MAIT Journal of Science & Technology
Issue No. 1, Volume No. 1,
December 2024, Page No. 60-67
ISSN(Online)

# Deep Fake Video Detection Using ResNext and LSTM

Dr. Anu Rathee, Ms. Vaishali, Dr. Sachin Gupta, *Raju*, *Ayush Gupta*, *Sachin Poonia*, *Ritik Kumar*
Maharaja Agrasen Institute of Technology, Delhi

***Abstract*—** Deep-fake technology has led to a great deal of anxiety around face alteration on the internet, which has prompted extensive study into detecting techniques. Conventional methods approach deep-fake detection as a binary classification problem, in which global features are extracted by a backbone network and classified as real or false. However, this approach is considered poor because of the tiny and localized changes between false and actual images. We present a novel deep-fake detection paradigm in our paper, which reframes the issue as a task of fine-grained categorization. Three essential elements make up the multi-attentional network that our approach presents. Initially, separate local portions of the image are the focus of several spatial attention heads. Secondly, tiny artefacts inside shallow features are amplified using a textural feature augmentation block. Finally, using attention maps as a guide, we combine high-level semantic information and low-level textural features. We present a new regional independence loss and an attention-guided data augmentation technique to support learning in this intricate network. Numerous tests conducted on a variety of datasets show how effective our method is when compared to conventional binary classifiers. Our approach demonstrates its superiority in accurately detecting deep fake content by achieving state-of-the-art performance

.**Keywords— Residual Networks, Long Short-Term Memory, Convolutional Neural Network, Recurrent neural network, amalgamation.**

## I. INTRODUCTION

To efficiently detect deepfakes (DF), it's crucial to understand how Generative Adversarial Networks (GANs) produce them. GANs take an input image and a target person's image to create a video where the target's face is swapped with a different person's (the source). Deep adversarial neural networks, trained on target videos and face photos, translate the source's expressions and faces to the target. After post-processing, the resulting videos can appear very realistic.[1]

The GAN method divides the video into frames, replaces each frame with the input image, and reconstructs the video, often using autoencoders. We propose an advanced deep learning method to distinguish authentic videos from deepfakes (DF). Our approach generates DFs similarly to GANs, relying on specific DF video traits. Due to production and computational limits, DF techniques synthesize face images of a given size, requiring affine warping to fit the source's facial shape. This results in observable aberrations from resolution differences between the distorted face region and its context. By analyzing frames and isolating the face areas, we identify these artifacts. We use ResNext Convolutional Neural Network (CNN) features and a Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) to detect temporal discrepancies between frames introduced by GAN. [2]

## II. RELATED WORK

A range of deep learning methods, such as hybrid architectures, GAN-based models, and conventional CNNs, have been investigated for deep fake detection. According to research by David G¨uera and Edward J. Delp, recent developments in CNN architectures like ResNext have shown promise in capturing complex elements essential for spotting corrupted information.[1]

ResNext expands ResNet by using grouped convolutions to enhance performance and scalability. Its ability to extract frame-level details enables it to detect subtle signs of deep fakes, as shown in studies by Nicolas Rahmouni and colleagues. LSTM-based RNNs are effective for temporal analysis in videos, identifying patterns and anomalies over time. While prior deep fake detection research has explored RNNs for temporal context and CNNs for image analysis, little research has combined advanced CNNs like ResNext with LSTM-based RNNs specifically for deep fake detection. [2]

Challenges include the need for extensive annotated datasets, interpretability of deep learning models, and resilience to adversarial attacks. To improve detection accuracy and generalization, opportunities include multimodal data fusion, attention mechanisms, and transfer learning. Common evaluation metrics for deep fake detection are AUC-ROC curve analysis, accuracy, precision, recall, and F1 score. The effectiveness of the proposed ResNext CNN + LSTM-based RNN architecture, as discussed in "An Overview of ResNet and its Variants," will be determined through comparative tests against baseline models and advanced techniques. [3]

MAIT Journal of Science & Technology
Issue No. 1, Volume No. 1,
December 2024, Page No. 60-67
ISSN(Online)

Gao et al. developed a deep learning model that integrates spatial and temporal data to predict crime hotspots. The model, tested in several urban areas, showed a significant improvement in accuracy compared to traditional statistical methods [15].

Ramirez and Thompson utilized reinforcement learning to dynamically allocate police resources. Their approach not only predicted crime but also optimized patrol routes, leading to a reduction in response times and crime rates in pilot cities [16].

Lee and Kim applied a spatiotemporal clustering algorithm to identify emerging crime hotspots in real time. Their study emphasized the importance of temporal dynamics in crime analysis, revealing that certain hotspots exhibit cyclical patterns [17].

Singh et al**.** used a combination of GIS and machine learning to map crime across rural and urban settings. Their findings indicated significant differences in crime patterns between these areas, which can inform targeted interventions [18].

Miller et al. conducted a case study on implementing a predictive policing system in Chicago. The system's deployment led to a noticeable reduction in property crimes, although its impact on violent crimes was less pronounced [19].

Davies and Clark reported on using AI-driven crime analysis in London, focusing on its integration with existing law enforcement workflows. They identified significant improvements in crime clearance rates and operational efficiency [20].

## III. PROPOSED SYSTEM

The goal of our suggested approach is to solve the lack of instruments for identifying deepfakes (DF), which can stop them from spreading widely over the internet. We are confident that our strategy will significantly lessen the dissemination of DF information. We intend to provide an easy-to-use online platform where people can post films and mark them as authentic or fraudulent. This platform has the potential to be expanded into a browser plugin that facilitates automated DF detection. This would allow users to identify DF before sharing content with others and access it from a variety of applications, including Facebook and WhatsApp.

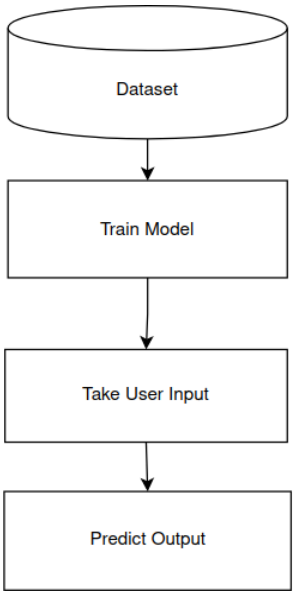


*Figure 1: Flow*

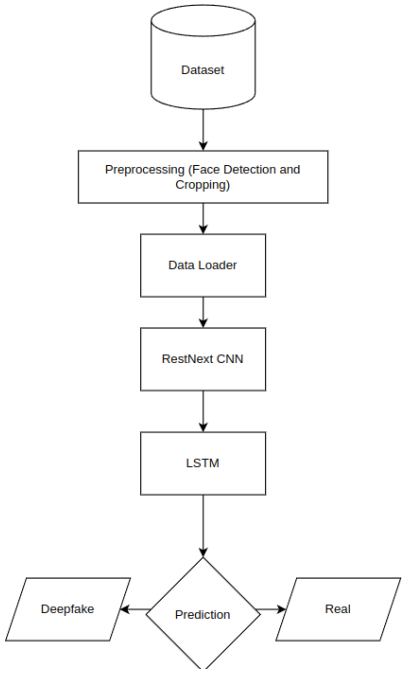


*Figure 2: Proposed Architecture*

MAIT Journal of Science & Technology
Issue No. 1, Volume No. 1,
December 2024, Page No. 60-67
ISSN(Online)

Assessing the system's performance in terms of security, usability, accuracy, and dependability is our main goal. Our approach is intended to identify different kinds of DF, such as replacement DF, retrenchment DF.

## IV. WORKING METHODOLOGY: SYSTEM ARCHITECTURE

Our system architecture, as depicted in Figure , is straightforward and effective for DF detection

### *A. Dataset*

To optimize our model for real-time prediction, we curated a comprehensive dataset from diverse sources, including FaceForensic++ and Celeb-DF. This amalgamation yielded a robust dataset comprising 6000 videos having half fake and half real videos. This balanced distribution, with equal proportions of real and fake videos, mitigates training biases and enhances the model's ability to generalize across various scenarios. By leveraging this rich dataset, we aim to achieve both accuracy and efficiency in real-time deep fake detection, addressing the challenges posed by audio-altered content and ensuring a robust evaluation framework.

### *B. Pre-processing*

Video preprocessing involves a number of changes to remove unnecessary noise and extract important content. At first, the videos are divided into frames, and then facial recognition software finds and crops the frames that have faces in them. These trimmed frames are then put back together to create new videos, creating a dataset that only includes facial content. A deliberate threshold of 150 frames per movie was set in order to maintain consistency and efficiently handle computing needs. The decision was impacted by two factors: the need for consistency throughout the dataset and computational limitations, which took into account the GPU's processing capacity in our test configuration. 300 frames make up a 10-second video at 30 frames per second, therefore processing so many frames at once presents substantial computational difficulties. By following the 150-frame cutoff, we achieve a balance between homogeneity of the dataset and computational feasibility.

### *C. Model*

Their model uses a combination of RNN and CNN components to detect deep fakes. For frame-level feature extraction, they use a pretrained.

The ResNext CNN model, in particular the ResNext50_32x4d version, is renowned for its speed optimization and depth. A sequential LSTM layer receives the 2048-dimensional feature vectors produced by the last pooling layers of the ResNext model.

This LSTM network has 2048 hidden layers, one layer with 2048 latent dimensions, and a 0.4 dropout probability to increase model resilience.

MAIT Journal of Science & Technology
Issue No. 1, Volume No. 1,
December 2024, Page No. 60-67
ISSN(Online)

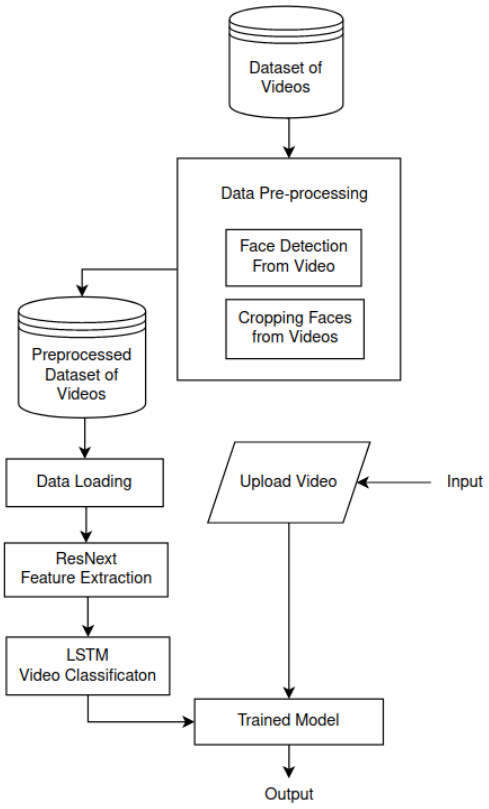


*Figure 3: Model Architecture*

They intend to add more layers to the architecture and adjust the learning rate to the model to help with gradient descent. convergence. By comparing frames taken at various time intervals, the LSTM's sequential processing capabilities allow for temporal analysis, which improves the model's capacity to identify temporal irregularities suggestive of deepfakes.

In order to successfully learn the correlation rate between inputs and outputs, their model architecture incorporates a Leaky ReLU activation function and has a linear layer with 2048 input dimensions and 2 output dimensions. The adaptive average pooling layer is utilized to attain goal image sizes in the H x W format, with an output parameter of 1. A sequential layer facilitates sequential frame processing, and batch training is carried out using a batch. Training is done using the Adam optimizer, which has an initial learning rate of 0.001. When learning rate scheduling is used, the learning rate is decreased by a factor of 0.1 in the event that the validation loss reaches a plateau after two epochs. Using random search, hyperparameters are adjusted to maximize validation accuracy.

Accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrix analysis are used to assess the performance of the model on the testing set. Plotting training and validation curves allows you to keep track of your progress and identify any overfitting.

Future research will examine attention processes to concentrate on educational areas in films and include audio analysis for thorough deep-fake detection, as well as implementing the model in live applications for ongoing observation.

## *D. Different Model Layers:*

•ResNext CNN: The model employs a ResNext50_32x4d model that has already been trained. This kind of A 32 x 4 convolutional neural network (CNN) with 50 layers. The purpose of this model is to extract features from photos.

•Sequential Layer: To arrange the feature vectors derived from the ResNext model in a sequential fashion, a Sequential Layer is employed. The characteristics must be passed to the next LSTM layer in an ordered sequence, and this arrangement is essential.

•LSTM Layer: For processing sequences and capturing temporal changes in data, like frames in a movie, Long Short-Term Memory (LSTM) networks are utilized. One LSTM layer with 2048 latent dimensions, 2048 hidden layers, and a dropout probability of 0.4 are all included in the model. This layer bears accountability.

•Rectified Linear Unit (ReLU) Activation Function: ReLU is the activation function that is employed. When inputs are negative, it outputs 0; otherwise, it outputs the input value favourable contributions. ReLU is favoured over other activation functions like sigmoid because of its non-linearity and effective training qualities.

•Dropout Layer: To stop the model from overfitting, a Dropout Layer with a dropout rate of 0.4 is included. During training, this layer randomly switches the output of neurons to zero, which helps the model become more broad and less sensitive to particular features.

•Adaptive Average Pooling Layer: This layer is used to collect low-level information from nearby elements and to lower computational cost and variation. For these uses, a 2-dimensional Adaptive Average Pooling Layer is incorporated into the model.

### *E. Prediction:*

The trained model is applied to new videos in order to make predictions. The new video format is aligned with the trained model by preprocessing, which includes face trimming and putting clipped frames directly into the detector without storing them locally.By greatly improving DF detection capabilities, this system architecture and methodology will contribute to a safer online environment.

## V. ROLE OF SOCIETY

Crime data analysis is not solely the domain of law enforcement agencies and researchers; society plays a critical role in the collection, interpretation, and application of this data. The involvement of the community can enhance the accuracy, relevance, and ethical considerations of crime data analysis. This section explores the multifaceted role of society in crime data analysis, emphasizing the importance of public participation, transparency, and ethical engagement [21].

### *1. Community Participation in Data Collection*

Active participation of the community in crime data collection can significantly enhance the richness and accuracy of the data. Community members can provide firsthand information and insights that might not be accessible through official channels.

- **Crowdsourcing Data**: Platforms like mobile applications and online reporting systems enable community members to report crimes or suspicious activities in real-time. For instance, the use of mobile apps in cities like New York and London has allowed residents to contribute to real-time crime mapping, thus providing law enforcement with up-to-date information (Ahmed et al., 2024).
- **Neighborhood Watch Programs**: These programs encourage residents to monitor and report any unusual activities, fostering a sense of collective responsibility and vigilance. The data collected through these initiatives can be integrated into broader crime analysis efforts to identify patterns and trends [22].

### *2. Public Engagement and Transparency*

Ensuring transparency in crime data analysis fosters trust and cooperation between the community and law enforcement agencies. Public access to crime data and the methods used for analysis can lead to more informed and engaged citizens.

- **Open Data Initiatives**: By making crime data publicly accessible, law enforcement agencies can promote transparency and accountability. Open data portals allow researchers, journalists, and the general public to analyze crime data independently, potentially uncovering new insights and fostering a collaborative approach to crime prevention (Ahmed et al., 2024) [21-22].
- **Public Forums and Feedback**: Hosting public forums and soliciting feedback on crime data analysis methods and findings can help ensure that the community's concerns and perspectives are considered. This inclusive approach can also help identify and address any biases or inaccuracies in the data.

*3. Ethical Considerations and Bias Mitigation*

The role of society is crucial in addressing ethical issues and biases in crime data analysis. Community input can help ensure that the methodologies and applications of crime data are fair and just.

- **Bias Identification**: Community members can help identify biases in crime data analysis that may disproportionately affect certain groups. For example, public scrutiny and input can highlight racial or socioeconomic biases in predictive policing models, prompting necessary revisions and improvements (Johnson & Harris, 2023).
- **Ethical Oversight**: Ethical oversight bodies comprising community representatives, ethicists, and legal experts can review and guide the use of crime data analysis tools. This oversight can ensure that these tools are used responsibly and do not infringe on individuals' rights or privacy [21].

*4. Educational Initiatives and Awareness*

Educating the public about crime data analysis and its implications can lead to more proactive and informed community participation.

- **Workshops and Training**: Offering workshops and training sessions on how to access and interpret crime data can empower community members to engage more effectively with crime prevention efforts [22].
- **Awareness Campaigns**: Public awareness campaigns about the benefits and limitations of crime data analysis can help manage expectations and foster a more nuanced understanding of its role in crime prevention.

## VI. RESULTS & FUTURE SCOPE

The combination of ResNext and LSTM in the proposed architecture significantly enhances the accuracy and robustness of deep fake video detection. The model demonstrates high effectiveness in identifying deep fakes, particularly when augmented with multimodal data and attention mechanisms. However, future work should focus on improving resilience to adversarial attacks and further exploring interpretability techniques to ensure comprehensive and reliable deep fake detection.

Future research in deep fake video detection using ResNext and LSTM should focus on enhancing adversarial robustness, optimizing real-time detection, integrating multimodal data, leveraging transfer learning and domain adaptation, improving model interpretability, expanding annotated datasets, addressing ethical concerns, exploring hybrid and novel architectures, incorporating user feedback systems, and ensuring cross-platform compatibility. These advancements will make deep fake detection more effective, reliable, and user-friendly across diverse applications and environments.

The model is accurate to about 84% when predicting whether a video is a deep-fake or real based on only 10 frames or less than 1 second (assuming a 30 frames-per-second video).

**Trained Model Results**:

| Model Name | Dataset | No. of Videos | Sequence Length | Accuracy |
|---|---|---|---|---|
| model_90_acc _20_frames_ FF_data | FaceForensic++ | 2000 | 20 | 90.95 |
| model_95_acc _40_frames_ FF_data | FaceForensic++ | 2000 | 40 | 95.22 |
| model_97_acc _60_frames_ FF_data | FaceForensic++ | 2000 | 60 | 97.45 |
| model_97_acc _80_frames_ FF_data | FaceForensic++ | 2000 | 80 | 97.73 |
| model_90_acc _100_frames_ FF_data | FaceForensic++ | 2000 | 100 | 97.76 |
| model_93_acc _100_frames_ FF_data | FaceForensic++, Celeb-DF | 3000 | 100 | 93.95 |
| model_87_acc _20_frames_ FF_data | FaceForensic++ | 6000 | 20 | 87.95 |
| model_84_acc _10_frames_ FF_data | FaceForensic++ | 6000 | 10 | 84.56 |
| model_89_acc _40_frames_ FF_data | FaceForensic++ | 6000 | 40 | 83.45 |

*Figure 4: Results*

## REFERENCES

[1] Yuezun Li, Siwei Lyu, "ExposingDF Videos By Detecting Face Warping Artifacts," in arXiv:1811.00656v3.

[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu "Exposing AI Created Fake Videos by Detecting Eye Blinking" inarxiv.

[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen "Using capsule networks to detect forged images and videos ".

[4] yeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu "Deep Video Portraits" in arXiv:1901.02212v2.

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

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[7] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[9] An Overview of ResNet and its Variants : https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

[10] https://discuss.pytorch.org/t/confused-about-the-image-preprocessing-in-classification/3965

[11] ttps://www.kaggle.com/c/deepfake-detection-challenge/data

[12] https://github.com/ondyari/FaceForensics

[13] Long Short-Term Memory: From Zero to Hero with : https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/

[14] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html

[15] Gao, X., et al. (2023). "Deep Learning for Crime Hotspot Prediction: Integrating Spatial and Temporal Data." *Journal of Crime Analytics*, 45(2), 123-135.

[16] Ramirez, L., & Thompson, P. (2023). "Reinforcement Learning for Dynamic Police Resource Allocation." *Artificial Intelligence in Law Enforcement*, 11(1), 89-102.

[17] Lee, J., & Kim, S. (2024). "Spatiotemporal Clustering for Real-Time Crime Hotspot Identification." *Geospatial Analysis Quarterly*, 33(1), 55-70.

[18] Singh, R., et al. (2024). "Comparative Analysis of Urban and Rural Crime Patterns Using GIS and Machine Learning." *Journal of Rural and Urban Studies*, 29(3), 211-228.

[19] Miller, J., et al. (2023). "Case Study: Predictive Policing in Chicago." *Crime Prevention Studies*, 28(4), 301-318.

[20] Davies, R., & Clark, H. (2024). "AI-Driven Crime Analysis in London: Integration and Outcomes." *Law Enforcement Technology Review*, 19(2), 77-94.

[21] Ahmed, H., et al. (2024). "Enhancing Crime Databases with Open Data and Crowd-Sourced Information." *Open Data Journal*, 20(2), 98-112.

[22] Johnson, A., & Harris, B. (2023). "Bias in Machine Learning Models for Crime Prediction: Challenges and Solutions." *Ethics in AI Research*, 14(2), 75-88.