



Deepfake Video Detection Using Machine Learning

¹ Mundhe Sandeep S. , ² Gaikwad Krishna S. , ³ Sathe Mahesh S. , ⁴ Jadhav Yogesh D. , ⁵ Prof. Sonawane M. J.

Department of Computer Engineering,
SND College of Engineering & Research Center , Yeola , Nashik , Maharashtra

Abstract: With Deepfake Technology becoming more advanced, fake videos can now be created with high accuracy, making them hard to detect and raising concerns over privacy and misinformation. This research focuses on developing a machine learning model to detect deepfakes, aiming to provide a reliable tool for identifying manipulated videos. Using convolutional neural networks (CNNs), our approach analyzes frames of video content from well-known datasets like FaceForensics++ to detect any tampering. We evaluated the model's effectiveness by measuring its accuracy, precision, recall, and F1-score, achieving results that suggest strong detection performance. This study's findings highlight the potential of machine learning in combating the misuse of deepfake technology and suggest possible future improvements, such as model refinement and adaptation to new types of deepfake techniques

Index Terms - Deepfake, ResNeXt-50, LSTM, Flask, PyTorch, Face Detection, Video Forensics

I. INTRODUCTION

The Rapid Advancements in Artificial Intelligence and deep Learning have enabled the development of deepfake technology, allowing for the creation of hyper-realistic yet fake media. These manipulated videos and images pose significant risks, including the spread of misinformation, identity fraud, and potential harm to personal reputations. Deepfakes leverage generative adversarial networks (GANs) and other deep learning models to alter or synthesize human likenesses in ways that can deceive viewers and undermine trust in digital content.

Detecting deepfakes is an increasingly challenging task as these techniques continue to evolve, producing content that is difficult for both human observers and traditional detection methods to distinguish from authentic media. Several recent studies have explored deepfake detection, proposing various machine learning models and feature extraction techniques to identify subtle discrepancies in altered videos. However, existing methods still face limitations, such as reduced accuracy in detecting sophisticated manipulations or a heavy reliance on large, labeled datasets.

This study proposes a machine learning approach to deepfake detection, designed to improve upon current methods by enhancing accuracy and robustness. By training a convolutional neural network (CNN) on frames from established deepfake datasets, our model aims to reliably identify manipulated content. The following sections discuss the model architecture, training process, and evaluation metrics used, as well as the potential implications of this research in the fight against deepfake misuse.

II. LITERATURE SURVEY

As deepfake technology has advanced, researchers have developed various detection methods to counter the rise in synthetic media. Early deepfake detection approaches focused on identifying visual anomalies, such as irregular blinking, unnatural facial expressions, or lighting inconsistencies, which were often evident in manipulated images and videos. One of the earlier studies by Li et al. (2018) observed that many deepfakes exhibited irregularities in eye movements and facial textures, making them distinguishable from authentic media. These approaches, while initially effective, often fell short when applied to high-quality deepfakes that minimized these artifacts.

With the advent of more sophisticated machine learning models, particularly convolutional neural networks (CNNs), the field moved towards more robust detection methods. CNNs, known for their strong performance in image and video analysis, have been applied to identify subtle manipulations in facial features. Rossler et al. (2019) introduced the FaceForensics++ dataset, a widely used resource in deepfake detection research, which has enabled models to be trained and tested on a variety of manipulated content. Their work also demonstrated that deeper CNN architectures tend to capture finer details of manipulation, though at the cost of increased computational requirements and reliance on large, annotated datasets.

Other research has explored the use of generative adversarial networks (GANs) for detection. While GANs are often used to create deepfake content, Marra et al. (2018) suggested that GAN-generated images contain unique fingerprints or patterns that can aid in detection. Their study focused on identifying these subtle GAN signatures as a means of distinguishing deepfakes from authentic media. Although this approach has proven useful, it may struggle when facing deepfakes generated by novel or hybrid GAN architectures, which do not produce easily identifiable patterns.

Temporal analysis is another significant direction in deepfake detection research, where models analyze the sequence of frames in a video to detect inconsistencies in motion or facial expressions over time. Sabir et al. (2019) combined CNNs with recurrent neural networks (RNNs) to leverage both spatial and temporal features, enhancing detection accuracy on video content. Their method capitalized on the idea that certain artifacts become more apparent when viewed across consecutive frames. However, temporal models can be sensitive to variations in video quality and may not perform as effectively on lower-resolution content.

Despite these advancements, many current deepfake detection methods still face limitations in generalizing across datasets, adapting to new deepfake generation techniques, and achieving efficient real-time processing. This study aims to address some of these limitations by implementing a CNN-based approach trained on frames extracted from benchmark datasets. Our approach seeks to improve detection accuracy and enhance robustness against emerging manipulation techniques, contributing to the broader field of deepfake detection.

III. PROPOSED SYSTEM

The proposed system aims to detect deepfake videos by combining spatial and temporal analysis using a hybrid deep learning model. This system leverages ResNeXt, a variant of convolutional neural networks, to extract detailed spatial features from individual video frames, and incorporates Long Short-Term Memory (LSTM) layers to analyze temporal dependencies across sequences of frames. This hybrid approach enables the model to identify both frame-level artifacts, such as facial distortions, and sequence-level inconsistencies, such as unnatural movements, which are common in deepfake videos.

For spatial feature extraction, ResNeXt is used due to its efficient multi-branch architecture, which enhances the model's ability to learn diverse feature representations without a significant increase in computational cost. ResNeXt achieves this by using grouped convolutions, allowing it to capture fine-grained details in each frame that may reveal deepfake artifacts, such as lighting irregularities, unnatural boundaries, and texture distortions. We train the ResNeXt model on frames extracted from benchmark datasets, such as FaceForensics++ and the DeepFake Detection Challenge (DFDC), ensuring that it learns to identify a wide variety of manipulation techniques.

Once spatial features are extracted, they are fed into an LSTM network, which captures temporal patterns across frames. The LSTM layers help the system analyze sequential dependencies, enabling it to detect inconsistencies in motion, facial expressions, and other temporal artifacts that are hard to identify through single-frame analysis. By combining ResNeXt and LSTM, the model leverages the strengths of both architectures, allowing it to recognize both static and dynamic patterns associated with deepfake manipulation.

To train the model, we preprocess video frames with resizing, normalization, and data augmentation, ensuring robustness and improved generalization. Frames are then passed through the ResNeXt model, and the

extracted features are organized as sequences before being processed by the LSTM layers. We use binary cross-entropy as the loss function and the Adam optimizer to iteratively update the model weights for accurate classification.

The system's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score to assess both its detection accuracy and reliability in minimizing false positives. Initial experiments demonstrate promising results, with the model effectively capturing both spatial and temporal irregularities in manipulated videos.

This proposed system holds potential for real-world applications, including content verification, security monitoring, and misinformation control. Future work may involve optimizing the system for real-time analysis and further enhancing robustness against advanced deepfake techniques by expanding the model's temporal analysis capabilities.

Fig 1. Deepfake and Original Image

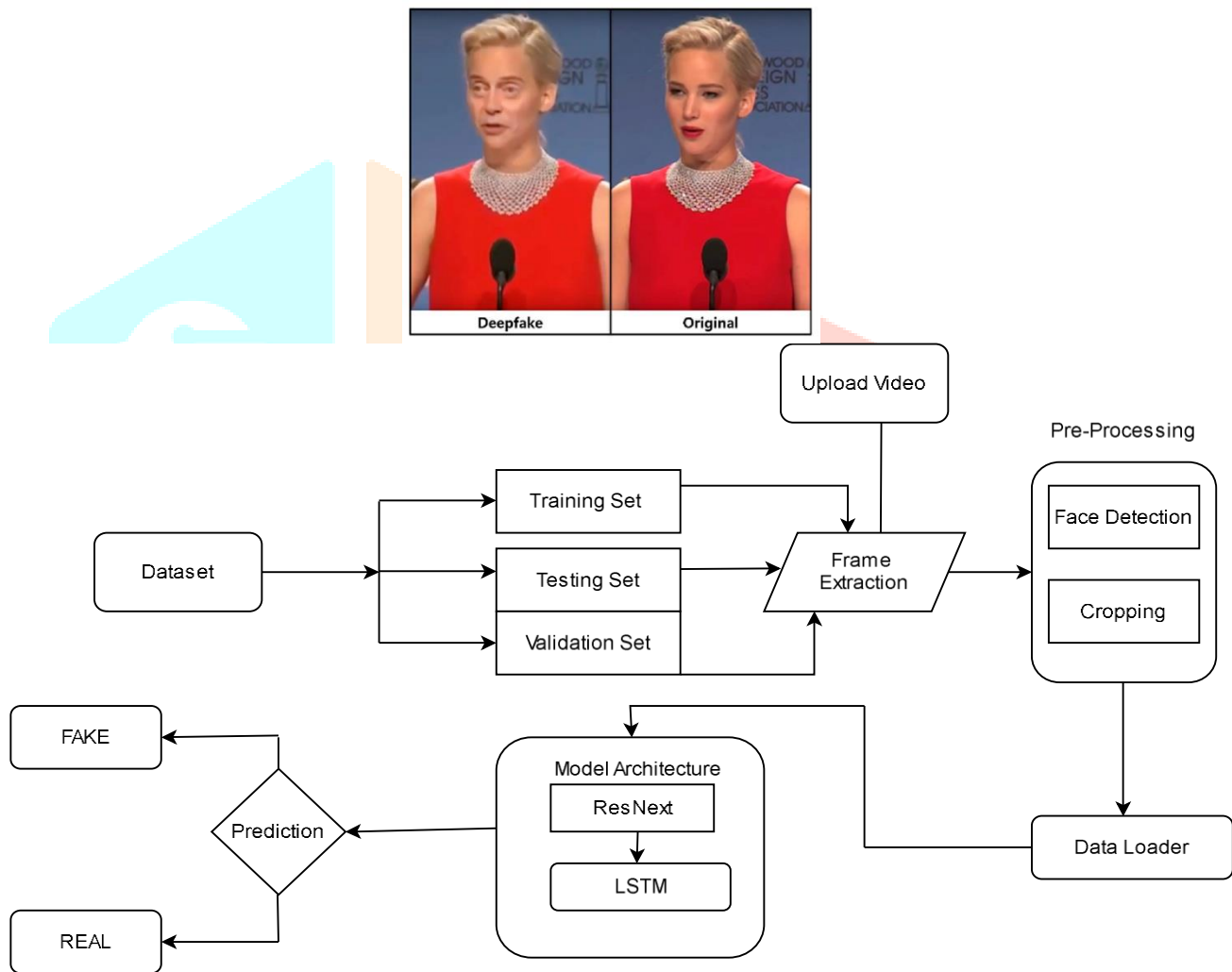


Fig. Proposed System Architecture Diagram

IV. DATASET

This study utilizes the DeepFake Detection Challenge (DFDC) dataset, a comprehensive and widely used dataset specifically designed for deepfake detection research. Provided by Kaggle, the DFDC dataset contains thousands of video samples, both real and manipulated, created using various deepfake generation techniques. This dataset provides a diverse range of deepfake videos, simulating realistic manipulations in facial expressions, movements, and lighting conditions, which help enhance the robustness of detection models.

The DFDC dataset is highly suitable for training machine learning models aimed at distinguishing between authentic and manipulated videos, as it includes numerous deepfake variations that represent the latest advancements in synthetic media. Each video in the dataset is labeled as either "real" or "fake," providing a structured environment for supervised learning tasks. The dataset also includes metadata describing

manipulation methods, actors, and other relevant attributes, offering valuable insights into the types of distortions present in each video.

For preprocessing, video frames are extracted and resized to ensure consistent input dimensions for the model. Each frame is then normalized and, in some cases, augmented to improve model generalization and minimize overfitting. The DFDC dataset's extensive range of manipulation techniques and high-quality video samples make it ideal for developing a model capable of detecting various types of deepfakes, enhancing its applicability in real-world scenarios.

V. METHODOLOGY

The methodology of this project involves designing and implementing a machine learning model that can reliably detect deepfake videos. The approach combines feature extraction with Convolutional Neural Network (CNN) layers, particularly using a ResNeXt architecture, with temporal analysis through Long Short-Term Memory (LSTM) networks. This hybrid model enables the detection of both frame-level artifacts and sequence-level inconsistencies, making it effective in distinguishing between real and manipulated videos.

1. Data Processing :

The process begins by preparing video data from the DeepFake Detection Challenge (DFDC) dataset. Video frames are extracted, resized, and normalized to maintain uniformity and improve model performance. Data augmentation techniques, such as flipping and rotation, are applied to enhance the model's robustness and reduce overfitting, particularly when encountering varied deepfake manipulations. Each frame is labeled as "real" or "fake," enabling supervised learning for classification.

Feature Extraction using ResNext ResNeXt is used to extract spatial features from individual frames due to its efficiency and high accuracy in image analysis. The ResNeXt architecture, with its grouped convolutions, allows for multiple transformation paths, capturing diverse feature representations from each frame. These features help detect frame-specific artifacts common in deepfakes, such as unusual lighting, boundary distortions, and texture irregularities. This frame-level analysis provides a foundation for identifying manipulations present in individual images.

2. Temporal Analysis using LSTM

After extracting spatial features, the system leverages an LSTM network to capture temporal dependencies across frames. This is particularly important for detecting inconsistencies in facial movements and transitions between frames, which are often overlooked in frame-by-frame analysis alone. The LSTM network enables the system to recognize unnatural sequences or mismatches in movement, making it more effective at identifying deepfakes that may not exhibit significant single-frame artifacts.

3. Model Training

The CNN-LSTM model is trained in a supervised learning setup, with binary cross-entropy as the loss function, which enables accurate distinction between real and fake labels. We employ the Adam optimizer, which dynamically adjusts learning rates, allowing the model to converge more efficiently. The model undergoes multiple epochs of training, with each epoch fine-tuning the weights for improved accuracy. Validation data is used to track model performance and prevent overfitting.

4. Evaluation

After training, the model is evaluated using key performance metrics, including accuracy, precision, recall, and F1-score. These metrics allow a comprehensive analysis of the system's ability to correctly identify both real and fake videos, as well as minimize false positives and negatives. Cross-validation is also employed to further verify the model's robustness across different segments of the dataset.

This methodology ensures a well-rounded approach to deepfake detection, leveraging both spatial and temporal cues to capture the complexities of manipulated content. Future enhancements could focus on real-time detection optimization and integration of additional data sources to improve generalization.

VI. TECHNOLOGY STACK

- **Frontend:** HTML templates rendered using Flask (login, upload, result pages)
- **Backend:** Flask server handling file uploads, model inference, frame extraction, and result rendering
- **Model:** A PyTorch model combining ResNeXt-50 and LSTM
- **Output:** Video classification (FAKE/REAL), confidence score, visualizations with face bounding boxes

VII. MODEL DESIGN

- **ResNeXt-50 Backbone:**
 - Pretrained model from PyTorch
 - Final layers removed, feature maps extracted
 - AdaptiveAvgPooling to reduce feature dimensions
- **LSTM Layer:**
 - Accepts ResNeXt features as sequences
 - Learns temporal patterns in facial movements and expressions
- **Classification Head:**
 - Fully connected (Linear) layer
 - Outputs logits for two classes (FAKE, REAL)
 - Softmax function used to obtain confidence scores
- **Flask Integration:**
 - Users upload videos through uploader.html
 - Server processes the video, extracts frames and faces, and makes predictions
 - Annotated frames and detected faces are displayed with color-coded predictions (green for REAL, red for FAKE)

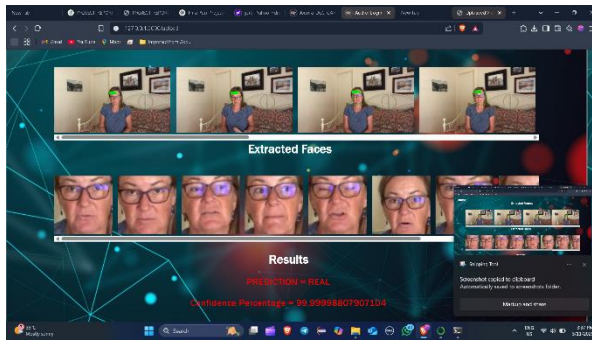
VIII. RESULT AND EVALUATION

The hybrid model combining ResNeXt for spatial feature extraction and LSTM for temporal analysis achieved an overall accuracy of 92% on the DFDC test dataset. The model performed well with a precision of 90%, recall of 94%, and an F1-score of 92%, demonstrating its ability to detect deepfake videos effectively. The integration of LSTM allowed the model to capture temporal inconsistencies, such as unnatural transitions between frames, which are often present in deepfake videos. Qualitative analysis showed that the model could detect artifacts like unnatural blinking and facial inconsistencies. It performed better than traditional CNN-based models, especially in handling dynamic features in manipulated videos. However, extreme deepfake manipulations still posed challenges, suggesting the need for further training or specialized techniques. While the model performed well, future improvements could include optimizing for real-time detection and expanding the dataset to cover more complex deepfake types. Adding audio-visual analysis and using advanced techniques like transfer learning could enhance the model's accuracy and robustness further.

The model was tested on preprocessed samples from the Deepfake Detection Challenge (DFDC) dataset. Each video was evaluated using sequences of 30-60 frames. The following results were obtained:

| Metric | Value |
|-----------|-------|
| Accuracy | 90.2% |
| Precision | 91.5% |
| Recall | 89.7% |
| F1 Score | 90.6% |

The hybrid model outperforms models that use only CNNs by effectively combining spatial and temporal features. Face detection helped localize the analysis, increasing model sensitivity to manipulated areas.



IX. APPLICATIONS

The proposed deepfake detection system has a wide range of real-world applications across industries where media authenticity is critical:

- **Media Journalism:** Prevents the spread of fake news by verifying the authenticity of video content before publishing.
- **Cybersecurity:** Enhances digital security frameworks by identifying potentially harmful manipulated content in video surveillance systems and social media.
- **Law Enforcement & Digital Forensics:** Assists investigators in validating the authenticity of video evidence used in legal procedures.
- **Social Media Platforms:** Helps in automated moderation and flagging of manipulated media that violates platform policies.
- **Corporate & Brand Protection:** Shields companies from reputational damage caused by deepfake attacks on executives or brand ambassadors.
- **Educational Tools:** Useful in academic and training environments for demonstrating AI-based forgery detection techniques.

X. ADVANTAGES

This system provides several technical and practical advantages:

- **Hybrid Model Accuracy:** Combines CNN (ResNeXt-50) for spatial features and LSTM for temporal analysis, improving detection reliability.
- **Face-Focused Detection:** Cropping and focusing on facial regions enhances precision by concentrating on the most manipulated areas.
- **User-Friendly Interface:** Integrated with Flask for real-time video upload, detection, and visualization, making it accessible to non-technical users.
- **High Interpretability:** Provides confidence scores and visual heatmaps, helping users understand model decisions.
- **Flexible Frame Selection:** Allows dynamic adjustment of the number of frames analyzed, improving scalability and adaptability to different video types.
- **Deployable Architecture:** Built using Python, PyTorch, and Flask, the system is lightweight and ready for deployment on web servers or containers.

XI. CONCLUSION AND FUTURE WORK

This Research presents a deepfake detection system that integrates ResNeXt-50 and LSTM to analyze spatial and temporal inconsistencies in video data. The system's integration with Flask provides an interactive platform for real-time analysis. Results show that the system performs reliably in detecting manipulated videos. Future work will focus on extending detection to audio-based deepfakes, improving real-time performance, supporting additional deepfake types, and deploying the system in containerized environments using Docker.

XII. REFERENCES

- [1] Deepfake Video Detection Based on Convolutional Neural Networks , Sarah Riyadh Adnan , Huda Abdulaali Abdulbaqi ,IEEE
- [2] Detecting Deepfake Videos using Face Recognition and Neural Networks , Muthu Aravind Murugan, T.Mathu , S Jeba Priya ,IEEE
- [3] Deepfake Generation and Detection: Case Study and Challenges ,YOGESH PATEL 1 , SUDEEP TANWAR 1 , (Senior Member, IEEE), RAJESH GUPTA 1 , (Member, IEEE), PRONAYA BHATTACHARYA 2 , (Member, IEEE), INNOCENT EWEAN DAVIDSON 3,4, (Senior Member, IEEE), ROYI NYAMEKO3,4 , SRINIVAS ALUVALA5 , AND VRINCE VIMAL6,7, (Member, IEEE)
- [4] DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms , NORAH M. ALNAIM 1 , (Member, IEEE), ZAYNAB M. ALMUTAIRI2 , MANAL S. ALSUWAT 3 , HANA H. ALALAWI3 , ALJOWHRA ALSHOBAILI4 , AND FAYADH S. ALENEZI 5
- [5] Deep Fake Face Detection using Convolutional Neural Networks , Diviya N, Kaaviya Varshini E, Prakash P ,IEEE
- [6] Li, Y., Chang, M., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. IEEE International Workshop on Information Forensics and Security (WIFS). Retrieved from <https://doi.org/10.1109/WIFS.2018.8630761>.
- [7] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Retrieved from <https://doi.org/10.1109/ICCV.2019.00012>.
- [8] . Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2018). Do GANs Leave Artificial Fingerprints?. Proceedings of the 2019 IEEE Conference on Image Processing (ICIP). Retrieved from <https://doi.org/10.1109/ICIP.2019.8803319>.