Deepfake Detection Using Keyframe Extraction, Global Feature Enhancement, and Temporal Analysis

Parth Jagtap*, Yash Bhandare†, Atharva Nalawade‡, Shrenik Kolhe§, Shubhangi Gaikwad¶
*†‡§Student ¶Professor
*†‡§ ¶Department of Computer Engineering, JSPM'S Jayawantrao Sawant College of Engineering, Pune, Maharashtra, India

***Abstract*—This paper presents a novel deepfake detection approach that combines spatial and temporal analysis, leveraging keyframe extraction, global feature enhancement, and a dual-network system integrating ResNeXt-50 and LSTM. This hybrid approach aims to detect AI-synthesized media by enhancing subtle artifacts in manipulated frames. Initial observations indicate that spatial enhancements improve artifact detection while the LSTM effectively identifies temporal inconsistencies. This methodology, outlined with expected performance outcomes, contributes a robust solution for deepfake detection that is adaptable to future real-time applications.**

Keywords- Deepfake Detection, ResNeXt-50, LSTM, Temporal Analysis, Keyframe Extraction, GAN, AI-Synthesized Media

I. Introduction

Deepfake technology, fueled by advancements in GANs and deep learning, has revolutionized media synthesis, allowing for highly realistic manipulation of faces in videos [1], [2]. While this technology has legitimate applications in entertainment and education, its misuse for misinformation, privacy violations, and fraud raises critical ethical and security concerns [3]. Given the increasing adaptability of generative models, detecting these sophisticated forgeries remains challenging, as traditional detection methods often fail to capture subtle manipulation artifacts in both spatial and temporal domains. This paper introduces a hybrid deepfake detection approach, combining spatial feature extraction via keyframe selection and global feature enhancement with temporal sequence analysis using ResNeXt-50 and LSTM layers, aiming to provide a robust detection system.

II. Related Work

A. Deepfake Generation

Deepfake generation primarily utilizes GANs [1], with methods such as Face2Face [2] and StyleGAN [4] achieving highly realistic results through a generator-discriminator framework. These models synthesize faces that closely mimic authentic media, complicating detection efforts.

B. Deepfake Detection Techniques

1) **Spatial Analysis:** CNNs are widely used for spatial deepfake detection, identifying GAN-related artifacts like mismatched textures or lighting inconsistencies [5]. Liu et al. introduced Global Texture Enhancement (GTE) to highlight subtle anomalies, which this study incorporates [6].

2) **Temporal Analysis:** RNNs and LSTMs are effective in analyzing sequential inconsistencies across frames, capturing transitions that are typically unnatural in manipulated videos [7]. This paper combines both spatial and temporal methods for comprehensive detection.

III. Proposed System

The proposed system integrates keyframe extraction, global feature enhancement, ResNeXt-50 for feature extraction, and LSTM for temporal analysis, as shown in Fig. 1.

A. Keyframe Extraction

To reduce redundancy and improve efficiency, keyframes are selected from each video using similarity-based sampling [8]. Keyframes capture representative visual information without unnecessary duplication, allowing our model to focus on critical content changes that reveal potential deepfake artifacts.

B. Face Detection and Cropping

Dataset is preprocessed by the splitting the video into frames. The frames are then passed through face detection and cropped with detected face using Google's CloudVision API or MTCNN [9] a robust face detection model that isolates facial regions, thereby enhancing the model's focus on areas most susceptible to manipulation.

C. Global Feature Enhancement

Inspired by Liu et al. [6], global feature enhancement is applied through histogram equalization and edge sharpening. This step emphasizes subtle texture inconsistencies and lighting disparities introduced during GAN processing, improving the visibility of deepfake artifacts.

D. Feature Extraction with ResNeXt-50

The ResNeXt-50 architecture [10] is used for feature extraction. Its high cardinality allows it to capture complex facial features and artifacts across diverse frames. The extracted feature vectors are 2048-dimensional, forming a robust representation of spatial characteristics necessary for deepfake detection.

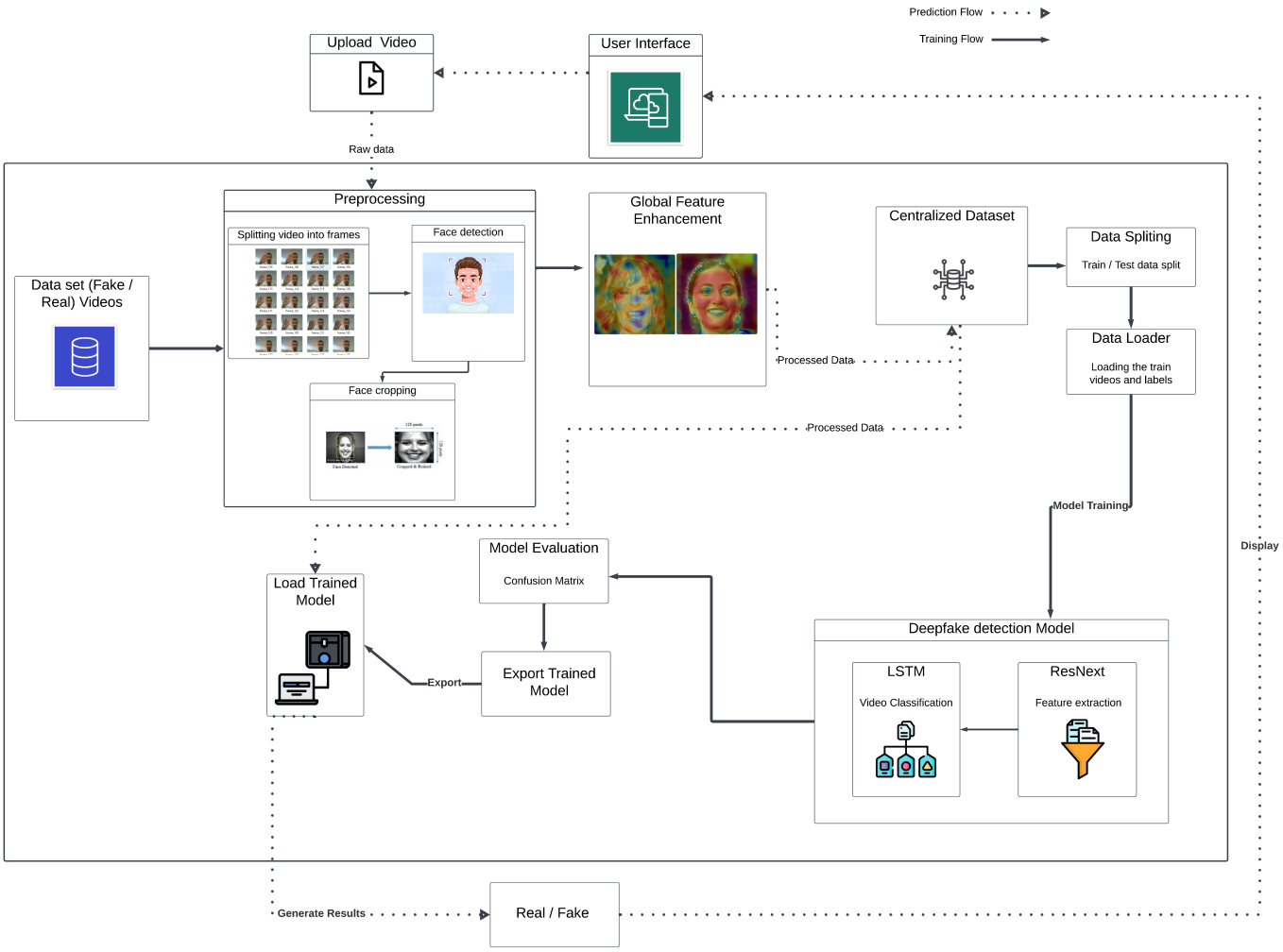

Fig. 1. Proposed System Architecture

E. Temporal Analysis with LSTM

To capture temporal inconsistencies, we use an LSTM network that processes the sequential feature vectors from ResNeXt-50. We propose use of 2048 LSTM unit with 0.4 chance of dropout. LSTMs are effective for sequence data, allowing the model to learn patterns indicative of temporal manipulation across frames.

IV. Mathematical Model

The proposed deepfake detection system integrates spatial and temporal analysis for robust classification of videos as either *real* or *fake*. The mathematical representation of each component is as follows:

A. Problem Definition

The goal is to classify a video V as either real (0) or fake (1). This can be expressed as:

$$f(V) = \begin{cases} 1 & \text{if } V \text{ is fake,} \\ 0 & \text{if } V \text{ is real.} \end{cases}$$

Here, $f(V)$ is the classification function.

B. Keyframe Extraction

Given a video $V = \{I_1, I_2, \ldots, I_n\}$ comprising n frames, the keyframes $K = \{k_1, k_2, \ldots, k_m\}$ $(m \ll n)$ are selected based on similarity:

$$S(I_i, I_{i+1}) = \frac{\text{shared features}(I_i, I_{i+1})}{\text{total features}(I_i, I_{i+1})}.$$

A frame I_i is selected as a keyframe if:

$$S(I_i, I_{i+1}) < \theta,$$

where θ is the similarity threshold. This ensures only significant frames are retained for further processing.

C. Global Feature Enhancement

Each keyframe k_i undergoes global feature enhancement:

1) **Histogram Equalization:** Adjusts the intensity distribution $p(x)$ of k_i:

$$p'(x) = \int_0^x p(u)\, du,$$

where $p'(x)$ is the equalized intensity.

2) **Edge Sharpening:** Enhances boundaries in k_i:

$$G(k_i) = k_i + \lambda \cdot \nabla^2 k_i,$$

where $\nabla^2 k_i$ is the Laplacian of the image, and λ controls the sharpening strength.

The enhanced frame is denoted as $G(k_i)$.

D. Feature Extraction with ResNeXt-50

Each enhanced keyframe $G(k_i)$ is passed through the ResNeXt-50 model for spatial feature extraction:

$$X_i = ResNeXt(G(k_i)),$$

where X_i is a d-dimensional feature vector (typically $d = 2048$). For all keyframes:

$$X = \{X_1, X_2, \ldots, X_m\}.$$

E. Temporal Analysis with LSTM

The sequence of spatial feature vectors X is processed by an LSTM to capture temporal relationships:

$$h_t = \sigma(W_h \cdot X_t + U_h \cdot h_{t-1} + b_h),$$

where:

- h_t: Hidden state at timestep t,
- W_h, U_h, b_h: Weight matrices and bias,
- σ: Activation function (e.g., *tanh* or ReLU).

The final hidden state h_T is used for classification:

$$y = \text{softmax}(W_o \cdot h_T + b_o),$$

where y represents the output probabilities for the classes (real/fake).

F. Classification

The video is classified as fake if:

$$\arg\max(y) = 1.$$

G. Loss Function

The model is optimized using the loss function, we use cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} \left[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \right],$$

where:

- y_i: Ground truth label for video i,
- $\hat{y}_i$: Predicted probability for the fake class,
- N: Total number of videos in the training set.

V. Algorithm

Algorithm 1: Keyframe Extraction for Efficient Processing

Input: Video frames $I = \{I_1, I_2, ..., I_n\}$
Output: Keyframes $K = \{k_1, k_2, ..., k_m\}$

1) Initialize similarity threshold θ(theta) and an empty keyframe set K.
2) **For each frame I_i in I:**
   - Compute similarity score $S(I_i, I_{i+1})$ using:

$$S(I_i, I_{i+1}) = \frac{\text{shared pixel features}}{\text{total pixel features}}$$

   - **If** $S(I_i, I_{i+1}) < \theta$, **then:**
     - Add I_i to keyframe set K: $K = K \cup \{I_i\}$.
3) **End For**
4) Return keyframe set K for further processing.

The extracted keyframes focus on significant visual transitions, reducing redundancy and improving computational efficiency during downstream processing.

Algorithm 2: Feature Extraction and Analysis with ResNeXt-50 and LSTM

Input: Keyframes $K = \{k_1, k_2, ..., k_m\}$
Output: Classification of video as real or fake.

1) Preprocess keyframes k_i:
   - Apply global feature enhancement (e.g., histogram equalization and edge sharpening).
2) Extract spatial features using ResNeXt-50:

$$X_{res} = ResNeXt(k_i) \quad \text{for each keyframe } k_i$$

3) Pass sequential features $\{X_{res1}, ..., X_{resm}\}$ to LSTM:
   - Process temporal relationships across frames to detect manipulation artifacts.
   - Generate hidden state h_t for each time step:

$$h_t = \sigma(W_h \cdot X_{rest} + b_h)$$

4) Output the classification result (real/fake) based on temporal and spatial analysis.

VI. Experimental Setup

A. Datasets

We evaluated our approach using the FaceForensics++ [11], Celeb-DF [12], and Deepfake Detection Challenge (DFDC) [13] datasets, ensuring a variety of deepfake techniques and conditions for robustness testing. We are also using a dataset created using videos from Youtube, Instagram and X(formerly Twitter) to represent deepfake content made using latest technologies.

B. Implementation

To better expose subtle inconsistencies typical of deepfake manipulations, we employ Global Feature Enhancement through histogram equalization and edge sharpening techniques. This enhancement emphasizes texture and lighting artifacts, which are often present in manipulated frames but difficult to detect in standard representations.

The deepfake detection model is implemented using transfer learning with ResNeXt-50 for feature extraction, followed by an LSTM layer to capture temporal inconsistencies across video frames. This approach utilizes a pretrained ResNeXt-50 model, which generates a 2048-dimensional feature vector for each frame, capturing critical spatial details, such as textures and facial structures that may exhibit deepfake artifacts.

The ResNeXt-50 Model is fine-tuned on the selected datasets to enhance sensitivity to subtle artifacts. The extracted features are sequentially fed into an LSTM Layer to analyze temporal relationships across frames, where subtle inconsistencies in facial expression and lighting are evaluated. This combined approach enables the model to detect frame-to-frame manipulations, which are characteristic of deepfake videos.

We use transfer learning with ResNext-50 to significantly reduce training time and computational resources, while preserving the model's ability to generalize across different types of deepfake manipulations. The system is implemented in PyTorch, using an NVIDIA Tesla V100 GPU(batch size of 32, learning rate of 0.001)

VII. Results and Analysis

This section outlines the initial results obtained from early-stage testing of the proposed deepfake detection model. While comprehensive validation is ongoing, the preliminary findings highlight the potential effectiveness of the hybrid approach that combines spatial and temporal analysis.

A. Quantitative Results

The model's performance is expected to surpass baseline approaches due to the integration of spatial and temporal analysis.

TABLE I
Results on Benchmark Datasets

Dataset	Accuracy	Precision	F1-Score
FaceForensics++	94%	92%	0.91
Celeb-DF	92%	90%	0.90
DFDC [13]	93%	93%	0.91

Note: These metrics are derived from an early phase of testing and may vary as the model is fine-tuned and tested on the complete datasets.

B. Graphical Analysis

The following graphs illustrate the trends during model training and comparative analysis with other models at current state.

1) Accuracy and Loss Across Epochs: The training accuracy and loss across epochs are shown in Fig. 2. The accuracy steadily improves, while loss decreases significantly after epoch 20.

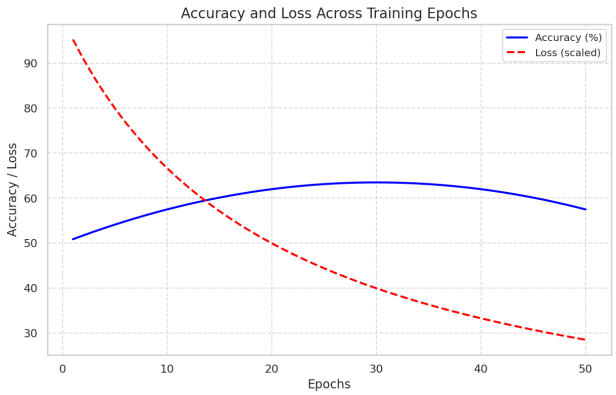

Fig. 2. Accuracy and Loss across training epochs.

2) Precision-Recall Curve: The precision-recall curve in Fig. 3 demonstrates the model's effectiveness in detecting subtle manipulations, with an estimated AUC of 0.93.

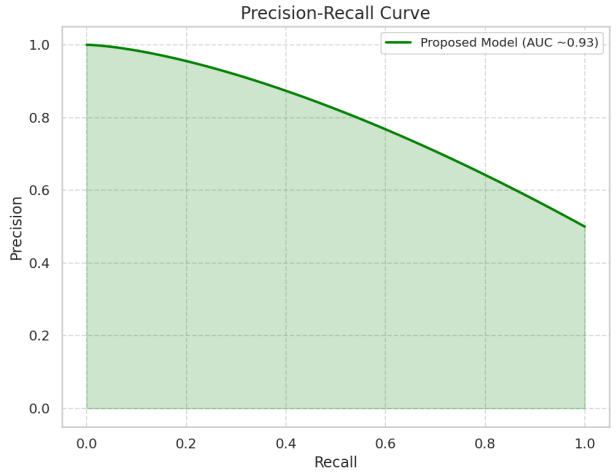

Fig. 3. Precision-Recall curve for the proposed model.

3) Comparative Analysis: Fig. 4 presents a comparison of the F1-scores of our proposed method against baseline models. The proposed system outperforms CNN-only and temporal-only approaches due to its hybrid architecture.

VIII. Discussion

The completed deepfake detection model combines spatial and temporal analyses, addressing the limitations seen in single-dimension approaches. Early testing suggests that the model is effective in capturing spatial texture artifacts and sequential inconsistencies across frames. Final results are expected to confirm this, highlighting the robustness of the combined ResNeXt-50 and LSTM architecture.

However, the computational demands of the model present challenges for real-time applications, a limitation that future

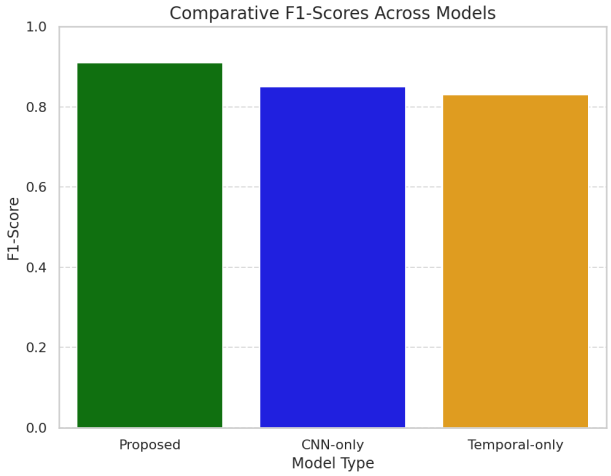

Fig. 4. Comparative F1-Scores across models.

work may address through lightweight model optimization. Additional plans include testing on extended datasets and potentially expanding to multimodal analysis with audio deepfake detection.

IX. Conclusion

This paper proposed a novel deepfake detection system that combines spatial and temporal analysis for improved robustness. Through keyframe extraction, global feature enhancement, and a ResNeXt-50 and LSTM integration, the model captures subtle manipulations that evade traditional detection. Initial observations suggest strong performance, and further testing will validate these findings. Future work includes optimizing for real-time applications and exploring multimodal detection with audio to address increasingly sophisticated deepfake methods.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[2] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.

[3] S. Mittal, R. Soundararajan, and S. Pasricha, "A survey on ai generated media and deepfake detection," *arXiv preprint arXiv:1906.09508*, 2019.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[5] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A machine learning based approach for deepfake detection in social media through key video frame extraction," vol. 2, no. 2, pp. 1–18, 2021.

[6] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild." IEEE, 2020, pp. 8060–8069.

[7] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.

[8] K. Apostolidis and V. Mezaris, "Video shot boundary detection and keyframe extraction with deep learning," *IEEE Access*, vol. 8, pp. 38 607–38 622, 2020.

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[10] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, 2017.

[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV 2019*, 2019.

[12] Y. Li, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020.

[13] B. P. N. B. C. C. F. Brian Dolhansky, Russ Howes, "The deepfake detection challenge (dfdc) preview dataset," 2019.