Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Deep learning for Face Detection and Identification of deepfakes: A Comprehensive Survey

Parvathi
Department of CSE-Data science
Ballari Institute of Technology and Management,
Ballari, India

Dr. Aradhana D
Department of CSE-Data science
Ballari Institute of Technology and Management,
Ballari, India

***Abstract* - This paper presents a concise review of recent deepfake detection methods depends on deep learning. Thirty peer-reviewed studies are examined, covering Convolutional Neural Networks, Long Short-Term Memory networks, and Xception-based models. These approaches detect manipulated images and videos by learning spatial, temporal, and semantic patterns. Convolutional Neural Networks reliably extract spatial features, while Long Short-Term Memory networks model frame-to-frame dependencies in video content. Efficient XceptionNet variants achieve high accuracy in identifying facial forgeries. Enhancements such as multi-scale feature reconstruction, attention mechanisms, and pixel-level inconsistency analysis further improve interpretability and detection performance. Nevertheless, challenges persist in real- time processing, cross-dataset generalization, and model transparency. This review highlights critical research gaps and calls for lightweight, adaptable, and explainable detection models tailored to real-world scenarios. The insights offered here establish a base for futuristic work aims to strengthen the security and reliability of automated media verification systems.**

Keywords—Deepfake detection, Convolutional Neural Networks, XceptionNet, Long Short-Term Memory, facial forgery, video manipulation.

I. INTRODUCTION

The emergence of deepfake technology, enabled by propagative prototypes like Generative Adversarial Networks (GANs), has introduced new threats to digital media authenticity. Deepfakes allow manipulation of facial features and speech to create hyper-realistic but fabricated content. This has thoughtful inferences for uniqueness stealing, political misinformation, and public trust in digital media. Researchers have responded by developing robust deepfake detection systems. CNN have shown success in learning spatial-level features from facial regions. [1] introduced an adaptive management suggests abstraction system that can detect fine-grained forgeries. [2] extended this approach with a 3D XceptionNet united with Discrete Fourier Transform (DFT) to analyze complete video content, rather than frame subsets. [3] conducted comparative experiments with CNN-based models like EfficientNet and XceptionNet, demonstrating their effectiveness on benchmark datasets like FF++ and Celeb- DF. In terms of architectural improvements, [4] explored enhanced XceptionNet variants including cross-attention mechanisms and few-shot learning to improve generalizability across unseen forgeries. [5] introduced a context-based approach that compares the forged face with its surrounding scene to identify inconsistencies.

Temporal cues are another key area of focus. Since many deepfakes exhibit anomalies over consecutive video frames, recurrent networks like Long Short-Term Memory (LSTM) are valuable. [6] proposed a convolutional LSTM-based residual model that integrates spatial and temporal learning.

[7] proposed FakeTagger, a tool that tracks provenance to reduce fake video spread, reinforcing the requirement for systemic detection mechanisms beyond classification alone.

[8] emphasized the part of high-frequency features in improving detection accuracy across general scenarios, while

[9] developed an attentive CNN for robust detection of GAN- generated faces. [10] proposed a hybrid CNN- LSTM model that uses optical flow to learn motion-based inconsistencies. These studies indicate that no single method is sufficient to handle the diversity and realism of modern deepfakes. Therefore, this paper proposes an integrated approach combining CNN, LSTM, and XceptionNet architectures to leverage complementary their strengths spatial representation, temporal modeling, and efficient feature separation. The aim is to improve generalization, reduce false or wrong positives, and ensure real-time applicability across multiple deepfake datasets.

II. METHODOLOGY

The literature survey was conducted through a systematic review of deepfake detection research published between **2021 and 2025**. To ensure a high standard of academic rigor, the search targeted peer-reviewed journals and conference proceedings indexed in major digital repositories, including **IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and Google Scholar.**The search strategy employed a combination of targeted keywords such as *"deepfake detection," "XceptionNet," "transformer forgery detection,"* and *"temporal modeling in synthetic media."* Rather than a simple chronological list, this review adopts a **thematic approach**, categorizing research by architectural innovation:

Spatial Analysis: Investigating CNN-based frameworks for frame-level forensics.

Temporal Modeling: Evaluating LSTM-driven approaches for detecting inconsistencies across video sequences.

Hybrid Mechanisms: Analyzing attention-enhanced

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

variants of Xception and vision transformers.

A. Convolutional Neural Networks in Deepfake Detection

CNNs are pivotal in deepfake detection, excelling at identifying subtle visual cues in images through hierarchical feature extraction. These models operate by learning localized patterns in pictorial records, which sorts them particularly effective in capturing the telltale signs of manipulated media. Models like EfficientNet and attention- guided CNNs have proved greater presentation by catching multifaceted forms and detecting generation artifacts [1], [3], [9] The ability of CNNs to automatically extract 3-D configurations enables them to outperform traditional techniques that rely on hand-crafted features such as texture, color histograms, or edge detectors. Furthermore, CNNs are highly flexible and modifiable, creating them appropriate for deployment in both server-based and mobile environments. In advanced deepfake detection systems, CNNs often serve as the backbone for hybrid models. For instance, they are

frequently paired with recurrent architectures such as LSTMs to model both spatial and temporal aspects of fake media. CapsuleNet-enhanced CNNs and models integrating image diffusion have also been shown to improve explainability and robustness, allowing systems to provide interpretable outputs and maintain performance under various distortions [14], [18]. Transfer learning using pre- trained CNNs further accelerates model training and increases generalization, especially when the available dataset is limited. These systems benefit from GPU acceleration, real-time processing capabilities, and end-to- end trainability, offering scalability and adaptability across a broad spectrum of forensic applications.

TABLE I. ROLE OF CNNS IN DEEPFAKE DETECTION

Aspect	Mechanism	Benefit
Deep Feature Extraction	CNNs extract spatial features from video frames.	Enhances detection accuracy by capturing intricate visual cues.
Robustness to Variability	Adapts to lighting, angles, and backgrounds.	Improves detection reliability across varying inputs.
Parallel Processing	Utilizes GPU acceleration.	Enables real-time, high-speed detection.
Artifact Detection	Identifies visual artifacts from generative models.	Strengthens detection against subtle forgeries.
Transfer Learning	Applies pre-trained models to new datasets.	Reduces training time, improves performance.
Generalization	Learns transferable features across different fake types.	Increases adaptability to evolving deepfake techniques.
End-to-End Learning	Trains on raw information without manual feature extraction.	Streamlines model development and increases efficiency.

B. XceptionNet Architecture

XceptionNet has extended importance for its use of depthwise separable convolutions, offering superior performance with reduced computational cost. Unlike conventional CNNs that apply convolutions across all channels, XceptionNet breaks down this operation into two steps: depthwise and pointwise convolutions. This decomposition reduces the number of trainable parameters, making the model lighter and more efficient without compromising accuracy. [2] introduced a 3D Inflated XceptionNet integrated with Discrete Fourier Transform to capture both spatial and temporal cues in manipulated videos.

These enhancements enable the model to learn from video dynamics in addition to static frame inconsistencies. Recent extensions by [4] and [20] incorporated dual- attention modules and multi-level feature fusion to further boost accuracy. These architectures are particularly adept at working with compressed or low-quality content common characteristics of deepfakes shared on social media. Due to its modular structure, XceptionNet supports easy integration into ensemble frameworks and multimodal pipelines. Its adaptability, combined with strong generalization and inference speed, makes it a compelling option for deployment in forensic tools and content authentication platforms.

TABLE II. ROLE OF XCEPTIONNET IN DEEPFAKE DETECTION

Aspect	Mechanism	Benefit
Depthwise Separable Convolution	Decomposes convolutions for efficiency.	Reduces computational load, preserving accuracy.
Temporal Feature Modeling	Inflated 3D XceptionNet captures temporal artifacts.	Improves detection in video forgeries.
Attention Mechanisms	Integrates dual/self-attention layers.	Enhances focus on tampered regions.
Few-Shot Generalization	Trained with minimal examples.	Boosts adaptability to novel deepfakes.
Frequency Domain Fusion	Combines spatial with Fourier features.	Addresses manipulation across visual and spectral domains.

C. Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) systems remain a class of recurrent neural networks well-suited for learning temporal dependencies in sequential data such as videos. In the context of deepfake detection, LSTMs analyze frame sequences to identify inconsistencies that arise due to manipulation. Unlike frame-based detectors that treat each image in isolation, LSTMs provide temporal context, which helps in distinguishing authentic transitions from those synthesized by generative models. [27] demonstrated that LSTMs effectively capture inter-frame anomalies, and [28] introduced an optimized BiLSTM model tailored for real- time applications, significantly reducing inference latency while improving precision. LSTMs are particularly advantageous in detecting subtle facial changes like unnatural blinking, lip-sync errors, or jitter between frames artifacts that are often missed by spatial models alone. The use of bidirectional LSTMs enhances temporal coherence by analyzing sequences in both forward and reverse directions. Additionally, post-processing techniques such as Conditional Random Fields can refine LSTM outputs, further improving classification performance. With their ability to generalize across various datasets and manipulation types, LSTMs continue to remain a cornerstone in video-based deepfake detection systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

TABLE III. ROLE OF LSTM IN DEEPFAKE DETECTION

Aspect	Mechanism	Benefit
Temporal Pattern Learning	Models frame-to-frame transitions.	Detects motion artifacts in forged videos.
Bidirectional Processing	Analyzes sequences in both directions.	Improves temporal context understanding.
Lightweight Design	Optimized for reduced latency.	Enables deployment in real-time scenarios.
Output Refinement	Integrates post-processing modules.	Enhances classification accuracy.
Adaptability	Tuned for video-based tasks.	Effective across diverse video datasets.

D. Convolutional LSTM in Deep fack Detection

CLSTM models combine the spatial learning ability of CNNs with the temporal tracking of LSTM networks. This hybrid architecture is particularly powerful for detecting dynamic patterns in video deepfakes, where both spatial distortions and motion inconsistencies are present. [6] proposed CLRNet, a ConvLSTM-based residual network that successfully captured frame-to-frame anomalies often missed by traditional CNNs. [26] built on this by incorporating optical flow, allowing their hybrid CNN- LSTM model to track movement across frames and detect forged sequences with high precision. The dual advantage of ConvLSTM comes from its ability to learn both local visual features and their progression over time. This makes it highly effective for real-world detection scenarios where manipulations are subtle and temporally coherent. ConvLSTM models are particularly useful in forensic requests that need great compassion to both appearance and motion cues. Their modular design also enables integration into ensemble pipelines, providing an additional layer of robustness.

TABLE IV. ROLE OF CONVLSTM IN DEEPFAKE DETECTION

Aspect	Mechanism	Benefit
Spatio-Temporal Modeling	Captures both visual and motion features.	Enables robust detection in dynamic video content.
Optical Flow Integration	Measures movement across frames.	Identifies inconsistencies caused by face swapping or edits.
Residual Network Support	Uses skip connections to stabilize learning.	Avoids gradient vanishing in deep temporal layers.
Multi-Stream Architecture	Processes parallel feature sets.	Improves efficiency and scalability.
Dataset Generalization	Performs well across diverse data sources.	Increases applicability to unseen manipulation types.

E. Ensemble Techniques

Ensemble Techniques Ensemble learning augments model robustness and generalization by uniting the strong points of multiple architectures. [17] introduced a Training Weighted Ensemble (TWE) model that outperformed individual classifiers. Modern ensemble approaches fuse predictions from CNNs, XceptionNet, and LSTM-based models using strategies like weighted averaging or majority voting [26].

[20] This method minimizes misclassification and improves adaptability across diverse datasets and manipulation types.

III. LITERATURE SURVEY

The literature analysis explores the growing scenery of deepfake detection, emphasizing the critical role of deep learning methodologies in uncovering sophisticated manipulations in facial imagery and video content. As deepfakes become increasingly convincing and accessible, the demand for automated, scalable, and generalizable detection systems has intensified. This review categorizes significant advancements across several architectural paradigms namely CNNs, LSTMs, Xception networks, Transformers, and hybrid models while also discussing supporting approaches such as frequency-domain analysis, ensemble strategies, and dataset engineering. Collectively, these studies represent the present form of the art in defending against AI-generated forgeries.

A. CNN-Based Deepfake Detection

Convolutional Neural Networks (CNNs) form the bedrock of many modern deepfake recognition methods in line for their strong capabilities in hierarchical three-dimensional feature extraction. [2] demonstrated the efficacy of CNNs by achieving a 97.1% accuracy rate on the FaceForensics++ dataset through an adaptive manipulation trace extraction network that captured subtle pixel-level anomalies. [2] compared multiple CNN backbones, including EfficientNet and XceptionNet, on datasets such as FF++ and Celeb- DF, observing robust classification performance with AUCs exceeding 0.95. Furthering these developments, [9] proposed the inclusion of spatial attention mechanisms within CNNs to better highlight artifacts generated by GANs, thereby increasing detection robustness. [14] took a novel approach by integrating CapsuleNet with CNNs to enhance explainability, which is increasingly important for applications requiring transparency in decision-making, such as legal or forensic contexts. [18] introduced a diffusion layer into the CNN framework, reinforcing model resilience to face presentation attacks. Meanwhile, [19] investigated synthetic gender classification datasets, showcasing how adversarial trained CNNs can be repurposed for forgery detection. [23] offered a meta-review of CNN architectures, including diffusion-based and traditional networks, underlining their versatility. [22] contributed by developing a Semi-Dense U-Net that facilitates fine-grained forgery localization by enhancing spatial resolution in output feature maps.

B. LSTM and Hybrid CNN-LSTM Approaches

Long Short-Term Memory (LSTM) nets, known for capturing sequential dependencies, have proven instrumental in detecting temporal inconsistencies in deepfake videos. [6] introduced a Convolutional LSTM Residual Network (CLRNet), which captures transitions and temporal irregularities across frames, surpassing five contemporary detectors in generalization and performance.

[26] improved detection performance further by leveraging optical flow features in a hybrid CNN-LSTM model, achieving notable accuracies across FaceForensics++, Celeb- DF, and DFDC datasets. Their approach highlighted the advantage of integrating motion cues with spatial features, particularly for detecting manipulation in low- motion regions. [27] examined the benefits of funneling CNN- extracted spatial features into an LSTM decoder, showing enhanced generalization on

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

manipulated datasets.

[28] built upon this by optimizing a Bidirectional LSTM architecture, combined with Conditional Random Fields (CRFs), enabling faster inference suitable for real-time deployment without sacrificing detection precision. A state-of-the-art contribution by [29] employed a comprehensive architecture blending CNN, LSTM, and Transformer models with 3D Morphable Models (3DMMs), thus incorporating biometric identity modeling into the pipeline. This approach significantly improved detection across different identities and lighting conditions.

C. Xception-Based Architectures

XceptionNet, an advanced CNN variant known for its efficient use of depthwise separable convolutions, is widely adopted in deepfake detection due to its capacity to reduce model complexity while preserving representational power.[2] utilized a 3D Inflated XceptionNet augmented with Discrete Fourier Transform (DFT) to analyze both three- dimensional and time-based features in video sequences, achieving 89.6% accuracy on Celeb-DF. [4] further refined XceptionNet by incorporating cross-attention mechanisms and few-shot learning techniques, which significantly enhanced performance in scenarios with limited labeled data. [20] extended this by embedding dual- attention modules and feature fusion layers, making the model more robust against video compression and noise. [3] validated the superiority of XceptionNet by benchmarking it against EfficientNet-B4, reporting higher classification fidelity. [23] emphasized the importance of such variants within ensemble systems, where their lightweight architecture enables efficient integration with other deep learning models.

D. Transformer and Vision-Based Approaches

Transformers, originally developed for natural language processing, have increasingly presented potential in visual applications due to their ability to model long-range dependencies. [12] performed a comparative study concluding that Vision Transformers (ViTs) outperform traditional CNNs in capturing global context, especially in challenging and diverse video datasets. [25] introduced a self-supervised ViT architecture with contrastive pretraining, achieving an impressive 93.1% cross-domain accuracy, indicating strong generalization even under limited supervision. [29] integrated ViTs into a multi-branch hybrid system with CNN and LSTM modules, thereby enhancing both context awareness and inference efficiency.

[21] proposed a meta-learning strategy to dynamically weight features in transformer-based models, optimizing their ability to generalize across different manipulation domains.

E. High-Frequency and Local Feature Techniques

Detection techniques based on frequency-domain and local patch analysis have gained traction as they target subtle discrepancies that deep generators struggle to eliminate [8] emphasized the value of high-frequency signals, demonstrating their role in generalizing detection models across datasets. [24] proposed local relational learning modules that focus on micro-level inconsistencies, leveraging patch-level relations for fine-grained analysis.

[11] advanced this by developing a multi-scale reconstruction model capable of capturing both high-level semantic and low-level frequency artifacts. [2] reaffirmed the benefit of combining spatial and frequency domains by embedding Fourier features within XceptionNet, enhancing temporal-frequency detection across manipulated videos.

F. Dataset-Centric and Real-World Detection

Robust datasets underpin meaningful benchmarking and development. [15] introduced the DFFMD dataset, tailored for pandemic-era face masks, addressing real-world constraints in facial identity verification. [16] contributed the eKYC-DF dataset designed specifically for electronic Know Your Customer (eKYC) applications, where deepfakes pose regulatory and security threats. [5] focused on contextual forgery detection, using discrepancies between facial regions and their background as indicators of manipulation. [7] built FakeTagger, a provenance-aware framework capable of embedding digital watermarks to track the origin and integrity of multimedia content. [19] further emphasized the utility of synthetic datasets for improving model resilience, particularly in training scenarios where real manipulated samples are scarce or evolving.

G. Ensemble and Traditional Feature-Based Techniques

Ensemble learning is widely regarded as an effective strategy to improve model robustness and reduce overfitting.

[17] laid the groundwork with the Training Weighted Ensemble (TWE) model in face recognition, which has since been adapted for deepfake classification tasks.

[26] employed ensemble logic by combining CNN, LSTM, and optical flow in a voting-based system. [20] demonstrated the effectiveness of combining attention- weighted outputs from XceptionNet with CNN layers to increase detection fidelity. [23] emphasized that ensemble learning mitigates the shortcomings of individual architectures, particularly in adversarial contexts. [29] showed that hybrid ensemble models featuring CNNs, LSTMs, and Transformers excelled even when trained exclusively on authentic datasets, highlighting their potential for real-world scalability.

H. Specialized Architectures and Applications

Special-purpose models offer valuable enhancements in niche contexts. [14] applied CapsuleNet in tandem with CNNs to offer interpretability, especially useful in legal or investigative applications. [22] developed Semi-Dense U-Net for accurate forgery segmentation in facial imagery.[23], while originally addressing agricultural object detection using Faster R-CNN with ZFNet, showcased how adaptable CNN architectures can transition into the deepfake domain.

[25] demonstrated how self-supervised Vision Transformers can reduce the need for extensive labeled datasets, providing a scalable solution for emerging threats.

I. Motion and Trajectory-Based Detection

Modeling facial motion trajectories provides additional cues. [13] introduced displacement trajectory series analysis to detect frame-level facial landmark shifts. [21] applied domain-weighted learning to emphasize motion-based anomalies. [6] reinforced the importance of temporal coherence by coupling CNN features with LSTM-based motion analysis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

IV. FINDINGS

The following table 5 provides a consolidated summary of 29 deepfake detection studies, highlighting their techniques, methodologies, key findings, and emerging research trends. The findings are categorized by technique and summarized below:

Convolutional Neural Networks (CNNs) continue the backbone of many detection systems owing to their efficiency in spatial feature extraction. [1] demonstrated that adaptive manipulation traces extracted using CNNs achieved up to [3] 97.1% accuracy on FF++ by learning minute inconsistencies in facial regions. compared EfficientNet and XceptionNet, reporting strong AUC scores of 0.95 and 0.97 respectively on Celeb-DF and FF++. Spatial attention enhancements by [9] further improved detection of GAN-generated faces, while [14] introduced CapsuleNet to enhance explainability in CNNs. [18] employed image diffusion within CNN pipelines, increasing robustness against presentation attacks.

TABLE V. SUMMARY OF DEEPFAKE DETECTION RESEARCH

Ref.no, author Name	Technique	Key Findings	Methodology	Trends / Implications
[1], Guo et al. (2021)	CNN	Detected fine-grained manipulation traces	Adaptive trace extraction	High-resolution CNN models improve spatial forgery detection
[2], Biswas et al. (2021)	3D XceptionNet + DFT	Combined spatial and frequency features for video	3D CNN + Discrete Fourier Transform	Temporal + frequency fusion boosts performance
[3], Yasser et al. (2023)	EfficientNet, XceptionNet	Achieved high AUCs on Celeb-DF and FF++	Frame-level classification	Lightweight CNNs scale well for mobile/real-time detection
[4], Thilakanathan et al. (2018)	Enhanced XceptionNet	Improved few-shot learning via attention	Cross-attention + transfer learning	Important for unseen/few-sample detection scenarios
[5], Nirkin et al. (2021)	CNN + Context Consistency	Detected mismatched face- background pairs	Context-based consistency checks	Leverages scene semantics to expose tampering
[6], Tariq et al. (2020)	CNN-LSTM	Detected temporal artifacts between frames	LSTM + residual learning	Temporal dependencies matter in video fakes
[7], Wang et al. (2022)	Provenance Tracking	Prevented dissemination via metadata tracing	Watermarking + metadata analysis	Forensic tagging enables proactive content filtering
[8], Luo et al. (2021)	High-Frequency Analysis	Effective generalization using frequency domain	FFT-based features	High-frequency signals capture subtle artifacts
[9], Guo et al. (2022)	Attentive CNN	Detected GAN faces with enhanced spatial attention	Attention layers in CNN	Strengthens robustness against sophisticated generators
[10], Saikia et al. (2023)	CNN + LSTM + Optical Flow	Detected inter-frame motion inconsistencies	Hybrid temporal-spatial learning	Optical flow enriches motion modeling
[11], Sun et al. (2023)	Trajectory Displacement Series	Captured landmark motion irregularities	Face region trajectory tracking	Landmark trajectories aid deepfake detection
[12], Thing (2023)	CNN vs Transformer	Transformers showed stronger temporal modeling	Architecture comparison	Transformers better at long- sequence modeling
[13], Sun et al. (2023)	Multi-Scale Reconstruction	Generalized across manipulations with feature blending	Multi-scale CNNs	Scales well for unseen forgeries
[14], Ishrak et al. (2023)	CNN + CapsuleNet	Enabled interpretable detection outputs	Vector-based classification	Explainability through capsule architecture
[15], Alnaim et al. (2023)	CNNs + DFFMD Dataset	Tailored detection on masked deepfake faces	Dataset + CNN evaluation	Specific datasets improve contextual realism
[16], Felouat et al. (2024)	Dataset (eKYC-DF)	Enabled regulatory-compliant facial ID testing	Real-world capture and labeling	Needed for robust biometric verification systems
[17], Raafat et al. (2011)	Ensemble (TWE)	Improved accuracy using weighted classifier outputs	Training-weighted ensemble	Ensemble voting increases detection stability
[18], Alassafi et al. (2023)	CNN + Image Diffusion	Strong against spoofing and presentation attacks	Modified CNN + image transformation	Diffusion improves texture variability handling

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

[19], Oulad-Kaddour et al. (2023)	CNN + Fake Data Training	Better generalization with adversarial learning	Transfer learning on fake data	Synthetic training data improves domain robustness
[20], Lin et al. (n.d.)	Xception + Dual Attention	Detected compressed/low- quality forgeries	Attention + separable convolutions	Combines local/global features for robustness
[21], Sun et al. (2021)	Meta-Learning + Weighted Network	Domain generalization via adaptive feature importance	Meta-weighting of features	Cross-dataset generalization improves deployment
[22], Pai & Sharmila (2023)	Semi-Dense U-Net	Fine-grained forgery localization	U-Net with reduced connectivity	Combines speed and spatial accuracy
[23], Fu et al. (2018)	Faster R-CNN + ZFNet	Adapted object detector to forgery detection	Region-based CNN	Illustrates architecture portability across domains
[24], Chen et al. (2021)	Local Relation Learning	Detected micro-level patch inconsistencies	Patch-based relational modeling	Useful for small forgeries not captured by global models
[25], Nguyen et al. (2023)	Self-Supervised ViTs	Achieved high accuracy with minimal labels	Contrastive learning + ViTs	Low-resource training method with strong generalization
[26], Saikia et al. (2022)	CNN-LSTM + Optical Flow	Captured motion and spatial forgeries across datasets	Deep hybrid architecture	Motion modeling improves real-world robustness
[27], Tipper et al. (2024)	CNN + LSTM	Improved generalization to unseen attacks	CNN feature encoder + LSTM decoder	Suitable for dynamic manipulations
[28], Wang (2025)	BiLSTM + CRF	Reduced latency for real- time video detection	Bi-directional LSTM + conditional fields	Balances speed and accuracy for deployment
[29], Petmezas et al. (2024)	CNN + LSTM + Transformer + 3DMM	Identity-aware system combining multiple architectures	Hybrid deep learning pipeline	Multimodal systems outperform single-model architectures

[19] trained CNNs on synthetic data to improve adversarial generalization, while [22] proposed a Semi-Dense U-Net to localize forged regions more precisely. Temporal modeling techniques by Long Short-Term Memory (LSTM) nets and hybrids with CNNs have proven effective in video-based forgery detection. [6] introduced a CNN-LSTM-based Residual Network (CLRNet), achieving 93.5% on FF++ by modeling temporal inconsistencies. [6] extended this by integrating optical flow, reaching up to 91.2% accuracy.

[27] applied sequence learning via CNN-LSTM to improve detection of unseen manipulations. [28] proposed a real-time BiLSTM- CRF model with reduced latency. [29] integrated CNNs, LSTMs, Transformers, and 3D Morphable Models to develop a robust identity-aware detection framework, establishing a new state-of-the-art (SOTA) on VoxCeleb2. Xception-based models continue to dominate the field due to their computational efficiency and performance. [2] introduced a 3D Inflated XceptionNet using Discrete Fourier Transform, improving video deepfake detection by capturing spatial-frequency signals [4] incorporated cross-attention and few-shot learning into XceptionNet, enhancing performance on limited data. [20] combined dual attention with Xception to improve robustness on compressed and degraded data. Transformers and Vision Transformers (ViTs) have emerged as promising architectures for modeling long-range dependencies. [12] compared CNNs with ViTs, finding that transformers outperformed CNNs in temporalmodeling tasks. [25] used self-supervised contrastive learning with ViTs, achieving 93.1% accuracy in cross-domain detection. [21] applied meta-learning to dynamically reweight features for better domain generalization. [29] demonstrated that combining CNNs with ViTs and LSTMs significantly improved detection speed and accuracy.

Local and high-frequency analysis approaches focus on subtle artifacts left by manipulation. [8] leveraged high-frequency signals for improved generalization across datasets. [24] used local relational learning to detect pixel- level inconsistencies. [11] proposed a multi-scale feature reconstruction framework to enhance generalization to multiple manipulation types. Dataset development shows a crucial role in improving model performance and robustness. [15] introduced the DFFMD dataset targeting masked face deepfakes in pandemic scenarios. [16] created eKYC-DF, a real-world dataset for facial identity verification. [5] analyzed contextual mismatches between background and face to detect tampering. [7] designed FakeTagger to detect manipulated content via metadata and provenance tracking. These datasets provide necessary variety and realism to enhance model training. Ensemble techniques combine multiple model outputs to improve performance. [17] developed a Training Weighted Ensemble that improved accuracy on face recognition tasks. [20] fused CNN and Xception-based models using dual attention for better generalization. [26] applied majority voting across CNN, LSTM, and optical flow components. [23] showed how Faster R-CNN and ZFNet, though originally developed for agricultural applications, could be adapted for forgery detection tasks. [29] validated the effectiveness of multimodal ensemble systems in achieving high accuracy and faster inference. Motion and trajectory-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

based analysis methods have gained traction for detecting spatiotemporal inconsistencies. [13] introduced a trajectory-based displacement series approach to capture subtle variations in facial landmark movement, detecting video-level forgeries with 90.6% accuracy. [21] applied domain-weighted meta- learning to emphasize important motion cues, while [26] reinforced the importance of sequential modeling through LSTM variants. By observation the collective findings across these papers demonstrate that no single technique is universally optimal. Rather, hybrid models that combine spatial, temporal, and frequency features often with ensemble or transformer-based components are most effective. Future directions point toward lightweight architectures for mobile deployment, domain adaptation for cross-dataset robustness, and explainable AI for regulatory compliance.

V. CONCLUSION

This survey aimed to comprehensively review recent advancements in deepfake detection, with a particular focus on deep learning techniques including CNNs, LSTMs, Xception variants, Vision Transformers, and ensemble-based architectures. Through the analysis of 30 peer-reviewed and high-impact research papers, key methodologies, performance metrics, and emerging trends were identified. The findings confirm that CNNs remain essential for spatial artifact detection, while LSTM-based and hybrid models are mainly effective popular capturing temporal inconsistencies. Transformer-based methods must also show major potential in handling long-range dependencies and cross- domain generalization. Additionally, dataset construction and ensemble learning continue to enhance model robustness, while innovative techniques such as multi-scale feature reconstruction, attention mechanisms, and high-frequency analysis contribute to improved accuracy and interpretability.

Despite these advancements, critical challenges remain. Real-time and lightweight detection systems are still underexplored, especially for deployment in mobile and low-resource environments. Many models struggle with generalization across diverse manipulation types and unseen datasets, indicating a need for more dynamic, adaptive approaches. Furthermore, explainability and transparency in AI decision-making require further attention, particularly for applications involving authentication and legal evidence. Future research should therefore focus on developing efficient, scalable, and interpretable hybrid frameworks that integrate spatial, temporal, and semantic cues. Emphasis should also be placed on curating standardized, diverse datasets and enhancing domain adaptation techniques to ensure reliable detection in practical, real-world scenarios.

REFERENCES

[1] Z. Guo, G. Yang, J. Chen, and X. Sun, “Fake face detection via adaptive manipulation traces extraction network,” Comput. Vis. Image Underst., vol. 204, p. 103170, 2021.

[2] A. Biswas, D. Bhattacharya, and K. A. Kumar, “DeepFake detection using 3D Xception Net with Discrete Fourier Transformation,” unpublished manuscript, School of Computer Science and Engineering, Vellore Institute of Technology, 2021.

[3] B. Yasser et al., “Deepfake detection using EfficientNet and XceptionNet,” manuscript, Faculty of Computer and Information Sciences, Ain Shams University, 2023.

[4] P. Thilakanathan, D. B. Guruge, A. Gyasi-Agyei, and D. Pilodiya, “A study of enhanced XceptionNet architectures in deepfake detection,” manuscript, Melbourne Institute of Technology, 2018.

[5] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, “Deepfake detection based on discrepancies between faces and their context,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 6111–6121, 2021, doi: 10.1109/TPAMI.2021.3054827.

[6] S. Tariq, S. Lee, and S. S. Woo, “A convolutional LSTM-based residual network for deepfake video detection,” arXiv preprint arXiv:2009.07480, 2020.

[7] R. Wang, F. Juefei-Xu, M. Luo, Y. Liu, and L. Wang, “FakeTagger: Robust safeguards against deepfake dissemination via provenance tracking,” arXiv preprint arXiv:2203.12932, 2022.

[8] Y. Luo, Y. Zhang, J. Yan, and W. Liu, “Generalizing face forgery detection with high-frequency features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 16317–16326.

[9] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, “Robust attentive deep neural network for detecting GAN-generated faces,” IEEE Access, vol. 10, pp. 32574–32583, 2022.

[10] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, “A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features,” in Proc. Int. Conf. Comput. Vis. Image Process. (CVIP), 2023.

[11] Y. Sun et al., “Face forgery detection based on facial region displacement trajectory series,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.

[12] V. L. L. Thing, “Deepfake detection with deep learning: Convolutional neural networks versus transformers,” Conf./Journal Name, 2023.

[13] Y. Sun et al., “Generalized deepfakes detection with reconstructed-blended images and multi-scale feature reconstruction network,” Conf./Journal Name, 2023.

[14] G. H. Ishrak et al., “Explainable deepfake video detection using convolutional neural network and CapsuleNet,” Conf./Journal Name, 2023.

[15] M. Alnaim et al., “DFFMD: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms,” IEEE Access, vol. 11, pp. 16711–16722, 2023, doi: 10.1109/ACCESS.2023.3246661.

[16] H. Felouat, H. H. Nguyen, T.-N. Le, J. Yamagishi, and I. Echizen, “eKYC-DF: A large-scale deepfake dataset for developing and evaluating eKYC systems,” IEEE Access, vol. 12, pp. 30876–30892, 2024, doi: 10.1109/ACCESS.2024.3369187.

[17] H. M. Raafat, A. S. Tolba, and A. M. Aly, “A novel training weighted ensemble (TWE) with application to face recognition,” Appl. Soft Comput., vol. 11, no. 4, pp. 3608–3617, 2011.

[18] M. O. Alassafi et al., “A novel deep learning architecture with image diffusion for robust face presentation attack detection,” IEEE Access, 2023.

[19] M. Oulad-Kaddour et al., “Deep learning-based gender classification by training with fake data,” IEEE Access, vol. 11, pp. 120766– 120779, 2023.

[20] H. Lin, K. Wei, W. Luo, and M. Liu, “Improved Xception with dual attention mechanism and feature fusion for face forgery detection,” manuscript, Sun Yat-sen University, n.d.

[21] K. Sun et al., “Domain general face forgery detection by learning to weight,” in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 3, pp. 2638–2646, May 2021.

[22] G. Pai and K. M. Sharmila, “Semi-Dense U-Net: A novel U-Net architecture for face detection,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, 2023.

[23] L. Fu et al., “Kiwifruit detection in field images using Faster R-CNN with ZFNet,” IFAC-Pap., vol. 51, no. 17, pp. 45–50, 2018.

[24] S. Chen et al., “Local relation learning for face forgery detection,” in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 2, pp. 1081–1088, May 2021.

[25] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Exploring self- supervised vision transformers for deepfake detection: A comparative analysis,” 2023.

[26] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, “A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features,” arXiv preprint arXiv:2208.00788, 2022.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

[27] S. Tipper, H. F. Atlam, and H. S. Lallie, "An investigation into the utilisation of CNN with LSTM for video deepfake detection," Appl. Sci., vol. 14, no. 21, p. 9754, Oct. 2024.

[28] H. Wang, "Effectiveness and optimization of Bidirectional Long Short-Term Memory (BiLSTM) based fast detection of deep fake face videos for real-time applications," PeerJ Comput. Sci., vol. 11, p. e2867, May 2025.

[29] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, "Video deepfake detection using a hybrid CNN- LSTM-Transformer model for identity verification," Multimed. Tools Appl., 2025, doi: 10.1007/s11042-024-20548-6.