Detecting Deepfake Videos: A Deep Learning and Neural Network Approach

Dinesh Choudhary[1] , Lalit Hinduja[2] , Omkar Pansare[3] , Parth Tilekar[4] , Dr. Shilpa Khedkar[5]
[1,2,3,4,5]*Computer Science, Modern Education Society's Wadia College of Engineering, Pune, India*

***Abstract*—Recent progress in GenAI has enabled the creation of deepfake videos that are strikingly realistic, capable of reconstructing visuals with minimal distortions, modifying voices, and seamlessly swapping identities. These manipulated videos introduce serious risks to public trust, digital security, political stability, and individual privacy. A major problem for law enforcement and cybersecurity systems is deepfake-based video fraud, which includes identity theft, impersonation schemes, and deceptive political propaganda. With a focus on deep learning architectures, multimodal detection techniques, transformer-based systems, and adversarially resilient frameworks, this review paper investigates cutting-edge deepfake detection techniques. Currently, existing methods such as hybrid CNN-LSTM models [1], Swin-BiLSTM architectures [6], and multimodal audio-visual fusion networks [10], [18] demonstrate considerable capabilities in identifying manipulated frames across spatial and temporal domains. Recent works additionally focus on transformer-based systems like EffiSwinNet [3] and lightweight models such as FakeFormer [17] to improve generalization and reduce computational load for real-time deployments.**

Furthermore, new research indicates that detectors are vulnerable to adversarial attacks and quality variations, prompting the development of robust systems such as identity masking frameworks and gray-box-resilient detectors. This review summarizes theoretical foundations, technology modeling, empirical performance insights, and highlights key research gaps necessary for reliable deepfake detection systems, addressing issues like cross-dataset robustness and real-time inference constraints.

***Index Terms*—Deepfake detection, deepfake video, generative adversarial networks (GANs), transformer models, computer vision, video forgery detection, multimodal analysis, CNN-LSTM architectures, temporal consistency, spatial feature extraction, adversarial robustness, perceptual fidelity.**

I. INTRODUCTION

AI has changed how digital media is made, and one of the most important results of this change is deepfake technology. Deepfakes let you make fake audio and video that sound and look like a person is saying or doing things that never happened. Using advanced generative methods like transformer-based models, encoder–decoder architectures, and generative adversarial networks (GANs), it is possible to make facial features, lip movements, voice characteristics, and even whole video sequences look very real. Originally, these methods were linked to good uses in entertainment, accessibility, and making movies, but now they are widely seen as tools that can be used for fraud, political manipulation, harassment, and spreading false information. For this reason, deepfakes are now a big problem in cybersecurity and other social and technical systems.

As generative models get better at copying small human cues like eye movement, facial geometry, lighting patterns, and speech consistency, the threat gets worse. Deepfake video fraud is no longer easy to ignore because it looks so real. Alrashoud [2] and Kaur et al. [21] say that reliable detection is now necessary to keep trust in places like social media, government communication channels, financial services, and organizational networks.

At the same time, deepfake videos are becoming harder to detect for a number of reasons. First, the quality of generated content keeps getting better. Systems like StyleGAN, diffusion models, and Meta's audio-driven reconstruction methods can generate highly realistic videos that follow changes in lighting, expression, movement, and emotion. These models are capable of producing plausible outputs even when trained on a relatively small amount of input data and

therefore the artifacts that previous detectors depended on are often either absent or suppressed. This limits the efficacy of traditional CNN based methods to detect the small visual anomalies that earlier revealed manipulated content [3].

In addition, deepfakes are harder to detect due to the way they are shared. Videos are often compressed or resized when sent on platforms like Twitter, Instagram, WhatsApp to save storage and bandwidth. This process removes or blurs fine details, including subtle traces that detection models use as evidence of manipulation.Ramadhani et al. [22] show that such platform-level compression can significantly reduce the accuracy of even strong detection systems, which creates a major challenge for real-world deployment.

Also, many previous methods were too frame-centric and did not consider temporal behavior over a video sufficiently. Deepfake creators have countered by implementing temporal smoothing techniques to reduce visible flicker or blinking anomalies. To overcome this, hybrid approaches have been proposed that combine CNNs with sequence models like LSTMs to better model the irregularities from frame to frame [1], [12].

To address these challenges, researchers have moved toward more advanced detection strategies. CNN-based models remain important because they are effective at identifying spatial artifacts such as texture mismatch, pixel blending, and inconsistent lighting. Improvements in dense CNN design and feature extraction have further strengthened this direction, as discussed by Singh et al. [15] and Patel et al. [25].

However, deepfake generation has also evolved beyond simple autoencoders into high-resolution transformer-based systems capable of modeling long-range dependencies. This has encouraged the use of transformer-based detectors such as Swin-BiLSTM [6] and FakeFormer [17], which rely on hierarchical attention to detect both local and global inconsistencies in video content. These models are often more robust against sophisticated GAN-based manipulations, especially when the forgery does not leave obvious convolutional artifacts.

Another important trend is multimodal detection, where visual and audio information are analyzed together. By comparing lip synchronization, voice features, and facial movement patterns, these systems can detect inconsistencies that may not be visible in a single modality. This is especially useful for compressed or low-quality videos, where individual visual or audio cues may be weak. Studies by Panja et al. [10] and Yıldırım and Aydın [18] demonstrate the value of combining audio-visual analysis for stronger detection performance.

Adversarial robustness has also become a major focus, since attackers may deliberately design deepfakes to evade existing detectors. Li et al. [5] proposed proactive detection using learnable hidden-face representations to improve generalization, while Guarnera et al. [4] examined gray-box attacks that reveal weaknesses in DCT-based detection methods. In a related direction, Peng et al. [11] introduced perceptual fidelity assessment to identify forgery characteristics that may be missed by conventional classifiers.

In addition to accuracy, practical deployment now requires models that are lightweight and efficient enough for real-time use. This is especially important for applications such as surveillance, mobile authentication, and banking verification. Models like DeepFaceGuard [16] and FakeFormer [17] aim to balance computational efficiency with reliable performance, making them more suitable for operational environments where speed and resource use matter.

Overall, deepfake video fraud represents a fast-growing and multidimensional threat that cannot be addressed by a single detection method. Effective solutions must combine spatial, temporal, multimodal, perceptual, and adversarially robust techniques to keep pace with rapidly improving generation methods. As deepfake technology continues to advance, the need for continued research in detection remains critical for maintaining digital trust.

II. RELATED WORK

With significant advancements in convolutional neural networks, hybrid spatio-temporal architectures, transformer-based systems, multimodal detection, feature-selection-driven frameworks, and adversarially robust detectors, deepfake detection research has advanced quickly in recent years. This section reviews key contributions from the literature,

focusing on the period from 2023 to 2025, and highlighting how the field has moved from static, image-centric approaches to more dynamic, transformer-oriented and multimodal frameworks. These newer techniques are more suited to deal with increasingly sophisticated deepfake generation pipelines.

A. Early CNN-Based and Static Image Approaches

During the early stages of deepfake detection research, convolutional neural networks (CNNs) were crucial because they are very good at finding spatial features. Dense CNNs, Xception variants, and EfficientNet models were often used in early detection systems. For example, Singh et al. [15] used CNN-based pipelines to find inconsistencies at the pixel level and blending artifacts that generative techniques caused. These models worked well with high-resolution data, but they didn't work as well with low-quality or compressed media.

An enhanced Dense CNN architecture designed especially for deepfake image identification was presented by Patel et al. [25]. By deepening convolution layers and altering the receptive field to better catch small texture irregularities brought forth by deepfake generation, their model improved detection accuracy. Despite the fact that these methods mostly focused on static images, their discoveries established fundamental spatial characteristics that subsequent temporal models also expanded upon.

One major problem with these methods is that they can't find temporal inconsistencies between video frames. Frame-by-frame analysis does not pick up on things like blinking patterns that aren't regular, motion discontinuities, and identity mismatches. These deficiencies ultimately spurred the shift towards spatio-temporal modeling methodologies.

B. Hybrid Spatio-Temporal Architectures

To get around the problems with purely spatial models, hybrid architectures that combine CNN-based feature extraction with temporal modeling components like LSTMs, GRUs, and Transformers have become more popular. The main idea is that deepfake systems can often keep spatial realism, but they have a hard time copying natural temporal dynamics between frames.

One of the most significant hybrid models was presented by Akshaya et al. [1]: the Xception-LSTM pipeline, which uses a CNN backbone to extract high-level visual information and feeds them into an LSTM to identify temporal inconsistencies such implausible facial movements. This model established a standard for hybrid deepfake detection with its impressive performance on FaceForensics++ and related datasets.

This concept was extended by Petmezas et al. [12], who suggested a CNN-LSTM-Transformer hybrid that integrated several sequence-modeling layers. Their approach was more resistant to sophisticated deepfake techniques because it caught both short-term and long-term spatiotemporal relationships. Sequence modeling was much enhanced by the use of transformers, especially when visual compression and quality degradation were present.

Using a Swin Transformer + BiLSTM + Attention architecture, Lukitania and Suryani [6] improved hybrid detection. This model captured significant frame-level anomalies using attention mechanisms and employed the Swin Transformer for in-depth spatial analysis and BiLSTM for improved temporal representation. In face-swap detection circumstances when temporal irregularities are more noticeable, their approach demonstrated efficacy.

By combining CNN-based feature extraction and transfer learning techniques, Pandey and Kushwaha [9] also investigated hybrid modeling, emphasizing that a mix of recurrent units and pre-trained models produces reliable detectors with enhanced generalization capabilities.

Overall, hybrid spatio-temporal architectures remain some of the best methods for detecting deepfakes, and perform well for various manipulation techniques owing to their joint modeling of spatial details and temporal dynamics.

C. Transformer Based Approaches for Deepfake Detection

While originally developed for natural language processing, transformers have demonstrated strong capability in modeling long-range dependencies through global attention mechanisms. This strength has made them increasingly valuable in tasks related to computer vision . As deepfake generation methods have advanced, traditional CNN-based approaches have struggled to capture high-level semantic

inconsistencies, prompting a transition toward transformer-based detection models.

Basu et al. [3] proposed EffiSwinNet, a hybrid model combining EfficientNet with Swin Transformer blocks for cross-architecture deepfake detection. This led to good performance in detecting fake faces generated by StyleGAN. The hierarchical attention mechanism of Swin Transformers makes the model robust to multiple manipulation techniques..

Ramadhani et al. [22] enhanced Video Vision Transformers (ViViT) by incorporating facial landmark information, depthwise separable convolutions, and improved self-attention mechanisms. These modifications helped preserve global contextual information while lowering computational cost, making the approach effective for deepfake video detection even under varying video quality conditions.

Wang et al. [17] introduced FakeFormer, a lightweight transformer architecture specifically designed for video deepfake detection. Emphasizing efficiency and deployability, FakeFormer achieves high detection performance with significantly fewer parameters compared to conventional transformer models. This makes it well-suited for real-time applications, especially in low-power or mobile environments.

Transformers represent a significant change in deepfake detection, since they provide stronger generalization across datasets, improved robustness, and greater adaptation to deceptive manipulation artifacts.

D. Multimodal Detection: Integrating Audio and Visual Cues

The increasingly realistic modification of visual content, as well as the growing frequency of voice-based deepfakes, have drawn significant interest to multimodal deepfake detection. Modern generative models frequently fail to fully synchronize speech and facial movements or imitate true voice timbre, creating potential for multimodal detection.

Panja et al. [10] suggested a multimodal paradigm that included CNN, Vision Transformer, and audio-visual synchronization analysis. Their methodology analyzed lip movement coherence, phoneme-viseme alignment, audio spectral patterns, and facial inconsistencies all at once. This method outperforms single-modality detectors, particularly when dealing with low-resolution or severely compressed data.

Yıldırım and Aydın [18] developed multimodal analysis to detect both audio deepfake and video manipulation. They demonstrated how audio spectrogram discrepancies, when linked with frame-level abnormalities, offer a powerful detection modality capable of handling real-world situations where manipulation may occur in only one modality.

These experiments suggest that the most reliable deepfake detectors in the future will be those that integrate several sensory streams and closely resemble human vision.

E. Adversarial Robustness and Gray-Box Attacks

As deepfake detection algorithms improve, attackers are developing adversarial tactics to trick them. This has resulted in an increased demand for detectors that can tolerate adversarial perturbations, compression artifacts, and quality degradation.

Guarnera et al. [4] described a novel adversarial gray-box assault that targets DCT-based detectors. Their findings demonstrated that many deepfake detection systems are extremely sensitive to targeted attacks that subtly alter frequency-domain components. The authors stressed the critical necessity for frequency-aware, robust, and adversarially protected detection models.

Li et al. [5] proposed the Learnable Hidden Face method, a proactive deepfake detection technique that covers identity-specific characteristics during training to avoid overfitting to specific face traits. This allows the model to focus on manipulation artifacts rather than individual facial features, increasing robustness and generality.

Peng et al. [11] created DeepFidelity, a perceptual fidelity assessment method that assesses photorealistic consistency and detects small manipulation traces that are imperceptible to existing models.

With the rise of adversarial deepfake production, robustness-focused research has emerged as an important and fast growing topic of deepfake detection.

F. Feature-Selection-Driven Approaches and Classical Enhancements

Some studies investigate the merging of deep learning and standard machine learning technologies. Mohiuddin et al. [7] developed a feature selection-

aided deepfake video identification algorithm, indicating that feature optimization strategies can improve accuracy in deep learning pipelines.

Alhaji et al. [20] proposed an ACO-PSO hybrid feature extraction and optimization strategy that combines ant colony and particle swarm optimization to optimize deep learning features. Their strategy enhanced detection accuracy while reducing unnecessary characteristics.

Ou et al. [8] aimed to unify face verification and deepfake detection by accounting for differences in video quality, proving that verification and detection tasks may be incorporated into a single holistic framework.

These approaches demonstrate that conventional feature engineering remains an important tool for improving the resilience and efficiency of deepfake detectors.

III. THEORETICAL FRAMEWORK

The idea behind deepfake detection is that fake media can't perfectly copy the natural patterns found in how real people act, how things look, and how sound and video sync up. Modern detection systems are based on ideas from computer vision, signal processing, biometrics, temporal analysis, and adversarial learning. Researchers can use these ideas to make models that can find small inconsistencies that happen when deepfake generation happens and make them stronger against more advanced ways of manipulating data.

A. Spatial-Level Analysis

At the frame level, deepfake detection focuses on identifying visual irregularities within individual images. Generative models often fail to reproduce fine-grained details such as skin texture, lighting balance, and boundary consistency.

1) Texture and Pixel Irregularities:

Images made by GANs may have an unnatural smoothness, distorted skin details, or blending artifacts near the edges. Convolutional filters in CNN-based models take advantage of these flaws to find spatial distortions and frequency-related problems, as Singh et al. [15] and Patel et al. [25] have shown..

2) Frequency-Based Artifact:

Deepfake synthesis changes frequency coefficients, particularly in the DCT domain used for JPEG compression. Guarnera et al. [4] explain how frequency-based anomalies can reveal manipulations that are invisible in spatial space. The gray-box adversarial approach investigated in their research demonstrates how deepfake detectors must evaluate frequency patterns to be resilient.

3) Facial Structure Theory

Natural faces follow stable geometric patterns. During movement or changes in expression, deepfake systems may change these patterns. Ramadhani et al. [22] and Yu et al. [19] talk about techniques that use facial landmarks and attention mechanisms to find these structural problems.

B. Temporal Consistency Theory

While spatial inconsistencies highlight frame-level errors, temporal inconsistencies disclose abnormal movement patterns that are more difficult for generative algorithms to resolve.

1) Motion Sequence Irregularities

Human movements follow predictable biological patterns, but deepfakes often fail to maintain consistent blinking, lip synchronization, or head motion. Hybrid models like CNN-LSTM and Xception-LSTM [1] detect such inconsistencies by analyzing frame sequences.

2) Long-Range Dependency Theory

Transformer-based models, including ViT and FakeFormer [17], capture long-term relationships across frames using attention mechanisms. This enables detection of subtle inconsistencies that may not appear in individual frames but emerge over time.sequentially.

C. Multimodal Coherence Theory

Deepfake detection goes beyond visual discrepancies to include multi-sensory alignment of audio, face expressions, and speech patterns. This theoretical concept assumes that deepfake generation rarely achieves full synchronization across modalities.

1) Audio-Visual Synchronization Theory

According to Panja et al. [10] and Yıldırım & Aydın [18], mismatches in lip movement timing, speech rhythm, and mouth geometry can reveal deepfake manipulations in human speech, which requires precise coordination between audio phonemes and visual visemes.

2) Cross-Modal Redundancy Theory

Authentic videos exhibit natural redundancy across modalities:

- audio reflects emotional state seen in facial expressions,
- breathing sounds match chest motion,
- Speech intensity aligns with facial strain.

Multimodal deepfake detectors take advantage of these natural correlations. When multimodal cues differ, the video has most likely been edited.

D. Perceptual Fidelity Theory

According to perceptual fidelity theory, deepfake forgeries affect perceptual realism in subtle ways that human eyes may not perceive but machine models can quantify.

Peng et al. [11] introduce DeepFidelity, which applies perceptual similarity metrics to assess:

- illumination consistency,
- identity distortion,
- unnatural texture,
- perceptual deviation from human-like photorealism.

This theory explains how deepfake detectors assess a face's "authenticity score" using many perceptual factors.

E. Adversarial Learning and Robustness Theory

Deepfake detection is fundamentally adversarial; as detectors improve, forgers modify their manipulation algorithms to avoid detection.

1) Adversarial Perturbation Theory

Guarnera et al. [4] demonstrate how tiny alterations in input domains can trick detectors, particularly those that are overfitted to certain datasets. Adversarial learning theory suggests that detectors should focus on learning manipulation features rather than dataset artifacts. Robustness should also be evaluated under quality degradation, compression, and adversarial noise.

2) Proactive Detection

Li et al. [5] introduce a "learnable hidden face" idea in which identification traits are veiled during training to prevent detectors from depending on biometrics rather than manipulation patterns. This theoretical understanding provides generalizability across identities and datasets.

F. Feature Optimization Theory

Deepfake patterns can be detected more effectively by improving feature sets to minimize redundancy. Mohiuddin et al. [7] proposed feature-selection-aided optimization.

Alhaji et al. [20] use ACO-PSO hybrid optimization to improve learnt representations.

These strategies follow the theoretical idea that eliminating irrelevant information improves model performance while preventing overfitting.

G. Integrated Detection and Verification

Recent works jointly execute deepfake detection and face verification tasks. Ou et al. [8] propose that both processes are based on evaluating identity consistency, and their combination may enhance robustness to quality variations and domain shift.

H. Summary of Theoretical Framework

Deepfake detection is facilitated by theoretical spatial, temporal, multimodal, perceptual, and adversarial perspectives. They form a foundation for the development of advanced models such as CNN-LSTM hybrids, transformer-based models, and multimodal approaches. As deepfake technology continues to improve, reinforcing these theoretical principles will be essential to build reliable and future-ready detection systems.

IV. CORE TECHNOLOGY MODELLING AND PERFORMANCE

Modeling of computer architectures, feature extraction algorithms, multimodal learning pipelines and resilience mechanisms needs to be stringent for deepfake detection systems. This section presents a comprehensive study of the primary technologies used in state-of-the-art deepfake detection research. It covers model architectures ranging from CNN-based spatial feature extractors to transformer-based spatiotemporal architectures, multimodal systems, lightweight detectors, adversarial robustness techniques, and advanced optimization frameworks. we focus on architectural reasons, underlying mechanisms and measurable performance on standard datasets.

A. CNN-Based Architectures for Spatial Analysis

CNNs remain fundamental in deepfake detection due to their shown capacity to extract fine-grained spatial data. These systems use convolution operations, batch normalization, and pooling layers to find patterns associated with manipulation artifacts.

1) Improved Dense CNN Architectures

Patel et al. [25] proposed an improved Dense CNN framework that deepens feature hierarchies for identifying subtle textural inconsistencies typical in deepfake-generation pipelines. The dense connectivity pattern promotes:

- feature reuse,
- enhanced gradient propagation,
- higher sensitivity to artifacts in boundary regions,
- improved generalization across datasets.

Dense CNNs perform well on datasets such as Celeb-DF and FaceForensics++, frequently exceeding 90% accuracy in controlled conditions.

2) Xception Models and Their Extensions

The Xception network, which is built on depthwise separable convolutions, is commonly used in deepfake detection because of its ability to capture separable spatial correlations. Akshaya et al. [1] used Xception as the major spatial backbone in their CNN-LSTM hybrid model, resulting in robust frame-level feature extraction before temporal processing.

Xception-based detectors demonstrate strong performance on:

- face-swap deepfakes,
- identity replacement forgeries,
- video-level manipulation where spatial consistency varies across frames.

Their performance advantage stems from efficient convolution operations and deeper representation learning.

B. Spatio-Temporal Hybrid Models

While CNNs excel at spatial analysis, many deepfakes rely on maintaining frame-by-frame coherence while ignoring natural temporal dynamics. Hybrid models capture these inconsistencies by combining spatial extractors with temporal modules like LSTMs or Transformers.

1) CNN + LSTM Detection Pipelines

The hybrid model by Akshaya et al. [1] is a prime example. Their architecture:

- uses Xception for frame-level feature extraction,
- feeds representations into LSTM layers,
- models motion continuity, blink rate, and lip movement,

detects inconsistencies arising from misaligned generative frames.

Empirical results show improved accuracy and reduced false positives in videos exhibiting subtle temporal distortions.

2) CNN + LSTM + Transformer Integration

Petmezas et al. [12] enhanced the hybrid approach by integrating a CNN feature extractor, LSTM for short-term temporal modeling, and Transformer blocks for long-range temporal attention. This architecture excels in detecting face reenactment, identity-driven deepfakes, and temporal smoothing by generative models. The use of Transformers minimizes information loss in extended video sequences, outperforming CNN-LSTM configurations.

C. Transformer-Based Vision Architectures

Transformer models have rapidly gained traction in deepfake detection due to their ability to capture global dependencies across spatial and temporal dimensions.

1) Swin Transformer Frameworks

Swin Transformers introduce a hierarchical attention mechanism that:

- partitions images into shifted windows,
- captures both fine (local) and global patterns,
- improves robustness to compression and generative smoothing.

Two major transformers in your reference list:

- EffiSwinNet by Basu et al. [3]
- Swin-BiLSTM with Attention by Lukitania & Suryani [6]

Both outperform standard models on datasets using StyleGAN, GAN-based synthesis, and neural reenactment approaches.

2) Video Vision Transformer (ViViT) Enhancement

Ramadhani et al. [22] enhanced ViViT by introducing:

- facial landmark integration,

- depthwise separable convolution,
- self-attention optimization.

These modifications boost the model's capacity to focus on biologically relevant parts like the eyes, nose, and mouth, where alterations are most common.

3) Lightweight Transformer Designs

FakeFormer by Wang et al. [17] represents a major push toward efficient deepfake detection. Its architecture:

- reduces parameter count,
- maintains high context-awareness,
- enables real-time deployment.

FakeFormer is ideal for mobile fraud detection, such as KYC verification or on-device content authentication.

D. Multimodal Audio-Visual Detection Systems

Deepfakes are increasingly manipulating visual imagery and audio signals. Thus, multimodal detectors perform substantially better than unimodal detectors.

1) Audio-Visual Synchronization Networks

Panja et al. [10] propose a comprehensive multimodal architecture combining:

- CNNs for spatial feature extraction,
- Vision Transformers (ViT) for global attention,
- audio-spectrogram CNN for voice anomaly detection,
- cross-modal consistency modules.

This system evaluates:

- lip-speech alignment,
- phoneme–viseme synchronization,
- voice-timbre consistency,
- audio spectral coherence.

2) Multi-signal Deepfake Detection

Yıldırım & Aydın [18] extend multimodal work to handle:

- audio deepfake detection (voice cloning),
- video deepfake detection (face manipulation),
- cross-modal coherence checking.

This dual-detection feature mimics real-world fraud scenarios in which attackers can alter both face and voice.

Multimodal models provide some of the highest real-world resilience, particularly for manipulated social-media videos.

E. Lightweight Deepfake Detection Models

As deepfake detection moves toward real-world applications, efficiency becomes essential. Lightweight models enable deployment in:

- mobile devices,
- IoT security systems,
- real-time content moderation networks,
- edge-based face verification.

1) DeepFaceGuard

Tauseef et al. [16] introduced DeepFaceGuard, a lightweight CNN optimized for:

- low-parameter processing,
- real-time inference,
- robustness to compression.

Its design makes it suited for applications with minimal computational resources.

2) FakeFormer (Lightweight Transformer)

As discussed earlier, FakeFormer [17] balances performance with speed. It is one of the few architectures that:

- achieves transformer-level accuracy,
- runs efficiently on constrained hardware,
- supports real-time fraud detection.

These lightweight systems provide an important step toward accessible deepfake security.

F. Adversarially Robust Detection Models

Deepfake detection techniques are vulnerable to adversarial assaults designed to circumvent categorization. Research has begun to incorporate adversarial robustness into core models.

1) Gray-Box Attack Defense

Guarnera et al. [4] expose vulnerabilities in DCT-based detectors. Their work motivated:

- adversarial training,
- frequency-domain consistency checking,
- robust spatial-frequency fusion networks.

Models incorporating these principles perform better against deliberate sabotage attempts.

2) Proactive Generalization Mechanisms

Li et al. [5] introduce the concept of "learnable hidden face features," which include identity-masking

modifications in training pipelines to prevent overfitting on identity cues.
This increases generalization, enabling detectors to:

- perform consistently across unseen identities,
- resist domain shifts,
- reduce identity bias.

G. Feature Selection, Optimization, and Classical Enhancements

Traditional optimization algorithms still contribute meaningfully to deepfake detection.

1) Feature-Selection-Driven Deepfake Detection

Mohiuddin et al. [7] leverage advanced feature selection techniques to refine CNN-based deepfake detection. Benefits include:

- reduced dimensionality,
- enhanced inference speed,
- decreased overfitting.

2) Hybrid Bio-Inspired Optimization (ACO + PSO)

Alhaji et al. [20] use Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) to improve deep learning feature spaces. This method is applied prior to classification, producing:

- highly discriminative features,
- improved accuracy,
- faster convergence.

Such hybrid optimization approaches simplify training while increasing detection robustness.

H. Summary of Core Technology Modeling

Across the reviewed literature, certain trends emerge:

- CNNs remain essential for spatial feature extraction.
- Hybrid CNN-LSTM and CNN-Transformer architectures dominate temporal detection.
- Transformer-based models (Swin, ViViT, FakeFormer) demonstrate superior generalization.
- Multimodal architectures provide the most robust real-world performance.
- Lightweight CNNs and transformers are crucial for practical deployment.
- Adversarial robustness is an emerging priority.
- Feature selection and optimization enhance efficiency and accuracy.

These technological breakthroughs contribute to the global effort to defend digital environments from deepfake-enabled video fraud.

V. SYSTEM VALIDATION STATUS AND EMPIRICAL PERFORMANCE

Deepfake detection systems must be rigorously validated in order to demonstrate their reliability, resilience, and generalizability in real-world circumstances. System validation consists of evaluating models on benchmark datasets, cross-dataset testing, compression robustness, adversarial stress testing, real-world generalization, multimodal consistency checks, runtime efficiency evaluation, and ablation investigations. This section summarizes the empirical performance outcomes presented in the surveyed literature and explains how these studies gain detection credibility.

A. Benchmark Datasets Used for Validation

i) Most evaluated papers evaluate their models on standard public datasets recognized for a variety of manipulation types:

- FaceForensics++: Used in [1], [12], [14], [25]. Contains manipulated videos created using:

1. DeepFake,
2. Face2Face,
3. FaceSwap,
4. NeuralTextures.

ii) Models typically achieve >90% accuracy on this dataset.

- Celeb-DF (v2): Used in [3], [6], [10], [11], [17], [18]. More realistic and demanding; detectors frequently lose 10 - 25% accuracy compared to FF++.
- DFDC (Deepfake Detection Challenge): Used in [3], [6], [10], [18], [22]. A large-scale real-world dataset curated by Facebook.
- Custom GAN, StyleGAN, and Reenactment Datasets: Used in [3, 4], [7], and [19]. These datasets test resilience against generative models that are not present in the training data.

B. Validation of CNN-Based Architectures

CNN models primarily undergo frame-based validation.

1) Xception and Dense CNN Performance
- Akshaya et al. [1] report 92-96% accuracy using Xception-LSTM on FF++.
- Patel et al. [25] shows improved Dense CNNs achieving 90-94% accuracy for high-quality images.
- Singh et al. [15] demonstrates 93% accuracy on hybrid image deepfake sets.

These findings demonstrate that CNNs excel at detecting spatial artifacts but suffer under compression or adversarial situations.

C. Spatio-Temporal Model Validation

1) CNN-LSTM Sequences

Hybrid architectures significantly outperform CNN-only detectors.
- Akshaya et al. [1] demonstrate a ≈5-8% improvement in accuracy over static CNN models due to temporal modeling.
- Petmezas et al. [12] show CNN-LSTM-Transformer models reduce false positives in multi-frame scenarios.

2) Effective Temporal Modeling Gains
- Temporal approaches specifically improve detection in:
- Deepfake talking-head videos,
- Face reenactment sequences,
- Videos with temporal smoothing,
- Low-quality mobile uploads.

Temporal modeling improves performance where static detection fails.

D. Transformer-Based Model Validation

Transformers show the best generalization ability.

1) EffiSwinNet
- Basu et al. [3] report top-tier performance against StyleGAN-generated faces.
- Achieves high accuracy even in cross-dataset scenarios.

2) Swin-BiLSTM (with Attention)
- Lukitania & Suryani [6] report competitive detection rates on FF++, Celeb-DF, and DFDC.
- Attention mechanisms help maintain performance under compression.

3) FakeFormer (Lightweight Transformer)
- Wang et al. [17] shown that FakeFormer achieves transformer-level accuracy with less parameters.
- Shows preparedness for real-time or mobile deployments.
- Transformers surpass traditional networks for generalization and robustness.

E. Multimodal Detector Validation

Multimodal systems provide the strongest defense in practical scenarios.

1) Audio-Visual Synchronization Models

Panja et al. [10] report:
- Improved detection under audio-visual manipulation,
- High accuracy on videos where lip and speech cues mismatch,
- Enhanced robustness to compression noise.

2) Multi-Signal Systems

Yıldırım & Aydın [18] verify:
- Strong results in detecting audio-only and video-only deepfakes,
- Best performance when both modalities are used jointly,
- Superior results in real-world videos from actors and politicians.

Multimodal detectors excel when deepfake creators focus manipulation on only one modality.

F. Lightweight Model Validation

Lightweight detectors emphasize accuracy–efficiency balance.

1) DeepFaceGuard
- Tauseef et al. [16] demonstrate:
- Real-time processing on low-power devices,
- Competitive accuracy (~85-90%) with minimal hardware.

2) FakeFormer
- Wang et al. [17] validate FakeFormer against:
- FF++: very high performance,
- DFDC: strong generalization,
- Mobile-based fraud cases: excellent runtime.

Lightweight models focus on deployment feasibility rather than pure accuracy.

G. Adversarial Stress-Testing and Robustness Validation

1) Gray-Box Attacks

Guarnera et al. [4] applied adversarial perturbations to show that:

- Many detectors fail under even mild DCT perturbations.
- Robust systems must incorporate frequency-domain analysis.

2) Hidden-Face Generalization

Li et al. [5] demonstrate:

- A learnable hidden-face strategy improves cross-dataset accuracy,
- Prevents overfitting on specific identities,
- Enhances robustness to unseen video qualities.

3) DeepFidelity Assessment

Peng et al. [11] validate DeepFidelity:

- Detects perceptually consistent forgeries that fool CNNs,
- Generalizes better across datasets.

H. Cross-Dataset Generalization Testing

Cross-dataset validation is essential because real-world deepfakes differ from academic datasets.

General Findings Across Studies

- CNN-only models typically lose 15-30% accuracy when tested on unseen datasets (e.g., FF++ → Celeb-DF).
- Transformer-based models show only 5-12% degradation, indicating stronger generalization.
- Multimodal detectors have the lowest drop, around 3-8%, due to multi-signal reinforcement.

This highlights the growing importance of multimodal and transformer-based architectures for real-world deployment.

I. Ablation Studies

Several works conduct ablation studies to evaluate component contribution:

1) With vs. Without attention mechanisms

Attention mechanisms (in [6], [12], [17]) consistently improve performance by:

- increasing focus on manipulated regions,
- reducing irrelevant feature noise,
- improving generalization.

2) With vs. Without multimodal integration

- Multimodal systems outperform by:
- capturing cross-modal inconsistencies,
- improving robustness against compression,
- enabling real-world reliability.

VI. CONCLUSION

The rapid development of deepfake technology powered by powerful generative models presents exciting opportunities but also poses serious risks to security, privacy, and digital trust. In this paper, we reviewed the main detection approaches: CNN-based, hybrid, transformer-based and multimodal methods. CNNs are good at extracting spatial features but poor in generalization. Hybrid and transformer-based models capture both spatial and temporal patterns to improve the performance. Multimodal approaches are very useful in real-world scenarios as they add robustness to detection by looking for inconsistencies between audio and visual data.

In conclusion, effective deepfake detection requires a combination of techniques for improved accuracy and robustness, focusing on lightweight models, adversarial resilience, and real-time deployment. Continued research and innovation are essential for upholding digital trust against the rising threat of synthetic media.

ACKNOWLEDGMENT

The authors would like to offer their heartfelt appreciation to all academics, scientists, and institutions whose pioneering work in deepfake detection helped shape the development of this review. The complete insights derived from research on CNN architectures, transformer-based models, multimodal analysis, feature optimization, and adversarial robustness have considerably increased understanding in this quickly growing subject. The authors also recognize the value of open-source datasets, academic benchmarking projects, and international conferences, which all contribute to continual innovation and enable researchers to analyze, compare, and improve detection approaches. Their contributions served as the foundation for the development of this review paper.

REFERENCES

[1] D. Akshaya, D. Sai Ram, S. Prathiba, and K. Jamal, "Hybrid Deepfake Video Detection using Xception and LSTM," in Proc. 2025 9th Int. Conf. on Inventive Systems and Control (ICISC), pp. 237-244, 2025.

[2] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," Alexandria Engineering Journal, vol. 125, pp. 265-277, 2025.

[3] S. Basu, P. Yanduri, S. Sharma, and A. Banerjee, "EffiSwinNet: A Cross-Architecture Deepfake Detection with Style GAN Generated Faces," in Proc. 2025 Int. Conf. on Computing, Intelligence, and Application (CIACON), 2025.

[4] F. Guarnera, L. Guarnera, A. Ortis, S. Battiato, and G. Puglisi, "A Novel Adversarial Gray-Box Attack on DCT-Based Face Deepfake Detectors," IEEE Access, vol. 13, pp. 172216-172229, 2025.

[5] H. Li, S. Yang, R. Xia, L. Yuan, and X. Gao, "Big Brother Is Watching: Proactive Deepfake Detection via Learnable Hidden Face," IEEE Signal Processing Letters, vol. 32, pp. 2284-2288, 2025.

[6] N. F. Lukitania and V. Suryani, "Swin-BiLSTM with Attention for Face Swap Detection in Deepfake Videos," in Proc. 2025 Int. Conf. on Information and Communication Technology (ICoICT), 2025.

[7] Sk Mohiuddin, A. Roy, S. Pani, S. Malakar, and R. Sarkar, "A feature selection-aided deep learning based deepfake video detection method," Multimedia Tools and Applications, vol. 84, pp. 40617-40636, 2025.

[8] F.-Z. Ou, H. Guo, and S. Kwong, "Towards Unified Face Verification Against Quality Variations and Deepfake," in Proc. 2025 Int. Symp. on Machine Learning and Media Computing (MLMC), 2025.

[9] R. Pandey and A. K. S. Kushwaha, "Hybrid Deep-Learning Model for Deepfake Detection in Video using Transfer Learning Approach," Natl. Acad. Sci. Lett., vol. 48, no. 3, pp. 337–341, 2025.

[10] A. Panja, D. Chakravorty, D. Saha, and A. Das, "Multimodal Framework for Deep Fake Detection and Content Moderation Using CNN, ViT, and Audio-Visual Analysis," in Proc. 2025 Int. Conf. on Computing, Intelligence, and Application (CIACON), 2025.

[11] C. Peng, H. Guo, D. Liu, N. Wang, R. Hu, and X. Gao, "DeepFidelity: Perceptual Forgery Fidelity Assessment for Deepfake Detection," IEEE Transactions on Circuits and Systems for Video Technology, 2025.

[12] 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," Multimedia Tools and Applications, vol. 84, pp. 40617-40636, 2025.

[13] G. Pranussha and V. Naveen Kumar, "Integrated Ensemble UNet Architecture for Deep Fake Detection in Videos," in Proc. 2025 9th Int. Conf. on Inventive Systems and Control (ICISC), pp. 527-534, 2025.

[14] T. Rajesh and S. Maruthuperumal, "An Innovative Model of Deepfake Detection in Video using 3D EfficientnetB7 with Spatial Attention," presented at an IEEE Conference, 2025.

[15] S. K. Singh, G. Kumar, A. K. Dubey, R. Yadav, V. Rai, and A. Sharma, "Image-based DeepFake Detection Using Artificial Intelligence," in Proc. 2025 IEEE 14th Int. Conf. on Communication Systems and Network Technologies (CSNT), pp. 927–931, 2025.

[16] Md. Tauseef, L. Sosa Viji, and J. Fenwick, "DeepFaceGuard: A Lightweight CNN Framework for Robust Face Manipulation Detection," in Proc. 2025 Third Int. Conf. on Networks, Multimedia and Information Technology (NMITCON), 2025.

[17] X. Wang, G. Zhong, and Q. Song, "FakeFormer: Transformer-Based Lightweight Deepfake Video Detection Model," in Proc. 2025 7th Int. Conf. on Information Science, Electrical and Automation Engineering (ISEAE), pp. 856-859, 2025.

[18] M. Yıldırım and İ. Aydın, "Detecting Deepfake Audio and Video Manipulation with Multimodal Deep Learning," in Proc. 2025 15th Int. Conf. on Advanced Computer Information Technologies (ACIT), pp. 863-867, 2025.

[19] Z. Yu, Y. Dai, W. Dai, and Y. Lin, "Deepfake Face Detection Algorithm Based on Multi-scale

Attention Reconstruction," in Proc. 2025 44th Chinese Control Conference (CCC), pp. 7989-7996, 2025.

[20] [20] H. S. Alhaji, Y. Celik, and S. Goel, "An Approach to Deepfake Video Detection Based on ACO-PSO Features and Deep Learning," Electronics, vol. 13, no. 12, p. 2398, 2024.

[21] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: challenges and opportunities," Artificial Intelligence Review, vol. 57, article 159, pp. 1-47, 2024.

[22] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention," IEEE Access, vol. 12, pp. 8932–8939, 2024.

[23] Y. Tian, W. Zhou, and A. U. Haq, "Detection of Deepfakes: Protecting Images and Vedios Against Deepfake," in Proc. 2024 21st Int. Comput. Conf. on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 20-24, 2024.

[24] B. Wang, Y. Tang, F. Wei, Z. Ba, and K. Ren, "FTDKD: Frequency-Time Domain Knowledge Distillation for Low-Quality Compressed Audio Deepfake Detection," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 32, pp. 4905-4918, 2024.

[25] 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. 22081-22095, 2023.