International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

Deep Fake Face Detection Using LSTM&CNN

Shaik Akramali
Department of ECE
Tirumala Engineering College
skakramaliskakramali63@gmail.com

Pagadala Ajay
Department of ECE
Tirumala Engineering College
pagadalaajay104@gmail.com

Bekkanti Akash
Department of ECE
Tirumala Engineering College
bekkantiakash@gmail.com

Shaik Moin
Department of ECE
Tirumala Engineering College
smoin0722@gamil.com

Syed Rabbani
Department of ECE
Tirumala Engineering College
syedrabbani@gamil.com

Abstract: The increasing sophistication of Deepfake technology necessitates robust detection methods. This paper proposes a Deepfake detection approach utilising Long Short-Term Memory (LSTM) networks to analyse temporal variations in facial geometric features. By extracting precise facial landmark coordinates over video frames, we capture subtle dynamic inconsistencies characteristic of manipulated content. These landmark sequences are transformed into feature vectors, which are then fed into an LSTM network designed to model temporal patterns and distinguish between genuine and forged videos. The efficacy of our method is demonstrated through experiments on publicly available datasets.

Keywords: Deepfake, LSTM Techniques, Temporal Feature Analysis, Video Manipulation Detection, Temporal Sequence Modelling

I. Introduction

With the rapid development of deep learning technologies, Deepfake videos have become increasingly realistic and difficult to identify with the human eye. These manipulated videos can spread misinformation and create serious issues in areas such as social media, journalism, and cybersecurity. Therefore, developing reliable methods to detect Deepfakes has become very important. In this project, a Deepfake face detection system is proposed using Long Short-Term Memory (LSTM) networks. The approach focuses on analysing facial geometric features by extracting facial landmark points from video frames. These landmarks represent important facial regions such as the eyes, nose, and mouth. By observing the changes in these landmarks across consecutive frames, temporal patterns of facial movements can be captured. Since LSTM networks are designed to handle sequential data, they are well-suited for learning these temporal dependencies. The extracted landmark sequences are converted into feature vectors and provided as input to the LSTM model to classify videos as real or fake. By detecting inconsistencies in facial motion patterns, the proposed method aims to effectively identify manipulated videos and improve the reliability of Deepfake detection systems.

Deepfake technology has grown rapidly with the advancement of artificial intelligence and deep learning techniques. These technologies can generate highly realistic fake videos by manipulating a person's facial expressions and appearance, making it difficult to distinguish between real and fake content. Such manipulated media can cause serious problems, including misinformation, identity misuse, and security threats. To address this issue, this project focuses on developing a Deepfake face detection system using Long Short-Term Memory (LSTM) networks. The proposed method analyses facial landmark points extracted from video frames to capture the geometric structure of the face. By tracking these landmarks over multiple frames, the system can observe temporal changes in facial movements. Since LSTM networks can learn patterns in sequential data, they help identify abnormal or inconsistent facial motion that may indicate manipulation. The extracted features are processed and fed into the LSTM model to classify the video as real or fake. This approach improves the ability to detect Deepfake videos by focusing on temporal facial behaviour rather than only visual appearance.

The increasing availability of powerful video editing tools and deep learning algorithms has made it easier to create highly convincing Deepfake videos. These manipulated videos can alter a person's face and expressions in a way that appears natural, which makes detection a challenging task. To address this challenge, this project focuses on detecting Deepfake faces by analysing the temporal behaviour of facial features using Long Short-Term Memory (LSTM) networks. The system first extracts facial landmark points from each frame of a video, which represent important facial regions such as the eyes, nose, and mouth. These landmarks are then arranged as a sequence of feature vectors that describe how the face changes over time. By learning these temporal patterns, the LSTM model can identify irregular movements or inconsistencies that are often present in manipulated videos. This method helps improve the accuracy of Deepfake detection by focusing on the natural motion patterns of facial features across video frames.

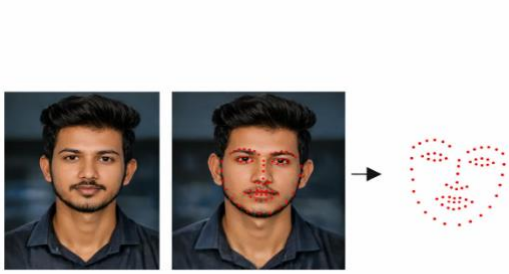

Fig. 1. Landmarks from video frames

II. Related Work

In recent years, Deepfake detection has become an important research area due to the rapid growth of manipulated digital media. Early detection techniques mainly relied on visual artefacts such as irregular blinking patterns, blurred regions, or image quality degradation to identify fake content. Although these approaches showed reasonable performance in controlled environments, they often failed to generalise well when different manipulation techniques or video compression methods were applied.

To overcome these limitations, recent studies have focused on analysing facial geometric features and temporal information. Facial landmarks, which represent key points on the face such as the eyes, nose, and mouth, provide valuable information about the spatial structure and movement of facial components. By tracking these landmark points across video frames, researchers can identify abnormal motion patterns that may indicate manipulation. Some studies have shown that analysing landmark trajectories improves detection performance under varying lighting and compression conditions.

Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have also been widely used for Deepfake detection because of their ability to process sequential data. LSTM networks can capture long-term dependencies in temporal sequences, making them effective for analysing changes in facial landmarks over time. Several research works have combined facial landmark extraction with LSTM-based architectures to detect temporal inconsistencies in manipulated videos. These approaches have demonstrated improved performance compared to traditional texture-based detection methods.

Despite these advancements, many existing models still struggle with issues such as computational complexity and limited generalisation across different datasets. Therefore, in this project, we focus on a Deepfake detection approach that uses facial landmark features combined with an LSTM network to analyse temporal facial movements. This method aims to effectively identify manipulated videos while maintaining good detection accuracy and robustness.

III. Methods

I. Proposed Method

Our Deepfake detection system is developed to identify subtle temporal irregularities in facial movements by analysing geometric facial features extracted from video sequences. The framework mainly consists of two key stages: **facial landmark extraction** and **temporal sequence analysis using a Long Short-Term Memory (LSTM) network**. The landmark extraction stage captures the spatial structure of the face, while the LSTM model studies the temporal variation of these features across consecutive frames to determine whether a video is authentic or manipulated.

A. Facial Landmark Extraction

The first stage of the proposed approach focuses on detecting facial landmarks from every frame of the input video. For this purpose, a reliable landmark detection technique is used. In this project, the **Dib 68-point facial landmark predictor** is adopted to locate important facial key points such as the eyes, eyebrows, nose, mouth, and jawline. Each detected landmark is represented using two-dimensional coordinates corresponding to its position in the image.

For a given frame at time step *t*, the facial landmarks are converted into a feature vector that contains the coordinates of all detected points. If the number of landmarks is *n*, the resulting feature vector consists of **2n values**, representing the horizontal and vertical positions of the landmarks.

To reduce the influence of variations in face size, camera angle, or distance from the camera, the extracted coordinates are normalised using the bounding box surrounding the detected face. This normalisation step improves the robustness of the model and helps maintain consistency across different video sources.

For a video containing **T frames**, the landmark vectors from all frames are arranged sequentially to form a temporal feature sequence. This sequence serves as the input for the LSTM network, allowing the model to analyse facial movements over time.

B. Long Short-Term Memory Network

To capture temporal relationships between facial landmark sequences, a **Long Short-Term Memory (LSTM)** network is employed. LSTM is a type of recurrent neural network that is specifically designed to process sequential data while preserving long-term dependencies.

Unlike traditional neural networks, LSTM models use internal memory cells and gating mechanisms to selectively store, update, and discard information during sequence processing. These mechanisms allow the model to learn important motion patterns in facial landmarks while ignoring irrelevant variations.

During training, the landmark sequence extracted from the video is fed into the LSTM model frame by frame. After processing the entire sequence, the final hidden state of the

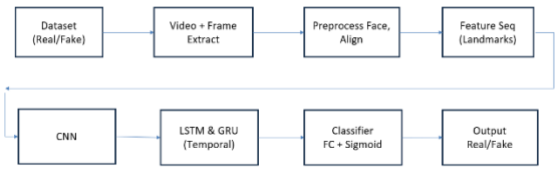

Fig: proposed methodology of the LSTM

The network represents a summarised description of the temporal facial dynamics present in the video.

This representation is then passed through a fully connected layer followed by a **SoftMax activation function**, which produces the probability of the video belonging to either the **real** or **fake** category.

The model is trained using **binary cross-entropy loss**, which measures the difference between the predicted output and the actual label of the video. Through this training process, the system learns to identify subtle temporal inconsistencies that commonly appear in manipulated videos.

II. Experiments and Evaluation

A. Dataset Description

To test the performance of the proposed Deepfake detection system, experiments were carried out using two widely used datasets: **UADFV** and **Face Forensics++ (FF++)**.

The **UADFV dataset** is one of the early datasets developed for Deepfake detection research. It contains a total of **98 videos**, including **49 real videos and 49 manipulated videos** created using Deepfake face-swapping techniques. Each video typically lasts around **11 seconds** and features a person speaking directly in front of the camera. Due to its relatively clean visual conditions and minimal compression artefacts, this dataset is suitable for evaluating lightweight detection methods based on geometric features.

The **Face Forensics++ dataset**, on the other hand, provides a much larger and more challenging collection of manipulated videos. It includes videos generated using multiple face manipulation techniques such as **Deepfakes, Face2Face, Face Swap, and Neural Textures**. In this work, a subset containing both real and Deepfake videos is used. The videos vary in length, generally ranging from **5 to 30 seconds**, and are available at different compression levels. This diversity makes FF++ useful for evaluating the robustness of detection models under realistic conditions.

B. Experimental Setup

All experiments were conducted using the **Pytorch deep learning framework** on a system equipped with an **NVIDIA RTX 3070 GPU with 8GB VRAM**.

The datasets were divided into **training and testing sets using an 80:20 split**, where 80% of the videos were used for training the model, and the remaining 20% were reserved for evaluation. This split ensures that the model learns from a wide range of examples while still being tested on previously unseen data.

The model training process uses **binary cross-entropy loss** as the objective function. Optimisation is performed using the **Adam optimiser** with a learning rate of **0.001**. The network is trained for **750 epochs**, and a batch size of **1024** is used to achieve stable learning and efficient convergence.

C. Performance Evaluation

To evaluate the effectiveness of the proposed method, its performance is compared with several existing deepfake detection models, including **LD-CNN, Capsule Network, and Maisonet**. Two commonly used evaluation metrics are used for comparison: **Accuracy (ACC)** and **Area Under the Curve (AUC)**.

Experimental results show that the proposed **LSTM-based detection approach** achieves strong performance on both datasets. The model reaches an accuracy of approximately **93.50% on the Face Forensics++ dataset** and **90.45% on the UADFV dataset**. Additionally, the AUC values achieved are **96.27% for FF++** and **94.90% for UADFV**.

These results demonstrate that analysing **temporal facial landmark patterns using LSTM networks** is an effective approach for identifying manipulated videos. The model shows good capability in distinguishing between genuine and forged content across different datasets.

IV. Existing System

Deep fake face recognition using a CNN-based approach involves leveraging the power of neural networks to detect manipulated media. The model is trained to distinguish between real and manipulated media. CNNs are particularly well-suited for image and video analysis tasks, making them a popular choice for deep fake detection. The data given as input and passed to the convolutional layers is in image format (pixels in a matrix).

A. Deep Fake Face Detection Model

The input image is loaded and pre-processed to standardise its size and format, and to remove any unwanted noise or artefacts that could interfere with subsequent processing steps. A set of features is extracted from the pre- processed image, such as texture, shape, and colour information.

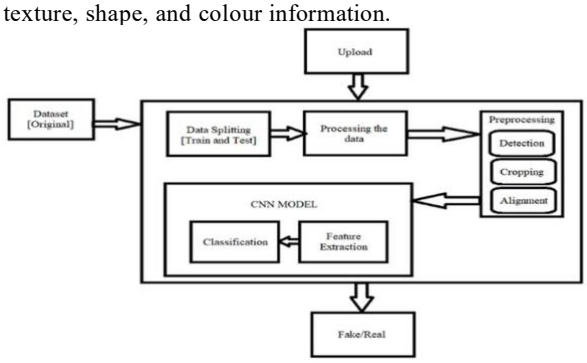

Fig. 1. Pipeline of Deep Fake Face Detection Model

These features are used to capture the characteristics of the image that are most relevant for distinguishing real and fake faces. The feature set is fed into a classification algorithm. Flickr's 140k dataset with 50K real faces and 50K deep fakes was used to train and validate the deep fake face detection model. Fig. 1 illustrates the processes such as data pre- processing, feature extraction and classification for detecting fake faces.

B System Architecture

The Deep Fake Face Detection model has five convolutional blocks and a classifier block. There are 13 convolution layers (Conv2D) followed by Pooling Layers, Activation Layers and Dropout Layers. Three Dense Layers and Fully Connected Layers are present in the classifier block. By applying convolutional filters to the input pictures, convolutional layers accomplish feature extraction, collecting regional patterns and structures. To make the feature maps less computationally difficult, pooling layers down samples them. Dropout layer randomly drops out a fraction of neurons in the hidden layers. the final

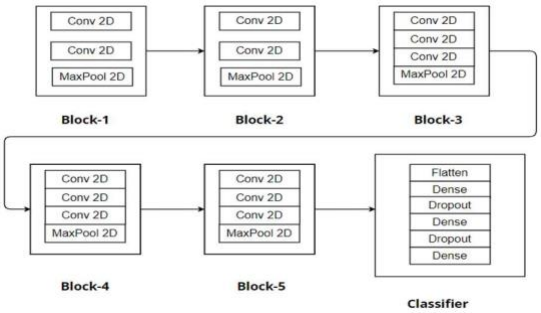

Fig. 2. Architecture of the Deep Fake Face Detection Model using CNN

The classification output is given by fully connected layers, which process the retrieved attributes. The CNN model gains non- linearity from activation functions, which enables it to understand intricate connections between input information and output predictions. Fig. 2 shows the architecture of the deep fake face detection model. The input and output are in image format. The preprocessed input data is fed into Block 1 for feature extraction, followed by dimensionality reduction and the introduction of non-linearities. Lower-level blocks detect simple and local features, such as edges and textures, while higher-level blocks combine these features to detect more complex patterns and objects. Each block captures more abstract and high-level features as we move deeper into the network.

.

1. Experiments And Analysis

A. Dataset Description

To assess the performance and reliability of the proposed Deepfake detection system, experiments were conducted using two well-known benchmark datasets: **UADFV** and **Face Forensics++ (FF++)**.

The **UADFV dataset** is one of the earlier datasets developed for Deepfake research. It contains **98 video clips**, including **49 authentic videos and 49 manipulated videos** generated using the Deepfake face-swapping technique. In these videos, individuals appear to speak directly toward the camera, mostly under frontal viewing conditions. Each clip has an approximate duration of **11 seconds**, which provides consistency across the dataset. The manipulation in this dataset mainly involves replacing the face of a target person with the synthesised face of another individual. Even though the dataset is relatively small, its clean recording conditions and limited compression artefacts make it suitable for evaluating geometry-based Deepfake detection methods.

On the other hand, the **Face Forensics++ (FF++) dataset** provides a more diverse and complex collection of manipulated videos. This dataset includes videos generated through several manipulation methods, such as **Deepfakes, Face2Face, Face Swap, and Neural Textures**. For our experiments, a subset containing both real and Deepfake videos was selected. These videos are available in multiple compression levels, and their durations vary between **5 and 30 seconds**, which helps simulate real-world video conditions. The variability present in FF++ makes it a valuable dataset for testing the robustness and generalisation ability of the proposed model.

B. Experimental Configuration

The implementation of the proposed Deepfake detection model was carried out using the **PyTorch deep learning framework**. All experiments were executed on a system equipped with an **NVIDIA GeForce RTX 3070 GPU with 8GB VRAM**.

To train and evaluate the model, the dataset was divided into **training and testing sets using an 80:20 ratio**. In this setup, 80% of the data is used to train the model, while the remaining 20% is used to evaluate its performance on unseen samples. This approach ensures that the model learns diverse facial patterns during training while maintaining fair evaluation during testing.

During training, **binary cross-entropy loss** is used as the objective function. The optimisation process is performed using the **Adam optimiser** with a learning rate of **0.001**. The network is trained for **750 epochs**, and a **batch size of 1024** is applied to improve the stability of the training process and achieve effective convergence.

C. Model Performance Evaluation

To measure the effectiveness of the proposed Deepfake detection approach, its performance was compared with several existing models, including **LD-CNN, Capsule Network, and Maisonet**. Two common evaluation metrics were used in this study: **Accuracy (ACC)** and **Area Under the Curve (AUC)**.

Experimental results indicate that the proposed **LSTM-based method** outperforms the other models on both datasets. The model achieves an accuracy of **93.50% on the FF++ dataset** and **90.45% on the UADFV dataset**. Similarly, the obtained AUC values are **96.27% for FF++** and **94.90% for UADFV**. These results demonstrate that the proposed approach is capable of identifying manipulated videos with high reliability.

The superior performance of the model can be attributed to its ability to capture **temporal relationships in facial landmark movements**. By focusing on dynamic facial patterns rather than relying only on static visual features, the model becomes more effective in identifying subtle inconsistencies introduced by Deepfake manipulation techniques.

Compared with traditional CNN-based methods that rely heavily on pixel-level textures, the proposed approach uses **geometric motion patterns derived from facial landmarks**. This makes the system less sensitive to variations such as lighting conditions, compression artefacts, or visual noise, allowing it to achieve better generalisation across different datasets.

D. Robustness Analysis

To further evaluate the stability of the proposed model, experiments were conducted under different video quality conditions: **RAW**, **High Quality (HQ)**, and **Low Quality (LQ)**. The experimental observations reveal that the LSTM-based detection model maintains strong performance across all quality levels. Even when the video quality is reduced, the model continues to achieve high accuracy compared to the other evaluated approaches.

In contrast, models that depend mainly on visual textures and pixel-level information tend to experience performance degradation as video quality decreases. Compression artefacts, blurring, or reduced resolution can significantly affect the reliability of such models.

The proposed approach addresses this issue by focusing on **facial landmark trajectories and motion-based geometric features**.

These high-level temporal patterns remain relatively stable even when visual quality is degraded. As a result, the model retains its ability to detect manipulated content under challenging conditions where appearance-based approaches may fail.

E. Ablation Study

An ablation experiment was conducted to understand the contribution of different components in the proposed system. The evaluation was performed on the **Face Forensics++ dataset under high-quality settings**.

Two modified versions of the model were created for comparison. In the first variant, detailed facial landmark extraction was replaced with simplified facial contour information. In the second variant, the LSTM network was replaced with a standard recurrent neural network (RNN).

The results of this study indicate that removing the landmark extraction stage leads to a noticeable reduction in detection performance. Similarly, replacing the LSTM network with a basic RNN also results in lower accuracy and AUC values.

These findings confirm that both **precise facial landmark extraction** and **LSTM-based temporal modelling** play a crucial role in improving the performance of the Deepfake detection system. When both components are combined, the model achieves the best overall performance, demonstrating the effectiveness of the proposed approach.

2. Results of Ablation Study

The results of the ablation experiments are presented in **Table III**. When the detailed facial landmark extraction stage is removed, and only basic facial contour information is used, the performance of the model decreases. In this case, the accuracy reduces to **90.21%,** and the **AUC score drops to 92.07%**. This shows that precise facial landmark points play an important role in capturing the geometric structure of the face, which helps the system detect manipulated content more effectively.

Similarly, when the **LSTM module** is replaced with a simple **RNN model**, the detection performance decreases further. The accuracy drops to **88.45%**, and the AUC value becomes **90.73%**. This result highlights the importance of the LSTM network, which is capable of learning long-term temporal dependencies in facial motion. These temporal patterns help identify subtle inconsistencies that often appear in manipulated videos.

The complete model that combines **facial landmark extraction with LSTM-based temporal analysis** achieves the best performance among all configurations. The model obtains an **accuracy of 93.50% and an AUC score of 96.27%**, demonstrating that both components significantly contribute to the overall effectiveness of the Deepfake detection system.

3. Conclusion

In this project, a Deepfake face detection approach based on **facial landmark analysis and Long Short-Term Memory (LSTM) networks** is presented. The system first extracts facial landmark coordinates from each frame of a video to represent the geometric structure of the face. These landmarks are then organised as a sequence and provided to an LSTM model, which learns the temporal relationships between facial movements across frames.

By analysing both spatial and temporal facial patterns, the proposed system is able to identify subtle inconsistencies that occur in manipulated videos. Experimental results show that the method achieves strong performance, reaching **93.50% accuracy and 96.27% AUC**, which demonstrates its capability in distinguishing between real and fake videos effectively.

Overall, the combination of **facial geometric features and temporal sequence modelling** improves the reliability of Deepfake detection and provides a promising solution for identifying manipulated media.

REFERENCES

[1] Zhiqing Guo, Gaobo Yang, Jiyou Chen, Xingming Sun (2021) “Fake face detection via adaptive manipulation traces extraction network” in Computer Vision and Image Understanding-Volume 204.

[2] Belhassen Bayar and Matthew C. Stamm (2016) “A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer” in IH & MM Sec '16: Proceedings of the 4thACM Workshop on Information Hiding and Multimedia Security.

[3] Richard Zhang et al. (2018), “Making Convolutional Neural Networks Shift-Invariant Again” in ICML 2019.

[4] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[5] Carlini, N. and Farid, H. (2020) “Evading deep-fake-image detectors with white-and black-box attacks” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[6] L. Tran, X. Yin, and X. Liu, “Representation learning by rotating your faces,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 41, pp. 3007–3021, 2018.

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “Star GAN v2: Diverse image synthesis for multiple domains,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2020, pp. 8188–8197.

[8] Zhou, P., Han, X., Morariu, V. I., and Davis, L. S. Learning rich features for image manipulation detection. In IEEE International Conference on Computer Vision (ICCV), 2018.

[9] Frank, Joel & Eisenhofer, Thorsten & Schoenherr, Lea & Fischer, Asja & Colossal, Dorothea & Holz, Thorsten. (2020). Leveraging Frequency Analysis for Deep Fake Image Recognition.

[10] McCloskey, S. and Albright, M. Detecting GAN-generated imagery using colour cues. arXiv preprint arXiv:1812.08247, 2018.

[11] Marra, F., Gragnani Ello, D., Verdolaga, L., and Poggi, G. Do GANs leave artificial fingerprints? In IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019.

[12] Yu, N., Davis, L. S., and Fritz, M. Attributing fake images to GANs: Learning and analysing GAN fingerprints. In IEEE International Conference on Computer Vision (ICCV), 2019.