WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DEEPFAKE DETECTION SYSTEM USING RESNET, LSTM AND CAPSNET

Ranjeet Maurya*[*1], Rohit Jaiswal*[*2], Priyanka Kumari*[*3], Anya Rajpoot*[*4], Mr. Kapil Verma*[*5]

*[1,2,3,4,5]Babu Babarsi Das Northern India Institute Of Technology, Lucknow, India.

ABSTRACT :

Deepfake technology, derived from deep learning methods, enables the creation of highly realistic yet synthetic media by altering an individual's appearance or voice in digital content. Initially developed for research, entertainment, and industrial applications, deepfakes have become increasingly convincing and accessible, raising serious ethical and security concerns. Generative Adversarial Networks (GANs) are among the most widely used techniques for producing such content. They leverage dual-model architecture to generate and refine fake imagery. Advancements in deepfake generation have made identifying manipulated media progressively more challenging. This research proposes a deepfake detection framework combining Convolutional Neural Networks (CNN), Capsule Networks (CapsNet), and Long Short-Term Memory (LSTM) models. The goal is to accurately identify manipulated video frames and enhance the robustness of deepfake detection systems.

INTRODUCTION

Over recent years, the line dividing real and artificially generated digital content has grown increasingly blurred. Breakthroughs in AI, machine learning, and computer graphics now allow the creation of ultra-realistic images and videos that can deceive even careful observers. While these advances have propelled innovation in entertainment and virtual environments, they have also introduced significant risks, particularly with the rise of deepfake technologies. Deepfakes employ complex neural network models, including GANs and autoencoders, to convincingly substitute one person's facial features or voice with another's in media content. Although originally developed for legitimate research and artistic expression, these techniques are frequently misused to produce deceptive material. Such misuse threatens personal privacy and societal trust. Prominent figures—ranging from public officials to celebrities—have found themselves subjects of fabricated media crafted to mislead, malign, or manipulate public opinion. In some cases, deepfakes have been weaponized for harmful purposes such as blackmail, revenge pornography, and disinformation campaigns, eroding confidence in digital communication channels. As deepfake generation methods grow more advanced, detecting forged content with traditional approaches becomes increasingly challenging. This underscores the critical need for robust, automated detection systems. This study introduces a hybrid model combining CNNs, Capsule Networks, and LSTMs to improve detection precision at the frame level, aiming to bolster the defense against the spread of synthetic media and protect the integrity of digital information.

LITERATURE REVIEW

The proliferation of deepfake videos has created significant concern due to their potential to mislead the public, erode trust in digital content, and cause harm in legal, political, and social contexts. Detecting these manipulations has become an urgent research focus, with several techniques emerging in recent years:

Face Warping Artifact Detection [1]:

This technique identifies inconsistencies introduced during face synthesis by analyzing discrepancies between the generated face regions and their adjacent areas. A specialized Convolutional Neural Network (CNN) is employed to identify these artifacts. The core insight lies in the resolution limitations of many deepfake generation algorithms, which often require the synthesized face to be resized and adjusted, leading to detectable distortions.

Eye Blink Detection [2]:

This method targets the unnatural blinking patterns in fake videos. Since many deepfake generators do not simulate eye movement accurately, a lack of blinking becomes a useful cue for detection. Although promising, this technique is limited, as it focuses solely on eye movement without accounting for other facial features such as wrinkles or lip movement.

Capsule Networks for Forgery Detection [3]:

Capsule networks have been explored for their ability to preserve spatial hierarchies, making them suitable for detecting facial manipulations. These models show potential in differentiating authentic and tampered videos, especially in scenarios involving synthetic content or replay attacks. However, using random noise during training can reduce effectiveness in real-world applications where data variability is high.

Biological Signal Analysis [4]:

This approach leverages biological cues—like pulse signals captured from facial skin regions—to detect manipulated videos. Features extracted from these signals are used to train models such as SVMs and CNNs. While it offers high accuracy, the method's complexity and sensitivity to video quality can affect performance.

FakeCatcher System:

Designed to detect deepfakes across various content types and qualities, FakeCatcher works independently of the generating model or resolution. Despite its accuracy, challenges persist in optimizing the signal-based detection pipeline, especially in designing differentiable loss functions aligned with signal processing logic.

Research Problem

Advancements in artificial intelligence have made it increasingly easy to fabricate highly realistic fake videos, often indistinguishable from real footage. While deepfake technologies were initially developed for positive use cases, such as film dubbing, character de-aging, and accessibility, they have since been misused in harmful ways.

One prominent example includes the misuse of deepfake tools to superimpose celebrity faces onto pornographic content, primarily for malicious intent. According to a Deep trace report, around 96% of all deepfake videos online are pornographic. These technologies have also been used in political misinformation campaigns. Notable cases include the manipulated video of President Barack Obama, where another actor's voice was used to synthesize a speech, and a fabricated clip of Belgium's Prime Minister during the COVID-19 pandemic, aimed at pushing environmental agendas. Such incidents demonstrate the growing threat deepfakes pose to political stability and public perception.

As deepfake generation methods become more advanced—using face-swapping, reenactment, and synthetic voice generation—the line between authentic and fabricated content continues to blur. Traditional CNN-based models struggle to detect these manipulations reliably due to information loss during pooling operations and difficulties handling rotated or misaligned faces. Additionally, their performance degrades with lower-quality or transformed video inputs.

These limitations necessitate the development of improved detection mechanisms that can both identify deepfakes accurately and explain their decision-making process. Our research seeks to address these challenges by introducing a hybrid model capable of robust and interpretable deepfake detection.

Example of deepfake images frame from video

Research Objectives

The primary aim of this research is to develop an advanced deepfake detection model that not only identifies manipulated videos but also explains the nature of the manipulations. Instead of merely classifying a video as fake, the model will identify specific facial regions that have been altered, enhancing the interpretability and trustworthiness of the system.

- To achieve this aim, the following specific objectives have been defined:
- To review existing techniques:
- Investigate the current methods used for both generating and detecting deepfakes, highlighting their strengths and limitations.
- To design an effective detection model:
- Develop a hybrid machine learning model capable of accurately identifying deepfakes using standard datasets.
- To improve model robustness:
- Ensure that the detection system performs reliably even under challenging conditions such as low video quality, facial movement, or varying head poses.
- To implement explainable AI features:
- Integrate interpretability into the model so that it can provide clear and understandable justifications for its decisions, making the detection process transparent.

In summary, this research seeks to contribute to the development of a reliable, explainable, and generalizable deepfake detection system that can effectively address real-world challenges..

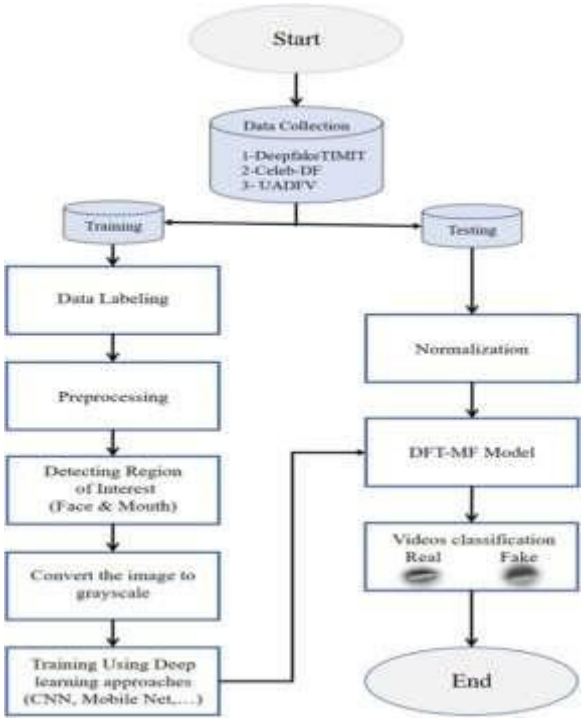
METHODOLOGY

The detection of deepfake videos necessitates the use of advanced deep-learning frameworks due to the intricacies involved in identifying manipulated facial patterns. Our approach centers around constructing a deep neural network model that combines the strengths of traditional Convolutional Neural Networks (CNNs), Capsule Networks, and Long Short-Term Memory (LSTM) units. While CNNs are widely used for feature extraction, their reliance on pooling layers can result in the loss of important spatial details. To overcome this limitation, we incorporate Capsule Networks, which are more adept at preserving spatial hierarchies and detecting subtle alterations in facial structure and orientation.

Since videos consist of sequential frames, integrating temporal analysis is essential. To this end, we utilize LSTM networks to model the temporal relationships across frames, capturing motion-based irregularities that often indicate tampering. Moreover, to enhance the trustworthiness and interpretability of our model, we integrate explainable AI techniques. This allows us to highlight and visualize the specific facial regions that influenced the model's decisions, thereby increasing transparency in the classification process.

The proposed methodology follows these steps:

- **Dataset Acquisition**: Collection of real and manipulated video datasets for training and evaluation.
- **Data Preprocessing**: Frame extraction, resizing, normalization, and augmentation to ensure consistency and improve model generalization.
- **Annotation and Labeling**: Manual or automated tagging of video frames as authentic or fake to prepare supervised learning datasets.
- **Data Partitioning**: Splitting the dataset into training, validation, and testing sets to facilitate unbiased evaluation.
- **Model Development**: Construction of the core architecture using Capsule Networks paired with LSTM to jointly model spatial and temporal features.
- **Hybrid Framework Integration**: Incorporation of pre-trained CNN backbones with the Capsule-LSTM network to enhance feature extraction capabilities.
- **Model Evaluation**: Performance testing of the hybrid model using accuracy, precision, recall, and F1-score metrics.
- **Benchmarking**: Comparative analysis between the proposed hybrid model and state-of-the-art deepfake detection approaches.
- This layered methodology ensures that both spatial inconsistencies and temporal anomalies are addressed, offering a more holistic solution to the complex challenge of deepfake detection.



Dataset and Preprocessing

DFDC Dataset Overview- For this study, we selected the *Deepfake Detection Challenge (DFDC)* dataset, which is among the most comprehensive datasets publicly available for deepfake detection. The DFDC dataset comprises over *100,000 video clips*, sourced from *3,426 paid actors*, encompassing both real and manipulated videos. Each manipulated clip represents a unique instance of source-target identity swap, created using a diverse set of *GAN-based*, *auto encoder-based*, and *heuristic face-swapping techniques.*

The dataset is divided into two versions:
Preview Dataset: Contains 5,000 videos and two facial manipulation algorithms.
Full Dataset: Contains approximately 124,000 videos and eight facial manipulation techniques.
For this work, we utilized the *Full Dataset*, leveraging its scale and algorithmic diversity to train and evaluate our hybrid deepfake detection model.
Data Preprocessing
Preprocessing is a critical step to ensure that the data conforms to the input requirements of deep learning architectures. The DFDC dataset includes a JSON metadata file in each data split, containing real/fake labels for each video.
Frame Extraction and Face Cropping
We extracted *1 frame per second* from each video using the *OpenCV (CV2)* library.
To train on both holistic and localized facial features, we created *two datasets*:
Full-frame dataset: Original frames as captured.
Cropped-face dataset: Using *DLIB* and *Deep Face* libraries, facial regions were extracted. Due to resolution inconsistencies in cropping, both frame types were retained for robustness.
Each frame was resized to *128 × 128* with 3 color channels (RGB). A temporal window of *5 consecutive frames* was used per input instance, resulting in the final input shape of *(5, 128, 128, and 3)*
Data Labeling.
Each segment of the DFDC dataset includes a metadata. Json file that annotates the videos with their respective labels: REAL or FAKE. These labels were parsed and associated with the extracted video frames.
Data Splitting
The entire dataset comprising *4,648 videos* was split using an *80:20* ratio for training and testing. Additionally, *20% of the training data* was reserved for validation. Each video contributed an average of *5 frames*, resulting in a total of *approximately 24,740 frames* for model training, validation, and testing.

Table 1: Data Preprocessing

Data Segment	Parameters	Value
Training	Rescale	1./255
	Validation Split	0.8
	Horizontal Flip	False
	Vertical Flip	False
Validation	Rescale	1./255
	Validation Split	0.2
Testing	Rescale	1./255

Table 2: Data Distribution

Segment	Percentage	Videos	Target Size	Batch Size	Class Mode
Training	80%	3,658	(5, 128, 128, 3)	4	Categorical
Validation	20% of training	762	(5, 128, 128, 3)	4	Categorical
Testing	20%	686	(5, 128, 128, 3)	4	Categorical

Proposed Model Architecture
We propose a hybrid deepfake detection architecture that combines *ResNet-50* for spatial feature extraction with *LSTM* for temporal dependency modeling. This fusion allows the model to capture both frame-level artifacts and sequential inconsistencies typically present in synthetic videos.
Spatial Feature Extraction using ResNet-50
ResNet-50, a deep Convolutional neural network pre-trained on *Image Net*, is employed to extract high-dimensional spatial features from each frame. Its skip connections and residual learning capabilities enable:
Extraction of fine-grained visual cues, such as unnatural lighting, pixelation, and warping.
Anomaly detection in facial symmetry and structure.
Transfer learning advantages, accelerating convergence, and enhancing performance on limited-label datasets.

Temporal Analysis using LSTM

Following spatial feature extraction, the sequence of frame-level features is fed into an *LSTM network*, which models temporal relationships and identifies sequential inconsistencies:

Memory of past frames enables the detection of unnatural motion transitions, poor blending, and frame-by-frame artifacts.

Improved understanding of dynamic facial expressions, which are often poorly replicated in deepfakes.

Sequential anomaly detection, enhancing overall classification accuracy.

Hybrid Integration

The model pipeline operates as follows:

ResNet-50 extracts spatial features from each frame in the 5-frame input clip.

LSTM processes the sequence of spatial features, learning temporal dynamics.

Dense layers with softmax activation output a binary classification (real or fake).

This hybrid CNN-LSTM framework effectively captures both *spatial irregularities* and *temporal artifacts*, thereby significantly improving the reliability of deepfake detection.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are widely utilized in image and video processing tasks due to their ability to effectively capture spatial features. These networks are particularly useful in the context of deepfake detection, where identifying subtle distortions such as irregular facial geometry, unusual lighting, or unnatural texture is critical.

In this study, we incorporate the ResNet50 model, a deep residual network composed of 50 layers. ResNet50 addresses the challenges of training deep networks by using skip connections, allowing for more efficient learning of complex patterns. This architecture is especially powerful in extracting detailed spatial characteristics from video frames, which is essential for detecting forged facial regions.

Advantages of Using ResNet50:

Detailed Feature Extraction: Capable of capturing intricate spatial structures, which aids in identifying manipulated facial features and inconsistencies.

Detection of Visual Anomalies: Sensitive to irregularities such as texture mismatches or unnatural shading, which are often signs of tampering?

Transfer Learning Capabilities: By initializing the model with pre-trained weights (e.g., from ImageNet), we reduce training time and enhance performance on smaller, domain-specific datasets.

High Classification Performance: The model's architectural depth contributes to strong predictive capabilities in binary classification tasks like real vs. fake video identification.

3.2 Long Short-Term Memory (LSTM) Networks

Although CNNs are proficient at recognizing spatial patterns in single frames, they do not inherently capture the dynamics across multiple frames. To address this, we integrate Long Short-Term Memory (LSTM) networks, which are designed to process and learn from sequential data.

LSTMs help model how facial expressions and features evolve, which is crucial in detecting deepfakes where unnatural or inconsistent movements may be present due to frame synthesis errors.

Benefits of LSTMs in Temporal Modeling:

Temporal Irregularity Detection: Capable of identifying inconsistencies in motion, such as abrupt or unnatural changes in expressions across frames.

Memory of Past Frames: Maintains context over time, allowing the model to distinguish between realistic and artificial transitions in video sequences.

Frame Relationship Modeling: Assesses whether facial features change in a natural progression, aiding in the detection of temporal discontinuities.

Synergy with CNNs: When combined with CNNs, LSTMs take spatial features from each frame and analyze their sequence, resulting in a more comprehensive understanding of the video content.

3.3 CNN-LSTM Hybrid Framework

To exploit both spatial and temporal features, we propose a hybrid approach combining ResNet50 with LSTM networks. In this framework, ResNet50 is responsible for extracting spatial representations from individual frames, while the LSTM processes these representations sequentially to capture temporal dependencies.

This hybrid design enables the system to identify both localized visual artifacts and unnatural motion across frames, thereby enhancing the model's capability to distinguish between authentic and manipulated videos with greater reliability.CAPSULE NETWORKS

Capsule Networks (CapsNet) provide a powerful alternative to traditional Convolutional Neural Networks (CNNs) by preserving complex spatial hierarchies within images and video frames. Unlike CNNs, which often lose spatial relationships due to pooling layers, Capsule Networks retain detailed positional and orientational information about facial features.

Enhanced Feature Representation: Capsules encode the spatial configuration and pose of facial elements, leading to a more nuanced understanding of facial structure and geometry.

Detection of Subtle Manipulations: CapsNet are particularly effective at identifying fine-grained inconsistencies and unnatural expressions often found in deepfake videos.

Improved Robustness: These networks demonstrate resilience to variations in lighting, pose, and adversarial noise, ensuring dependable performance across diverse scenarios.

Effective with Occlusions: Capsule Networks can still accurately detect facial features even when parts of the face are obscured, making them highly suitable for real-world video data.

Seamless Integration: CapsNet integrate well with CNN and LSTM architectures, offering a hybrid approach that captures both spatial and temporal features for enhanced deepfake detection.

IMPLEMENTATION'S RESULTS

The proposed hybrid architecture, integrating ResNet, CapsuleNet, and LSTM, demonstrates strong performance and enhanced resilience when compared to standalone models used for deepfake detection. On the DFDC test set, the hybrid model surpasses the single-frame CapsuleNet by approximately 5% in accuracy and exceeds the multi-frame averaged CapsuleNet approach by 2–2.5%. This performance gain is primarily attributed to the LSTM component, which augments the spatial analysis of CapsuleNet with temporal context, enabling the detection of subtle motion inconsistencies and artifact patterns over sequential frames.

In comparison with the XceptionNet architecture, the hybrid model reports a slightly lower accuracy—by approximately 3.3%—on the DFDC test set. However, when evaluated on a public test set, both models yield nearly identical results, achieving an accuracy of 78.38%. This parity highlights the hybrid model's strong generalization ability, particularly when faced with unfamiliar manipulations and complex augmentations.

Moreover, the hybrid model and CapsuleNet-based systems exhibit increased robustness relative to XceptionNet in cross-dataset evaluations. The public test set, which includes heavily augmented videos and novel deepfake generation methods not seen during training, causes a more significant performance drop for XceptionNet than for the hybrid model. This suggests that the hybrid model is more adaptable to real-world variability and less sensitive to distribution shifts.

The architectural combination—ResNet for detailed spatial feature extraction, CapsuleNet for 3D spatial relationship modeling, and LSTM for temporal pattern recognition—offers a comprehensive solution to the multifaceted challenges of deepfake detection. Although the hybrid system may slightly trail XceptionNet in controlled benchmarks, its competitive performance and superior adaptability across diverse and unseen data environments make it a promising candidate for robust, real-world applications.

Despite advancements in deepfake generation, current models like the proposed hybrid system can still effectively identify falsified content, reinforcing the importance of integrated spatial-temporal analysis in combating increasingly sophisticated manipulation techniques.

CONCLUSION

This research proposes an innovative hybrid architecture that fuses spatial and temporal modeling through the integration of ResNet34, LSTM, and Capsule Networks to enhance the reliability of deepfake video detection. The ResNet34 module is employed for extracting spatial-level features, effectively identifying visual inconsistencies in key facial landmarks often manipulated in synthetic content. Capsule Networks contribute by modeling spatial hierarchies and preserving pose-related information, which is crucial for distinguishing natural expressions from tampered ones. Complementing this, the LSTM component captures dynamic facial transitions across frames, helping the system detect unnatural motion patterns.

When benchmarked against the DFDC dataset, the hybrid framework attained an accuracy of 96.85%, closely approaching that of XceptionNet while utilizing significantly fewer parameters. On a separate public dataset containing more diverse and heavily augmented content, both models achieved comparable accuracy (~95%). However, the proposed architecture exhibited superior adaptability to unseen data variations, suggesting a stronger capacity for generalization compared to XceptionNet, which struggled with novel alterations not present during training.

The frame sampling strategy also emerged as a vital factor influencing performance. Among the tested approaches, sampling frames at equal intervals yielded the most consistent and accurate results. Moreover, the proposed model maintains a lightweight structure with around 4 million parameters, enhancing its viability for real-time deployment in resource-constrained environments such as content moderation pipelines on social media.

Future investigations could examine ensemble-based architectures and fusion techniques that combine diverse deep learning backbones with optimized frame selection heuristics. This would further enhance the model's resilience to emerging deepfake methods and support scalable solutions for live video analysis and content verification.

VII. REFERENCES

[1] Y. Li, M.-C. Chang, and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv preprint arXiv:1811.00656, 2018.

[2] Y. Li, M.-C. Chang, and S. Lyu, "Exposing AI-created fake videos by detecting eye blinking," in Proc. IEEE Int. Workshop Inf. Forensics Secur., 2018.

[3] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in IEEE ICASSP, 2019, pp. 2307–2311.

[4] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[5] "Deepfake Detection Challenge Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/competitions/deepfake-detection-challenge/data

[6] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," arXiv preprint arXiv:2006.07397, 2020.

[7] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in IEEE AVSS, 2018, pp. 1–6.

[8] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in Int. Conf. Artificial Neural Networks, 2011.

[9] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in IEEE ICASSP, 2018.

[10] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," in IEEE ICIP, 2016.

[11] U. Ozbulak, "CNN visualizations," 2016. [Online]. Available: https://github.com/utkuozbulak/pytorch-cnn-visualizations

[12] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in IEEE ICCV, 2019.

[13] E. Sabir et al., "Recurrent convolutional strategies for face manipulation detection in videos," in IEEE WACV, 2016, pp. 80–87.

[14] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in Advances in Neural Information Processing Systems, 2017, pp. 3857–3867.