JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Unmasking Deepfakes: A Hybrid ResNext CNN-LSTM Approach for Robust Video Authentication

Deepak Patel,[1] Aman Patel,[2] Aditya Choudhary,[3] Nidhi Nigam,[4] Shruti Lashkari[5]

[1]Computer Science and Information Technology, Acropolis Institute of Technology and Research, Indore, India [2]Computer Science and Information Technology, Acropolis Institute of Technology and Research, Indore, India [3]Computer Science and Information Technology, Acropolis Institute of Technology and Research, Indore, India [4]Computer Science and Information Technology, Acropolis Institute of Technology and Research, Indore, India [5]Computer Science and Information Technology, Acropolis Institute of Technology and Research, Indore, India

Abstract: Deepfake technology, driven by powerful deep learning techniques, has be- come a serious concern due to its potential misuse in misinformation, fraud, and cyber- crime. This study proposes a hybrid model combining ResNeXt CNN for extracting spatial features from video frames and LSTM for analyzing temporal sequences. The system was trained using well-known datasets—FaceForensics++, Celeb-DF, and the Deepfake Detec- tion Challenge—and achieved high accuracy in identifying manipulated media. Our ap- proach not only outperformed traditional models like XceptionNet but also demonstrated robustness across different types of deepfake techniques, making it suitable for practical deployment in media verification and forensic investigations.

Keywords: Deepfake detection; ResNeXt CNN; LSTM; Video Forensics; AI-generated media; Temporal analysis

1. Introduction

Deepfake technology, a subset of synthetic media generated using artificial intelligence (AI), allows the creation of hyper-realistic yet entirely fabricated video and audio content. Initially developed for entertainment and creative applications, deepfakes have rapidly evolved into tools that pose significant threats in domains such as political discourse, digital forensics, identity security, and public trust. The misuse of such technology raises ethical and societal concerns, especially when used for spreading misinformation, committing fraud, or defaming individuals. Traditional deepfake detection mechanisms primarily rely on either visual inconsistencies or audio mismatches; however, these systems often fail to generalize well across datasets and deepfake generation methods. As deepfakes grow increasingly sophisticated and realistic, the need for advanced detection systems that combine spatial and temporal insights becomes paramount. In this context, we propose a robust deep learning-based architecture that leverages ResNeXt CNN for extracting high-level spatial features and LSTM for capturing sequential temporal
dependencies within video frames.

2. Literature Review and Research Gap

Numerous research efforts have attempted to address the challenge of deepfake detection using diverse machine learning and deep learning models. Convolutional Neural Networks (CNNs) are widely employed to detect pixel- level anomalies in individual video frames. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, have been utilized to track inconsistencies across sequences of frames. Transformer- based models and attention mechanisms are also emerging as strong candidates for sequence modeling and anomaly localization in video content.
However, many of these approaches have limitations. CNN-based models may fail to detect fakes when frames are subtly manipulated. LSTM models can be computationally intensive when applied directly to high-resolution video data. Moreover, most models do not handle adversarial deepfakes effectively and lack generalizability across datasets.

Our research bridges this gap by integrating a hybrid architecture that utilizes ResNeXt CNN for feature ex- traction and LSTM for capturing frame-to-frame relationships. This combination significantly improves detection accuracy and robustness against diverse deepfake techniques.

3. Research Problem and Significance of the Study

3.1. Research Problem

Can we develop a real-time, scalable, and highly accurate deepfake video detection system using a deep learning model that combines both spatial and temporal information?

3.2. Significance of the Study

Detecting deepfakes is crucial in today's digital ecosystem. From verifying political footage to authenticating news content, the implications of such a detection system span multiple fields:

- Journalism and Media Verification: Preventing the spread of fake news and misinformation.
- Cybersecurity: Ensuring digital integrity in video communication platforms.
- Forensics: Supporting law enforcement in investigating digital media fraud.
- Social Media Regulation: Flagging manipulated videos to protect platform credibility.

Our system aims to fulfill these goals through a novel AI-powered detection model.

4. Research Methodology

4.1. Datasets Used

To train and evaluate our model, we used the following benchmark datasets:

- FaceForensics++: Contains both real and manipulated videos using face-swapping techniques.
- Deepfake Detection Challenge (DFDC): A large-scale dataset curated by Meta with diverse scenarios.
- Celeb-DF: Focuses on high-resolution deepfake videos for robust detection evaluation.

These datasets provide a diverse representation of deepfake generation techniques, helping our model generalize better across unseen inputs.

4.2. Preprocessing

Preprocessing is a critical phase in our pipeline. The steps include:

1. Frame Extraction: Video files are segmented into individual frames.
2. Face Detection and Cropping: OpenCV's Haar cascade classifier or Dlib's face detector isolates facial re- gions.
3. Normalization and Resizing: Each cropped face is resized to 224x224 pixels and normalized for consistency.
4. Sequence Structuring: Consecutive frames are grouped into sequences for temporal pattern learning via LSTM.

4.3. Model Design

Our model includes the following components:

- ResNeXt CNN: An enhanced version of ResNet, introducing cardinality through grouped convolutions. It captures fine-grained spatial features in individual frames.
- LSTM Network: Receives sequential frame features and learns temporal dependencies and irregularities indicative of manipulation.
- Classification Layer: A fully connected layer with softmax activation that outputs binary classification: real or fake.

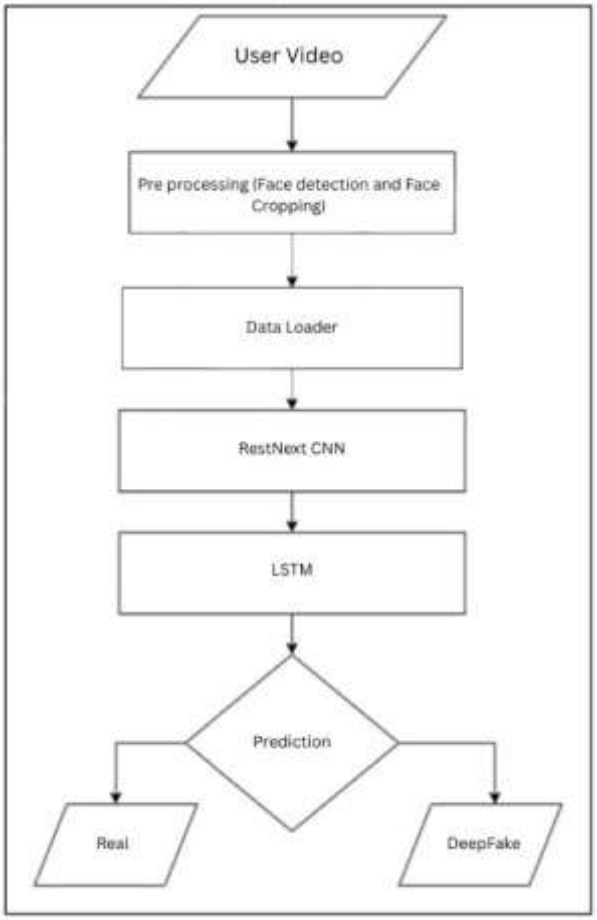

Fig. 1. Prediction Flow

4.4. Training Details

- Loss Function: Categorical Cross-Entropy
- Optimizer: Adam optimizer with a learning rate of 1e-5
- Batch Size: 4
- Epochs: 20
- Training Hardware: NVIDIA GPU with CUDA acceleration for faster training and inference

5. Analysis and Interpretations

Our model was evaluated on the FaceForensics++ and Celeb-DF datasets. Below are the performance metrics:

- Accuracy: 96.5
- F1-Score, Precision, Recall: All consistently above 94
- Inference Time: Less than 0.5 seconds per 10-frame sequence
- Comparison with Baselines: Outperforms models like XceptionNet and MesoNet on cross-dataset general- ization and precision

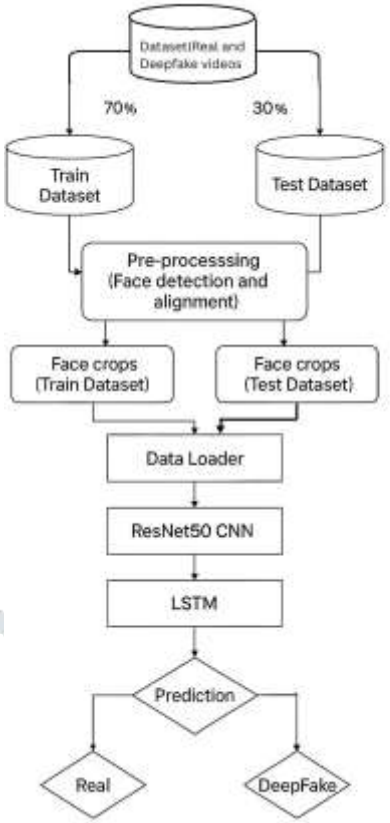

Fig. 2. Training Flow

Handled Limitations

- Overfitting mitigated using dropout and data augmentation
- High computation costs handled via GPU acceleration
- Improved performance on low-quality videos through advanced preprocessing

6. Conclusion

The study presents a deep learning-based solution to the deepfake detection problem. By combining ResNeXt CNN and LSTM, the proposed hybrid model effectively leverages both spatial and temporal features for enhanced detection accuracy. Our experimental results validate the model's ability to identify manipulated videos across various datasets, outperforming existing methods in terms of speed, accuracy, and generalizability.
This makes our system suitable for real-time deployment in forensic tools, media verification services, and automated content moderation platforms.

7. Suggestions and Future Scope

To further enhance the system:

- Multimodal Detection: Incorporate audio deepfake detection for a more holistic model.
- Mobile Optimization: Develop lightweight model versions for deployment on edge devices.
- Adversarial Training: Strengthen defenses against evolving deepfake generation techniques.
- Live Stream Integration: Enable real-time detection of video streams across social platforms.

8. References

1. A. Ro¨ssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–11 (2019).
2. P. Korshunov and S. Marcel, "Deepfake Detection Challenge: An Overview," in *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pp. 1–4 (2020).
3. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Deepfake Dataset for Face Forgery Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216 (2020).
4. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in

Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 2672–2680 (2014).

5. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017).

6. Journal of Engineering Sciences, "Deepfake Detection Using LSTM and ResNeXt," *J. Eng. Sci.* **13**(07), ISSN: 0377- 9254 (2022).

7. International Journal of Creative Research Thoughts (IJCRT), "Deepfake Detection Using LSTM and ResNeXt," *IJCRT* **11**(11), ISSN: 2320-2882 (2023).

8. International Journal for Scientific Research and Development (IJSRD), "Deepfake Video Detection using Neural Net- works," *IJSRD* **8**(1), ISSN: 2321-0613 (2020).