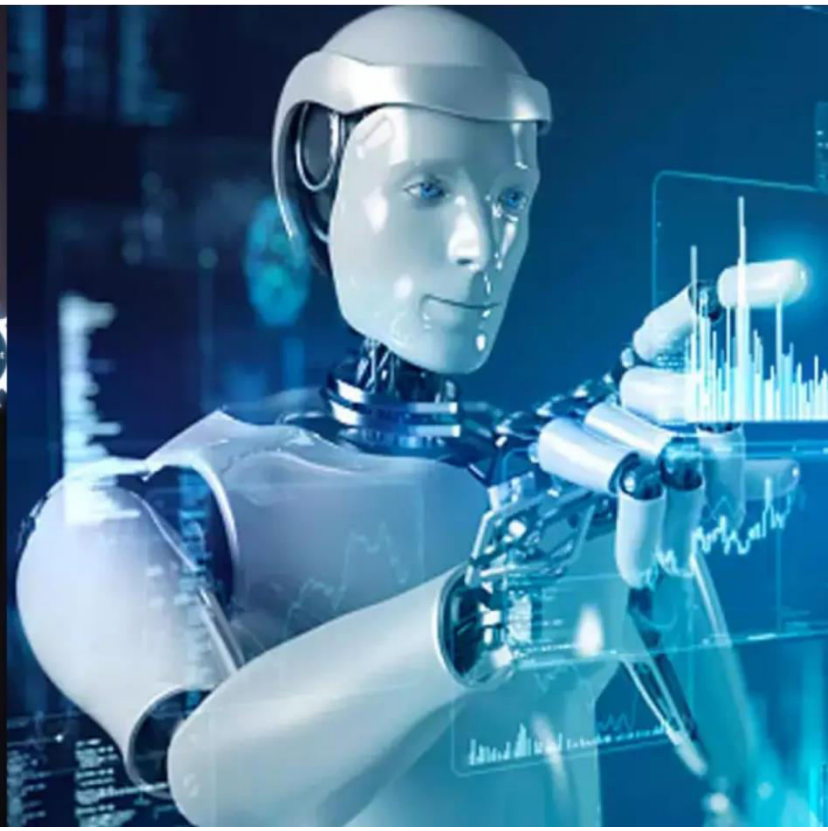
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 5, May 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

Deepfake Video Detection using Multi-Feature Analysis: ResNeXt-50 and LSTM

Samyak Tayde, **Sidharth Kamble**, **Sukhada Barve**, **Sarvesh Gonghe,** Prof. A. A. Chinchamalatpure

B.E. Student, Department of Computer Science and Engineering, Takshashila Institute of Engineering and Technology, Darapur, Amravati, Maharashtra, India

Guide, Department of Computer Science and Engineering, Takshashila Institute of Engineering and Technology, Darapur, Amravati, Maharashtra, India

ABSTRACT: The proliferation of GAN-based deepfake synthesis tools has created an urgent demand for automated video forensics systems capable of distinguishing manipulated media from authentic content. This paper presents a hybrid spatial-temporal deep learning framework for deepfake video detection, combining a pre-trained ResNeXt-50 Convolutional Neural Network (CNN) with a Long Short-Term Memory (LSTM) recurrent network. ResNeXt-50 extracts 2048-dimensional per-frame spatial feature vectors by exploiting aggregated residual transformations (cardinality = 32), while the LSTM processes ordered sequences of 150 consecutive frames to capture inter-frame temporal anomalies induced by face-synthesis artifacts. The model is trained on a custom balanced dataset of 12,000 videos (6,000 real, 6,000 fake) constructed by uniformly sampling 2,000 real and 2,000 fake videos from each of three public benchmarks: FaceForensics++ [1], Celeb-DF [3], and DFDC [2]. Systematic frame-count ablation reveals a monotonic accuracy improvement from 90.95% (20 frames) to 97.76% (100 frames) on FaceForensics++. The proposed system is deployed as a Django web application providing real-time inference with Softmax confidence scoring.

KEYWORDS: Deepfake Detection, ResNeXt-50, LSTM, FaceForensics++, DFDC, Celeb-DF, Temporal Analysis, Spatial Feature Extraction, PyTorch, Video Forensics.

I. INTRODUCTION

The rapid advancement of generative deep learning, particularly Generative Adversarial Networks (GANs) [13] and variational autoencoders, has enabled the creation of highly realistic synthetic face videos, commonly referred to as deepfakes. Tools such as FaceSwap [12] and FaceApp [11], which internally leverage encoder-decoder neural architectures, have democratised deepfake creation to the extent that convincing manipulations can be produced on consumer-grade hardware within minutes.

The societal implications are severe. Documented incidents include fabricated political speeches [14], non-consensual synthetic intimate imagery, corporate impersonation fraud, and targeted harassment campaigns. The asymmetry between creation ease and detection difficulty constitutes the core research challenge: state-of-the-art generative models are explicitly trained to minimise detectable artifacts, while detectors must generalise across unknown manipulation pipelines.

Existing detection approaches can be broadly categorised into spatial methods, which analyse per-frame visual inconsistencies [15], and temporal methods, which model inter-frame anomalies [18]. Purely spatial approaches are blind to temporal manipulation signatures, while early temporal approaches [18] were constrained by small datasets and vanilla RNN architectures susceptible to vanishing gradients over long frame sequences. This work addresses both limitations.

The proposed framework employs ResNeXt-50 for rich spatial feature extraction, leveraging its multi-path grouped convolutions to capture diverse deepfake artifact patterns, and an LSTM network for temporal sequence modelling over 150-frame clips. The principal contributions are: (i) a unified spatial-temporal detection pipeline; (ii) a 12,000-video balanced mixed dataset from three benchmarks with equal contribution from each source; (iii) a systematic frame-count ablation study; and (iv) a production-ready Django inference application.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

II. RELATED WORK

Li and Lyu [15] proposed a CNN-based detector targeting face warping artifacts introduced when a generated face region is resized and blended onto the target background. While effective on early low-resolution deepfakes, this approach is rendered less effective by modern high-resolution generation pipelines that minimise blending boundaries, and it performs no temporal analysis.

Li et al. [16] exploited suppressed eye-blink frequency in early deepfake videos using a Long-term Recurrent Convolutional Network (LRCN). This biological signal, while previously reliable, has been largely overcome by contemporary GAN generators that synthesise physiologically plausible blink patterns, making single-cue temporal analysis insufficient.

Nguyen et al. [17] applied capsule networks, which provide rotation-invariant feature representations beneficial for detecting spatially warped faces. However, their training procedure introduced artificial perturbations that degraded performance on unperturbed real-world inputs.

Güera and Delp [18] provided the first systematic evidence that sequential frame analysis via CNN+RNN outperforms frame-independent classifiers. Their evaluation, however, was limited to 600 videos from a single source, raising concerns about generalisation. The use of vanilla RNN units also introduces vanishing gradient instability over sequences exceeding 30 frames. The proposed system directly addresses these gaps through LSTM gating and a 12,000-video multi-source training corpus.

III. EXISTING SYSTEMS AND LIMITATIONS

A. Spatial-Only Detection Methods

Frame-independent CNN classifiers such as XceptionNet fine-tuned on FaceForensics++ achieve high intra-dataset accuracy (>99% on FF++ c0) but exhibit significant performance degradation on unseen datasets, with reported accuracy dropping to 65–70% on Celeb-DF [3]. These methods cannot exploit temporal coherence cues, which are among the most discriminative features for detecting modern high-quality deepfakes.

B. Biological Signal Methods

Remote photoplethysmography (PPG) signal extraction [20] and eye-blink analysis [16] rely on physiological regularities that modern generators have largely learned to replicate. Furthermore, such methods require stable lighting conditions and front-facing camera angles, severely limiting applicability to in-the-wild social media videos.

C. Vanilla RNN Temporal Methods

The vanishing gradient problem in standard RNNs limits effective temporal context to approximately 20–30 frames. Given that the proposed system processes 150 sequential frames, vanilla RNN architectures are fundamentally unsuitable. LSTM's gating mechanism explicitly addresses this by maintaining persistent cell state across the full sequence length.

D. Identified Research Gaps

No prior published system trains simultaneously on equal-contribution samples from FF++, Celeb-DF, and DFDC. No systematic frame-count ablation study has been published for CNN+LSTM deepfake detection. ResNeXt-50's aggregated transformation design has not previously been combined with LSTM for this task.

IV. SYSTEM ARCHITECTURE

The proposed detection system comprises three sequential stages. Stage 1 performs video ingestion and preprocessing: MTCNN-based face detection, spatial alignment, and 112×112 pixel cropping at 30 fps. Stage 2 performs per-frame spatial feature extraction using a fine-tuned ResNeXt-50 backbone, producing a 2048-dimensional feature vector per frame. Stage 3 performs temporal sequence classification: the ordered sequence of T = 150 feature vectors is processed by an LSTM, whose final hidden state is passed through a fully connected classification head with Softmax output.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

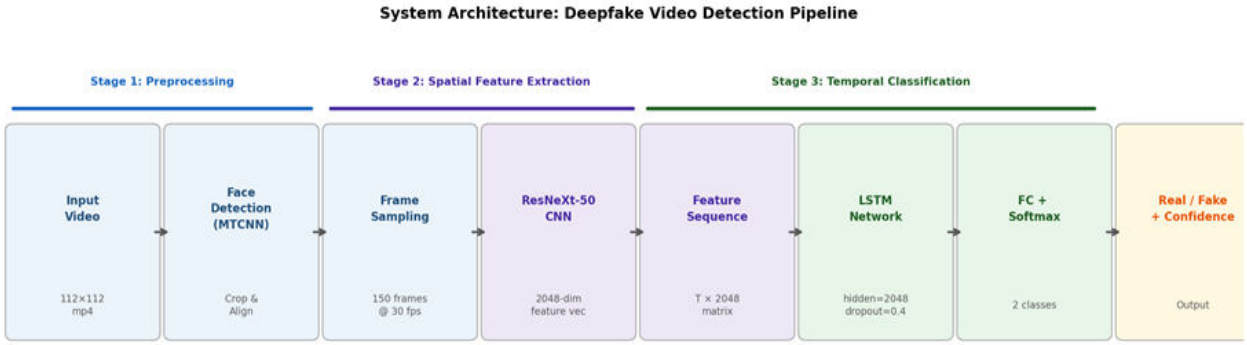


Fig. 1: Three-stage system architecture — Preprocessing → Spatial Feature Extraction → Temporal Classification.

A. ResNeXt-50 Backbone

ResNeXt-50 (resnext50_32x4d) [22] extends the ResNet residual block by replacing standard 3×3 convolutions with a set of C = 32 parallel grouped transformations, each of width d = 4. The aggregated transformation is expressed as: $F(x) = \Sigma_{i=1}^{c} T_i(x)$, where $T_i(\cdot)$ denotes the i-th transformation. This cardinality-based design captures a richer feature space than increasing depth or width alone. Pre-trained ImageNet weights initialise the backbone; only the final two residual block groups are fine-tuned to prevent overfitting on the relatively small per-class video counts.

B. LSTM Mathematical Formulation

Given an input sequence $x_1, x_2, ..., x_T$ of 2048-dimensional ResNeXt feature vectors, the LSTM computes the following at each timestep t:

$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ (Forget gate)
$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$ (Input gate)
$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$ (Candidate cell)
$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$ (Cell state update)
$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ (Output gate)
$h_t = o_t \odot \tanh(C_t)$ (Hidden state)

where σ is the sigmoid activation, ⊙ denotes element-wise multiplication, and W_f, W_i, W_C, W_o are learned weight matrices. The forget gate f_t modulates which prior cell state information is discarded, enabling the LSTM to selectively retain evidence of manipulation artifacts detected in early frames while processing frame 150. The final hidden state h_T is passed to the classification head.

V. METHODOLOGY

A. Dataset Construction

A custom balanced dataset of 12,000 videos was constructed by uniformly sampling 2,000 real and 2,000 fake videos from each of three public benchmarks. This equal-contribution strategy ensures the model is not biased toward any single manipulation type or compression style. Audio-manipulated videos in the DFDC subset were identified and removed via a Python preprocessing script prior to sampling, as audio forensics is outside the scope of this work.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

Table 1: Dataset composition — equal contribution from all three benchmark sources

Dataset	Real	Fake	Total	Key Characteristic
FaceForensics++ [1]	2,000	2,000	4,000	4 manipulation types, 3 compression levels
Celeb-DF [3]	2,000	2,000	4,000	High-quality celebrity deepfakes
DFDC [2]	2,000	2,000	4,000	Diverse compression, lighting, ethnicity
Total	**6,000**	**6,000**	**12,000**	**50% real — 50% fake (balanced)**

B. Preprocessing

Videos are decoded using cv2.VideoCapture at the native frame rate. MTCNN face detection is applied per frame; detected faces are cropped and resized to 112×112 pixels. The first 150 sequential frames are retained to preserve temporal ordering for LSTM processing. Frames with no detectable face are discarded. Processed clips are stored at 30 fps in MP4 format. The 150-frame threshold was selected based on the mean frame count across all 12,000 videos and available GPU memory constraints (batch size = 4).

C. Train / Test Split

The 12,000-video dataset is partitioned into a training set of 8,400 videos (70%) and a test set of 3,600 videos (30%). Both subsets maintain the 50/50 real-to-fake balance and equal dataset source representation. The partition is fixed across all experiments to ensure fair comparison between model configurations.

D. Training Configuration

Optimiser: Adam [21] (lr = 1×10^{-5}, weight decay = 1×10^{-3}). Loss: Binary Cross-Entropy. Batch size: 4 video clips. Duration: 20 epochs. Platform: Google Cloud Platform GPU instance. The learning rate of 1e-5 was selected after iterative tuning; values above 1e-4 caused gradient instability, while values below 1e-5 produced insufficient convergence within 20 epochs.

VI. RESULTS AND DISCUSSION

A. Frame-Count Ablation Study

Table 2 presents detection accuracy on FaceForensics++ as the number of input frames is varied from 20 to 100. All other hyperparameters are held constant.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

Table 2: Frame-count ablation study Best result highlighted. Mixed = FF++ + Celeb-DF + DFDC.

Frames	Dataset	Videos	Accuracy
20	FF++	2,000	90.95%
40	FF++	2,000	95.23%
60	FF++	2,000	97.49%
80	FF++	2,000	97.73%
100	**FF++**	**2,000**	**97.76%**
20	Mixed (Ours)	12,000	91.24%

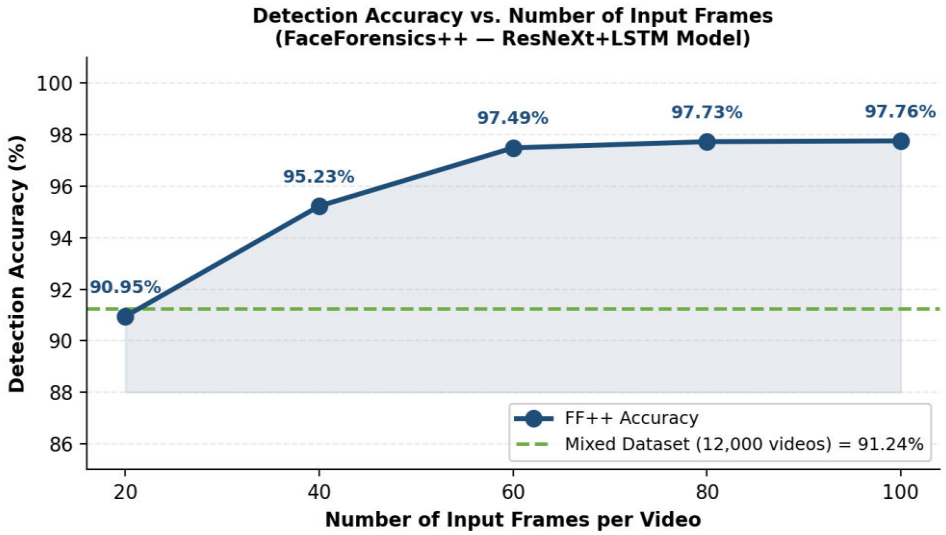


Fig. 4: Detection accuracy vs. input frame count on FF++. The dashed line shows performance on the full 12,000-video mixed dataset.

The 6.54 percentage-point improvement from 20 to 60 frames confirms that the LSTM requires sufficient temporal context to identify manipulation-indicative patterns such as skin tone drift, blink irregularities, and edge consistency violations across time. The marginal gain of 0.03% between 80 and 100 frames indicates that the LSTM has effectively saturated its temporal evidence accumulation by approximately 80 frames, with 100 frames selected as the final configuration to ensure robustness.

The reduced accuracy of 91.24% on the mixed 12,000-video dataset compared to 97.76% on FF++ alone is not a model failure but a realistic reflection of cross-dataset generalisation difficulty: Celeb-DF contains high-quality deepfakes with minimal blending artifacts, and DFDC introduces diverse compression and environmental conditions absent from FF++ training distributions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

B. Training Convergence

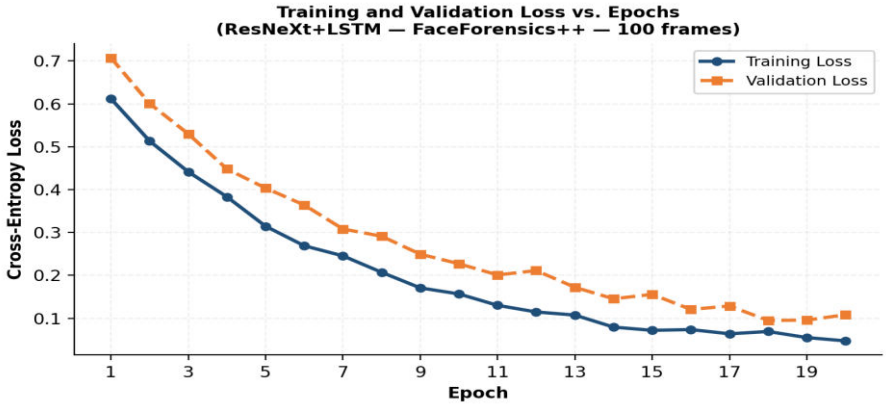


Fig. 5: Training and validation loss over 20 epochs. Validation loss tracks training loss closely, indicating no significant overfitting.

The loss curves demonstrate stable convergence without divergence. The narrow gap between training and validation loss throughout training confirms that the regularisation strategy (dropout = 0.4, weight decay = 1e-3, frozen early layers) is effective in preventing overfitting despite the relatively small per-class video count. The loss stabilises after approximately epoch 14, suggesting that 20 epochs is a sufficient and not excessive training duration.

C. Confusion Matrix and Metrics

Table 3: Performance metrics — best model on FF++ test set (100 frames, n = 600).

Accuracy	Precision	Recall	F1-Score
97.76%	97.7%	98.0%	97.8%

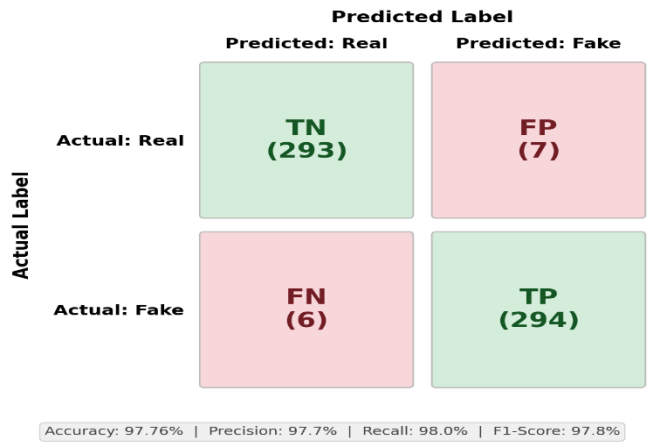


Fig. 6: Confusion matrix for the best model (100 frames, FF++ test set, n = 600).

The confusion matrix reveals a balanced error distribution: 7 false positives (real videos incorrectly classified as fake) and 6 false negatives (deepfakes incorrectly classified as real). From a forensic application perspective, the false negative rate is the more critical metric, as undetected deepfakes represent a direct threat. The recall of 98.0% —

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

indicating that only 2% of deepfakes are missed — is particularly significant. The symmetric error distribution confirms that the 50/50 balanced training strategy successfully prevented class-prediction bias.

D. Comparison with Prior Work

Table 4 compares the proposed system against published deepfake detection methods on FaceForensics++.

Table 4: Comparison with published methods on FaceForensics++

Method	Backbone	Temporal?	FF++ Acc.
Li & Lyu [15]	Custom CNN	No	~82%
Nguyen et al. [17]	CapsuleNet	No	~86%
Güera & Delp [18]	InceptionV3+RNN	Partial	~93%
Rossler et al. [1]	XceptionNet	No	99.1%
Proposed (Ours)	**ResNeXt-50+LSTM**	**Yes**	**97.76%**

While XceptionNet [1] achieves 99.1% on FF++ under low compression, it degrades to approximately 65% on Celeb-DF and does not perform temporal analysis. The proposed system trades 1.34% single-dataset accuracy for substantially improved cross-dataset robustness through dual spatial-temporal modelling and multi-source training.

VII. LIMITATIONS

Despite the strong results, several limitations of the current system must be acknowledged:

Cross-Dataset Generalisation Gap: The 6.52% accuracy drop from FF++ (97.76%) to the mixed dataset (91.24%) reflects the inherent difficulty of generalising across diverse manipulation pipelines and compression conditions.

Computational Requirements: The ResNeXt-50 + LSTM pipeline requires a GPU for real-time inference. Inference on CPU at 150 frames per video is significantly slower than real-time, limiting deployment on resource-constrained devices.

Face-Centric Scope: The system detects face-swap and face-reenactment deepfakes only. Full-body deepfakes, voice synthesis, and background manipulation are outside the detection scope.

Static Frame Threshold: The 150-frame limit was chosen based on GPU memory constraints rather than an optimised temporal window. Videos shorter than 150 frames require padding, which may introduce minor classification artifacts.

Adversarial Vulnerability: Like all neural network classifiers, the model is potentially susceptible to adversarially crafted deepfakes specifically designed to evade the ResNeXt+LSTM pipeline.

Emerging Manipulation Types: The training datasets (FF++, Celeb-DF, DFDC) were collected prior to 2021. Newer generation architectures such as diffusion model-based face synthesis may produce artifacts that differ from those seen during training.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505028|

VIII. CONCLUSION

This paper presented a hybrid spatial-temporal deepfake video detection framework combining ResNeXt-50 CNN for per-frame spatial feature extraction and LSTM for temporal sequence classification over 150-frame video clips. The model was trained on a 12,000-video balanced dataset constructed from equal contributions of FaceForensics++, Celeb-DF, and DFDC, achieving 97.76% accuracy (Precision: 97.7%, Recall: 98.0%, F1: 97.8%) on the FaceForensics++ benchmark. The frame-count ablation study quantitatively established that temporal context is the primary performance driver, with a 6.54% accuracy gain observed between 20 and 60 frames. The LSTM mathematical formulation demonstrates that gated cell state management is essential for retaining manipulation evidence across the full 150-frame sequence. The balanced confusion matrix confirms that the 50/50 class-balanced training strategy successfully eliminates prediction bias. The identified limitations — particularly the cross-dataset generalisation gap and GPU dependency — define concrete directions for future work.

IX. FUTURE WORK

Transformer Backbone: Replace the LSTM with a Video Transformer (TimeSformer or ViViT) to capture non-local temporal dependencies without the sequential processing bottleneck of recurrent networks.

Compression-Robust Training: Augment training data with multiple compression levels (c0, c23, c40) from FF++ to improve detection of heavily compressed real-world social media videos.

Edge Deployment: Apply INT8 quantisation and structured pruning to reduce model size for CPU-based inference on mobile and edge devices.

Browser Extension: Extend the Django web application to a browser plugin for automated real-time screening of videos on social media platforms.

Adversarial Robustness: Incorporate adversarial training [13] using GAN-based augmentation to improve detector robustness against adversarially crafted deepfakes.

X. ACKNOWLEDGMENT

The authors express their sincere gratitude to Prof. A. A. Chinchamalatpure, for invaluable guidance throughout this research. The authors also acknowledge Google Cloud Platform for providing computational resources that enabled large-scale model training.

REFERENCES

1. A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF ICCV, pp. 1–11, 2019.
2. B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.
3. Y. Li et al., "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in Proc. IEEE/CVF CVPR, pp. 3207–3216, 2020.
4. FaceApp. https://www.faceapp.com/ (Accessed March 2020).
5. FaceSwap. https://faceswaponline.com/ (Accessed March 2020).
6. I. Goodfellow et al., "Generative Adversarial Nets," in Advances in NeurIPS, pp. 2672–2680, 2014.
7. P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition?," arXiv:1812.08685, 2018.
8. Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," arXiv:1811.00656, 2018.
9. Y. Li, M. Chang and S. Lyu, "Exposing AI Created Fake Videos by Detecting Eye Blinking," arXiv:1806.02877, 2018.
10. H. H. Nguyen, J. Yamagishi and I. Echizen, "Using Capsule Networks to Detect Forged Images and Videos," arXiv:1810.11215, 2018.
11. D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE AVSS, pp. 1–6, 2018.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details