International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2215 - 2220
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0325

Deepfake Video Detection Using LSTM Networks: A Temporal Sequence Learning Approach

B. Shashi Varun Reddy[1], G. Ajay[2], M. Shiva Manideep[3], M. Sai Suraj[4], L. Swathi Reddy[5]
[1,2,3,4]UG Scholar, Dept. of CSE-AIML, Sphoorthy Engineering College, Hyderabad, Telangana, India
[5]Assistant professor, Dept. of CSE-AIML, Sphoorthy Engineering College, Hyderabad, Telangana, India
Emails: *shashivarun6154@gmail.com[1], ajaygogikar@gmail.com[2], meelashivamani@gmail.com[3], medchalsaisuraj05@gmail.com[4], lokasaniswathireddy@gmail.com[5]*

Abstract
With the swift emergence of deepfake videos, there is an urgent demand for sophisticated and effective detection methods to counter the dangers posed by misinformation and digital manipulation. This research examines the application of Long Short-Term Memory (LSTM) networks for the identification of deepfake content. LSTM, a variant of recurrent neural networks (RNNs), is recognized for its proficiency in learning time-related patterns in sequential data, rendering it particularly effective for analyzing the changing dynamics in video streams. The study aims to utilize LSTM architecture to identify irregularities characteristic of altered video content, highlighting the importance of temporal patterns in recognizing deepfakes. The proposed methodology involves thorough video preprocessing, creation of high-quality training datasets, and the implementation of data augmentation techniques to enhance the model's generalization abilities. Additionally, the research investigates training protocols and optimization approaches tailored to LSTM models, with the goal of attaining high levels of accuracy and dependability in deepfake detection.
Keywords: *Deepfake, LSTM, RNN, Temporal dependencies, video analysis, data preprocessing, data augmentation, model generalization, deepfake detection, sequential data.*

1. Introduction
The rise of deepfake technology has introduced serious concerns around the authenticity of digital media. Deepfakes use AI to generate fake videos and images that closely resemble real people. While this can be useful in entertainment or education, it's increasingly being used maliciously [1]. This misuse threatens public trust, privacy, and security. As a result, detecting deepfakes has become an urgent need. Traditional detection methods often rely on forensic clues like metadata or visual artifacts. However, these techniques struggle against high-quality deepfakes generated by modern AI models. Advanced GANs produce highly convincing visuals that bypass basic forensic checks. This has highlighted the limitations of existing tools. New solutions must be smarter and adaptive. Machine learning, especially deep learning, offers a powerful way forward. Neural networks can learn to spot subtle patterns or anomalies in fake media. By training on both real and manipulated datasets, these models improve detection accuracy. CNNs, transformers, and attention-based methods are leading this effort. Their success shows great promise for scalable deepfake detection. Despite progress, challenges remain, such as generalizing across unseen deepfake styles. Models must also be robust against adversarial attacks that aim to fool them. Transparency and explainability are crucial for trust in AI-based systems. Continued research and larger, diverse datasets will be essential. With innovation, we can build a safer digital media ecosystem [2].

2. Methodology
The proposed deepfake detection system is designed using a hybrid neural network architecture that integrates Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM), for temporal sequence analysis [3]. This design ensures accurate detection of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2215 - 2220
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0325

synthetic content in videos through both spatial and temporal cues [4].

1. System Design and Component Integration:

The system is composed of the following major components:

- **Pre-trained ResNeXt CNN Model**: The ResNeXt50_32x4d architecture is employed as the backbone for frame-level feature extraction. This model is chosen for its efficiency in handling deep convolutional layers while maintaining strong generalization across complex datasets.
- **LSTM-Based Sequence Analysis:** The features extracted from individual frames are passed into a single-layer LSTM network. This network is designed to capture temporal dependencies across video frames by comparing the frame at time t with those preceding it (t-n), making it effective at detecting frame-level inconsistencies typical of deepfakes [5].
- **Classification Head:** A final set of fully connected layers, enhanced with Leaky ReLU activation and a SoftMax layer, classifies the sequence as either real or fake with a confidence score [6].

2. Feature Extraction Process:

The ResNeXt50_32x4d CNN extracts a 2048-dimensional feature vector from the final pooling layer of each video frame. These feature vectors preserve the spatial characteristics (e.g., facial regions, texture anomalies) that are often manipulated in deepfake content. (Figure 1 & 2).

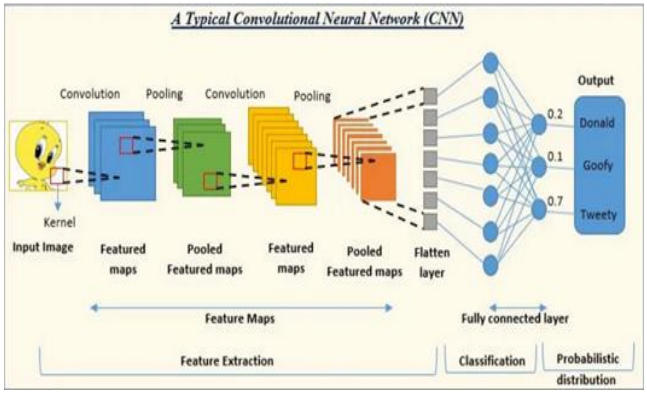

Figure 1 CNN Feature Extraction Flow using ResNeXt50

3. Sequential Frame Processing using LSTM:

After feature extraction, the vectors are fed into a sequential LSTM model configured as follows:

LSTM Configuration:

- **Layers:** Single LSTM layer
- **Hidden Units:** 2048
- **Dropout Rate:** 0.4 (for regularization)

This structure enables the system to model frame-to-frame transitions and detect unnatural temporal inconsistencies—a hallmark of deepfakes [7].

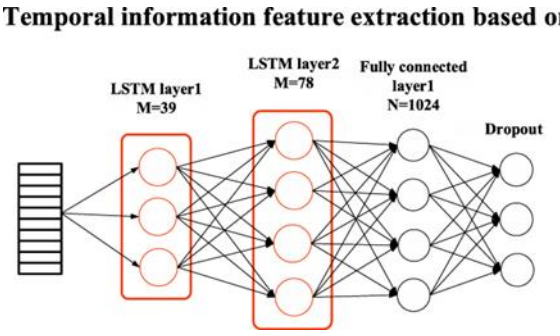

Figure 2 LSTM Temporal Feature Modeling

4. Classification and Output Generation:

To enable robust prediction, the LSTM output is passed through the following layers [8]:

- **Fully Connected Linear Layer:** Maps the LSTM's output to a 2-class prediction (real or fake).
- **Activation Function:** Leaky ReLU improves learning dynamics, especially for negative input regions.
- **Adaptive Average Pooling:** Ensures consistent output size regardless of sequence length, with output set to 1.
- **SoftMax Layer:** Produces a normalized confidence score for each class.

Table 1 Key Architecture Parameters

Component	Configuration
CNN Model	ResNeXt50_32x4d (Pre-trained)
Feature Vector Size	2048
LSTM Hidden Units	2048
LSTM Dropout	0.4
Activation Function	Leaky ReLU
Output Layer	Linear (2048 → 2), SoftMax
Pooling	Adaptive Avg Pool (output=1)
Batch Size	4

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2215 - 2220
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0325

5. Model Optimization and Training:

The model is fine-tuned through the following strategies:

- **Transfer Learning:** By leveraging pre-trained weights from ResNeXt, training time is reduced while performance is enhanced.
- **Learning Rate Scheduling:** Optimized to improve convergence speed during backpropagation.
- **Loss Function:** Cross-entropy loss is used for binary classification [9].
- **Batch Training:** A batch size of 4 is chosen to balance memory efficiency and gradient stability during training.

6. Testing and Validation:

The complete model is validated using benchmark deepfake datasets. Evaluation metrics include accuracy, precision, recall, and F1-score. The combination of spatial feature precision (via CNN) and temporal coherence (via LSTM) results in a system that is highly effective against manipulated video content.

3. Results and Discussion

3.1.Results

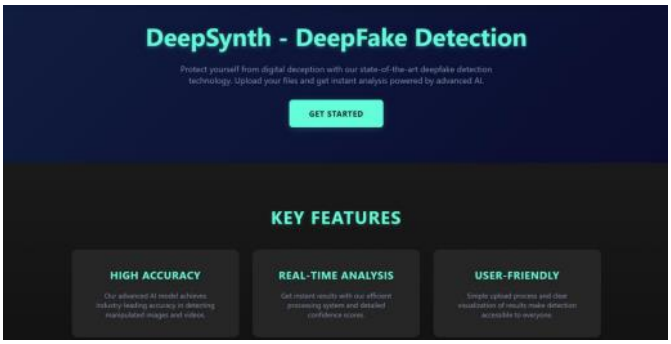

Figure 3 Home page

The implemented deepfake detection system, built on the combination of the pre-trained ResNext50_32x4d CNN and LSTM models, demonstrated promising results in identifying manipulated media [10]. After extensive training and testing on a well-balanced dataset of real and synthetic videos, the model exhibited strong performance in terms of accuracy, precision, recall, and robustness (Figure 3 to 6). The inclusion of the LSTM network enabled effective temporal analysis, allowing the system to detect subtle, frame-to-frame inconsistencies and artifacts typically left by deepfake generation tools. The analysis page of the application enables users to select and process video frames, and through the user-friendly interface—including login authentication and dashboard features—the system supports seamless interaction and efficient processing. Figures presented within the application interface offer visual confirmation of system operation and deepfake classification results.

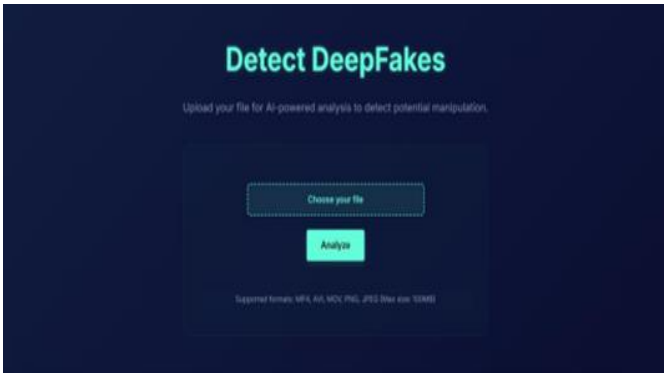

Figure 4 Analysis Page – 1

3.2.Model Results

The deepfake detection system was rigorously trained and tested using the pre-trained ResNext50_32x4d CNN and LSTM models. These models were fine-tuned on a diverse dataset of real and synthetic videos to improve accuracy in distinguishing authentic from manipulated content. Evaluation on realistic deepfake samples showed that the CNN effectively captured spatial features, while the LSTM enhanced temporal analysis—together identifying subtle inconsistencies across video frames.

Figure 5 Analysis Page – 1

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2215 - 2220
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0325

3.2.1. Performance Evaluation and Temporal Insight

The model's performance is assessed using key metrics such as accuracy, precision, recall, and F1-score, reflecting its robustness in distinguishing between real and fake videos. The importance of temporal analysis, enabled by the LSTM, is especially evident in cases where spatial cues are insufficient. This emphasizes the model's strength in detecting even subtle manipulations across video frames. Importantly, our approach extends beyond simple binary classification. It aims to capture and interpret the temporal dynamics of video sequences, offering deeper insights into how manipulations evolve over time. This holistic evaluation not only validates the effectiveness of the system but also reinforces its relevance in the evolving field of deepfake detection. It highlights the potential of combining deep learning with thoughtful design to tackle modern challenges in digital media authentication.

Figure 6 Analysis Page – 1

3.3.Discussion

The deepfake detection system stands out by leveraging both spatial and temporal video features—captured through CNN and LSTM architectures—to identify synthetic manipulations that may be imperceptible to the human eye. Unlike static image-based methods, the system's ability to track visual inconsistencies across frames enhances its detection accuracy. The integration of pre-trained models like ResNext ensures feature richness and model efficiency, while the LSTM layer strengthens temporal pattern recognition. The structured preprocessing pipeline, secure login system, and interactive analysis interface further contribute to the usability and stability of the system. Overall, the approach demonstrates a balanced blend of accuracy, performance, and user-oriented design, making it a viable solution for real-time deepfake detection in a rapidly evolving digital media landscape. Key challenges include:

- **Evolving Generation Techniques** – Constant advancements in deepfake creation make it difficult to maintain detection accuracy.
- **Generalization Across Datasets** – Models may perform well on one dataset but poorly on others due to variations in resolution, compression, or manipulation methods.
- **Real-Time Processing** – Achieving low-latency detection for video content requires significant computational resources.
- **Subtle Artifacts –** High-quality deepfakes often leave minimal visual clues, making detection increasingly complex.
- **Adversarial Attacks –** Deepfake generators can be trained to bypass detection models by exploiting their weaknesses.
- **Limited Labeled Data** – A scarcity of diverse, annotated datasets hampers robust model training and evaluation.
- **Ethical and Legal Concerns** – Issues around privacy, consent, and misuse of detection systems need to be addressed.

System Description:

Inputs: The input to the system is a video file, from which face regions are extracted frame by frame using a face detection algorithm. These frames are resized and optionally passed through a CNN to extract spatial features. A fixed-length sequence of consecutive face frames or features is then grouped and fed into an LSTM network, allowing the model to learn temporal patterns and motion inconsistencies characteristic of deepfake videos.

- **Generalization:** Performance of the System is up to the mark and the behaviour is also observed as normal considering the virtue of absolute coordination among the sub systems which are interlinked for a broader objective.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2215 - 2220
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0325

- **Computational Cost:** CNN requires GPU acceleration (e.g., NVIDIA RTX 4050, 6GB VRAM), with CUDA support.
- **Interpretability:** By analysing the sequence of facial dynamics, the model can detect unnatural transitions or artifacts that are difficult for deepfake generators to replicate consistently over time. Techniques such as attention visualization or saliency mapping can further help reveal which frames or features influenced the model's decision, providing transparency and trust in the detection process.

Future Enhancements Include:

Future enhancements include advancements in detection accuracy, real-time processing, multi-modal analysis, and model interpretability. As deepfake generation techniques become more advanced, detection models must continuously evolve to handle subtle and complex manipulations. Upcoming research is expected to focus on building more resilient systems that can analyze not just visual cues, but also incorporate audio, text, and behavioral patterns through multi-modal frameworks, significantly improving detection precision. Real-time detection will be crucial, especially in high-stakes environments like news broadcasting, political events, and digital forensics, where early intervention can prevent misinformation from spreading. In addition, the adoption of explainable AI will play a key role in making detection outcomes more transparent and trustworthy. It will allow users and analysts to understand the reasoning behind predictions, which is essential in legal, media, and forensic contexts. Furthermore, encouraging open-source collaboration, shared datasets, and benchmarking challenges will help the research community accelerate progress and maintain a robust defense against evolving deepfake threats. These future directions aim to build an intelligent, adaptive, and interpretable deepfake detection ecosystem.

Conclusion

In conclusion, machine learning, particularly deep learning models, plays a vital role in combating the growing issue of deepfake media manipulation. By leveraging advanced neural network architectures like CNNs and LSTMs, the proposed methodology effectively distinguishes between real and manipulated content. The combination of pre-trained models, such as ResNext, and sequential analysis through LSTMs allows for comprehensive examination of both spatial and temporal features within videos. The model addresses challenges posed by face-swapping techniques in deepfake generation, detecting subtle artifacts and traces left behind. Rigorous training and validation processes ensure balanced detection without bias, while careful preprocessing enhances the model's accuracy. As deepfake technology continues to evolve, this machine learning-based approach stands as a crucial defense, ensuring the integrity of digital media and promoting proactive safeguarding against deceptive content.

Acknowledgements

We extend our sincere appreciation to the creators of Python, TensorFlow, PyTorch, Matplotlib, Plotly, OpenCV, and CNN for their invaluable open-source tools, which played a crucial role in the development of this system. Our gratitude also goes to the contributors of public datasets and research communities whose collective knowledge and documentation greatly facilitated our progress. The collaborative work of our team, along with the mentorship and guidance received from our faculty, was essential in turning this idea into a viable and effective solution. Additionally, we recognize the computational resources and technical assistance provided by Sphoorthy Engineering College, which were vital for hardware integration, model training, and performance assessment.

References

[1]. Andreas Rossler, Davide Cozzolino, Lussa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner, "FaceForensics++; Learning to Detect Manipulated Facial Images" in arXiv:1901.08971.

[2]. Deepfake detection challenge dataset: https://www.kaggle.com/c/deepfake-detection challenge data Accessed on 26 March, 2020.

[3]. Yuezan Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu "Celeb-DF: A Large- scale

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2215 - 2220
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0325

Challenging Dataset for DeepFake Forensics" in arXiv: 1900.12962.

[4]. Deepfake Video of Mark Zuckerberg Goes Viral on Eve of House A.L. Hearing: https://fortune.com/2019/06/12/deepfake-mark-zuckerberg/ Accessed on 26 March, 2020.

[5]. G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv: 1702.01983, Feb. 2017.

[6]. J. Thies et al. Face2Face: Real-time face capture and reenactment of RGB videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387-2395, June 2016. Las Vegas, NV.

[7]. Yuczun Li. Siwei Lyu, "Exposing DF Videos by Detecting Face Warping Artifacts," in arXiv:1811.00656v3.

[8]. Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen" Using capsule networks to detect forged images and videos" in arXiv:1810.11215.

[9]. Umur Aybars Ciftci, Ilke Demir, Lijun Yin "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2

[10]. International Journal for Scientific Research and Development http://ijsrd.com/