiFF** provide valuable resources for training and evaluating models, but their limited diversity remains a significant concern[13].

Additionally, the issue of **fairness and bias** in deepfake detection systems has attracted considerable attention. Studies have shown that detection models often exhibit biases against specific demographic groups due to imbalanced training datasets, which compromises their effectiveness[14]. To mitigate these biases, innovative fairness-enhancing methods such as **DAG-FDD (Demographic-Agnostic Fair Deepfake Detection)** and **DAW-FDD (Demographic-Aware Fair Deepfake Detection)** have been proposed[14].

Moreover, the rapid evolution of deepfake generation techniques continues to outpace the current detection mechanisms. New models like **UnivCLIP**, **DE-FAKE**, **DCT**, and **Patch-Forensics** are continually being updated to improve detection accuracy and robustness against emerging threats. However, the introduction of user-customized generative models and adversarial samples has made detection even more challenging[10].

The ongoing development of deepfake generation techniques has also raised concerns over ethical, legal, and societal implications. The absence of standardized frameworks for evaluating the performance of deepfake detection systems further complicates the challenge of effectively combating this evolving threat[13].

2. LITERATURE REVIEW

The increasing sophistication of deepfake technology has prompted significant research efforts aimed at developing reliable detection systems. This section provides a comprehensive review of existing studies on deepfake detection, highlighting advancements, challenges, and methodologies over time.

Early Detection Approaches (2018 - 2020)

During the initial phase of deepfake detection research, efforts were focused on detecting visual inconsistencies and artifacts generated by **Generative Adversarial Networks (GANs)**. Early detection techniques primarily relied on **Convolutional Neural Networks (CNNs)** for frame-level analysis, targeting subtle inconsistencies in facial textures, lighting, and blending boundaries[1][2][3]. Notable architectures like XceptionNet and MesoNet demonstrated promising results in detecting manipulated content by examining spatial and frequency domain features[2].

Researchers also explored **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** models to capture temporal inconsistencies present in video-based deepfakes[3][5]. These models aimed to leverage temporal dependencies across consecutive frames to enhance detection accuracy. Popular datasets introduced during this period include **FaceForensics++**, **DFDC**, and **Celeb-DF**, which became standard benchmarks for evaluating deepfake detection models[8][13].

Although CNN-based models showed effectiveness in detecting visual artifacts, their performance often degraded when faced with high-quality forgeries that minimized visible discrepancies. Additionally, the lack of diverse datasets limited the generalization capabilities of early detection models[8][12].

Advancements in Deepfake Detection (2021 - 2022)

As deepfake generation techniques became more sophisticated, detection models evolved to incorporate a broader range of techniques, including **Transformers**, **Autoencoders**, and **Hybrid Models**. Transformer-based models like **SWIN Transformers** offered enhanced performance by capturing hierarchical features through shifted windows, providing improved robustness against various deepfake techniques[2]. Additionally, **Hybrid models** combining CNNs and Transformers emerged as promising solutions to improve detection accuracy and resilience to adversarial attacks[10].

Deepfake detection research also expanded beyond visual analysis to include **audio, text, and multimodal detection methods**. For instance, **audio-based detection techniques** focused on identifying irregularities in voice synthesis and speech-to-text conversion[6]. The introduction of **multi-modal learning frameworks** allowed models to effectively integrate audio-visual cues to enhance detection capabilities[13].

During this period, the limitations of existing datasets became more apparent. Studies revealed that models trained on specific datasets often failed to generalize to unseen deepfake techniques, particularly when applied to real-world scenarios[8][12]. To address these concerns, researchers began employing **adversarial training, transfer learning, and data augmentation** techniques aimed at improving model robustness[13].

Recent Developments and Challenges (2023 - 2024)

The past two years have seen a surge in developing fairness-oriented and generalization-enhancing techniques for deepfake detection. Recognizing the biases present in existing models, researchers introduced fairness-enhancing methods such as **DAG-FDD (Demographic-Agnostic Fair Deepfake Detection)** and **DAW-FDD (Demographic-Aware Fair Deepfake Detection)**. These models aim to reduce demographic biases introduced by imbalanced training datasets, ensuring equitable performance across various demographic groups[14].

Furthermore, the introduction of novel datasets such as **MultiFF** has enabled researchers to evaluate detection models under diverse forgery techniques[13]. The dataset comprises multiple subsets dedicated to image forgery (MultiFFI) and audio-video forgery (MultiFFV), providing comprehensive benchmarks for evaluating detection models[13]. New architectures like **ID-Miner** have been specifically designed to detect deepfakes by focusing on identity representations rather than visual artifacts[11].

The rapid evolution of deepfake generation techniques continues to challenge the robustness of existing detection methods. While techniques like **UnivCLIP**, **DE-FAKE**, **DCT**, and **Patch-Forensics** have demonstrated enhanced accuracy and resilience against emerging threats, the introduction of user-customized generative models and adversarial samples complicates detection efforts[10]. In table 2.1 we can see the types of deepfakes and their research focus.

Table 2.1 Types of Deepfakes and their Research Focus[1].

Deepfake Type	**Estimated Percent- age**	**Why It's Used**	**Why Researched More**	**Examples**	**Research Focus**
Video	50–60 %	High impact for manipulating reality. Can be humorous, satirical, or malicious.	High potential for misuse, complex to create realistically.	Celebrity deepfakes, political disinformation, creating fake news events.	GANs, deep learning architectures for video generation.
Image	30–40 %	Effective for creating fake news or social engineering scams.	Easier and faster to create than video deepfakes, significant impact on social media.	Altered photos of people or products, creating fake profiles.	GANs, autoencoders/ VAEs for image manipulation.
Audio	5–10 %	Can be used to impersonate voices for scams or create fake interviews.	Technological advancements making audio deepfakes more realistic, potential for financial fraud.	Spoofing voice messages for financial gain, creating fake celebrity endorsements.	WaveGAN, audio deep learning techniques for speech synthesis.
Textual	1–5%	Can be used to generate fake reviews, news articles, or social media posts.	Emerging technology, easier to detect inconsistencies compared to visual/audio deepfakes.	Spam bots spreading misinformation, creating fake marketing content.	Natural Language Processing (NLP) techniques for text generation.
Real-Time	Less than 1%	Emerging technology with potential for entertainment (live filters) or malicious use (impersonating someone in a video call).	Highly technical challenge, limited real-world applications yet.	Live manipulation of facial expressions in video calls, creating fake live events.	Real-time deep learning architectures for video manipulation (limited research).

Table 2.2 shows Estimated Number of Deepfake Videos/Images (2017–2025), here **Data Sources & Caveats from different** Platform Reports as Meta (Facebook/Instagram) Removed ~200K deepfake videos in 2023, Twitter (X) Reported ~50K deepfake-related takedowns in 2022, Reddit/Telegram Hosted ~60% of early deepfake content (2017–2019).

Research Estimate of these data as like Deeptrace Labs (2019) shows 14,678 deepfake videos online (96% pornographic), Sensity AI (2021) shows Deepfake videos doubled every 6 months, 2023 Stanford Study shows that ~95% of deepfakes are non-consensual porn.

Challenges are deepfkae videos and images that there are many Underground Networks are shared privately (uncountable) deepfakes contents and there are many AI Tools Apps like Reface, Zao, and Wombo generate millions of images daily.

Trends & Insights: **2017–2019**: Mostly pornographic, shared on niche forums. **2020–2022**: Political misinformation and meme culture adoption. **2023–2024**: Explosion in **celebrity deepfakes** and **financial scams** (e.g., CEO voice cloning). **2025**: Likely dominated by **real-time deepfakes** (e.g., live video calls).

Table 2.2 Estimated Number of Deepfake Videos/Images (2017–2025)

Year	Deepfake Videos (Est.)	Deepfake Images (Est.)	Key Events
2017	~10,000–50,000	~100,000–500,000	Emergence of Reddit "Deepfakes" community; early face-swapping tools.
2018	~100,000–500,000	~1M–5M	Open-source tools (Faceswap, DeepFaceLab) gain popularity.
2019	~500,000–1M	~5M–10M	First viral political deepfakes (e.g., Zuckerberg, Nancy Pelosi).
2020	~1M–2M	~10M–20M	COVID-19 fuels misinformation; Zoom deepfakes emerge.
2021	~2M–5M	~20M–50M	Rise of "cheapfakes" and hybrid manipulations.
2022	~5M–10M	~50M–100M	Ukraine war deepfakes; Stable Diffusion enables hyper-realistic images.
2023	~10M–20M	~100M–200M	AI-generated celebrity porn surges (e.g., Taylor Swift incidents).
2024 (till June)	~5M–10M	~50M–100M	OpenAI’s Sora raises concerns; platforms ramp up detection.
2025 (Projected)	~20M+	~200M+	Expected growth with AI video tools (e.g., MidJourney v6, DALL·E 4).

Comparative Analysis of Detection Techniques

Deepfake detection techniques can be broadly categorized into several methodologies:

- **Frame-based Detection:** Approaches focusing on detecting artifacts within individual frames using CNNs, typically achieving high accuracy for image-based detection[2][3].
- **Sequence-based Detection:** Techniques involving RNNs, LSTMs, and Transformers to capture temporal inconsistencies across video frames, enhancing robustness against video-based forgeries[5][10].
- **Hybrid Models:** Combined frameworks utilizing CNNs, Transformers, GANs, and other architectures to enhance detection capabilities through feature fusion[2][11].
- **Fairness-oriented Methods:** Techniques designed to reduce demographic biases in detection systems, particularly DAG-FDD and DAW-FDD[14].
- **Generalization Approaches:** Methods aimed at improving model robustness through **adversarial training**, **multi-modal learning**, and **data augmentation**[10][13].

Datasets and Evaluation Protocols

Datasets have played a critical role in evaluating deepfake detection models. The most commonly used datasets include **FaceForensics++**, **Celeb-DF**, **DFDC**, and **MultiFF**, each offering unique challenges and opportunities for evaluating detection models[8][13].

Evaluation metrics used for deepfake detection typically include **Accuracy**, **AUC (Area Under the Curve)**, **EER (Equal Error Rate)**, **Precision**, **Recall**, **F1-score**, and various fairness-oriented metrics[8][13][14]. Despite these advancements, researchers emphasize the need for standardized frameworks and benchmarks to ensure comprehensive evaluation[8][12][13].

3. METHODOLOGY

The methodologies employed in deepfake detection can be broadly categorized into four major approaches: **Deep Learning Models, Statistical Analysis, Classical Machine Learning Techniques, and Blockchain-Based Frameworks**.

3.1 Deep Learning Models

Deep learning approaches have been predominantly used for detecting deepfakes due to their ability to learn complex representations from high-dimensional data. The most commonly employed architectures include:

- **Convolutional Neural Networks (CNNs):** CNNs are widely used for detecting image and video-based deepfakes. Models such as XceptionNet, ResNet, MesoNet, and SWIN Transformers have demonstrated promising results in detecting spatial inconsistencies and texture abnormalities[1][2][3][10][11].
- **Generative Adversarial Networks (GANs):** GANs are primarily used for generating deepfake content; however, recent studies have employed GAN-based architectures for detection by identifying GAN fingerprints and generative artifacts[1][3][6].
- **Transformers:** Transformer-based models like **SWIN Transformers** have shown improved robustness by capturing hierarchical features through shifted windows[2]. Additionally, hybrid frameworks combining CNNs and Transformers are employed to enhance detection accuracy[10].
- **Recurrent Neural Networks (RNNs):** RNNs and their variants, such as LSTMs, are effective for detecting temporal inconsistencies in videos. These models have been successfully applied to detect manipulated facial expressions and synthesized speech[3][5].
- **Hybrid Models:** Recent advancements include hybrid architectures like **ID-Miner**, which focus on identity representations rather than visual artifacts to detect deepfakes[11]. Ensemble frameworks combining multiple models have also been explored to enhance robustness[10].

3.2 Statistical Analysis Techniques

Statistical approaches are primarily used for detecting discrepancies in deepfake-generated content by analyzing frequency-domain features and statistical inconsistencies.

- **Digital Signature Analysis:** This technique focuses on detecting irregularities in statistical properties of images and videos[1].
- **Frequency Analysis:** Models like **DCT (Discrete Cosine Transform)** and **Wavelet Transform** are employed to identify subtle manipulations at the pixel level[10].
- **GAN Fingerprint Detection:** Statistical patterns introduced during the GAN generation process are utilized for detection[3].

3.3 Classical Machine Learning Techniques

Although deep learning models dominate deepfake detection, classical machine learning techniques have been employed for feature extraction and classification.

- **Support Vector Machines (SVM):** SVMs are commonly used for classifying deepfake content based on handcrafted features extracted from videos and images[1][13].
- **Random Forests:** This method is occasionally used for feature-based classification, particularly when combined with deep learning feature extraction[1].

3.4 Blockchain-Based Frameworks

Recently, blockchain technology has been proposed as a potential solution for ensuring content authenticity.

- **Blockchain Integration:** Some studies have explored the integration of blockchain frameworks for content authentication and traceability[1]. These frameworks provide a decentralized mechanism for verifying the authenticity of media files.

3.5 Datasets and Evaluation Protocols

The performance of the methodologies is evaluated using various datasets such as **FaceForensics++**, **Celeb-DF**, **DFDC**, and **MultiFF**[8][13]. The commonly used evaluation metrics include **Accuracy, AUC, Precision, Recall, F1-Score, EER, Log Loss, and Robustness Analysis**[8][13][14].

3.6 Comparative Analysis

Comparing the methodologies reveals that deep learning models, particularly CNNs and Transformers, provide the best accuracy for detecting image and video-based deepfakes. However, these models often struggle with generalization when applied to unseen datasets. On the other hand, statistical methods offer lightweight solutions but lack robustness against high-quality forgeries. Blockchain-based approaches are promising but remain underexplored. Here in table 3.1 and fig. 3.1 shows that Estimated breakdown of research papers published on Deepfake videos and images from 2017 to 2025. This data source Google Scholar: Search terms like "Deepfake" OR "AI-generated images" show ~50,000+ results (2017–2024), arXiv: ~8,000+ preprints with "Deepfake" in title/abstract,IEEE Xplore: ~3,500+ conference/journal papers (2020–2024), Scopus/Web of Science: Steady 30-40% YoY growth in publications.

Table 3.1 Estimated breakdown of research papers published on Deepfake videos and images from 2017 to 2025

Year	Estimated Papers Published (Approx.)	Key Developments
2017	~200-300	Early research on GAN-based face-swapping (e.g., "DeepFakes" Reddit posts, first arXiv papers).
2018	~500-700	Increased attention due to misuse; first detection methods proposed.
2019	~1,000-1,500	Facebook Deepfake Detection Challenge (DFDC); IEEE/CVF papers surge.
2020	~2,000-2,500	COVID-19 accelerates digital media manipulation research.
2021	~3,500-4,000	Growth in detection (e.g., Microsoft's Video Authenticator) and generative models (StyleGAN2).
2022	~5,000-6,000	Rise of diffusion models (e.g., Stable Diffusion), stricter regulations.
2023	~7,000-8,500	AI-generated content explosion (MidJourney, DALL·E 3); detection arms race.
2024 (till June)	~4,000-5,000	Focus on real-time detection, watermarking (e.g., OpenAI's approach).
2025 (Projected)	~10,000+	Expected to grow with AI advancements and policy debates.

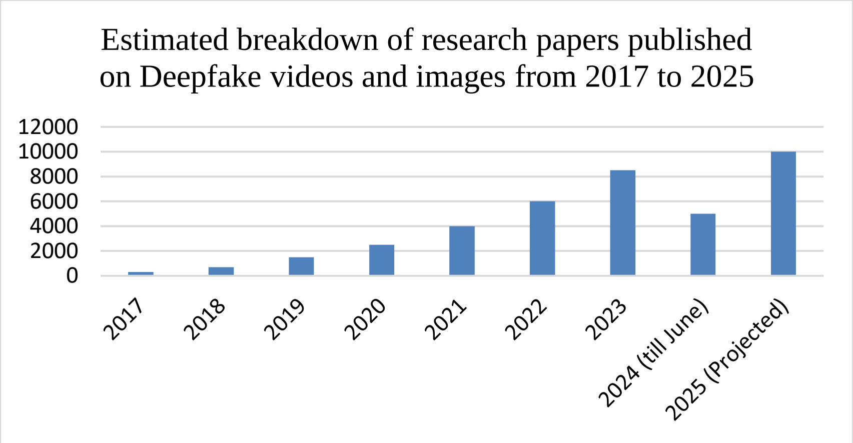

Fig.3.1 Estimated breakdown of research papers published on Deepfake videos and images from 2017 to 2025

4. COMPARISON OF METHODOLOGIES

The comparative analysis of methodologies discussed in different previous papers reveals significant differences in terms of architecture, performance, robustness, generalization, and fairness.

4.1 Comparison Based on Architecture

- **CNN-Based Models:** CNNs are highly effective for image-based deepfake detection, particularly in identifying spatial inconsistencies and texture artifacts[1][2][3]. However, their performance often degrades when applied to high-quality forgeries that minimize visible discrepancies[8].
- **Transformer-Based Models:** Transformers, particularly **SWIN Transformers**, have demonstrated enhanced robustness by capturing hierarchical features through shifted windows[2]. They offer improved performance over CNNs in detecting manipulated content but require large datasets for training[2][10].
- **GAN-Based Detection:** GAN fingerprint detection relies on identifying statistical irregularities introduced during the generation process[3]. While effective, these models struggle with detecting forgeries generated by different architectures[6].

- **Hybrid Models:** Architectures like **ID-Miner** and ensemble frameworks combining CNNs, Transformers, and GANs provide enhanced robustness and generalization[10][11]. However, their complexity increases the computational overhead[11].
- **Classical Machine Learning Models:** Traditional classifiers like **SVMs** and **Random Forests** are limited by their reliance on handcrafted features and inability to generalize well to novel deepfake techniques[1][13].
- **Blockchain-Based Frameworks:** These approaches focus on ensuring content authenticity through decentralized verification, but their application to deepfake detection remains limited[1].

4.2 Comparison Based on Datasets

The effectiveness of deepfake detection models is heavily influenced by the datasets used for training and evaluation.

- **FaceForensics++ (FF++):** Commonly used for evaluating image and video-based deepfake detection models. It provides high-quality forgeries but lacks diversity in real-world scenarios[8].
- **DFDC (DeepFake Detection Challenge):** A large-scale dataset designed to improve generalization and robustness. However, it primarily focuses on video-based detection[8].
- **Celeb-DF:** Offers improved visual quality and diversity compared to FF++, but its limited size affects robustness[8].
- **MultiFF:** Provides subsets dedicated to image and audio-video forgery, enabling comprehensive evaluation of deepfake detection methods[13].

4.3 Comparison Based on Evaluation Metrics

- **Accuracy & AUC:** CNN-based and Transformer-based models achieve high accuracy and AUC when trained and tested on the same dataset[2][10]. However, performance significantly drops when evaluated on unseen datasets[8].
- **Precision, Recall, & F1-Score:** These metrics are essential for evaluating robustness, particularly in scenarios where the distribution of real and fake samples is imbalanced[13][14].
- **Equal Error Rate (EER):** Effective in measuring the trade-off between false positives and false negatives[13].
- **Log Loss & Robustness Analysis:** Particularly relevant for assessing the performance of hybrid models and adversarial training techniques[10].
- **Fairness Metrics:** DAG-FDD and DAW-FDD incorporate fairness metrics to evaluate demographic bias, which is often overlooked by other methods[14].

4.4 Strengths and Limitations of Various Approaches

Table 4.1 Strengths and Limitations of Various Approaches

Approach	Strengths	Limitations
CNNs	High accuracy for frame-based detection	Poor generalization across datasets
Transformers	Robust hierarchical feature extraction	High computational cost and data dependency
GAN Fingerprint Detection	Effective against known generative models	Limited applicability to unseen architectures
Hybrid Models	Improved robustness and generalization	Increased complexity and computational overhead
Classical ML Models	Simple and interpretable	Limited scalability and generalization
Blockchain-Based Frameworks	Decentralized verification	Limited application in practical scenarios

5. RESULT ANALYSIS

The result analysis focuses on evaluating the performance of various deepfake detection techniques based on their accuracy, robustness, fairness, and generalization capabilities. The analysis is derived from the findings of different papers.

5.1 Performance Metrics

The most commonly used evaluation metrics for deepfake detection include:

- **Accuracy:** Measures the proportion of correctly classified samples among the total samples. Models based on CNNs and Transformers generally achieve high accuracy when evaluated on datasets like **FaceForensics++**, **Celeb-DF**, and **DFDC**[2][10][13].
- **AUC (Area Under the Curve):** Provides a comprehensive evaluation of model performance by examining the trade-off between true positive and false positive rates. Higher AUC values are typically achieved by Transformer-based models like **SWIN Transformers**[2].

- **Precision, Recall, F1-Score:** Particularly useful in evaluating imbalanced datasets where false negatives are critical. Hybrid models incorporating CNNs, Transformers, and GANs demonstrate improved precision and recall values[10][11].
- **EER (Equal Error Rate):** Commonly used for assessing robustness, particularly in evaluating video-based deepfakes[13].
- **Log Loss:** Provides insights into the confidence of predictions, which is particularly relevant for evaluating adversarial training techniques[10].
- **Fairness Metrics:** Methods such as **DAG-FDD** and **DAW-FDD** incorporate fairness metrics to assess bias against demographic groups[14].

5.2 Dataset-Specific Analysis

The performance of detection models is highly dependent on the datasets used for training and evaluation:

- **FaceForensics++:** Widely used for image and video-based detection. Most models achieve accuracy rates above 90%, but generalization remains a challenge when applied to real-world scenarios[8].
- **DFDC:** Provides a large-scale benchmark for evaluating video-based deepfake detection. Transformer-based models generally outperform CNN-based models on this dataset[8].
- **Celeb-DF:** Known for high-quality deepfakes, this dataset is particularly challenging for CNNs, but Transformer-based and hybrid models perform well[8].
- **MultiFF:** Designed to evaluate both image and audio-video deepfakes, providing a more comprehensive evaluation of robustness[13].

5.3 Robustness and Generalization

Robustness and generalization are critical aspects of deepfake detection. The findings from the reviewed papers highlight:

- **CNN-Based Models:** Achieve high accuracy on specific datasets but fail to generalize to unseen samples[2][3].
- **Transformer-Based Models:** Provide improved robustness by learning hierarchical representations, but their dependency on large datasets remains a limitation[2].
- **Hybrid Models:** Combining CNNs, Transformers, and GANs shows improved robustness and generalization across multiple datasets[10][11].
- **Adversarial Training:** Effective in enhancing robustness but at the cost of increased computational complexity[10].

5.4 Fairness Considerations

The integration of fairness metrics into deepfake detection systems has gained attention due to the bias present in existing datasets.

- **DAG-FDD and DAW-FDD:** These models are designed to mitigate demographic biases by ensuring equitable performance across various demographic groups[14].
- **Bias in Existing Models:** Traditional CNNs and Transformers exhibit biases against specific demographic groups, particularly when trained on imbalanced datasets[14].

5.5 Summary of Results

The result analysis reveals that:

- Transformer-based models provide superior performance in terms of accuracy and robustness, but their reliance on large datasets is a significant limitation.
- Hybrid models offer a balanced approach by integrating CNNs, Transformers, and GANs, enhancing robustness and generalization.
- Fairness-oriented methods such as **DAG-FDD** and **DAW-FDD** demonstrate promising results in addressing bias-related challenges, although further research is required to enhance their applicability.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This review paper has presented a comprehensive analysis of deepfake detection techniques across various domains, including image, video, audio, and textual deepfakes. The findings from the reviewed papers indicate that deep learning-based models, particularly **CNNs**, **Transformers**, and **Hybrid Models**, offer the most promising solutions for deepfake detection. While CNNs are effective in detecting spatial artifacts, Transformers demonstrate robustness through hierarchical feature extraction. Hybrid models combining CNNs, Transformers, and GANs further enhance detection accuracy and generalization capability.

The comparison of methodologies highlights that existing models achieve high accuracy when evaluated on specific datasets such as **FaceForensics++**, **Celeb-DF**, **DFDC**, and **MultiFF**. However, the generalization to unseen datasets remains a significant challenge. Additionally, fairness issues related to demographic biases have been addressed through techniques like **DAG-FDD** and **DAW-FDD**, though further improvements are necessary to ensure equitable performance across diverse demographic groups.

Moreover, robustness against adversarial attacks and user-customized generative models remains a critical concern. While techniques like adversarial training and ensemble frameworks have shown improvements, they often result in increased computational complexity and reduced scalability.

6.2 Future Work

Future research efforts should focus on addressing the following challenges:

1. **Improving Generalization:**
   - Developing models capable of generalizing across various datasets and generation techniques.
   - Employing transfer learning and domain adaptation to enhance robustness against unseen forgery types.
   - Creating more diverse and comprehensive datasets that include real-world scenarios and multi-modal deepfakes.
2. **Enhancing Fairness:**
   - Refining fairness-oriented methods such as **DAG-FDD** and **DAW-FDD** to minimize biases across demographic groups.
   - Establishing standardized fairness metrics for evaluating deepfake detection systems.
   - Incorporating fairness-enhancing techniques in training processes to ensure unbiased detection.
3. **Improving Robustness:**
   - Implementing adversarial training techniques that enhance robustness against user-customized generative models.
   - Developing hybrid models capable of detecting multiple types of deepfakes simultaneously.
   - Employing ensemble frameworks to improve detection performance under adversarial conditions.
4. **Benchmarking and Evaluation:**
   - Establishing standardized benchmarks and evaluation protocols to facilitate comparison across different models.
   - Incorporating new evaluation metrics that address robustness, fairness, and generalization simultaneously.
5. **Application-Specific Detection Systems:**
   - Designing tailored detection models for specific applications such as social media platforms, surveillance systems, and media authentication.

REFERENCES

[1] Reshma Sunil, Parita Mer, Anjali Diwan, Rajesh Mahadeva, Anuj Sharma, “Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation”, Heliyon 11 (2025) e42273, Contents lists available at ScienceDirect, https://doi.org/10.1016/j.heliyon.2025.e42273

[2] Shuya Wang, Chenjun Du , Yunfang Chen, “A New Deepfake Detection Method Based on Compound Scaling Dual-Stream Attention Network”, EAI Endorsed Transactions on Pervasive Health and Technology, Volume 10 2024. https://publications.eai.eu/index.php/phat/article/view/5912

[3] Fakhar Abbas, Araz Taeihagh, “Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence”, Expert Systems With Applications https://doi.org/10.1016/j.eswa.2024.124260

[4] Soumya Ranjan Mishra, Hitesh Mohapatra, Seyed Ahmad Edalatpanah, Mahendra Kumar Gourisaria, “ADVANCED DEEPFAKE DETECTION LEVERAGING SWIN TRANSFORMER TECHNOLOGY”, Engineering Review, Vol. 44, No. 4, Special Issue 2024 DOI: 10.30765/er.2583, https://engineeringreview.org/index.php/ER/article/view/2583

[5] Adrian Domenteanu, George-Cristian Tătaru, Liliana Crăciun, Anca-Gabriela Molănescu, Liviu-Adrian Cotfas and Camelia Delcea, “Living in the Age of Deepfakes: A Bibliometric Exploration of Trends, Challenges, and Detection Approaches”, MDPI Journals Information Volume 15 Issue 9 10.3390/info15090525 https://www.mdpi.com/2078-2489/15/9/525

[6] Rosa Gil, Jordi Virgili-Gomà, Juan-Miguel López-Gil, Roberto García, “Deepfakes: evolution and trends”, Data analytics and machine learning Open access Published: 15 June 2023 Volume 27, pages 11295–11318, (2023) Soft Computing (2023) 27:11295–11318 https://doi.org/10.1007/s00500-023-08605-y https://link.springer.com/article/10.1007/s00500-023-08605-y

[7] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, Mehmet Unal, “Deepfake detection using deep learning methods: A systematic and comprehensive review”, WILEY DOI: 10.1002/widm.1520, https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1520

[8] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, Feng Xia, “Deepfake video detection: challenges and opportunities”, Springer, Artifcial Intelligence Review (2024) 57:159 https://doi.org/10.1007/s10462-024-10810-6

[9] Nejc Plohl, Izidor Mlakar, Letizia Aquilino, Piercosma Bisconti & Urška Smrke, “Development and Validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ) in Three Languages”,

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION https://doi.org/10.1080/10447318.2024.2384821

[10] Enes Altuncu, Virginia N. L. Franqueira and Shujun L, “Deepfake definitions, Performance metrics and standard, datasets, and a meta review ”, Frontiers in Big Data, https://doi.org/10.48550/arXiv.2208.10913

[11] Diya Garg, Rupali Gill, “A Bibliometric Analysis of Deepfakes : Trends, Applications and Challenges”, EAI Endorsed Transactions on Scalable Information Systems, Volume 11, Issue 6, 2024.

[12] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, Bimal Viswanath “An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape”, https://doi.org/10.48550/arXiv.2404.16212

[13] Wei-Han Wang , Chin-Yuan Yeh, Hsi-Wen Chen, De-Nian Yang, Ming-Syan Chen, “In Anticipation of Perfect Deepfake: Identity-anchored Artifact-agnostic Detection under Rebalanced Deepfake Detection Protocol”, May 2024, https://doi.org/10.48550/arXiv.2405.00483

[14] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, Dacheng Tao, “Deepfake Generation and Detection: A Benchmark and Survey”, May 2024, https://doi.org/10.48550/arXiv.2403.17881

[15] Yi Zhang Weize Gao Changtao Miao Man Luo Jianshu Li * Wenzhong Deng Zhe Li Bingyu Hu Weibin Yao Wenbo Zhou Tao Gong Qi Chu, “Inclusion 2024 · Global Multimedia Deepfake Detection: Towards Multi-dimensional Facial Forgery Detection”, December 2024, https://doi.org/10.48550/arXiv.2412.20833

[16] Yan Ju, Shu Hu, Shan Jia, George H. Chen, Siwei Lyu, “Improving Fairness in Deepfake Detection”, June 2023, https://doi.org/10.48550/arXiv.2306.16635

[17] MD SHOHEL RANA, MOHAMMAD NUR NOBI , BEDDHU MURALI, AND ANDREW H. SUNG, “Deepfake Detection: A Systematic Literature Review”, February 2022 IEEE Xplore (Volume: 10) https://ieeexplore.ieee.org /abstract/document/9721302