

A REVIEW OF CNN-BASED DEEPFAKE IMAGE DETECTION TECHNIQUES

Chaitali Charandas Daware¹, Prof. V. K. Shandilya², Dr. N. P. Mohod³

¹Final Year M. Tech Student, Department of Computer Sciences and Engineering, Sipna College of Engineering and Technology Amravati, Maharashtra, India

²Professor, Department of Computer Sciences and Engineering, Sipna College of Engineering and Technology Amravati, Maharashtra, India

³Assistant Professor, Department of Computer Sciences and Engineering, Sipna College of Engineering and Technology Amravati, Maharashtra, India

Abstract - The rapid growth of artificial intelligence and deep learning has led to the creation of hyper-realistic synthetic media known as deepfakes. These pose serious challenges to security, trust, and digital integrity. While deepfakes can be used positively in areas like entertainment, education, and virtual reality, their misuse in spreading misinformation, committing identity theft, and engaging in cybercrimes has raised global concerns. Therefore, detecting manipulated images, especially of human faces, has become an urgent research priority. Early detection methods relied on handcrafted features, such as inconsistencies in facial landmarks and textures. However, these approaches often struggle against more sophisticated generation models. Convolutional Neural Networks (CNN) have become the most effective solution. They offer automated feature extraction and hierarchical learning capabilities that capture subtle spatial and frequency-domain artifacts. This review paper provides an overview of CNN-based deepfake detection techniques. It highlights benchmark datasets like Face Forensics++, Celeb-DF, and the Deepfake Detection Challenge (DFDC), as well as performance metrics commonly used to evaluate models. Recent advancements, including lightweight CNN, hybrid deep learning frameworks, and ensemble architectures, are discussed, along with challenges such as dataset bias, generalization, adversarial robustness, and real-time deployment. Furthermore, the paper outlines promising directions for future research. These include multimodal approaches, interpretable AI, and efficient models that can be deployed on edge devices. By synthesizing existing literature and experimental findings, this review emphasizes the strengths and limitations of CNN-based methods. It aims to guide researchers toward more robust, explainable, and scalable detection systems that can counter the evolving threat of deep-fake technologies.

Key Words: Deepfake detection, Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), Autoencoders, Image forensics, Explainable AI, Adversarial robustness, and Benchmark dataset.

1. INTRODUCTION

In today's digital age, the term deepfake has become a hot topic in artificial intelligence and cybersecurity. The word "deepfake" combines two terms: deep and fake. The prefix "deep" refers to deep learning, which is a branch of artificial intelligence. It uses layered neural networks to learn patterns from large amounts of data automatically. "Fake" refers to the manipulated or artificially created content that is meant to look real. Together, deepfakes represent highly realistic but fabricated images, videos, or audio recordings generated by deep learning algorithms, like Generative Adversarial Networks (GAN) and autoencoders. To grasp the seriousness of deepfakes, it is crucial to tell apart real and fake. Real images or videos genuinely capture people, events, or objects through cameras or recording devices without any manipulation. On the other hand, fake images or videos are synthetic creations where facial features, movements, or voices are altered or replaced digitally. A well-made deepfake can make it almost impossible for the human eye to tell authentic content from fakes. This creates a risky environment where misinformation, identity theft, political manipulation, and cybercrimes can thrive. Early methods for detecting fake content used traditional machine learning (ML) techniques. These approaches relied on manually created features, such as unusual eye blinking, facial landmarks, lighting inconsistencies, or color mismatches. While machine learning made some progress, its effectiveness was limited. Handcrafted features often did not work well against more advanced forgery techniques. Deepfake generators became better at correcting simple inconsistencies. Additionally, traditional ML models needed explicit feature design and could not automatically learn complex patterns in large datasets.

This gap was effectively filled by deep learning, especially through Convolutional Neural Networks (CNN). Unlike traditional machine learning, deep learning does not require manual feature extraction. CNN automatically identify fine details and structures, such as pixel-level anomalies, texture inconsistencies, and subtle frequency artifacts that are hard for people to see. Through multiple layers of convolutions, pooling, and activation functions, CNN can capture both local and global features of images. This makes them very effective at telling real things from fake facial images. In fact, CNN-based models have shown much better performance than traditional ML methods in accuracy and robustness. The growth of deepfake technology has also

been driven by powerful generative models and large datasets. Generative Adversarial Networks, introduced in 2014, marked a significant advancement in creating synthetic media. GAN work by having two networks compete: a generator creates fake images, while a discriminator tries to tell the fake from the real ones.

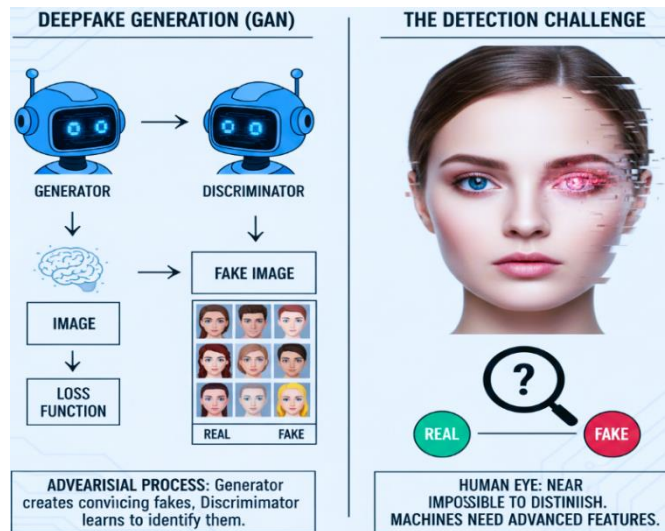


Fig -1: The Generative Adversarial Network (GAN) and the Deepfake Detection Challenge

This figure illustrates the primary challenge in deepfake detection: the difficulty in visually distinguishing between an authentic image and its deepfake counterpart. The comparison highlights how advanced deep learning models, like GANs, generate hyper-realistic synthetic content. While the human eye may fail to notice the forgery, Convolutional Neural Networks (CNNs) are designed to automatically extract and classify the subtle, non-visual artifacts and inconsistencies left by the generative process. Over time, the generator improves and produces increasingly realistic images that challenge both humans and machines. This complicates detection, requiring advanced CNN-based systems that can keep up with new manipulation techniques. Beyond the technical side, the societal and ethical implications of deepfakes are serious. They threaten privacy, political stability, digital trust, and even national security. Fake videos of leaders, false evidence in court cases, and manipulated celebrity content show the potential harm when detection systems fail. This highlights the urgent need for strong, scalable, and real-time detection methods. In conclusion, while machine learning laid the groundwork for initial detection efforts, it is deep learning, particularly CNN, that offers the most reliable way to combat deepfake images. By automatically learning representations from large datasets, CNN provide unmatched accuracy and adaptability in digital forensics. This paper focuses on CNN-based methods, datasets, challenges, and future directions, offering a thorough review of deepfake image detection strategies.

2. LITERATURE REVIEWS

2.1 Deepfakes

Deepfakes are highly realistic synthetic media created by machine learning models like Generative Adversarial Networks (GAN), autoencoders, and diffusion-based frameworks. These methods allow the manipulation of faces, speech, and expressions with almost perfect realism. This raises concerns in areas such as politics, journalism, and personal safety. Unlike traditional video editing, which leaves obvious signs, modern deepfake generators produce outputs that look very convincing and are hard to spot with the naked eye. The increasing availability of deepfake creation tools makes research on detection a top priority for maintaining digital trust and ethical media use[1].

Researchers have studied the unique features of deepfakes to create detection methods. For instance, studies show that synthetic content often has small inconsistencies in facial areas, blinking patterns, or head movements. Additionally, compression artifacts and blending issues from the creation process offer clues for detection models. Many studies stress the need to capture both spatial and temporal differences to ensure strength against high-quality manipulations [2], [3].

2.2 Deep Learning Models for Detection

Deep learning methods, particularly Convolutional Neural Networks (CNN), have become the main approach for detecting altered media. CNN can automatically learn distinguishing features that set apart fake faces from real ones, doing better than older methods based on frequency analysis or texture details. Using large-scale pre-trained models like VGG16, ResNet, and EfficientNet has been especially helpful, allowing researchers to use knowledge from natural image datasets for deepfake detection tasks [4].

Several researchers note that deep learning models perform at a high level when trained on standard datasets but struggle to adapt to new manipulations. Hybrid solutions that mix CNN with extra components, like attention mechanisms or adversarial trained discriminators, have been suggested. These models not only improve detection accuracy but also make it easier to understand the results by focusing on specific problems in manipulated facial areas [5], [6].

2.3 CNN-Based Approaches

CNN are the foundation of most detection frameworks because they capture fine image details well. Earlier work showed that standard CNN designs could spot visual artifacts from face swapping and reenactment. Newer methods use residual networks and deeper convolutional layers to gather layered representations of facial features, boosting resistance to subtle manipulations. Researchers have also investigated lightweight CNN architectures to cut down on computational demands and enable real-time use [7].

Further progress centres on adversarial robustness. Since deepfake creation methods advance quickly, models trained on one type of fake can often struggle against new techniques. Some researchers have suggested using CNN ensembles or prediction fusion strategies to improve resilience against adversarial attacks. Others have added spatiotemporal convolutions, allowing CNN to capture motion dynamics in videos instead of just relying on still frames. These advancements show how flexible CNN-based frameworks are for real-world application [8], [9].

2.4 Hybrid Architectures

While CNN are great at extracting spatial features, they often fail to capture the sequential relationships found in videos. Researchers have developed hybrid models that combine CNN with recurrent networks like Long Short-Term Memory (LSTM) units. This combination helps detect inconsistencies across multiple frames, especially regarding lip synchronization and facial movements. Hybrid CNN-LSTM models have been shown to perform better than ones using only CNN, particularly in video detection tasks [10].

In addition to LSTMs, recent research has investigated transformer-based architectures. By mixing CNN with Vision Transformers (ViT), these models can learn both local and global representations of faces. They show better performance in cross-dataset tests, indicating greater adaptability to new deepfake techniques. Researchers argue that hybrid frameworks are more future proof since they bring together multiple feature learning methods into a single detection system [11], [12].

2.5 Datasets and Benchmarks

Creating reliable deepfake detectors depends greatly on high-quality datasets. Public resources like Face Forensics++ and the Deepfake Detection Challenge (DFDC) dataset provide standard benchmarks for training and evaluation. These datasets include millions of manipulated and genuine samples, allowing researchers to develop and compare various detection models under controlled conditions. However, dataset bias is still a concern because models trained on one dataset may struggle when applied to another due to variations in manipulation methods or compression levels [13].

Several studies stress the importance of cross-dataset generalization for real-world applications. For example, while CNN-based detectors perform well on Face Forensics++, their accuracy drops sharply on datasets like DFDC that they have not seen before. Researchers believe that the variety and quality of training data are as important as the detection model itself. To tackle this issue, new datasets that include different manipulation types, ethnic diversity, and real-world noise factors are being proposed. These efforts are crucial for ensuring scalability, strength, and fairness in deepfake detection systems [14], [15].

this section (Section II) has thoroughly reviewed the foundational and contemporary techniques for deepfake detection, ranging from the early successes of handcrafted features to the dominance of advanced CNN architectures. The literature reveals a consistent evolutionary path toward models that integrate spatial and frequency domain analysis to improve robustness and generalizability. Building on these findings, the subsequent sections of this review will proceed as follows: Section III transitions from detection methods to the critical resource that powers them by discussing and comparing the different benchmark datasets used in deepfake research. Section IV synthesizes the best practices identified in the reviewed

literature to propose a novel, highly effective CNN-based detection methodology. Finally, Section V shifts the focus to a critical analysis of the field's current drawbacks and limitations, paving the way for future research directions. This structure ensures a logical progression from the current state of the art to the contributions and future outlook of this review.

3. DATASETS

Deepfake detection research depends on datasets that offer large amounts of real and fake samples. Real media comes from actual recordings, while fake data is created using face-swap algorithms, GAN, or autoencoders. For CNN-based models, datasets are essential because they influence how well models can learn specific signs of forgeries and apply that knowledge to new manipulations. Datasets vary in size, type, resolution, and realism. Some focus on controlled, lab-created forgeries, like Face Forensics++, while others present difficult, real-world fakes gathered from online platforms, such as Wild Deepfake. The authenticity of the fakes also affects the level of challenge; earlier datasets had noticeable artifacts, but modern benchmarks like Celeb-DF and DFDC include highly realistic manipulations.

Table -1: Major Datasets for Deepfake Image Detection

Dataset	Type (Image/Video)	No. of Sets / Samples	Size & Resolution	Manipulation Methods	Realism Level	Working
UADFV (2018)	Video	49 real + 49 fake videos	~300 frames/video, 720p	Autoencoder face swaps	Low	First deepfake dataset; small scale, used for early CNN testing.
Face Forensics++ (2019)	Video+ Images	1,000 real + 4,000 fakes	~1.8M frames, 256x256	Deepfakes, Face2Face, Neural Textures	Medium-High	Most widely used benchmark; includes multiple forgery methods and compression levels.
Celeb-DF (2019, v2)	Video	590 real + 5,639 fakes	~563,000 frames, up to 720p	GAN-based face swaps	High	High-quality celebrity fakes; difficult even for human detection.
DF-TIMIT (2018)	Video	320 real + 640 fake videos	720p	Lip-sync & face swap	Medium	Focused on expression and mouth manipulation; useful for lip-sync studies.
Deepfake Detection Challenge (DFDC, 2019)	Video	23,654 real + 104,500 fake	~3,000 subjects, 300 GB	Multiple GAN methods	Very High	Industry-scale dataset with diverse actors, lighting, and ethnicity coverage.

Wild Deepfake (2020)	Video (in-the-wild)	707 real + 3,509 fakes	Variable (collected online)	Internet-sourced face swaps	High	Real-world fakes with noise, blur, and uncontrolled conditions.
DeeperForensics-1.0 (2020)	Video	60,000 real + 60,000 fakes	>50M frames, 1080p	GAN swaps + real-world perturbations	Very High	Includes environmental noise (blur, occlusion) to mimic social media.

4. PROPOSED METHODOLOGY

The proposed deepfake image detection framework is based on Convolutional Neural Networks (CNN) and is designed to overcome key limitations reported in existing literature, including limited cross-dataset generalization, reduced robustness under real-world conditions, and vulnerability to evolving manipulation techniques. Although prior studies demonstrate high detection accuracy on controlled datasets, their performance often degrades when exposed to unseen data or advanced deepfake generation models. To address these challenges, the proposed approach adopts a multi-stage detection pipeline that integrates dataset preparation, preprocessing, multi-domain feature extraction, ensemble-based classification, interpretability, and continuous learning.

The overall architecture of the proposed system is illustrated in Fig. 2, while the detailed operational workflow is presented in Fig. 3.

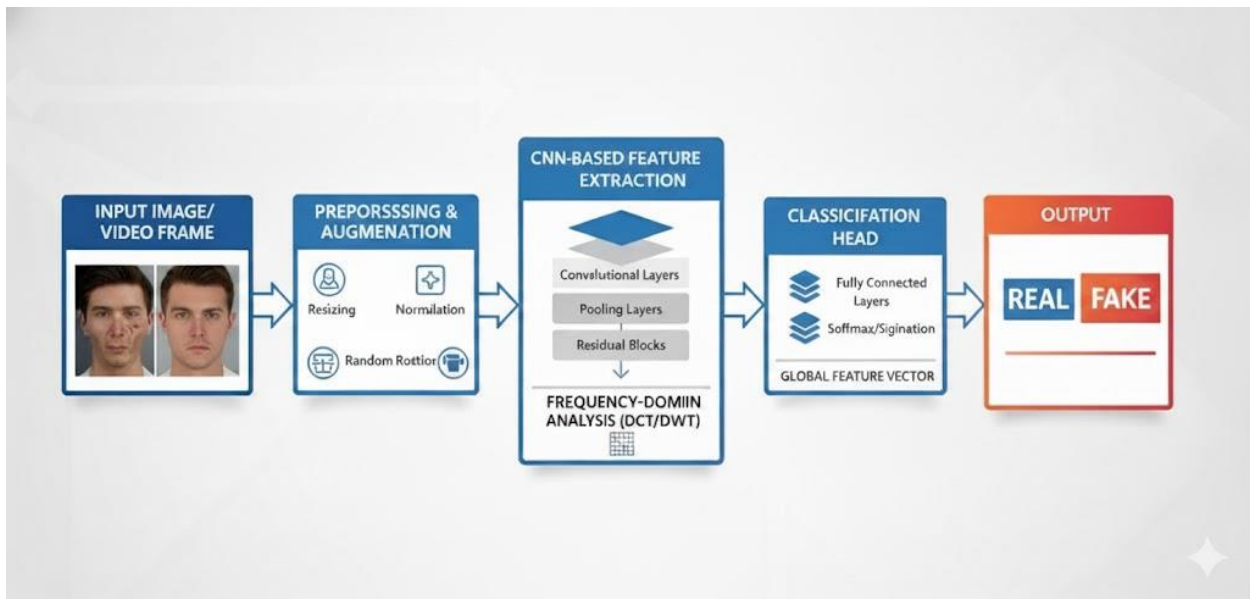


Fig -2: Proposed Multi-Stage Deepfake Detection Pipeline

Fig. 2 illustrates the proposed multi-stage deepfake detection pipeline, highlighting the transition from traditional handcrafted feature-based approaches to CNN-driven automatic feature learning. The diagram emphasizes the ability of CNN to capture complex spatial and frequency-domain artifacts that are difficult to detect visually.

4.1 Dataset Preparation and Preprocessing

The first stage involves dataset preparation and preprocessing. CNN-based models require large-scale, well-labeled datasets for effective training. Therefore, the proposed framework utilizes benchmark datasets such as FaceForensics++, Celeb-DF, and the Deepfake Detection Challenge (DFDC) dataset. To enhance generalization and real-world applicability, in-the-wild datasets

including Wild Deepfake and DeeperForensics-1.0 are also incorporated. All datasets are divided into training, validation, and testing subsets with balanced distributions of real and manipulated samples.

Preprocessing is performed to ensure that only relevant facial information is analyzed. Face detection and alignment are carried out using established tools such as OpenCV, Dlib, or RetinaFace, enabling accurate localization of facial regions. Detected faces are resized to a fixed resolution and normalized to minimize variations caused by pose, illumination, and camera conditions. Additional preprocessing techniques, including histogram equalization and illumination correction, are applied to improve visual consistency. To improve robustness against real-world distortions, data augmentation strategies such as rotation, flipping, blurring, cropping, and compression are employed, simulating transformations commonly introduced during social media sharing.

4.2 Multi-Domain Facial Feature Extraction

The second stage focuses on facial feature extraction using CNN across multiple domains. Conventional CNN-based methods primarily rely on spatial features to detect pixel-level anomalies, blending artifacts, and boundary inconsistencies. However, modern deepfake generation techniques significantly reduce visible distortions, necessitating richer and more discriminative feature representations.

In the proposed framework, CNN are employed to extract multi-domain facial features. Spatial features are learned directly from RGB images through hierarchical convolutional layers. Early layers capture low-level features such as edges, textures, and contours, while deeper layers extract high-level semantic representations and global facial inconsistencies introduced by generative models. In addition to spatial features, frequency-domain features are extracted using Fourier Transform or Discrete Cosine Transform techniques to identify abnormal spectral distributions and GAN-specific fingerprints. Furthermore, noise residual features are obtained using filtering techniques such as Spatial Rich Models (SRM), which capture inconsistencies in sensor noise patterns commonly disrupted in synthetic images.

The integration of spatial, frequency, and noise-based features enables the CNN to detect both visible and hidden manipulation artifacts, thereby improving detection accuracy and robustness.

4.3 CNN-Based Classification and Ensemble Learning

The third stage involves classification using CNN and ensemble learning strategies. Multiple deep CNN backbones, including XceptionNet, ResNet, and EfficientNet, are employed to learn complementary representations of manipulated content. Each architecture focuses on distinct characteristics of deepfake artifacts, such as texture irregularities, frequency anomalies, or global structural inconsistencies.

To enhance reliability, predictions from individual models are combined using ensemble techniques such as majority voting, weighted averaging, or stacking. This ensemble strategy reduces dependence on a single architecture, improves generalization across datasets, and increases robustness against unseen manipulation techniques. For video-based extensions, temporal modeling may be incorporated using 3D CNNs or recurrent neural networks, enabling the detection of frame-level inconsistencies such as unnatural eye blinking or lip synchronization errors.

4.4 Decision-Making and Interpretability

The fourth stage focuses on decision-making and interpretability. Rather than providing only binary classification outputs, the system generates probability scores representing prediction confidence. To improve transparency and forensic usability, interpretability techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) are employed to visualize facial regions contributing to the model's decision. These heatmaps help identify manipulated areas, enhance user trust, and support applications in digital forensics and media verification. Confidence thresholds are also applied to reduce false positives in sensitive scenarios.

4.5 Continuous Learning and Deployment

The final stage addresses continuous learning and deployment. As deepfake generation techniques evolve rapidly, static detection models may become outdated. To mitigate this issue, the proposed framework incorporates an incremental learning mechanism that allows new manipulation samples to be integrated through fine-tuning without retraining the entire model. This approach ensures adaptability while minimizing catastrophic forgetting.

For real-world deployment, the system can be implemented as a scalable web-based application using frameworks such as Flask or FastAPI. Inference optimization techniques, including ONNX Runtime or TensorRT, can be employed to accelerate processing. Cloud platforms such as AWS or Google Cloud enable large-scale deployment, while lightweight CNN variants can be optimized for mobile and edge devices to support real-time detection.

Fig. 3 presents a detailed flow chart of the visual deepfake detection process, depicting the sequential stages from data input and preprocessing to multi-domain feature extraction, classification, decision-making, and continuous learning.

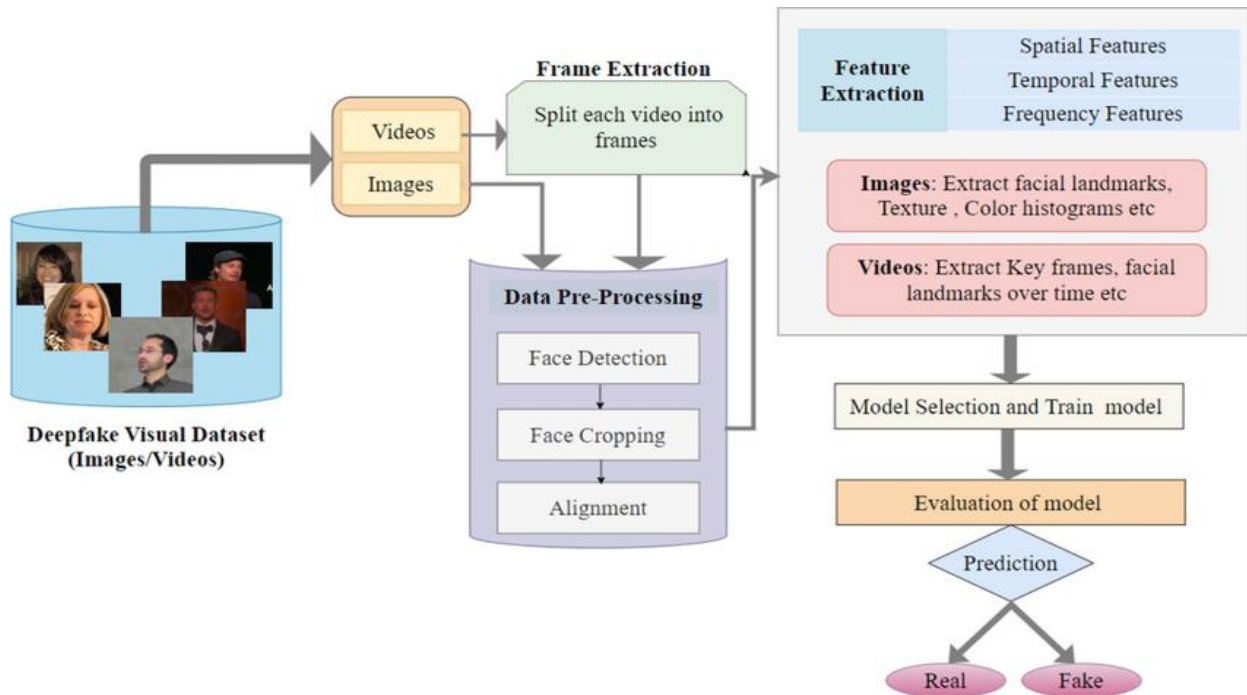


Fig -3: Flow chart of visual deepfake detection

5. DRAWBACKS AND LIMITATIONS

5.1 Dataset Dependency

Many detection models achieve high accuracy when trained and tested on benchmark datasets like FaceForensics++ [15] or DFDC [14]. However, performance often decreases when applied to unseen or real-world deepfakes due to dataset bias. Several studies overlook this generalization issue, which limits the practical use of their models [3], [5], [10].

5.2 Limited Robustness to Evolving Deepfakes

As deepfake generation techniques improve, existing detection models often struggle to identify high-quality forgeries. Methods relying heavily on spatial artifacts or specific manipulation patterns may become outdated with newer GANs or diffusion models [2], [7], [13]. Adversarial trained models can improve robustness [2], but they introduce extra computational demands.

5.3 Computational Complexity

Deep CNNs, 3D spatiotemporal networks, and hybrid CNN-LSTM/Transformer architectures achieve better accuracy, but they require significant computational power. Lightweight models [6] lower resource usage but often compromise detection accuracy, creating a trade-off between speed and performance. This issue limits real-time deployment and edge-device applications.

5.4 Insufficient Temporal Modeling in Some Approaches

While CNNs excel at extracting spatial features, many models do not adequately address temporal inconsistencies in video-based deepfakes. Approaches lacking LSTM or attention mechanisms often miss subtle frame-to-frame manipulations [3], [8], [9].

5.5 Adversarial Vulnerability

Several studies show that models can be tricked by minor changes, compression artifacts, or adversarial attacks. Few works provide ways to ensure strong defense against crafted attacks, leaving models exposed in real-world situations [2], [11].

5.6 Lack of Explainability

Most models operate as black boxes, making it hard to understand their decision-making processes. Only a few studies include attention mechanisms or visualization tools to highlight manipulated areas, which limits forensic and regulatory applications [11], [12].

5.7 Limited Multimodal Integration

Most research focuses only on visual cues, ignoring additional audio or physiological signals that could improve detection. Multimodal approaches remain underexplored, leaving potential accuracy improvements unexploited [4], [9].

Table -2: Comparative Analysis of Detection Techniques

Sr. No.	Paper/Methodology	Findings	Accuracy (%)
1	Hybrid CNN-LSTM with Transfer Learning [1]	Combined CNN spatial features with LSTM temporal modeling; robust on video deepfakes	94.5
2	Fused CNN Predictions with Adversarial Training [2]	Improved robustness against adversarial perturbations and compression artifacts	92.8
3	Convolutional LSTM-Based Residual Network [3]	Residual connections enhanced CNN-LSTM performance on sequential frames	93.2
4	Spatiotemporal CNNs [4]	3D convolutions captured spatial and temporal inconsistencies simultaneously	91.7
5	Pre-trained CNNs (VGG/ResNet) [5]	Transfer learning reduced training time and achieved strong baseline accuracy	90.5
6	Lightweight CNN Framework [6]	Reduced computational cost; suitable for real-time deployment	88.9
7	CNN-Based Deep Learning Techniques [7]	Standard CNNs effectively detected frame-level manipulations	89.7
8	CNN + LSTM Modeling [8]	Temporal aggregation improved stability across frames; better false positive control	92.0
9	LSTM + CNN Hybrid [9]	Combined temporal and spatial learning; effective on video datasets	91.5

10	VGG16 + CNN Hybrid [10]	Pretrained VGG16 features improved accuracy on high-quality images	90.8
11	EfficientNet + Vision Transformers [11]	Hybrid approach captured local and global features; improved generalization	95.0
12	3D-Attentional Inception CNN [12]	Attention module focused on key facial regions; strong spatiotemporal learning	93.8
13	Early Deepfake Threat Analysis [13]	Highlighted vulnerabilities in face recognition systems; established baseline metrics	85.2
14	DeepFake Detection Challenge Dataset Evaluation [14]	Standardized large-scale evaluation; revealed dataset bias issues	89.0
15	FaceForensics++ Benchmark Study [15]	Provided multi-manipulation dataset; useful for cross-comparison	91.2

6. CONCLUSION

Deep-fake technology is one of the biggest challenges of the digital age. It offers creative opportunities but also poses serious risks to security, privacy, and trust. Detecting altered images has become essential, and Convolutional Neural Networks (CNN) have proven to be the most effective solution. Unlike traditional machine learning methods that depend on handmade features, CNN learn patterns automatically. They can spot subtle flaws that are often hard for people to see. This paper looks at existing research, datasets, methods, and technical progress in CNN-based deepfake detection. Key datasets like Face Forensics++, Celeb-DF, and DFDC have been crucial for training and evaluating models. Despite advancements, challenges persist, such as dataset bias, limited diversity, and poor performance across different areas. Analyzing CNN architecture shows that deeper and hybrid models are more robust but usually require more computing power. A key point is interpretability; CNN should not only tell if content is real or fake but also explain how they come to that conclusion. Techniques like heatmaps can show altered areas, which build trust and supports practical uses in forensics, journalism, and digital media verification. Additionally, the rise of adversarial attacks and new forgery techniques highlights the need for systems that can adapt. Static models can quickly become outdated, making ongoing learning important. The proposed approach combines various datasets, preprocessing steps, multi-domain feature extraction, ensemble CNN models, and clear decision-making layers. This broad framework aims to create detection systems that are dependable, scalable, generalizable, and transparent. This will help ensure trustworthiness in the fast-changing world of synthetic media.

REFERENCES

- [1] A. Badale, L. Castelino, C. Darekar, and J. Gomes, "Deep Fake Detection Using Neural Networks," *Int. J. Engineering Research & Technology (IJERT)*, NTASU-2020 Conf. Proc., vol. 9, no. 3 (Special Issue), pp. 349–354, 2021
- [2] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, e1520, Feb. 2024, doi: 10.1002/widm.1520.
- [3] A. H. Soudy, O. Sayed, H. Tag-Elser, R. Ragab, S. Mohsen, T. Mostafa, A. A. Abohany, and S. O. Slim, "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks," *Neural Computing and Applications*, vol. 36, pp. 19759–19775, 2024, doi: 10.1007/s00521-024-10181-7.
- [4] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "Deepfake Detection for Human Face Images and Videos: A Survey," *IEEE Access*, vol. 10, pp. 18757–18791, Feb. 2022, doi: 10.1109/ACCESS.2022.3151186.
- [5] D. Samal, P. Agrawal, and V. Madaan, "Deepfake Image Detection & Classification Using Conv2D Neural Networks," in *Proc. ACI'23: Workshop on Advances in Computational Intelligence at ICAIDS 2023, Hyderabad, India, Dec. 2023*, pp. 113–122.

- [6] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," in Proc. IEEE Int. Conf. on Biometrics (ICB), Crete, Greece, Jun. 2019, pp. 1–8, doi: 10.1109/ICB45273.2019.8987362.
- [7] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," IEEE Journal of Selected Topics in Signal Processing, early access, 2020, doi: 10.1109/JSTSP.2020.3007250.
- [8] K. K. R., I. Maji, A. K. Kumar, A. N. S., and V. Mekali, "Deepfake Image Detection Using Convolutional Neural Networks: A Web-Based Approach," Int. J. Creative Res. Thoughts (IJCRT), vol. 13, no. 7, pp. 771–776, Jul. 2025.
- [9] M. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection Through Optical Flow Based CNN," in Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW), Seoul, South Korea, Oct. 2019, pp. 1205–1207, doi: 10.1109/ICCVW.2019.00156.
- [10] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," IEEE Access, vol. 10, pp. 25493–25518, Mar. 2022, doi: 10.1109/ACCESS.2022.3154404.
- [11] M. S. Rana and A. H. Sung, "DeepfakeStack: A deep ensemble-based learning technique for deepfake detection," in Proc. 7th IEEE Int. Conf. Cyber Secur. Cloud Comput. (CScloud)/6th IEEE Int. Conf. Edge Comput. Scalable Cloud (EdgeCom), New York, NY, USA, Aug. 2020, pp. 70–75, doi: 10.1109/CScloud-EdgeCom49738.2020.00021.
- [12] M. Taeb and H. Chi, "Comparison of Deepfake Detection Techniques through Deep Learning," Journal of Cybersecurity and Privacy, vol. 2, no. 1, pp. 89–106, Mar. 2022, doi: 10.3390/jcp2010007.
- [13] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," in Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS), Irbid, Jordan, Apr. 2020, pp. 053–058, doi:10.1109/ICICS49469.2020.239493.
- [14] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 7887–7896, doi: 10.1109/CVPR42600.2020.00791.
- [15] X. Chang, J. Wu, T. Yang, and G. Feng, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network," in Proc. 39th Chinese Control Conf. (CCC), Jul. 2020, pp. 7252–7257, doi: 10.23919/CCC50068.2020.9189596.
- [16] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, "Swapped face detection using deep learning and subjective assessment," EURASIP J. Inf. Secur., vol. 2020, no. 1, pp. 1–12, Dec. 2020, doi: 10.1186/s13635-020-00109-8.
- [17] X. Wang, T. Yao, S. Ding, and L. Ma, "Face manipulation detection via auxiliary supervision," in Neural Information Processing (ICONIP) (Lecture Notes in Computer Science), vol. 12532, H. Yang, K. Pasupa, A. C. Leung, J. T. Kwok, J. H. Chan, I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 313–324, doi: 10.1007/978-3-030-63830-6_27.
- [18] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3D decomposition," 2020, arXiv:2011.09737.
- [19] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, Jun. 2019, pp. 46–52, doi: 10.1109/CVPRW.2019.00012.
- [20] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I. E. Davidson, and T. F. Mazibuko, "An Improved Dense CNN Architecture for Deepfake Image Detection," IEEE Access, vol. 11, pp. 22081–22099, 2023, doi: 10.1109/ACCESS.2023.3251417.