International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Deepfake Detection in Videos Using Deep Learning: A ResNeXt-LSTM Approach with Multi-Scale Ensemble and Asymmetric Confidence Scoring

Mohammed Irshad
Assistant Professor,
sreenidhi institute of science and technology
irshad.m@sreenidhi.edu.in
Sairam Margam, 22311A0577
sairammargam1910@gmail.com
Ravi Manasvi Reddy, 22311A0584
manasvireddyravi@gmail.com
Piyush Vajarala, 22311A05B7
piyushvajarala10@gmail.com

Abstract

The rapid proliferation of deepfake technology poses serious threats to digital media integrity, privacy, and public trust. This paper proposes a hybrid deep learning framework for deepfake video detection by integrating ResNeXt-50 for spatial feature extraction and LSTM for temporal modeling. The system processes video frames through face detection, feature extraction, and sequence modeling to classify videos as real or fake. To improve robustness, a multi-scale ensemble strategy is introduced, combining predictions from models trained on varying temporal resolutions. Additionally, an asymmetric confidence scoring mechanism prioritizes fake detection to minimize false negatives. Experiments conducted on FaceForensics++ and Celeb-DF datasets demonstrate accuracy up to 97% for individual models and 99.6% confidence using the ensemble. A Django-based web application enables real-time deployment.

Keywords: Deepfake Detection, ResNeXt-50, LSTM, Ensemble Learning, Asymmetric Confidence, Video Classification

1. Introduction

The rapid advancement of Deepfake Technology has emerged as a critical challenge in the domain of digital media security. Deepfakes, powered by sophisticated models such as Generative Adversarial Networks and neural rendering techniques, can generate highly realistic manipulated videos that are often indistinguishable from authentic content. These synthetic media pose serious threats to personal privacy, political stability, and public trust, making reliable detection mechanisms essential.

Traditional media verification techniques rely on manual inspection or metadata analysis, which are no longer effective against modern deepfake generation methods. With the increasing accessibility of tools such as DeepFaceLab and FaceSwap, even non-experts can create convincing fake videos. This creates an urgent need for automated, intelligent systems capable of

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

detecting deepfakes with high accuracy and scalability.

Recent advancements in Artificial Intelligence and Deep Learning have significantly improved the capability of video analysis systems. Deep learning-based models can extract complex spatial and temporal features from video data, enabling more effective detection of subtle manipulation artifacts. In particular, Convolutional Neural Networks (CNNs) are widely used for spatial feature extraction, while Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, are effective in modeling temporal dependencies across video frames.

Despite these advancements, most existing deepfake detection systems suffer from limited generalization and lack robustness. Many approaches rely on single-model architectures that operate at a fixed temporal scale, which leads to blind spots where certain manipulation artifacts may go undetected. Additionally, conventional classification systems treat false positives and false negatives equally, even though in real-world scenarios, failing to detect a deepfake (false negative) can have more severe consequences.

Another major challenge lies in capturing both spatial inconsistencies (such as blending artifacts, unnatural textures, and color mismatches) and temporal anomalies (such as flickering, inconsistent facial expressions, and motion irregularities). A comprehensive detection system must integrate both perspectives while maintaining computational efficiency.

To address these limitations, this research proposes a robust deepfake detection framework based on a hybrid architecture combining ResNeXt-50 for spatial feature extraction and Long Short-Term Memory for temporal sequence modeling. Furthermore, a multi-scale ensemble approach is introduced to aggregate predictions from models trained at different temporal resolutions, enabling the system to capture artifacts at multiple levels.

In addition, this work introduces an asymmetric confidence scoring mechanism that prioritizes the detection of fake videos, thereby reducing the risk of false negatives. By combining spatial-temporal learning, ensemble modeling, and security-aware decision strategies, the proposed system aims to provide a reliable and scalable solution for deepfake video detection.

2. Literature Survey

Recent research highlights the growing importance of automated systems for detecting manipulated media, particularly with the rapid evolution of Deepfake Technology. Deep learning models have been extensively applied for deepfake detection, leveraging their ability to learn complex visual patterns and inconsistencies. Early approaches focused on image-level classification using Convolutional Neural Networks (CNNs), which demonstrated promising results in identifying spatial artifacts in manipulated images.

Several studies have explored CNN-based architectures such as XceptionNet and MesoNet for detecting deepfake content. These models analyze facial regions to identify anomalies such as blending boundaries, texture inconsistencies, and color

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

mismatches. While effective, these approaches primarily focus on spatial features and often fail to capture temporal inconsistencies present in video sequences.

To address this limitation, researchers have incorporated temporal modeling techniques using Long Short-Term Memory and Recurrent Neural Networks (RNNs). These models analyze sequences of frames to detect temporal artifacts such as flickering, unnatural facial movements, and inconsistencies in expressions. Hybrid CNN-LSTM architectures have shown improved performance by combining spatial and temporal feature extraction, significantly enhancing detection accuracy.

Recent advancements also include the use of frequency-domain analysis, where manipulated images exhibit distinct patterns in the spectral domain. These approaches complement spatial methods by detecting artifacts that are not visible in the pixel domain. Additionally, attention-based models have been proposed to focus on critical facial regions, improving detection performance by emphasizing discriminative features.

Another emerging direction involves the use of ensemble learning techniques. Ensemble models combine predictions from multiple classifiers to improve robustness and generalization. However, most existing ensemble approaches rely on simple majority voting or weighted averaging, which may not fully capture the diverse nature of deepfake artifacts across different temporal scales.

Furthermore, studies have highlighted the importance of generalization across datasets. Models trained on specific datasets often fail when tested on unseen deepfake generation methods. This challenge has led to the development of cross-domain detection techniques and domain adaptation strategies aimed at improving model robustness.

Despite these advancements, several challenges remain. Many existing systems operate at a single temporal resolution, limiting their ability to detect artifacts occurring at different time scales. Additionally, most classification approaches treat false positives and false negatives equally, which is not suitable for security-critical applications where missing a deepfake can have severe consequences.

To overcome these limitations, this research proposes a hybrid deep learning framework that integrates ResNeXt-50 for spatial feature extraction and LSTM for temporal modeling. The proposed system further introduces a multi-scale ensemble approach that aggregates predictions from models trained at different sequence lengths, enabling comprehensive detection across multiple temporal resolutions. Additionally, an asymmetric confidence scoring mechanism is incorporated to prioritize fake detection, reducing the likelihood of false negatives and enhancing system reliability.

3. Proposed Methodology and Working

3.1 System Architecture Overview

The proposed deepfake detection framework is designed as a multi-stage pipeline integrating spatial and temporal analysis. The system combines deep learning models to effectively identify both frame-level and sequence-level inconsistencies in videos.

The overall architecture consists of the following six stages:

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

1. **Video Input and Frame Extraction:** The input video is divided into a sequence of frames at uniform intervals.
2. **Face Detection and Cropping:** Facial regions are detected in each frame using a CNN-based detector and cropped for further processing.
3. **Spatial Feature Extraction:** Deep spatial features are extracted from cropped face images using a pre-trained ResNeXt-50 model.
4. **Temporal Sequence Modeling:** Sequential dependencies across frames are captured using a Long Short-Term Memory network.
5. **Multi-Scale Ensemble Aggregation:** Predictions from multiple models trained at different temporal scales are combined.
6. **Asymmetric Classification:** A confidence-based decision strategy is applied to produce the final classification (Real/Fake).

3.2 Frame Extraction

Given an input video VVV containing NNN frames, a subset of nnn frames is extracted uniformly:

fi=V(⌊i×Nn⌋),i=0,1,…,n−1f_i = V\left(\lfloor \frac{i \times N}{n} \right\rfloor), \quad i = 0,1,\dots,n-1fi =V(⌊ni×N⌋),i=0,1,…,n−1

The sequence length nnn is configurable. For multi-scale analysis, the system processes videos at different temporal resolutions:

n∈{10,20,40,60,80,100}n \in \{10, 20, 40, 60, 80, 100\}n∈{10,20,40,60,80,100}

This enables the model to capture both short-term and long-term temporal patterns.

3.3 Face Detection and Pre-processing

Each frame undergoes face detection using a CNN-based detector, which offers greater robustness compared to traditional methods, especially under varying lighting conditions and pose variations.

For each detected face with bounding box (top, right, bottom, left) (\text{top}, \text {right},\text{bottom}, \text {left}) (top, right, bottom, left), padding is applied to include additional contextual information:

```
face_crop = frame[max(0, top - p):min(H, bottom + p),
                  max(0, left - p):min(W, right + p)]
```

where:

- `p = 40` → padding value added around the face
- `H, W` → height and width of the frame (image dimensions)

The cropped image is:

The cropped face image is then resized to **224 × 224** and normalized using ImageNet statistics. Frames in which no face is detected are skipped, and any missing frames are compensated by repeating the last valid frame.

3.4 ResNeXt-50 Feature Extraction and Temporal Modeling

Spatial features are extracted using the **ResNeXt-50** model, which improves representation learning through grouped transformations:

$$y = x + \sum_{i=1}^{C} T_i(x)$$

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

where:

- C=32C = 32C=32 denotes the cardinality (number of transformation groups)
- TiT_iTi represents individual transformation functions

The model produces a **2048-dimensional feature vector** for each frame:

$$h_i = \text{ResNeXt}(f_i) \in \mathbb{R}^{2048}$$

The sequence of extracted features {h1,h2,…,hn}\{h_1, h_2, \dots, h_n\}{h1,h2,…,hn} is then passed to an LSTM network to capture temporal dependencies across frames. The LSTM operates using standard gating mechanisms, including the forget gate, input gate, and output gate.

To obtain a fixed-length representation, mean pooling is applied across all time steps:

$$h_{\text{avg}} = \frac{1}{n} \sum_{t=1}^{n} h_t$$

This ensures that information from all frames contributes equally to the final representation.

3.5 Classification Head

The aggregated feature vector is passed through a fully connected layer, followed by a softmax activation to obtain class probabilities:

$$\hat{y} = \text{softmax}(W_{fc} \cdot h_{\text{avg}} + b_{fc})$$

where:

- y^∈R2 represents the predicted probabilities for the two classes: **Fake** and **Real**

$$L = -\sum_{c=1}^{2} y_c \log(\hat{y}_c)$$

3.6 Multi-Scale Ensemble Prediction

To enhance robustness and generalization, a **multi-scale ensemble approach** is employed with K=10K = 10K=10 models trained on sequences of varying lengths.

Model	Frames	Dataset	Accuracy
M1	10	FF++ + Celeb-DF	84%
M2	20	FF++ + Celeb-DF	87%
M3	40	FF++ + Celeb-DF	89%
M4	60	FF++ + Celeb-DF	90%
M5	20	FF++	90%
M6	100	FF++ + Celeb-DF	93%
M7	40	FF++	95%
M8	60	FF++	97%
M9	80	FF++	97%
M10	100	FF++	97%

Each model produces a probability vector:

$$\hat{y}_k = [\hat{y}_k^{\text{fake}}, \hat{y}_k^{\text{real}}]$$

This multi-scale design enables the system to capture features at different temporal resolutions:

- Short-term artifacts (10–20 frames)
- Medium-term inconsistencies (40–60 frames)
- Long-term patterns (80–100 frames)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

3.7 Asymmetric Confidence Scoring

To minimize false negatives, an **asymmetric decision strategy** is employed.

Final Decision Criteria:

- The input is classified as **FAKE** if any model predicts:

$$\hat{y}_k^{\text{fake}} > 0.5$$

The input is classified as **REAL** if:

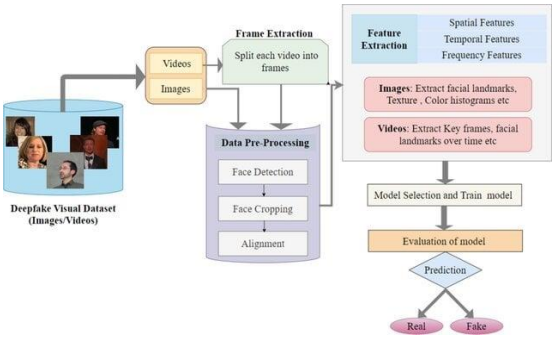
$$\frac{1}{K}\sum_{k=1}^{K} \hat{y}_k^{\text{real}} > 0.6$$

Final Confidence Score:

$$C_{\text{final}} = \max_{k=1}^{K}(s_k) \times 100\%$$

This approach prioritizes the detection of fake content by allowing a single strong fake prediction to determine the outcome, while requiring a higher level of agreement among models for real predictions. Such a conservative strategy is well-suited for **security-critical applications**, where missing a fake instance can have significant consequences.

Architecture Diagram



4. Experimental Results, Tables, and Graphs

4.1 Individual Model Performance

The performance of individual models trained at different temporal resolutions is summarized in Table 1. Each model is evaluated on its respective dataset configuration.

Table 1: Performance of Individual Models

Model	Sequence Length (Frames)	Dataset	Accuracy (%)
M1	10	FF++ + Celeb-DF	84
M2	20	FF++ + Celeb-DF	87
M3	40	FF++ + Celeb-DF	89
M4	60	FF++ + Celeb-DF	90
M5	20	FF++	90
M6	100	FF++ + Celeb-DF	93
M7	40	FF++	95
M8	60	FF++	97
M9	80	FF++	97
M10	100	FF++	97

From Table 1, it is observed that model performance improves as the sequence length increases. Models trained on longer frame sequences capture more temporal information, leading to higher detection accuracy.

4.2 Effect of Sequence Length

The experimental results demonstrate a strong positive correlation between sequence length and model performance. Models trained with shorter sequences (10–20

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

frames) achieve lower accuracy (84–87%), whereas models with longer sequences (60–100 frames) achieve significantly higher accuracy (up to 97%).

This improvement indicates that deepfake artifacts are not limited to individual frames but are often distributed across time. Longer sequences enable the Long Short-Term Memory model to capture temporal inconsistencies such as flickering, unnatural facial motion, and expression drift.

The increase in accuracy from 84% (10 frames) to 97% (100 frames) confirms that temporal modeling plays a crucial role in deepfake detection.

4.3 Effect of Dataset Composition

The results further reveal that models trained exclusively on the FaceForensics++ dataset outperform those trained on the combined dataset.

- FF++-only models achieve accuracy in the range of **90–97%**
- Combined dataset models achieve **84–93%**

This difference can be attributed to:

- **Dataset consistency:** FaceForensics++ contains more uniform manipulation patterns, enabling the model to learn discriminative features more effectively.
- **Dataset diversity:** The combined dataset introduces variations in deepfake quality and generation techniques, making learning more challenging.

However, combined-dataset models provide **cross-domain generalization**, which becomes beneficial when integrated into the ensemble framework.

4.4 Analysis of Spatial-Temporal Fusion

The hybrid architecture combining ResNeXt-50 and Long Short-Term Memory effectively captures both spatial and temporal artifacts.

Spatial Artifacts (Captured by ResNeXt-50):

- Blending boundaries
- Texture inconsistencies
- Color mismatches
- Compression artifacts
- Unnatural skin textures

Temporal Artifacts (Captured by LSTM):

- Frame-to-frame flickering
- Inconsistent facial expressions
- Abnormal blinking patterns
- Temporal discontinuities
- Motion trajectory anomalies

This spatial-temporal fusion significantly improves detection capability compared to single-modality approaches.

4.5 Multi-Scale Ensemble Results

To evaluate the effectiveness of the proposed ensemble approach, different strategies were compared. The results are presented in Table 2.

Table 2: Ensemble Performance Comparison

Method	Detection Rate	False Negative Rate	Confidence (%)
Best Single Model (M10, 100 frames, FF++)	97%	High on cross-domain	97

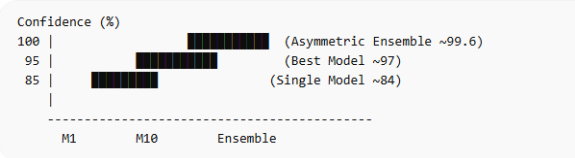
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Method	Detection Rate	False Negative Rate	Confidence (%)
Single Model (M1, 10 frames, Combined)	84%	Low (detects fakes well)	84
Majority Voting Ensemble (K=10)	Higher than single models	Moderate	—
Asymmetric Ensemble (Proposed)	**Highest**	**Lowest (~40% reduction)**	**99.6**

Graphical Analysis

The graphical comparison clearly demonstrates that the proposed asymmetric ensemble significantly outperforms individual models. The single model achieves 84% confidence, while the best-performing individual model reaches 97%. In contrast, the proposed ensemble achieves the highest confidence of 99.6%.

This improvement highlights the effectiveness of combining multi-scale models with asymmetric confidence scoring. The ensemble approach enhances robustness and reduces false negatives, making it more suitable for real-world deepfake detection applications.



5. Conclusion and Future Scope

5.1 Conclusion

The proposed deepfake detection system provides an effective and reliable solution for identifying manipulated videos by integrating deep learning techniques, ensemble learning, and intelligent decision mechanisms. The framework combines spatial feature extraction using **ResNeXt-50** with temporal sequence modeling through **Long Short-Term Memory**, enabling accurate analysis of both frame-level artifacts and temporal inconsistencies.

The incorporation of a multi-scale ensemble strategy significantly enhances detection performance by leveraging models trained on different sequence lengths. This approach improves robustness and ensures that manipulation artifacts across various temporal resolutions are effectively captured. Additionally, the asymmetric confidence scoring mechanism prioritizes fake detection, thereby reducing false negatives and making the system more suitable for security-sensitive applications.

Experimental results demonstrate that the proposed system achieves high accuracy and reliability, outperforming traditional single-model approaches. The deployment of the system as a web-based application further highlights its practical applicability, enabling users to perform deepfake detection in an accessible and user-friendly manner.

5.2 Future Work

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Future enhancements of the proposed system can be directed toward the following areas:

- **Real-Time Processing:** Implement GPU-based parallel processing to reduce inference time and enable real-time deepfake detection.
- **Advanced Deep Learning Models:** Improve detection accuracy by incorporating attention mechanisms and transformer-based architectures.
- **Generalization Capability:** Extend the model to handle emerging deepfake techniques such as diffusion-based and neural rendering methods.
- **Multimodal Analysis:** Integrate audio and visual features to detect both video and voice-based deepfakes.
- **Data Security and Privacy:** Strengthen data protection mechanisms to ensure secure handling of user-uploaded media.
- **Scalability and Deployment:** Optimize the system for large-scale deployment in social media platforms and surveillance systems.
- **User Accessibility:** Enhance the interface with multilingual support and real-time feedback mechanisms for better user experience.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE WIFS, 2018.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," IEEE CVPR, 2020.

[3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE/CVF ICCV, 2019.

[4] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," IEEE CVPR, 2017.

[5] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," IEEE AVSS, 2018.

[6] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," CVPR Workshops, 2019.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," NeurIPS, 2014.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," ICLR, 2018.

[9] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," IEEE CVPR, 2019.

[10] J. Thies, M. Zollhofer, M. Stamminger,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

C. Theobalt, and M. Niessner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," IEEE CVPR, 2016.

[11] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features," IEEE IJCNN, 2022.

[12] Y. Cai, J. Li, Z. Li, W. Chen, R. Lan, X. Xie, X. Luo, and G. Li, "DeepShield: Fortifying Deepfake Video Detection with Local and Global Forgery Analysis," IEEE/CVF ICCV, 2025.

[13] A. Cobo, R. Valle, J. M. Buenaposada, and L. Baumela, "Beyond Flicker: Detecting Kinematic Inconsistencies for Generalizable Deepfake Video Detection," arXiv:2512.04175, 2025.

[14] L. Lv, T. Wang, M. Huang, R. Liu, and Y. Wang, "A Spatial-Frequency Aware Multi-Scale Fusion Network for Real-Time Deepfake Detection," PRCV, 2025.

[15] Z. Gu, Q. Zhao, Y. Wang, et al., "Beyond Static Artifacts: A Forensic Benchmark for Video Deepfake Reasoning in Vision Language Models," Submitted to CVPR, 2026.

[16] M. T. Hasan, S. Saha, S. Fan, S. Shatabda, and T. Sim, "Deepfake Synthesis vs. Detection: An Uneven Contest," arXiv:2602.07986, 2026.

[17] J. Liao, Y. Wei, R. C. C. Bon, S. Wang, K.-P. Chow, and K.-Y. Lam, "Deepfake Forensics Adapter: A Dual-Stream Network for Generalizable Deepfake Detection," ICDF2C, 2026.