A Survey on Deep Learning Techniques for Deepfake Image/Video Detection

Samarth N. Tambe[1], Chetan R. Gajula[2], Raj S. Kubal[3], Harsh S. Hanchate[4], Archana Gopnarayan[5]
[1,2,3,4] *Department of Information Technology(IF) Vidyalankar Polytechnic* Wadala, Mumbai, India
[5]*Sr.Lecturer Department of Information Technology(IF) Vidyalankar Polytechnic* Wadala, Mumbai, India

***Abstract*— Deepfakes are synthetically generated media created using advanced deep learning techniques that manipulate a person's facial appearance or speech to produce highly realistic forged content. The rapid advancement of generative models has raised serious concerns regarding digital media authenticity, making reliable detection mechanisms essential. This paper presents a deep learning-based framework for image and video level deepfake detection, focusing on identifying spatial manipulation artifacts in facial regions.**

The proposed system follows a structured processing pipeline that includes dataset collection, preprocessing, facial frame extraction, and feature analysis using convolutional neural network architectures. Pretrained models such as XceptionNet, MesoNet, and EfficientNet are integrated to extract discriminative spatial features indicative of forgery. These architectures were selected to maintain a practical balance between detection performance and computational efficiency. The current implementation includes model integration and preprocessing modules, with ongoing work dedicated to fine-tuning, evaluation, and improving robustness under compressed and real-world conditions. The framework aims to provide an accurate, scalable, and practically deployable solution for deepfake detection.

Keywords— Deepfake Detection, Deep Learning, Video Analysis, Neural Networks, Digital Forensics

I. INTRODUCTION

The rapid growth of artificial intelligence and deep learning has enabled the creation of highly realistic synthetic media commonly referred to as *deepfakes* [1], [2]. Deepfake images/videos are digitally manipulated media in which a person's appearance, facial expressions, or speech are altered to appear authentic, often using advanced neural network architectures such as Generative Adversarial Networks (GANs) and autoencoders [2], [3]. While this technology has legitimate applications in areas such as entertainment, visual effects, and virtual reality [4], its misuse poses serious risks related to misinformation, identity fraud, political manipulation, and erosion of trust in digital media [5], [6].

Detecting deepfake images/videos has become increasingly challenging due to the rapid improvement in generative models and the availability of large-scale datasets that enable the creation of highly realistic forgeries [7], [8]. Traditional digital forensic techniques, which rely on handcrafted features and statistical inconsistencies, are often insufficient to handle the complex and subtle manipulations present in modern deepfake images/videos [9]. As a result, deep learning–based approaches, particularly convolutional neural networks (CNNs), have gained significant attention for identifying spatial and temporal artifacts introduced during the manipulation process [10], [11].

This paper presents a survey of deep learning–based approaches used for deepfake video detection. Existing detection methods, commonly used datasets, and performance evaluation techniques are reviewed and compared to identify their strengths and limitations. The study also highlights current challenges and open research gaps, emphasizing the need for robust and generalizable detection systems suitable for real-world deployment [12]. Based on the reviewed literature, the development of a deep learning–based deepfake video detection system has been initiated as part of this project.

Overview of Existing Research

Deepfake detection has gained significant research attention in recent years due to rapid advancements in

generative models such as Generative Adversarial Networks (GANs) and autoencoders [1], [2]. Early detection approaches focused on identifying visual artifacts in facial regions, including inconsistencies in eye blinking, lighting conditions, and facial movements [9]. However, these approaches demonstrated limited robustness as deepfake generation techniques continued to evolve.

Recent studies have increasingly adopted deep learning–based architectures to enhance detection accuracy. Convolutional neural network models such as XceptionNet [11], MesoNet [10], and EfficientNet have been widely used for detecting manipulated images and images/videos. These architectures are capable of learning discriminative spatial features from facial regions and capturing subtle inconsistencies introduced during the manipulation process. Some approaches further integrate temporal modeling techniques, such as recurrent neural networks and long short-term memory (LSTM) networks, to improve video-level detection performance [12], [13].

Several benchmark datasets, including FaceForensics++ [3], the Deepfake Detection Challenge (DFDC) dataset [7], and Celeb-DF [8], are commonly used for training and evaluating these architectures. Although high detection accuracy is achieved on benchmark datasets, many architectures struggle to generalize to real-world scenarios involving compressed or low-quality images/videos. These observations motivate the need for robust and computationally efficient deepfake detection architectures, which form the basis of the proposed system.

II. LITERATURE SURVEY

Deepfake detection methods have evolved significantly in recent years. MesoNet, introduced by Afchar et al., is a lightweight convolutional neural network designed to detect facial manipulations in images/videos [10]. It focuses on mesoscopic features rather than fine-grained details, enabling efficient detection with lower computational cost. Experimental results demonstrate good performance on early deepfake datasets.

However, MesoNet exhibits limitations when handling high-quality or heavily compressed images/videos, which can reduce its detection accuracy. Its design prioritizes efficiency over robustness, making it less effective for large-scale real-world scenarios [10].

XceptionNet, proposed by Chollet, replaces standard convolutions with depthwise separable convolutions to improve efficiency and accuracy [11]. Although originally developed for image classification, it has been widely adopted in deepfake detection due to its strong feature extraction capabilities.

Despite its high performance on benchmark datasets, XceptionNet requires significant computational resources, which may hinder real-time deployment. Additionally, it may struggle with generalization when exposed to unseen deepfake generation techniques [11].

FaceForensics++ by Rössler et al. provides one of the most widely used datasets for deepfake detection [14]. It evaluates multiple detection methods under various compression levels, enabling standardized benchmarking.

However, detection accuracy significantly drops under strong compression or low-quality images/videos, highlighting challenges for real-world applications. The dataset focuses primarily on facial manipulations, limiting the scope for non-facial deepfake detection [14].

The Deepfake Detection Challenge (DFDC) dataset, introduced by Dolhansky et al., offers a large variety of realistic deepfake images/videos to promote robust detection research [7]. It emphasizes generalization across different manipulation techniques and video sources.

Yet, models trained on DFDC may still underperform on completely unseen manipulations or in low-light and low-resolution conditions. The dataset also requires substantial storage and computational resources for training [7].

Li and Lyu proposed detecting deepfakes by identifying face warping artifacts [15]. Their method uses convolutional neural networks to learn spatial inconsistencies introduced during manipulation.

While effective for early deepfakes, this approach becomes less reliable as deepfake generation techniques improve. It also focuses solely on facial warping, neglecting other manipulation cues [15].

Wang et al. showed that CNN-generated images contain characteristic artifacts detectable by deep learning models [16]. This provides insights into general forensic cues in synthetic media.

However, as generative models improve, these artifacts may diminish, reducing detection effectiveness. The

approach may require frequent retraining to keep up with new GAN architectures [16].

Nguyen et al. provided a comprehensive survey of deepfake creation and detection techniques [1]. They categorized detection approaches into spatial, temporal, and hybrid methods, highlighting the ongoing arms race between generation and detection.

The survey, while broad, does not provide detailed experimental evaluation or comparative analysis, leaving open questions about real-world robustness [1].

Demir et al. proposed detecting deepfakes using biological signals such as heart rate inferred from facial images/videos [17]. This alternative leverages features difficult to synthesize accurately.

The method, however, is sensitive to video quality and lighting, limiting its applicability in uncontrolled environments. It also requires additional preprocessing to extract reliable signals [17].

Celeb-DF, introduced by Li et al., addresses limitations of earlier datasets by providing higher-quality deepfake images/videos [8]. It reveals that many existing detection models perform poorly on these images/videos.

Despite the improved quality, the dataset still mainly focuses on celebrity faces, limiting diversity. Detection models still struggle with generalization to non-celebrity images/videos [8].

Agarwal et al. analysed vulnerabilities of deepfake detection systems when applied to public figures [5]. They highlighted biases and robustness issues, emphasizing ethical and societal concerns.

However, this study focuses more on ethical implications than technical evaluation, providing limited guidance on improving detection models [5].

Nawaz et al. explored LSTM networks for capturing temporal inconsistencies in deepfake images/videos [13]. By analysing frame sequences, their approach improves video-level detection accuracy.

Temporal models, however, increase computational complexity and may not be suitable for real-time applications or large-scale datasets [13].

Marra et al. investigated whether GAN-generated images leave artificial fingerprints [18]. They showed that these traces can be used for forensic detection.

The approach may require retraining with new GAN architectures, and it primarily addresses image data, limiting applicability to images/videos [18].

III. COMPARATIVE ANALYSIS OF EXISTING METHODS

Several deep learning–based methods have been proposed for deepfake image and video detection, which can be broadly categorized based on their underlying architecture and feature modeling strategy, including convolutional neural networks, recurrent neural networks, hybrid models, and transformer-based approaches [1], [4]. Table I presents a comparative analysis of these major deepfake detection methods, highlighting their representative models, key characteristics, datasets used, and limitations.

METHOD CATEGORY	REPRESENTATIVE MODELS / PAPERS	KEY FEATURES	DATASETS USED	LIMITATIONS
CNN-BASED METHODS	MesoNet [10], XceptionNet [11]	Learn spatial artifacts from facial regions	FaceForensics++, Celeb-DF	Poor generalization under compression
RNN / LSTM-BASED METHODS	Güera & Delp [18], Wang et al. [13]	Capture temporal inconsistencies across frames	DFDC, Custom datasets	High computational cost
HYBRID CNN + RNN METHODS	CNN + LSTM models [13]	Combine spatial and temporal features	FaceForensics++, DFDC	Complex architecture
ARTIFACT-BASED METHODS	Li & Lyu [15], Marra et al.[18]	Detect warping & GAN fingerprints	FaceForensics	Ineffective on newer deepfakes
BIOLOGICAL SIGNAL-BASED METHODS	FakeCatcher [17]	Uses heart-rate & physiological cues	Custom datasets	Sensitive to lighting & motion
TRANSFORMER-BASED METHODS	Vision Transformer [9]	Long-range dependency modeling	FaceForensics++	High training cost, data-hungry

IV. PROPOSED SYSTEM AND IMPLEMENTATION STATUS

Based on the comparative analysis of existing deepfake detection methods, a deep learning–based framework for image and video-level deepfake detection is proposed. The system focuses on identifying spatial manipulation artifacts within facial regions using convolutional neural network architecture. Emphasis is placed on achieving a balance between detection accuracy and computational efficiency to ensure practical applicability.

A. System Architecture

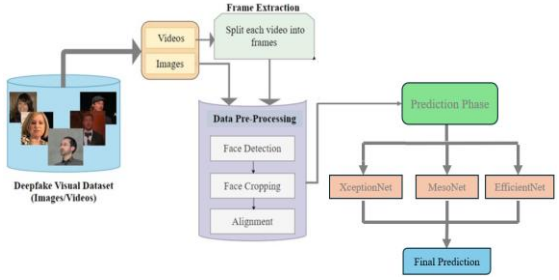

Fig. 1.1

The overall architecture of the proposed system is illustrated in Fig. 1.1. The system follows a sequential processing pipeline beginning with input acquisition in the form of an image or video. For video inputs, frames are extracted at predefined intervals to reduce redundancy and computational overhead.

Subsequently, face detection and cropping are performed to isolate relevant facial regions. The extracted facial frames undergo preprocessing steps including resizing and normalization to ensure compatibility with the selected deep learning models.

The preprocessed frames are then passed through CNN-based architectures for feature extraction and classification. The models analyze spatial inconsistencies and forgery artifacts introduced during deepfake generation. Finally, a binary classification output is generated, indicating whether the input media is authentic or manipulated.

B. Selected Deep Learning Models

Model Name	*Original Author*	*Year*	*Key Contribution*	*Limitations*
XceptionNet	Chollet et al.	2017	Depthwise separable convolutions enabling efficient feature extraction[11]	High computational cost for video-level analysis[11]
MesoNet	Afchar et al.	2018	Lightweight CNN designed for deepfake detection[10]	Limited performance on highly realistic forgeries[10]
EfficientNet	Tan and Le	2019	Compound scaling for improved accuracy-efficiency balance[19]	Requires fine-tuning for domain-specific tasks[19]

The proposed system integrates multiple CNN-based architectures to improve detection robustness. Each selected model contributes distinct advantages in feature extraction and computational efficiency.

XceptionNet [11]

XceptionNet utilizes depthwise separable convolutions to efficiently learn discriminative spatial features. Its strong representational capability makes it suitable for detecting subtle manipulation artifacts in facial regions. However, the model requires significant computational resources, especially for video-level analysis.

MesoNet [10]

MesoNet is a lightweight CNN architecture designed specifically for deepfake detection. It focuses on mesoscopic features rather than fine-grained pixel details, enabling faster inference with lower computational overhead. While efficient, its detection performance decreases for highly realistic and heavily compressed deepfakes.

EfficientNet [19]

EfficientNet employs a compound scaling strategy that balances network depth, width, and resolution. This results in improved accuracy–efficiency trade-offs compared to traditional CNNs. In the proposed system, EfficientNet is used to enhance detection performance

while maintaining reasonable computational cost. Domain-specific fine-tuning is required for optimal results.

C. Use of Pre-trained Models:

The proposed system utilizes pre-trained deep learning models obtained from publicly available third-party implementations consistent with architectures reported in prior literature. Models such as XceptionNet, MesoNet, EfficientNet were adopted based on their proven effectiveness in deepfake detection tasks as demonstrated in existing studies[10], [11], [19].

These pre-trained weights were originally trained on large-scale deepfake datasets and shared through open-source repositories for research purposes. While the weights were not released by the original architecture authors, they follow the same network designs and training protocols described in the corresponding research works.

The use of such pre-trained models enables faster experimentation and reduces computational overhead during the initial phase of development. This approach is commonly adopted in applied deep learning research. Further validation, fine-tuning, and performance analysis are planned to improve robustness and generalization across diverse real-world scenarios.

D. Implementation Status

The implementation of the proposed system is currently in progress. The completed phases include dataset collection, preprocessing pipeline development, and model integration. Frame extraction and face isolation modules have been partially implemented.

Ongoing work focuses on fine-tuning the selected architectures, conducting performance evaluation across benchmark datasets, and optimizing computational efficiency. Future development will emphasize improving robustness under compressed video conditions and enhancing cross-dataset generalization

V. FUTURE SCOPE AND EMERGING TECHNIQUES

Despite significant progress in deepfake image and video detection, the rapid evolution of generative models continues to pose serious challenges to existing detection systems. While convolutional neural network (CNN)–based architectures have demonstrated strong performance on benchmark datasets, their limitations in terms of generalization and robustness highlight the need for exploring emerging detection paradigms [1], [4]. Future research in deepfake detection is expected to focus on the following directions.

Recent studies have shown growing interest in transformer-based architectures for deepfake detection. Vision Transformers (ViTs) and hybrid CNN–Transformer models are capable of capturing long-range dependencies and global contextual information, which are often missed by traditional CNNs [20], [21]. These models have demonstrated improved generalization across datasets and robustness against compression artifacts, making them a promising alternative to purely convolutional approaches.

Another important research direction involves multimodal deepfake detection. Instead of relying solely on visual cues, future systems may integrate multiple modalities such as audio signals, lip synchronization, facial motion patterns, and textual context [22], [23]. Multimodal approaches improve detection reliability by exploiting inconsistencies between modalities, which are significantly harder to synthesize convincingly in deepfake generation pipelines.

Frequency-domain analysis has also emerged as an effective complementary technique. Methods based on Fourier Transform or Discrete Cosine Transform (DCT) analyze spectral artifacts introduced during image and video synthesis [24], [25]. These frequency-based features have shown resilience against high-quality manipulations and aggressive compression, suggesting their potential for improving real-world robustness when combined with spatial-domain CNN features.

Self-supervised and continual learning approaches represent another promising future direction. Most existing deepfake detection models rely heavily on labelled datasets and require retraining when new manipulation techniques emerge. Self-supervised learning can reduce dependence on annotated data, while continual learning frameworks can help detection systems adapt dynamically to evolving deepfake generation methods without catastrophic forgetting [26], [27].

Finally, improving real-world deployment readiness remains a key challenge. Future work should focus on developing lightweight and computationally efficient models capable of operating under real-world constraints such as low-resolution videos, varying illumination conditions, and social media compression [12], [28]. Enhancing cross-dataset generalization and robustness against unseen manipulation techniques is critical for deploying deepfake detection systems in practical forensic and security applications.

Incorporating these emerging techniques in future iterations of the proposed system can significantly enhance detection accuracy, robustness, and scalability, ensuring long-term relevance in the rapidly evolving deepfake landscape.

VI. CONCLUSION

This paper presented a comprehensive survey of deep learning–based approaches for deepfake video detection and outlined the design of a proposed detection system. Existing state-of-the-art methods were critically analyzed to identify their strengths, limitations, and applicability to real-world scenarios, which directly informed the architectural and dataset choices for the proposed approach.

The project implementation is currently in progress, with dataset collection, preprocessing, and initial integration of pre-trained deep learning models completed. Future work will focus on model fine-tuning, extensive performance evaluation across multiple datasets, and system optimization to improve robustness, generalization, and computational efficiency. The final objective is to develop a practical and scalable deepfake video detection system suitable for real-world deployment.

REFERENCES

[1] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, Cuong M. Nguyen, Deep learning for deepfakes creation and detection: A survey, Computer Vision and Image Understanding, https://doi.org/10.1016/j.cviu.2022.103525.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (November 2020), 139–144. https://doi.org/10.1145/3422622

[3] Korshunova, I., Shi, W., Dambre, J. and Theis, L., 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 3677-3685).

[4] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C. and Tao, D., 2024. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

[5] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H., 2019, June. Protecting world leaders against deep fakes. In *CVPR workshops* (Vol. 1, No. 38).

[6] Chesney, B. and Citron, D., 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, p.1753.

[7] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C., 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

[8] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S., 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).

[9] Wodajo, D. and Atnafu, S., 2021. Deepfake video detection using convolutional vision transformers. *arXiv preprint arXiv:2102.11126*.

[10] Afchar, D., Nozick, V., Yamagishi, J. and Echizen, I., 2018, December. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE.

[11] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258. 2017.

[12] Medsker, Larry R., and Lakhmi Jain. "Recurrent neural networks." *Design and applications* 5, no. 64-67 (2001): 2.

[13] Nawaz, M., Javed, A. and Irtaza, A., 2024. Convolutional long short-term memory-based approach for deepfakes detection from videos. *Multimedia Tools and Applications*, *83*(6), pp.16977-17000.

[14] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1-11. 2019.

[15] Li, Yuezun, and Siwei Lyu. "Exposing deepfake videos by detecting face warping artifacts." *arXiv preprint arXiv:1811.00656* (2018).

[16] Wang, Sheng-Yu, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. "CNN-generated images are surprisingly easy to spot... for now." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695-8704. 2020.

[17] Ciftci, U.A., Demir, I. and Yin, L., 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.

[18] Marra, Francesco, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. "Do gans leave artificial fingerprints?." In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 506-511. IEEE, 2019.

[19] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In *International conference on machine learning*, pp. 6105-6114. PMLR, 2019.

[20] Deligiannis, N., 2018. SNIPPET: A Framework for Subjective Evaluation of Visual Explanations Applied to DeepFake Detection Yang, Yuqing; Joukovsky, Boris; Oramas Mogrovejo, Jose Antonio; Tuytelaars, Tinne.

[21] Li, Z., Chen, G. and Zhang, T., 2020. A CNN-transformer hybrid approach for crop classification using multitemporal multisensor images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, pp.847-858.

[22] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. and Manocha, D., 2020, October. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2823-2832).

[23] Gu, Y., Zhao, X., Gong, C. and Yi, X., 2020, November. Deepfake video detection using audio-visual consistency. In *International Workshop on Digital Watermarking* (pp. 168-180). Cham: Springer International Publishing.

[24] Qian, Y., Yin, G., Sheng, L., Chen, Z. and Shao, J., 2020, August. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision* (pp. 86-103). Cham: Springer International Publishing.

[25] Hongmeng, Z., Zhiqiang, Z., Lei, S., Xiuqing, M. and Yuehan, W., 2020, July. A detection method for deepfake hard compressed videos based on super-resolution reconstruction using CNN. In *Proceedings of the 2020 4th high performance computing and cluster technologies conference & 2020 3rd international conference on big data and artificial intelligence* (pp. 98-103).

[26] Agarwal, S., Farid, H., El-Gaaly, T. and Lim, S.N., 2020, December. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.

[27] Marra, F., Saltori, C., Boato, G. and Verdoliva, L., 2019, December. Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.

[28] Nguyen, H.H., Yamagishi, J. and Echizen, I., 2019. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*.