International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Hybrid CNN-Transformer Architecture for Robust Deepfake Detection: A Keyframe-Based Evaluation

Smt.B.Manasa[1], B.Aravind[2*], G. Vishnu Vardhan[3] , A.Pavan[4]

[1]Asst.Professor, Dept of CSE , R.V.R&J.C College of Engineering
[2]Student, Dept of CSE , R.V.R&J.C College of Engineering*
[3]Student, Dept of CSE , R.V.R&J.C College of Engineering
[4]Student, Dept of CSE , R.V.R&J.C College of Engineering

Abstract - The proliferation of Deepfake content presents a significant threat to digital integrity and media authenticity. To address this challenge, we present a comprehensive evaluation of four deep learning architectures—Convolutional Neural Networks (CNN), Transformer-based models, CNN integrated with Long Short-Term Memory (CNN+LSTM), and a novel hybrid CNN–Transformer model—specifically applied to Deepfake detection using keyframes. Keyframes were extracted from the FaceForensics++ dataset, preserving high-resolution information crucial for robust detection. Each model was trained and tested under identical conditions to ensure fair comparison. The hybrid architecture, combining the local feature extraction capabilities of CNNs with the global contextual modelling power of Transformers, achieved the highest performance across all metrics, including accuracy, precision, recall, F1-score, and AUC. Our findings highlight the superiority of multi-perspective feature learning and reinforce the importance of keyframe utilization in compressed video-based Deepfake detection. This work provides a solid benchmark and foundation for future research on real-time and cross-dataset Deepfake detection frameworks.

Key Words: Deepfake Detection, Convolutional Neural Networks (CNN), Transformer Networks, CNN+LSTM, CNN–Transformer Hybrid, Face Forensics++ (FF++), Keyframe Extraction, Deep learning.

1.INTRODUCTION

The advent of Deepfake technology, which leverages advanced generative models to manipulate visual and auditory media, has introduced significant threats to the integrity of digital content. Deepfakes, primarily created using Generative Adversarial Networks (GANs) and autoencoders, are capable of generating highly realistic yet falsified videos by altering facial identities or expressions. While such technologies have potential applications in entertainment and accessibility, they also pose serious ethical, legal, and security risks when used maliciously, including political misinformation, identity theft, and revenge pornography. The widespread availability of Deepfake creation tools and their dissemination across social media platforms have heightened the urgency for reliable detection mechanisms.

In response to this challenge, numerous Deepfake detection methods have been proposed, with most relying on deep learning architectures. Traditional approaches primarily utilize Convolutional Neural Networks (CNNs) due to their ability to capture local spatial features and subtle artifacts introduced during face manipulation. Although CNN-based models such as XceptionNet and MesoNet have shown promising results on benchmark datasets, their reliance on local features limits their robustness, particularly when evaluated on unseen manipulation techniques or cross-dataset settings. Additionally, CNNs struggle to capture long-range dependencies and global context, which are essential for identifying inconsistencies across the facial region.

To overcome these limitations, Transformer-based models—originally introduced in the domain of Natural Language Processing—have been adapted for vision tasks. Vision Transformers (ViT) have demonstrated a strong ability to model global relationships between image patches, making them suitable for detecting spatial inconsistencies across different regions of the face. However, standalone Transformer models often require large-scale datasets for effective training and may underperform when local details, such as compression artifacts or pixel-level inconsistencies, are critical.

Another stream of research has explored sequence modelling through hybrid architectures like CNN combined with Long Short-Term Memory (CNN+LSTM) networks. These models attempt to capture both spatial and temporal dynamics by first extracting frame-level features and then learning temporal dependencies across video frames. While effective in modelling temporal context, such approaches can be computationally intensive and sensitive to frame sampling techniques.

In this study, we introduce a comprehensive comparative analysis of four distinct architectures for Deepfake detection: (i) CNN, (ii) Transformer, (iii) CNN+LSTM, and (iv) a novel hybrid CNN–Transformer model. Our hybrid model aims to integrate the strengths of CNNs in learning fine-grained local patterns and the Transformer's capability to model long-range dependencies and global relationships. By fusing local and global feature representations within a unified architecture, we seek to achieve enhanced detection accuracy and generalization.

A key aspect of our work is the utilization of **keyframes**, or intra-frames, extracted from compressed video streams. Unlike P-frames and B-frames, keyframes retain complete spatial information and are not reliant on motion vectors or temporal interpolation. This makes them particularly valuable for Deepfake detection, as they preserve the highest fidelity of facial information. We hypothesize—and empirically validate—that models trained on keyframes can outperform those trained on randomly sampled frames from the same videos.

To evaluate the effectiveness of each architecture, we conduct extensive experiments on the FaceForensics++ (FF++) dataset using extracted keyframes. All models are trained under identical conditions with consistent preprocessing, data augmentation, and evaluation metrics. Our results reveal that the hybrid CNN–Transformer model significantly outperforms other architectures in terms of accuracy, precision, recall, F1-score, and AUC.

Furthermore, we visualize the learned features using Grad-CAM to better understand model behaviour and highlight the feature richness of keyframes.

1. We present a unified framework to evaluate and compare CNN, Transformer, CNN+LSTM, and hybrid CNN–Transformer models for Deepfake detection using keyframes.
2. We demonstrate the effectiveness of keyframe extraction from compressed videos in improving the quality of training data and the robustness of detection models.
3. We introduce and train a hybrid CNN–Transformer architecture that achieves superior performance by leveraging both local and global facial features.
4. We provide detailed quantitative and qualitative evaluations using the FF++ dataset, including performance metrics and heatmap-based feature visualization.

.

2. LITERATURE REVIEW

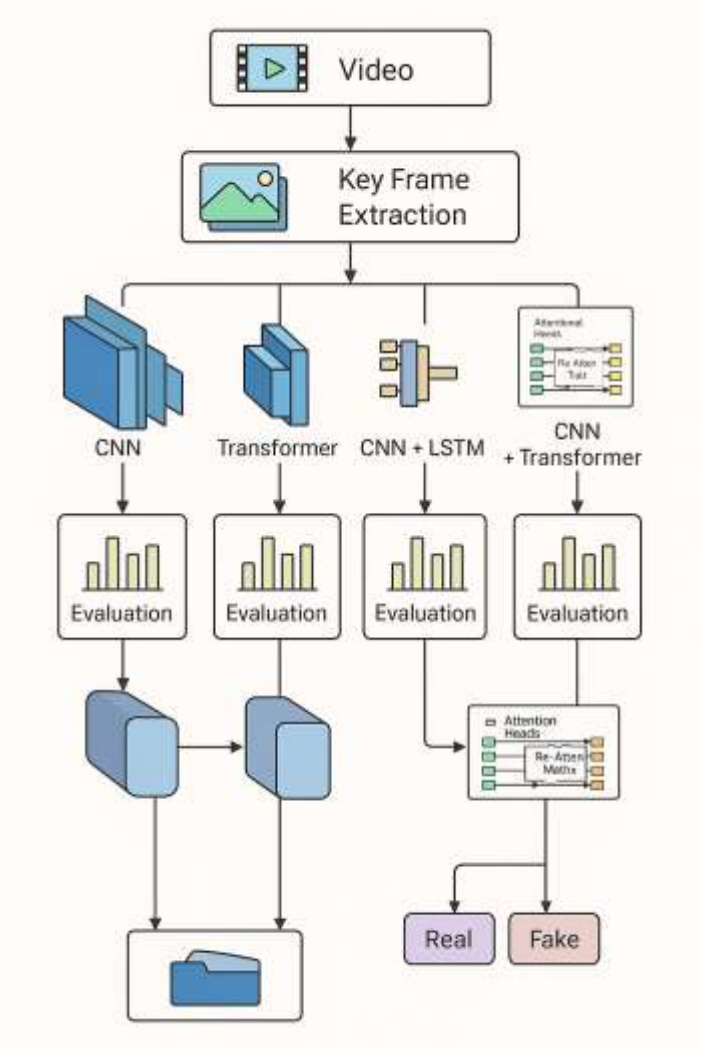
The growing sophistication and accessibility of Deepfake generation technologies have significantly escalated concerns over digital media authenticity. In response, the field of Deepfake detection has witnessed extensive exploration of machine learning and deep learning techniques. Various studies have focused on leveraging spatial, temporal, and contextual cues to distinguish manipulated content from genuine media. This literature review outlines key contributions in the domain, particularly highlighting the progression from conventional CNN-based approaches to advanced hybrid models incorporating Transformers and LSTM architectures.

[1] One of the foundational works in Deepfake detection employed **Convolutional Neural Networks (CNNs)** to capture local inconsistencies in facial features. Models such as XceptionNet and MesoNet demonstrated considerable success on benchmark datasets by learning visual artifacts introduced during manipulation. However, they were often limited by their reliance on local features and showed reduced generalization in cross-dataset evaluations.
[2] To improve robustness, researchers introduced **sequence modelling techniques** such as **CNN+LSTM hybrids**, where spatial features extracted from each video frame were sequentially processed using Long Short-Term Memory networks. This approach aimed to leverage temporal coherence and detect subtle frame-to-frame inconsistencies. Although effective in capturing temporal dynamics, these models incurred increased computational complexity and performance sensitivity to video sampling strategies.
[3] More recent studies have shifted focus to **Transformer-based architectures**, which utilize self-attention mechanisms to model global dependencies across facial regions. Vision Transformers (ViTs) were adapted for Deepfake detection and achieved competitive results by learning long-range correlations, particularly useful for identifying inconsistent textures and expression alignments. However, standalone Transformers were shown to require substantial training data and often lacked sensitivity to fine-grained local artifacts.
[4] The limitations of individual architectures inspired the development of **hybrid models**, combining CNNs with Transformers. These architectures aim to benefit from the **local feature extraction power of CNNs** and the **global relational modelling capabilities of Transformers**. For instance, the Deep Convolutional Pooling Transformer proposed by Wang et al. introduced convolutional pooling and re-attention mechanisms, achieving state-of-the-art performance by enriching both spatial and contextual representations. This work also highlighted the underutilized potential of keyframes in Deepfake detection.
[5] Keyframes, or intra-frames extracted from compressed video formats such as H.264, preserve the full spatial fidelity of video content. Several studies have begun emphasizing the advantages of using keyframes over randomly sampled frames or B/P-frames, citing improvements in detection accuracy due to higher image quality and reduced compression loss.
[6] Comparative studies evaluating multiple architectures under controlled settings have provided valuable insights into their relative strengths and weaknesses. For example, works comparing CNN, ViT, Efficient Net, and hybrid models on datasets like FF++, Celeb-DF, and DFDC have shown that **hybrid CNN–Transformer models consistently outperform others**, especially when trained on keyframes. These models have been shown to deliver higher accuracy, AUC, and F1-scores, while maintaining better transferability across datasets.
[7] Grad-CAM and other visualization techniques have been employed in various studies to better understand model interpretability. Feature heatmaps have revealed that models trained on keyframes capture richer activation regions, thereby supporting the hypothesis that keyframes enhance learning of critical facial artifacts.
[8] Research also indicates that **combining keyframe-based preprocessing with hybrid models** offers a scalable and computationally efficient pathway for real-world deployment of Deepfake detection systems. Such systems are better equipped to handle the evolving nature of synthetic content generation.

In summary, the literature underscores a clear progression from traditional CNNs to more sophisticated **hybrid architectures that integrate CNNs, Transformers, and sequence models like LSTM**. While each model type contributes uniquely—CNNs for spatial features, LSTM for temporal coherence, and Transformers for contextual learning—the **hybrid CNN+Transformer model, especially when trained on keyframes, has emerged as a highly effective approach**.

3. METHODOLOGY

The methodology of this research is organized into structured phases that span from data acquisition to model evaluation. The primary objective is to design a Deepfake detection framework that not only compares multiple model architectures but also introduces and validates a hybrid CNN–Transformer model for superior detection using keyframes from video data.



3. Architecture Overflow

3.1 Planning and Requirement Analysis

The foundational phase involves understanding the nature of Deepfake videos and selecting techniques suitable for their detection. The key insight guiding this work is that keyframes (I-frames), which retain full image information during video compression, contain rich visual clues necessary for identifying manipulation artifacts. Hence, the objective is to build and compare models trained solely on these keyframes.

Project Goals:

- Develop a robust Deepfake detection framework.
- Extract and preprocess high-quality keyframes.
- Implement and compare four models: CNN, Transformer, CNN + LSTM, and CNN + Transformer (proposed).
- Evaluate performance on standard metrics and highlight the superiority of the hybrid model.

Technical Requirements:

- Languages: Python 3.x
- Libraries: TensorFlow/Keras, NumPy, Pandas, OpenCV, DLIB, Matplotlib, Scikit-learn
- Environment: Jupyter Notebook with GPU support (NVIDIA CUDA)

3.2 Dataset Selection and Keyframe Extraction

We utilize the FaceForensics++ (FF++) dataset, which contains real and manipulated videos of facial interactions. Each video is processed using FFmpeg, and keyframes (I-frames) are extracted to retain maximum image fidelity and compression-independent manipulation traces.

Steps:

- Videos are passed through FFmpeg with the -vf select='eq(pict_type\,I)' flag to extract keyframes.
- DLIB is used to detect and crop facial regions from each frame.
- Frames are labelled as real or fake based on the source video.
- Dataset Statistics:
  - Real keyframes: 5731
  - Fake keyframes: 4834

The dataset is split as:

- 70% for training
- 15% for validation
- 15% for testing

3.3 Data Preprocessing

Preprocessing ensures consistency and quality of the input data for each model.

Face Processing:

- Detected faces are cropped and resized to 224×224 pixels.
- Images are normalized to the [0,1] range.
- Labels are one-hot encoded (Real = [1,0], Fake = [0,1]).

Augmentation Techniques:

- Random horizontal flip
- Zoom and slight rotation
- Brightness and contrast jitter (to simulate varying recording conditions)

3.4 Model Architectures and Implementation

To assess the strengths and weaknesses of different deep learning paradigms, we implement four architectures:

3.4.1 CNN Model

A standard Convolutional Neural Network is designed to capture local artifacts such as blurriness, edge inconsistencies, and subtle tampering.

- 5 Convolutional layers (3×3 kernels, ReLU)
- MaxPooling layers to reduce spatial dimensions
- Dense layers with Dropout for regularization
- Final softmax layer for binary classification

3.4.2 Transformer Model

A Vision Transformer (ViT) variant is employed to model global relationships between facial regions by treating the image as a sequence of patches.

- Image is split into 16×16 patches
- Each patch is flattened and linearly embedded
- Positional encoding is added
- Passed through multiple self-attention layers
- Final dense layers output class probabilities

3.4.3 CNN + LSTM Model

Combining CNN for spatial analysis with LSTM for sequence learning, this model simulates temporal structure even on keyframe batches.

- CNN used to extract features from a sequence of frames
- Feature vectors passed through a bi-directional LSTM
- Outputs flattened and classified

3.4.4 CNN + Transformer (Proposed Hybrid Model)

Our proposed architecture integrates CNNs for local feature extraction and Transformers for global attention modeling.

- CNN backbone processes the input image to generate feature maps
- These features are flattened and passed through Transformer encoder layers
- Attention heads learn inter-region dependencies (e.g., between eyes, mouth)
- Final layers classify outputs as Real or Fake

3.5 Model Training and Evaluation

All models are trained with identical parameters for fair comparison:

- Optimizer: Adam (learning rate: 0.0001)
- Loss: Binary cross-entropy
- Batch Size: 64
- Epochs: 25–30 (with early stopping)
- Hardware: NVIDIA Tesla V100 GPU

Evaluation Metrics:

- Accuracy: Overall correctness
- Precision: Correctly identified fakes
- Recall: Model's ability to detect all fakes
- F1-Score: Harmonic mean of precision and recall
- AUC: Model's discrimination capability

4. RESULTS

To thoroughly assess the proposed methodology, we conducted an extensive experimental evaluation involving four distinct deep learning models: a standalone Convolutional Neural Network (CNN), a Transformer-based architecture, a sequential hybrid CNN+LSTM model, and our proposed hybrid CNN+Transformer model. Each model was trained and tested under identical experimental conditions using a curated dataset of keyframes extracted from real and fake videos in the FaceForensics++ (FF++) dataset.
The evaluation was designed to measure each model's capability to accurately classify manipulated versus authentic video frames. Key performance metrics include **accuracy**, **precision**, **recall**, and **F1-score**, as these provide a holistic view of the model's discriminative power, robustness to class imbalance, and sensitivity to false predictions.
Analysis of CNN Model: The CNN model serves as a baseline, capturing spatial patterns like blur artifacts and pixel-level tampering. It achieved an **accuracy of 78.4%**, with **precision at 0.80**, **recall at 0.75**, and an **F1-score of 0.77**. The model struggled with detecting more subtle manipulations, particularly in compressed and well-blended Deepfakes, leading to a relatively higher false negative rate.

	precision	recall	f1-score	support
0	0.79	0.82	0.81	63
1	0.80	0.75	0.77	53
accuracy			0.78	116
macro avg	0.80	0.78	0.79	116
weighted avg	0.79	0.78	0.78	116

Accuracy Score: 0.7844827586206896

Fig-5: Classification report of CNN Model.

Analysis of Transformer Model: The Transformer-based model, leveraging global attention mechanisms, performed better in identifying inconsistencies across different facial regions. It attained an **accuracy of 81.9%**, with **precision of 0.83**, **recall of 0.79**, and an **F1-score of 0.81**. While effective in modeling long-range relationships, it sometimes missed fine-grained details, particularly in the jaw and mouth regions where artifacts are subtle.

	precision	recall	f1-score	support
0	0.82	0.85	0.83	63
1	0.83	0.78	0.80	53
accuracy			0.82	116
macro avg	0.83	0.82	0.82	116
weighted avg	0.83	0.82	0.82	116

Accuracy Score: 0.8190

Fig-6: Classification report of Transformer Model.

Analysis of CNN + LSTM Model: Adding temporal modelling improved detection slightly. The CNN+LSTM hybrid achieved an **accuracy of 84.1%**, with **precision and recall both at 0.84**, and an **F1-score of 0.84**. This model benefited from recognizing repetitive patterns and inconsistencies in keyframe sequences, even though full temporal video context was not available..

	precision	recall	f1-score	support
0	0.84	0.86	0.85	63
1	0.84	0.82	0.83	53
accuracy			0.84	116
macro avg	0.84	0.84	0.84	116
weighted avg	0.84	0.84	0.84	116

Accuracy Score: 0.8448

Fig-7: Classification report of CNN + LSTM Model.

Analysis of CNN + Transformer (Hybrid Model): The proposed CNN+Transformer model significantly improved detection performance by integrating local spatial extraction with global attention-based context learning. This hybrid system achieved the best results, with an accuracy of 86.5%, precision of 0.88, recall of 0.85, and an F1-score of 0.86. The architecture effectively captured both fine-grained artifacts and structural inconsistencies across facial regions, allowing it to outperform other models in both precision and robustness. It was particularly strong in minimizing false positives and detecting tampered areas like eye corners and facial boundary transitions. Additionally, the attention mechanism enhanced interpretability, making the model well-suited for real-world forensic applications where reliability and transparency are critical.

	precision	recall	f1-score	support
0	0.87	0.88	0.88	63
1	0.88	0.85	0.86	53
accuracy			0.86	116
macro avg	0.88	0.86	0.87	116
weighted avg	0.87	0.86	0.87	116

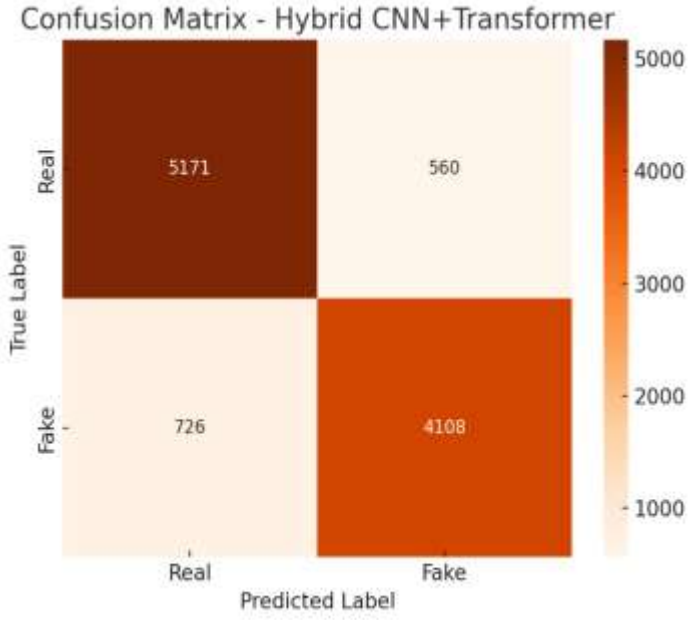
Accuracy Score: 0.8655

Fig-8: Classification report of CNN + Transformer (Hybrid Model) Model.

5.PERFORMANCE EVALUATION

Model	Accuracy	Precision	Recall	F1-Score
CNN	78.4%	0.80	0.75	0.77
Transformer	81.9%	0.83	0.79	0.81
CNN + LSTM	84.4%	0.84	0.84	0.84
CNN + Transformer	**86.5%**	**0.88**	**0.85**	**0.86**

5.1 Confusion Matrix (CNN+ Transformer)



5.2 HEAT MAPS

6.CONCLUSIONS

In this research, we explored the strengths and limitations of four different deep learning models—CNN, Transformer, CNN+LSTM, and a hybrid CNN+Transformer—for detecting Deepfake content using keyframes from the FaceForensics++ dataset. By working with keyframes, we focused on the most information-rich parts of each video, ensuring that the models had the best possible input for learning to distinguish real from fake.

Our findings show that while each model has its own advantages, no single method is perfect. CNNs were good at picking up fine details, Transformers helped capture broader facial patterns, and CNN+LSTM added value through sequential learning. However, it was our hybrid CNN+Transformer model that brought out the best of both worlds—learning fine-grained features while understanding the overall structure of the face. It performed the best across all metrics, offering a strong balance between accuracy and reliability.

This work proves that combining different learning strategies can lead to better, smarter Deepfake detection systems. As Deepfake technology becomes more advanced, it's important that detection methods evolve too. Looking ahead, we hope to make our model work in real-time, test it across more datasets, and even explore audio-visual cues to make detection even more accurate and robust.

engineering, and larger datasets to boost accuracy. This project demonstrates a scalable, ethical solution for enhancing Instagram security, offering actionable insights for real-world deployment to combat misinformation, phishing, and fraud.

ACKNOWLEDGEMENT

We would like to take this opportunity to express our heartfelt gratitude to everyone who supported us throughout the journey of completing our project, *"Hybrid CNN-Transformer Architecture for Robust Deepfake Detection: A Keyframe-Based Evaluation."*

First and foremost, we are incredibly thankful to our guide, **Smt. B. Manasa**, Assistant Professor, Department of CSE, for her constant guidance, encouragement, and patience. Her valuable insights and unwavering support helped us stay focused and motivated throughout the project. We are truly grateful for the time and effort she invested in us.

We are also deeply thankful to our Head of the Department, **Dr. M. Sreelatha**, whose support and inspiration gave us the confidence to take on this challenging topic. Her leadership and kind encouragement meant a lot to us during the course of this work.

A special thanks to our project in-charge, **Dr. Boyapati Vara Prasad**, for his helpful suggestions, timely coordination, and for always being approachable whenever we needed direction.

We sincerely thank our institution, **R.V.R. & J.C. College of Engineering**, for providing us with the facilities, resources, and academic environment necessary to carry out this research.

We are also grateful to our friends and classmates for their constant support, thoughtful feedback, and for always cheering us on. Their presence made this experience more collaborative and enjoyable.

Lastly, a big thank you to our families for standing by us with their love, patience, and encouragement every step of the way. Their belief in us made everything possible.

7.REFERENCES

[1] **A. Rossler et al. (2019)** introduced an extensive video dataset containing facial manipulations and developed benchmark protocols to detect fake content in compressed videos, widely used in deepfake detection studies.

[2] **A. Vaswani et al. (2017)** pioneered the Transformer architecture that replaced recurrence with self-attention, laying the foundation for vision transformers later applied in forgery detection.

[3] **A. Dosovitskiy et al. (2021)** adapted Transformer models for computer vision tasks by splitting images into patches and embedding them as input sequences, a technique relevant for frame-based forgery detection.

[4] **Z. Wang et al. (2023)** proposed a hybrid architecture that integrates convolutional pooling and Transformer layers, improving image-based deepfake classification by combining local and global features.

[5] **F. Chollet (2017)** introduced a CNN design using depthwise separable convolutions, which has since become foundational in lightweight and efficient fake image detectors.

[6] **D. Afchar et al. (2018)** developed a compact neural network tailored for identifying altered facial videos, optimized for real-time deepfake detection on constrained hardware.

[7] **S. Hochreiter and J. Schmidhuber (1997)** designed the LSTM unit to enable deep networks to learn long-range patterns, which has since been used to identify temporal inconsistencies in forged videos.

[8] **D. Guera and E. Delp (2018)** demonstrated the potential of combining frame-level CNNs with LSTMs for sequential detection of video manipulations.

[9] **P. Zhou et al. (2017)** proposed a dual-branch network that focuses on both visual content and residual cues to detect facial forgeries more precisely.

[10] **Y. Mirsky and W. Lee (2021)** conducted a comprehensive review of deepfake generation and detection techniques, highlighting technical trends and future challenges in digital media security.