

Multimodal Deepfake Detection using Meso4 CNN for Spatial Image Analysis and LRCN for Temporal Video Analysis

Mahalakshmi K.¹, Lavanya A.², Kiruthika R.³, Kannamma R.⁴

^{1,2,3}Student, ⁴Assistant Professor, Department of Artificial Intelligence and Data Science, Prathyusha Engineering College, Tamil Nadu, India.

E-mail: ¹kmahalakshmi2005@gmail.com, ²lavanyalavanya5080@gmail.com, ³kiruthikarosu10@gmail.com,

⁴hod.ai.ds@prathyusha.edu.in

Abstract

Generative artificial intelligence technology has resulted in the creation of deepfakes, highly realistic forged images and videos, posing major automated media authentication challenges. This research proposes a multimodal approach for deepfake recognition based on Meso4 CNN and Long Short-Term Convolutional Recurrent Network (LRCN). The proposed approach includes input validation, resizing, normalization, and data augmentation followed by feature extraction and real-fake classification for each modality. While Meso4 CNN is responsible for extracting mesoscopic facial features from images, LRCN analyzes the temporal inconsistencies in a sequence of video frames. The proposed system was tested on the FaceForensics++ dataset and analysed the performance metrics such as accuracy, precision, and recall. The proposed system has resulted in 94.1% accuracy, 93.5% precision, and 94.6% recall.

Keywords: Deepfake Forensics, MesoNet Convolutional Neural Network, Long Short-Term Convolutional Recurrent Network, Facial Manipulation Detection, Mesoscopic Feature Learning, Spatiotemporal Analysis, Sequential Video Frame Analysis

1. Introduction

The rapid advancement in generative artificial intelligence technology has allowed for the production of very realistic modified pictures and videos called deepfakes. Deepfake

content can manipulate facial identity, expressions, and appearance while maintaining realistic features. This has led to increased fears about deepfake content being used for purposes of misinformation, impersonation, and fraud [3], [4]. Deepfake detection involves the detection of small manipulations in the visual medium. In case of images, the artifacts result from spatial inconsistencies in the facial textures as well as local features, whereas in videos, there is an added feature of temporal inconsistency across the frame sequences. Deep neural networks have shown their strength in detecting spatial manipulation artifacts, whereas temporal modeling helps detect temporal inconsistencies [1], [2].

In this research, an approach to detect deepfakes based on Meso4 CNN for image detection and LRCN for video detection is developed. It consists of media verification, pre-processing, modality-specific feature extraction, classification, confidence score calculation, and result presentation. Such a combination allows the system to handle specific spatial and temporal features of fake images and videos.

2. Related Works

Deepfake detection has emerged as a popular research field, where the existing research works have concentrated on spatial artifacts, temporal inconsistencies, facial features, and traces of manipulation. Convolutional networks have revealed that synthetic images have unique artifacts, which are useful for the purpose of classification. In fact, convolutional network traces have revealed the success in analyzing the intrinsic features of generated images using CNN-based deepfake detection algorithms [1].

Temporal and spatial properties have been considered in video-based detection as well. Facial landmark detection utilizes the inconsistency in face geometry and movements in video sequences and is thus resilient to some distortions [2]. Video forgery detection using capsules has extended the study on recognizing forged images and videos via the hierarchy of manipulated facial information [7]. The above methods show that both spatial features and high-level facial information can be used as evidence for image/video manipulation recognition.

Various researchers have conducted studies on deepfake generation and detection approaches, pointing out issues such as dataset biases, manipulation diversities, cross-dataset generalization, and the changing nature of deepfake generation methods [3], [4], [5], [8]. Small

CNNs have also been developed to detect fake faces with the purpose of showcasing that mesoscopic features can be learned from fake faces [6]. Spatiotemporal methods proposed recently have proven the importance of using convolutional feature extraction and recurrent modeling for temporal inconsistencies in deepfake videos [9].

In spite of these advancements, reliable detection has been difficult because manipulation techniques, compression settings, and visual attributes vary from the ones that were seen during training. Thus, frequency domain analysis has been explored as a way of increasing the generalization capability of deepfake detectors by enabling changes in the manipulation settings [10]. The research analysis shows that learning spatial features is useful for detecting deepfakes in the image domain, whereas learning both spatial features and modeling temporally is useful for videos.

3. Proposed Methodology

The proposed framework adopts the modality-specific deepfake detection pipeline, where the input media undergoes validation and pre-processing prior to being classified for image detection and video detection. Images are subjected to Meso4 CNN for learning the manipulation features, while videos are processed via LRCN to recognize the spatial features and temporal dependency. Finally, the output is given to the user in terms of the classification result along with the confidence score. This entire process is illustrated in Figure 1.

3.1 User Input and Input Validation

The program receives the image or video as the first input from the user interface. The input validation module checks whether the received file is in the accepted media format and meets the constraints of the file. Such validation guarantees that the inappropriate input will not enter the detection system and later selects the proper processing path depending on the nature of the input.

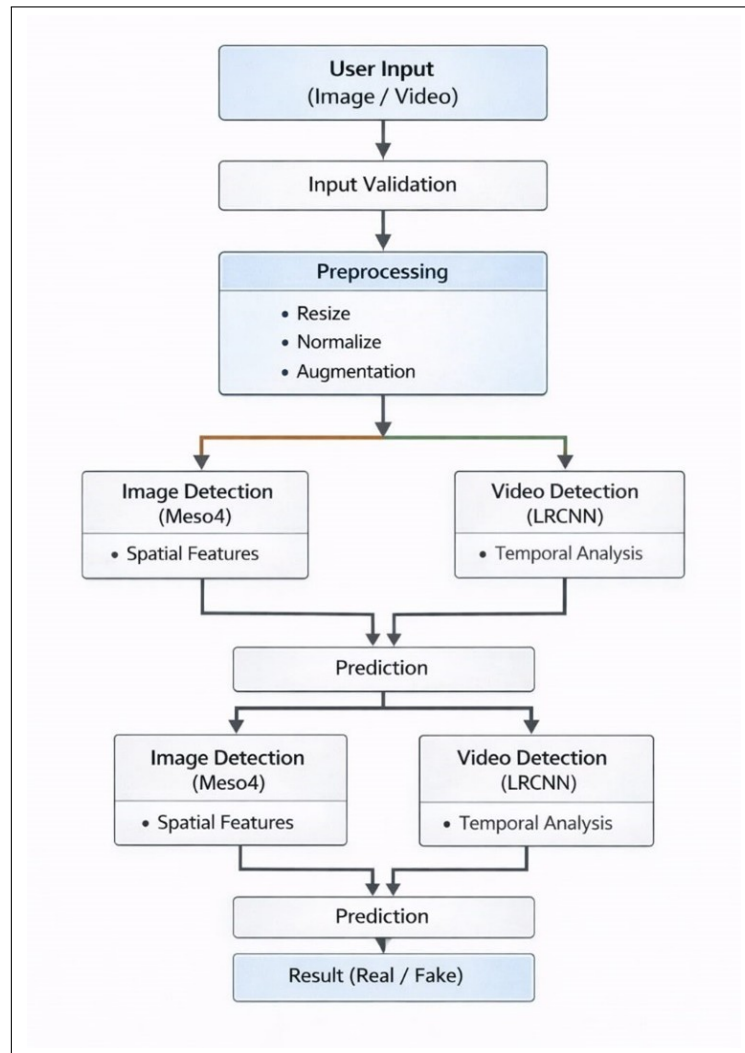


Figure 1. Workflow of the proposed deepfake detection system for image and video analysis

3.2 Dataset Description and Preprocessing

FaceForensics++ [11] dataset is chosen to serve as the benchmark dataset for the developed deepfake detection system. This dataset consists of 1,000 videos featuring facial videos extracted from YouTube and manipulated with DeepFakes, Face2Face, FaceSwap, and NeuralTextures. The dataset offers a built-in division of 700 videos for training, 150 for validation, and 150 for testing which ensures reliable modeling and testing process. The video format of the dataset corresponds to the proposed dual-modality method in which frames are extracted to perform spatial analysis with Meso4 CNN and the order of consecutive frames is used to perform temporal analysis with LRCN. The main attributes of the used dataset are presented in Table 1.

Table 1. Faceforensics++ dataset description

Dataset Component	Description
Dataset	FaceForensics++
Data Source	Original facial videos sourced from YouTube
Data Type	Image frames and video sequences
Original Videos	1,000
Manipulated Data	DeepFakes, Face2Face, FaceSwap, NeuralTextures
Training Split	70% (700 videos)
Validation Split	15% (150 videos)
Testing Split	15% (150 videos)
Image Branch	Extracted frames used for Meso4 spatial analysis
Video Branch	Sequential frames used for LRCN temporal analysis
Compression	Raw, high-quality (C23), and low-quality (C40) versions are available

Before proceeding with model processing, there are three steps of media preprocessing performed, which are demonstrated in Figure 1 below. These are resizing, normalization, and augmentation. The images and frames of the video are resized. If I denotes the original image or frame, the resized representation can be expressed as

$$I_r = \mathcal{R}(I, H, W) \quad (1)$$

where $\mathcal{R}$ denotes the resizing operation and H and W represent the target dimensions. Pixel normalization is subsequently applied to provide a consistent numerical range for model processing. A general normalized representation is given by

$$I_n = \frac{I_r - \mu}{\sigma} \quad (2)$$

where I_r is the resized input, and μ and σ denote the corresponding mean and standard deviation. Before the model processing, the media undergoes three preprocessing procedures,

which are illustrated in Figure 1 below. These are resizing, normalization, and augmentation. The images and frames of the video are first resized.

3.3 Image Deepfake Detection Using Meso4 CNN

For the image inputs, the preprocessed image goes into the Meso4 CNN that learns the mesoscopic spatial features of the face manipulations. The convolution layers capture the local patterns from the facial images, allowing the model to detect any anomalies caused by the deepfakes generation process.

The convolution operation for a feature map can be represented as

$$F_k = \sigma(I_n * W_k + b_k) \quad (3)$$

where I_n is the normalized input, W_k and b_k represent the k -th convolution kernel and bias, respectively, $*$ denotes convolution, and σ represents the activation function. The extracted feature maps serve as the spatial representation for the following classification layers.

3.4 Video Deepfake Detection Using LRCN

For video inputs, the temporal structure of the media is represented as an ordered sequence of frames:

$$V = \{I_1, I_2, \dots, I_T\} \quad (4)$$

where I_t denotes the t -th frame and T is the number of frames processed from the video. Each frame is processed individually at the convolutional feature extraction stage in order to convert its visual elements into the corresponding spatial feature representation. As a result, the feature vectors that are generated retain frame-based data concerning facial texture, appearance, blending, and other visual properties caused by Deepfake. Then, these spatial features are reorganized based on their original temporal order to form the sequence of features for recurrent processing.

The LSTM component of the LRCN receives the extracted feature vector x_t at each temporal step and updates its hidden representation according to

$$h_t = LSTM(X_t, h_{t-1}) \quad (5)$$

where X_t represents the spatial feature, vector extracted from frame t , h_{t-1} is the previous hidden state, and h_t represents the updated temporal representation. Due to this recurrent approach, the model is capable of retaining the necessary information from previous frames and at the same time considering the spatial features of the present frame.

The hidden representations successively encode the development of facial characteristics within the sequence and form the basis of the sequence discrimination. The resulting temporal representation is sent to the classification layer that predicts the Real/Fake label of the entire video. The confidence of the prediction is calculated based on the probability of the chosen class. Therefore, the LRCN model creates a sequential processing flow wherein spatial features are extracted by CNNs before the temporal dependencies are learned by LSTM, enabling Deepfake detection by both spatial and temporal discrepancies.

4. Results and Discussion

4.1 Experimental Setup

For testing of the system, the FaceForensics++ database was employed [11] by considering both real and artificial face videos. Python environment is used for implementation; for images, the Meso4 CNN architecture was used, whereas for videos, LRCN was utilized. In the preprocessing step, resizing, normalization, and augmentation were performed, and classification was done on the Real-Fake basis. Performance metrics like accuracy, precision, and recall were considered for the evaluation of the classification model. The experimental setup of the proposed model is presented in Table 2. However, due to limited hardware capabilities, only a part of the dataset was used.

4.2 Image Deepfake Detection Results

The Meso4 CNN has attained an accuracy of 94.1%, with precision at 93.5% and recall at 94.6%. This accuracy means that the model was able to classify most of the images presented as either fake or non-fake. The precision shows that most of the classified fake images were indeed manipulated, whereas recall shows that most of the manipulated images were detected.

Table 2. Experimental setup of the proposed system

Parameter	Configuration
Dataset	FaceForensics++ [11]
Programming Environment	Python
Input Modalities	Image and Video
Image Detection Model	Meso4 CNN
Video Detection Model	LRCN
Preprocessing	Resize, Normalization, Augmentation
Image Feature Analysis	Spatial Features
Video Feature Analysis	Sequential/Temporal Features
Classification Classes	Real, Fake
Evaluation Metrics	Accuracy, Precision, Recall

Figure 2 shows the image-based Deepfake detection procedure and the classification result for the given example. In this particular case, the Meso4 CNN classifies the input data as REAL with 56.41% confidence. As mentioned before, since the detection problem is treated as a binary classification, the model compares the scores of the two classes and assigns the data point to the one whose score is larger. Hence, REAL is assigned since the score for the REAL class is greater than the score for the FAKE class in this particular example.

Table 3. Comparison of existing and proposed deepfake detection approaches

Approach	Detection Model	Analysis Type	Key Feature Representation	Data Type	Reported Outcome
Afchar et al. [6]	MesoNet	Spatial	Mesoscopic facial features	Facial images/ videos	Compact CNN-based facial forgery detection
Chen et al. [9]	Xception-LSTM with spatiotemporal attention and ConvLSTM	Spatial + Temporal	Intra-frame and inter-frame features	Deepfake videos	Improved detection performance across three benchmark datasets
Proposed System	Meso4 CNN + LRCN	Spatial + Temporal	Spatial facial features and sequential frame features	Images + Videos	94.1% accuracy, 93.5% precision, 94.6% recall

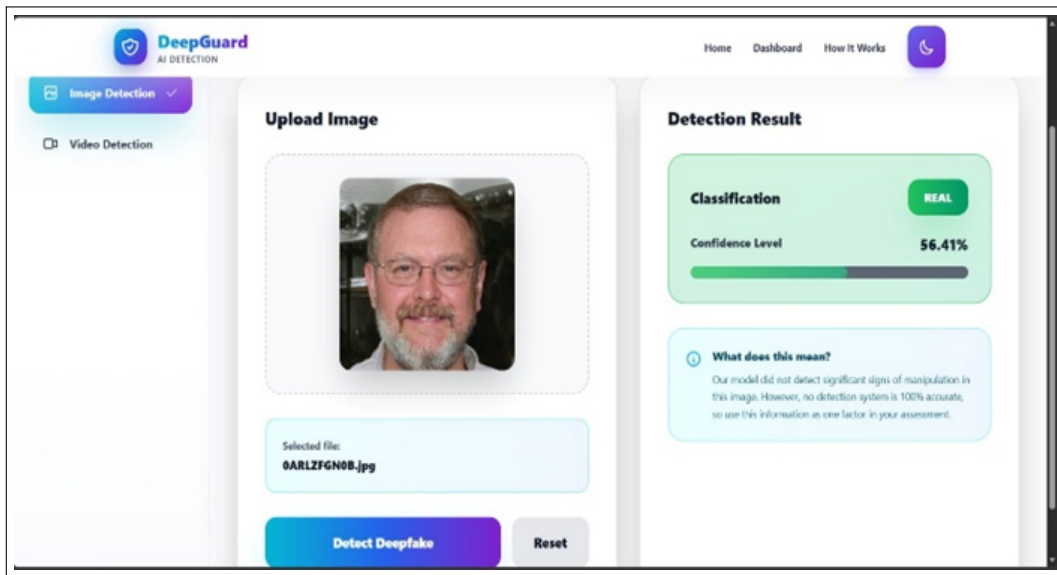


Figure 2. Meso4 CNN-based image deepfake detection interface and classification result

4.3 Video Deepfake Detection Results

The LRCN obtained an accuracy score of 93.2%, a precision of 92.6%, and a recall score of 93.8% for the task of detecting fake videos. The confusion matrix provided in Figure 3 presents a class-by-class assessment of the video classification results. From the confusion matrix, it is clear that out of a total of 151 Real videos, 76 were classified as Real, whereas out of a total of 150 Fake videos, 75 were classified as Fake. On the other hand, six Real videos were classified as Fake, and five Fake videos were classified as Real.

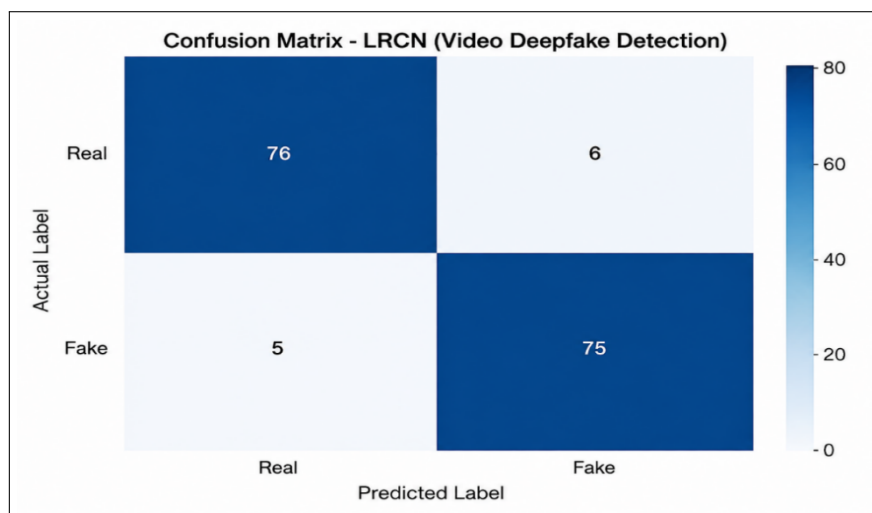


Figure 3. Confusion matrix of LRCN-based video deepfake detection

The accuracy will be obtained from the correctly classified instances, which are represented by the values along the main diagonal of the confusion matrix. In this particular case study, the correctly predicted instances are $76 + 75 = 151$ out of 162 instances used in the experiment for testing purposes. As such, the achieved accuracy value equals approximately 93.2%, which is coherent with the LRCN performance. In turn, the precision of approximately 92.6% means that the most of the predicted instances as Fake are actually Fake. In its turn, the recall of approximately 93.8% means that the most of the actual Fake instances were recognized.

In relation to Meso4 image branch, which obtained 91.5%, 90.8%, and 92.3% of accuracy, precision, and recall respectively, the LRCN branch delivered better results in all three-evaluation metrics in the experiment. This shows that the sequence of frame helps provide more evidences on Deepfake videos. Confusion Matrix below shows that Real and Fake classes are mainly classified correctly except a few predictions off-diagonal. The previous video detection results graph is also presented along with Figure 3 to show the working output of the LRCN video branch.

4.4 Overall Performance and Discussion

The integrated Meso4 CNN+LRCN network model has an accuracy of 94.1%, 93.5% precision, and 94.6% recall score, representing the best performance from all the models used in this study. These findings suggest that this system can handle both images and videos, with feature extraction for each modality separately.

This is evident by the roles played by the different branches. Meso4 concentrates on the spatial attributes of the facial images while LRCN operates on sequential data from video images. The combined architecture thus offers a single detection pipeline without compromising the separate approaches used in each type of media.

According to Table 3, the proposed model incorporates the analysis of space and time into one image-video detection approach. Nevertheless, it should be noted that the mentioned methods have been tested based on different datasets; thus, their results should be regarded as a comparison rather than a numerical benchmark for the present research.

5. Conclusion and Future Work

The proposed deepfake detection framework offers a comprehensive framework for the processing of both image and video manipulations by means of modality-specific deep learning algorithms. Meso4 CNN is used for spatial analysis of facial content, whereas the Long Short-Term Convolutional Recurrent Network extracts temporal information across sequential video frames. Input validation, preprocessing, spatial feature extraction, temporal modeling, and classification comprise an effective pipeline for deepfake analysis. The system exhibits the benefits of using complementary visual and temporal representations in the same detection framework. Possible directions for future research include generalizing across different datasets, considering a variety of manipulation methods including the newly emerging ones, developing better robustness to data compression and adversarial attacks, and implementing lightweight network architectures for real-time implementation. Additional research on multimodal and frequency-domain representations can further improve the performance of deepfake detection systems.

References

- [1] Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. "Deepfake detection by analyzing convolutional traces." In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2020, 2841-2850.
- [2] Li, Meng, Beibei Liu, Yongjian Hu, Liepiao Zhang, and Shiqi Wang. "Deepfake Detection using Robust Spatial and Temporal Features from Facial Landmarks." In 2021 IEEE International Workshop on Biometrics and Forensics (IWBF), IEEE, 2021, 1-6.
- [3] Rana, Md Shohel, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. "Deepfake Detection: A Systematic Literature Review." IEEE Access, 2022, vol. 10: 25494-25513.
- [4] Patel, Yogesh, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvala, and Vrince Vimal. "Deepfake Generation and Detection: Case Study and Challenges." IEEE Access, 2023, vol. 11: 143296-143323.
- [5] Edwards, Peter, Jean-Christophe Nebel, Darrel Greenhill, and Xing Liang. "A Review of Deepfake Techniques: Architecture, Detection, and Datasets." IEEE Access, 2024, vol. 12:

154718-154742.

- [6] Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "Mesonet: A Compact Facial video Forgery Detection Network." In 2018 IEEE international workshop on information forensics and security (WIFS), IEEE, 2018, 1-7.
- [7] Nguyen, Huy H., Junichi Yamagishi, and Isao Echizen. "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, 2307-2311.
- [8] Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes And Beyond: A Survey of Face Manipulation and Fake Detection." *Information Fusion*, 2020, vol. 64: 131-148.
- [9] Chen, Beijing, Tianmu Li, and Weiping Ding. "Detecting Deepfake videos Based on Spatiotemporal Attention and Convolutional LSTM." *Information Sciences*, 2022, vol. 601: 58-70.
- [10] Tan, Chuangchuang, Yao Zhao, Shikui Wei, Guanhua Gu, Ping Liu, and Yunchao Wei. "Frequency-Aware Deepfake Detection: Improving Generalizability Through Frequency Space Domain Learning." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 5: 5052-5060.
- [11] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to Detect Manipulated Facial Images." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 1-11. <https://github.com/ondyari/FaceForensics>