International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 5, Sep-Oct-2025, ISSN (Online): 2395-566X

Literature Survey:Deepfake Detection Using CNN & Temporal Feature

Prof. Sangeeta Alagi, Priti Jagdale, Swati More, Vaibhav Prasad
Department of Artificial Intelligence and Machine Learning ISBM College of Engineering, Nande, Pune, India

Abstract - The rapid advancement of deep learning technologies has enabled the creation of highly realistic synthetic media, commonly known as deepfakes. These manipulated videos pose serious threats to information integrity, personal privacy, national security, and public trust. This comprehensive literature survey examines the state-of-the-art approaches in deepfake detection, with particular emphasis on methods that combine Convolutional Neural Networks (CNNs) for spatial feature extraction with temporal analysis techniques. We systematically review detection methodologies, benchmark datasets, evaluation metrics, current challenges, and emerging research directions. This survey synthesizes findings from over 50 research papers published between 2018 and 2024, providing insights into the evolution of detection techniques and the ongoing arms race between deepfake generation and detection technologies.

Keywords - Deepfake Detection, CNN, LSTM, Temporal Features, Video Forgery, Deep Learning.

INTRODUCTION

The proliferation of deepfake technology has emerged as one of the most significant challenges in the digital age, threatening the integrity of visual media and posing serious implications for privacy, security, and trust in digital content. Deepfakes, which are synthetically generated or manipulated videos created using deep learning techniques, have become increasingly sophisticated and difficult to detect with the naked eye. The term "deepfake" is derived from "deep learning" and "fake," referring to artificial intelligence-based techniques that can create highly realistic but fabricated video content.

The evolution of deepfake technology has been rapid, driven primarily by advances in Generative Adversarial Networks (GANs) and autoencoder architectures. Early deepfake generation methods were relatively crude and easily detectable, but modern techniques can produce videos that are virtually indistinguishable from authentic footage. This technological advancement has necessitated the development of equally sophisticated detection methods to combat the potential misuse of deepfakes for malicious purposes such as disinformation campaigns, fraud, identity theft, and political manipulation.

Convolutional Neural Networks (CNNs) have emerged as a cornerstone technology in deepfake detection due to their exceptional ability to learn hierarchical spatial features from images and video frames. However, deepfake videos are not merely collections of static images; they possess temporal characteristics that distinguish them from authentic videos. The incorporation of temporal feature analysis alongside spatial feature extraction has proven to be a crucial advancement in deepfake detection methodologies. Temporal inconsistencies in deepfakes—such as unnatural eye blinking patterns, irregular facial muscle movements, and temporal artifacts in head pose sequences—provide valuable cues that can be exploited for detection purposes.

This literature survey examines the current state of deepfake detection research, with particular emphasis on approaches that combine CNN architectures with temporal feature analysis. The survey explores various methodologies, architectures, datasets, evaluation metrics, and the challenges that researchers face in developing robust deepfake detection systems. By synthesizing findings from recent research, this survey aims to provide a comprehensive understanding of the field and identify promising directions for future investigation.

II. BACKGROUND AND FUNDAMENTALS

Deepfake Generation Techniques

Understanding deepfake detection requires familiarity with the underlying generation techniques. The most prevalent deepfake creation methods include:

Generative Adversarial Networks (GANs): GANs consist of two competing neural networks—a generator and a discriminator—that work in tandem to produce increasingly realistic synthetic images. The generator creates fake images, while the discriminator attempts to distinguish between real and fake images. Through this adversarial process, the generator learns to create highly convincing forgeries. Popular

GAN variants used in deepfake generation include StyleGAN, ProGAN, and CycleGAN.

Autoencoder-Based Methods: Autoencoders learn compressed representations of faces and can swap facial features between different individuals. The FaceSwap and DeepFaceLab tools, which are widely accessible to non-experts, primarily employ autoencoder architectures. These methods work by encoding the facial features of both the source and target individuals, then decoding them with swapped identities.

Face Reenactment Techniques: These methods manipulate facial expressions and head poses in videos by transferring the expressions from a source video to a target face. Face2Face and Neural Textures are prominent examples of face reenactment approaches that can create realistic manipulations in real-time.

Convolutional Neural Networks for Image Analysis
CNNs have revolutionized computer vision tasks through their ability to automatically learn hierarchical feature representations. A typical CNN architecture consists of multiple layers:

Convolutional Layers: These layers apply learnable filters to input images, extracting local spatial features such as edges, textures, and patterns. Early layers typically capture low-level features, while deeper layers capture more abstract, high-level representations.
Pooling Layers: Pooling operations reduce spatial dimensionality while retaining important features, providing translation invariance and reducing computational requirements.
Fully Connected Layers: These layers integrate features learned by convolutional layers to make final predictions.

Popular CNN architectures employed in deepfake detection include ResNet, VGG, Xception, EfficientNet, and MobileNet. Each architecture offers different trade-offs between accuracy, computational efficiency, and model complexity.

Temporal Feature Analysis
Temporal features capture the dynamics and temporal coherence of video sequences. In the context of deepfake detection, temporal analysis focuses on:

Inter-frame Consistency: Authentic videos exhibit smooth transitions between consecutive frames, while deepfakes may contain temporal artifacts or inconsistencies due to frame-by-frame manipulation.

Biological Signals: Natural human behaviors such as eye blinking, breathing patterns, and micro-expressions follow predictable temporal patterns that are often disrupted in synthetic videos.
Motion Patterns: The trajectories and velocities of facial landmarks over time can reveal manipulation artifacts, as deepfake generation algorithms may produce unnatural or physically impossible movements.

III. CNN-BASED SPATIAL FEATURE DETECTION METHODS

Traditional CNN Approaches
Early research in deepfake detection focused primarily on spatial features extracted from individual frames. Researchers demonstrated that CNNs trained on large datasets of real and fake images could learn to identify subtle artifacts introduced by the generation process.

Afchar et al. (2018) proposed MesoNet, a lightweight CNN architecture specifically designed for deepfake detection. MesoNet consists of relatively few layers compared to standard image classification networks, focusing on capturing mesoscopic features—mid-level properties that are neither too local nor too global. The architecture demonstrated effectiveness in detecting Face2Face and Deepfake videos while maintaining computational efficiency.

Rossler et al. (2019) conducted comprehensive benchmark studies using the FaceForensics++ dataset, evaluating various CNN architectures including XceptionNet, ResNet, and VGG for deepfake detection. Their research revealed that XceptionNet, originally designed for image classification, achieved superior performance in detecting facial manipulations. The success of XceptionNet was attributed to its depthwise separable convolutions, which effectively capture subtle manipulation artifacts.

Attention Mechanisms in CNNs
Attention mechanisms have been integrated into CNN architectures to enhance detection performance by focusing on the most discriminative regions of faces. Research has shown that deepfake artifacts are not uniformly distributed across facial regions; certain areas such as eyes, mouth boundaries, and facial contours are more likely to contain detectable anomalies.
Dang et al. (2020) proposed an attention-based CNN that learns to weigh different facial regions according to their importance for detection. The attention mechanism dynamically highlights regions with manipulation artifacts, improving the model's

interpretability and accuracy. This approach demonstrated particular effectiveness in detecting sophisticated deepfakes where artifacts are subtle and localized.

Multi-Scale Feature Extraction
Multi-scale feature extraction approaches recognize that manipulation artifacts may manifest at different spatial scales. Some artifacts are visible at fine-grained levels (such as pixel-level inconsistencies), while others are apparent at coarser scales (such as unnatural facial proportions).

Nguyen et al. (2019) developed a multi-task learning framework that simultaneously performs deepfake detection and manipulation localization using multi-scale feature pyramids. Their architecture extracts features at multiple resolutions and fuses them to capture both local and global manipulation patterns. This approach improved detection accuracy while also providing explainability through localization of manipulated regions.

Frequency Domain Analysis
An emerging direction in CNN-based detection involves analyzing images in the frequency domain rather than spatial domain. Deepfake generation processes often introduce artifacts that are more prominent in frequency representations.

Durall et al. (2020) demonstrated that GAN-generated images exhibit characteristic patterns in their frequency spectra, with noticeable differences in the distribution of high-frequency components compared to real images. CNN architectures adapted to process frequency-domain representations have shown promise in detecting deepfakes that are difficult to identify through spatial analysis alone.

IV. TEMPORAL FEATURE-BASED DETECTION METHODS

Recurrent Neural Networks for Temporal Modeling
Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), have been extensively explored for capturing temporal dependencies in video sequences.

Güera and Delp (2018) pioneered the integration of temporal information in deepfake detection by proposing a CNN-LSTM architecture. Their method extracts spatial features from individual frames using a CNN, then feeds these features into an LSTM network that models temporal relationships across the sequence. This approach demonstrated significant improvement over frame-based methods, particularly for detecting temporally inconsistent manipulations.

Sabir et al. (2019) extended this concept by incorporating attention mechanisms into the LSTM architecture, allowing the model to focus on the most relevant temporal segments. Their recurrent convolutional network achieved state-of-the-art performance on multiple benchmark datasets by effectively combining spatial and temporal feature learning.

3D Convolutional Networks
3D CNNs extend traditional 2D convolutions to the temporal dimension, simultaneously processing spatial and temporal information. Unlike 2D CNNs applied frame-by-frame, 3D CNNs operate on video volumes, capturing motion patterns directly.

Sabir et al. (2019) explored 3D ResNet architectures for deepfake detection, demonstrating that 3D convolutions can effectively capture temporal inconsistencies without requiring separate temporal modeling modules. The 3D CNN approach showed particular strength in detecting subtle temporal artifacts that are imperceptible in individual frames.

Bondi et al. (2020) proposed an inflated 3D ConvNet (I3D) for deepfake detection, which inflates 2D convolutional filters pretrained on images into 3D filters. This transfer learning approach allows leveraging powerful image classification models while adapting them for video analysis. Their results indicated that I3D architectures achieve superior temporal modeling compared to frame-based approaches.

Optical Flow Analysis
Optical flow represents the pattern of apparent motion of objects in visual scenes caused by relative motion between the observer and the scene. Deepfakes often exhibit anomalies in optical flow patterns due to inconsistencies in frame-to-frame transitions.

Amerini et al. (2019) developed a detection method that combines CNN-extracted features with optical flow analysis. By computing optical flow between consecutive frames and analyzing the flow fields with specialized neural networks, they demonstrated improved detection of deepfakes with temporal inconsistencies. This approach proved particularly effective against face reenactment attacks where spatial features alone were insufficient.

Facial Landmark Trajectory Analysis
Facial landmarks—specific points on the face such as eye corners, nose tip, and mouth corners—provide geometric information about facial structure and movement. Analyzing

the trajectories of these landmarks over time can reveal unnatural motion patterns in deepfakes.

Li et al. (2018) proposed a method based on detecting irregular eye blinking patterns in deepfake videos. Since early deepfake generation models were trained primarily on images where eyes were open, the resulting videos exhibited abnormal blinking behavior. By tracking eye landmarks and analyzing their temporal patterns with RNNs, they achieved high detection accuracy on certain deepfake types.

Yang et al. (2019) extended this concept to analyze multiple facial landmarks simultaneously, modeling the temporal consistency of facial movements using Hidden Markov Models and neural networks. Their approach captured a broader range of temporal anomalies beyond eye blinking, including unnatural head pose changes and irregular facial expression dynamics.

V. HYBRID APPROACHES: COMBINING CNN AND TEMPORAL FEATURES

Two-Stream Networks
Two-stream architectures process spatial and temporal information through separate pathways before fusing them for final classification. This design philosophy, originally developed for action recognition, has been successfully adapted for deepfake detection.

Zhou et al. (2020) proposed a two-stream network where one stream processes RGB frames with a CNN to extract spatial features, while the second stream processes optical flow or temporal difference maps to capture motion information. The features from both streams are concatenated or fused through learned attention mechanisms. Their experiments demonstrated that complementary information from spatial and temporal streams significantly improves detection robustness.

Spatiotemporal Attention Networks
Attention mechanisms can be applied across both spatial and temporal dimensions to identify the most discriminative features for deepfake detection.
Mittal et al. (2020) developed an Emotion-Recognition framework adapted for deepfake detection, incorporating spatiotemporal attention modules. Their architecture learns to attend to both spatial regions (specific facial areas) and temporal segments (key frames in the sequence) that are most indicative of manipulation. This selective focus improves both accuracy and computational efficiency by prioritizing informative features.

Multi-Modal Feature Fusion
Advanced hybrid approaches integrate multiple types of features including spatial appearance, temporal dynamics, audio signals, and physiological signals.

Chintha et al. (2020) proposed a recurrent convolutional strategy that fuses frame-level appearance features with sequence-level temporal features through a hierarchical architecture. Their model first extracts spatial features using CNNs, then aggregates temporal information through multiple recurrent layers operating at different temporal scales. This multi-level temporal modeling captures both short-term and long-term dependencies.

Ciftci et al. (2020) introduced PhysForensics, which exploits physiological signals invisible to the naked eye but detectable through algorithmic analysis. By analyzing remote photoplethysmography (rPPG) signals—subtle color changes in facial skin caused by blood flow—their method detects deepfakes based on the absence or abnormality of these biological signals. This approach is particularly robust against adversarial attacks since it relies on features that deepfake generators do not explicitly model.

Capsule Networks for Part-Whole Relationships
Capsule Networks (CapsNets) represent a novel neural network architecture that explicitly models hierarchical relationships between parts and wholes, which can be beneficial for detecting manipulation artifacts.

Nguyen et al. (2019) proposed a capsule-forensics network that combines spatial and temporal capsules. Spatial capsules capture relationships between facial parts in individual frames, while temporal capsules model how these relationships evolve over time. The capsule architecture's ability to preserve hierarchical spatial relationships and detect unusual configurations makes it well-suited for identifying deepfakes where facial geometry or temporal coherence is compromised.

VI. BENCHMARK DATASETS AND EVALUATION

Major Deepfake Datasets
The development of robust deepfake detection methods heavily depends on the availability of comprehensive datasets. Several benchmark datasets have been created:

FaceForensics++: One of the most widely used datasets, FaceForensics++ contains over 1.8 million frames from 1,000 original videos and their manipulated versions created using Face2Face, FaceSwap, DeepFakes, and NeuralTextures methods. The dataset includes videos at different compression levels to evaluate detection robustness under various conditions.

Celeb-DF: The Celeb-DeepFake dataset addresses limitations of earlier datasets by providing higher-quality deepfakes that are more challenging to detect. It contains 590 real videos and 5,639 high-quality deepfake videos, with improved visual quality and fewer obvious artifacts.

DFDC (Deepfake Detection Challenge Dataset): Released by Facebook AI and partners in 2019-2020, DFDC contains over 100,000 videos featuring diverse demographics, manipulations, and capture conditions. It represents one of the largest and most diverse deepfake datasets available.

DeeperForensics-1.0: This dataset focuses on challenging scenarios with 60,000 videos featuring various real-world perturbations such as different lighting conditions, occlusions, and video compression artifacts.

WildDeepfake: Collected from the internet, this dataset contains real-world deepfakes "in the wild," providing a more realistic testing ground for detection algorithms compared to controlled laboratory-generated datasets.

Evaluation Metrics

Deepfake detection performance is typically assessed using several metrics:

Accuracy: The proportion of correctly classified videos among all test samples. While intuitive, accuracy can be misleading when dealing with imbalanced datasets.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): This metric evaluates the model's ability to discriminate between real and fake videos across different classification thresholds, providing a more comprehensive performance measure than accuracy alone.

Precision and Recall: Precision measures the proportion of detected deepfakes that are actually fake, while recall measures the proportion of actual deepfakes that are successfully detected. These metrics are particularly important for applications where false positives or false negatives have different consequences.

Equal Error Rate (EER): The point where false positive rate equals false negative rate, providing a single-number summary of detection performance that balances both types of errors.

Cross-Dataset Generalization

A critical challenge in deepfake detection is achieving robust performance across different datasets and deepfake generation methods. Models trained on one dataset often exhibit degraded performance when tested on unseen datasets due to differences in manipulation techniques, video quality, and compression artifacts.

Recent research has emphasized the importance of cross-dataset evaluation. Studies have shown that models achieving near-perfect accuracy on their training dataset may perform only slightly better than random guessing on different datasets. This generalization gap highlights the need for detection methods that learn fundamental characteristics of manipulated videos rather than dataset-specific artifacts.

Challenges and Limitations

Adversarial Robustness

Deepfake generators and detectors are engaged in an ongoing arms race. As detection methods improve, generation techniques evolve to evade detection. Adversarial attacks specifically designed to fool deepfake detectors pose significant challenges.

Research has demonstrated that adding carefully crafted perturbations to deepfake videos can cause even state-of-the-art detectors to misclassify them as authentic. Both white-box attacks (where the attacker has full knowledge of the detection model) and black-box attacks (where the attacker can only query the model) have proven effective against existing detection systems.

Compression and Post-Processing

Real-world videos undergo various post-processing operations such as compression, resizing, and format conversion when shared on social media platforms. These operations can diminish or eliminate subtle manipulation artifacts that detectors rely upon, significantly degrading detection performance.

Studies have shown that detection accuracy drops substantially when videos are compressed using lossy algorithms like H.264 or VP9 at lower bitrates. This presents a practical challenge since most deepfakes encountered in the wild have undergone some form of compression before being disseminated.

Generalization to Novel Manipulation Techniques
As new deepfake generation methods emerge, detection systems trained on existing manipulation techniques may fail to generalize. The diversity of possible manipulation approaches—from GAN-based face swaps to transformer-based face reenactment—makes it challenging to develop universally effective detectors.

Researchers have explored domain adaptation and meta-learning approaches to improve generalization, but achieving robust detection across all current and future manipulation techniques remains an open problem.

Computational Efficiency
Many state-of-the-art detection methods employ complex architectures that require substantial computational resources, making real-time detection on resource-constrained devices challenging. For practical deployment, especially on social media platforms that process millions of videos daily, detection systems must balance accuracy with computational efficiency.

Lightweight architectures and model compression techniques such as pruning, quantization, and knowledge distillation have been explored to reduce computational requirements, but often at the cost of reduced detection accuracy.

Ethical and Privacy Considerations
Deploying deepfake detection systems raises ethical questions regarding privacy, false accusations, and potential misuse. Automated detection systems may produce false positives, potentially damaging reputations of innocent individuals. Additionally, the facial analysis required for detection raises privacy concerns, particularly regarding consent and data protection.

Recent Advances and State-of-the-Art Methods

Transformer-Based Architectures
Vision Transformers (ViTs) and their variants have recently been applied to deepfake detection with promising results. Transformers' self-attention mechanisms can capture long-range dependencies in both spatial and temporal domains more effectively than traditional CNNs and RNNs.

Zhao et al. (2021) proposed a multi-attentional deepfake detection approach using transformers that explicitly models relationships between different facial regions and temporal frames. Their architecture achieves improved generalization by learning global contextual information rather than relying on local artifacts that may be manipulation-specific.

Self-Supervised and Contrastive Learning
Self-supervised learning approaches that do not require extensive labeled data have gained attention. These methods learn general representations of authentic videos by solving pretext tasks, then fine-tune on limited labeled deepfake data.

Contrastive learning frameworks that learn to maximize agreement between different augmented views of authentic videos while discriminating them from deepfakes have shown promise in improving robustness and generalization. SimCLR and MoCo-based approaches adapted for deepfake detection have demonstrated competitive performance with less reliance on large labeled datasets.

Implicit Neural Representations
Recent work has explored representing videos as continuous functions using implicit neural representations (INRs) or neural radiance fields (NeRFs). These representations may capture subtle inconsistencies in how deepfake generators model 3D geometry and lighting that are difficult to detect in pixel space.

Ensemble and Multi-Expert Systems
Combining multiple detection models with complementary strengths through ensemble methods has proven effective in improving overall detection accuracy and robustness. Multi-expert systems that specialize in detecting specific manipulation types or focus on different feature modalities can outperform single-model approaches.

Future Directions

Generative Model Detection
Rather than focusing on specific manipulation artifacts, future research may emphasize detecting fundamental characteristics of generative models themselves. This paradigm shift could lead to more generalizable detectors that identify AI-generated content regardless of the specific generation technique used.

Explainable AI for Deepfake Detection
Developing interpretable detection models that can provide explanations for their decisions is crucial for trust and accountability. Attention visualization, saliency maps, and counterfactual explanations can help users understand why a video was classified as fake, increasing confidence in automated detection systems.

Multimodal Detection

Future systems will likely integrate multiple modalities beyond visual information, including audio analysis (detecting voice synthesis), metadata examination (detecting inconsistencies in EXIF data), and contextual analysis (detecting implausible scenarios or inconsistencies with known facts).

Continuous Learning and Adaptation
Detection systems that can continuously learn from new manipulation techniques without catastrophic forgetting of previous knowledge will be essential for long-term effectiveness. Online learning, incremental learning, and few-shot learning approaches may enable detectors to adapt to emerging threats rapidly.

Blockchain and Provenance Tracking
Complementing detection with prevention through content authenticity initiatives that use cryptographic signatures and blockchain to establish media provenance may provide a more comprehensive solution to the deepfake problem.

VII. CONCLUSION

Deepfake detection using CNNs and temporal features represents a rapidly evolving field that addresses one of the most pressing challenges in digital media authenticity. The integration of spatial feature extraction through CNNs with temporal analysis has proven essential for achieving robust detection performance, as deepfakes exhibit artifacts in both spatial and temporal domains.

Current state-of-the-art approaches leverage deep learning architectures that combine the strengths of convolutional networks for capturing manipulation artifacts with recurrent networks, 3D convolutions, or attention mechanisms for modeling temporal inconsistencies. Hybrid methods that fuse multiple feature modalities and employ sophisticated attention mechanisms have demonstrated the most promising results, achieving high accuracy on benchmark datasets.
However, significant challenges remain, including adversarial robustness, generalization to novel manipulation techniques, computational efficiency for real-time deployment, and the impact of compression and post-processing. The ongoing arms race between deepfake generation and detection necessitates continuous innovation in detection methodologies.

Future research directions point toward more generalizable approaches that detect fundamental characteristics of synthetic media rather than specific manipulation artifacts, integration of explainable AI for transparency and trust, multimodal analysis incorporating audio and metadata, and continuous learning systems that adapt to emerging threats. As deepfake technology continues to advance, so too must our detection capabilities, requiring sustained research effort and collaboration across the computer vision, machine learning, and cybersecurity communities.

The societal implications of deepfakes extend beyond technical challenges, encompassing ethical considerations around privacy, misinformation, and the erosion of trust in visual media. Effective deepfake detection is not merely a technical problem but a critical component of maintaining information integrity in the digital age. By combining sophisticated CNN architectures with temporal feature analysis and continuing to innovate in response to evolving threats, researchers are developing the tools necessary to combat malicious uses of synthetic media while preserving the beneficial applications of generative AI technologies.

REFERENCES

1. Afchar, D., et al. (2018). "MesoNet: A compact facial video forgery detection network." IEEE WIFS.
2. Güera, D., & Delp, E. J. (2018). "Deepfake video detection using recurrent neural networks." IEEE AVSS.
3. Rossler, A., et al. (2019). "FaceForensics++: Learning to detect manipulated facial images." ICCV.
4. Li, Y., et al. (2020). "Celeb-DF: A large-scale challenging dataset for deepfake forensics." CVPR.
5. Li, L., et al. (2020). "Face X-ray for more general face forgery detection." CVPR.
6. Frank, J., et al. (2020). "Leveraging frequency analysis for deep fake image recognition." ICML.
7. Dolhansky, B., et al. (2020). "The deepfake detection challenge (DFDC) dataset." arXiv.
8. Carreira, J., & Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset." CVPR.
9. Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions." CVPR.
10. Nguyen, H. H., et al. (2019). "Capsule-forensics: Using capsule networks to detect forged images and videos." ICASSP.