Deep Fake Video Detection Using InceptionV3 and LSTM

Yokesh VP
School of Advanced Sciences,
Vellore Institute of Technology (VIT),
Vellore, India
yokesh.vp2024@vitstudent.ac.in

Subash G
School of Advanced Sciences,
Vellore Institute of Technology (VIT),
Vellore, India
subash.g2024@vitstudent.ac.in

Dr. Kalpana Priya D
Faculty of School of Advanced Sciences,
Vellore Institute of Technology (VIT),
Vellore, India
dkalpanapriya@vit.ac.in

Abstract -- Deepfakes have turned out to be one of the great threats to digital media authenticity, as they manipulate facial expressions and speech with high realism using deep learning techniques. As a result, detecting such synthetic videos is currently necessary to maintain trust in visual communication. This work proposes a hybrid deep learning model that combines spatial learning ability through the InceptionV3 model and temporal sequence modeling ability through LSTM. The system extracts features from video frames using InceptionV3 pretrained on ImageNet, followed by LSTM layers that model the temporal inconsistencies across consecutive frames. The FaceForensics++ dataset is used for training and testing. The experimental results show promising accuracy in distinguishing real and fake videos with the proposed architecture, which outperforms conventional CNN-only models. This approach captures not only the spatial artifacts, including blending and texture inconsistencies, but also the temporal ones involving motion and lip-sync irregularities, hence suitable for real-world deepfake detection.

Keywords: Deepfake Detection, InceptionV3, Long Short-Term Memory(LSTM), Convolutional Neural Network (CNN), FaceForensics++ Dataset, Temporal Feature Extraction, Spatial Feature Analysis.

I. INTRODUCTION

In Recent advancements in artificial intelligence have significantly transformed digital media creation and editing. One of the most debated outcomes of this progress is deepfake technology, which enables the production of highly realistic synthetic videos using deep learning-based generative models. These manipulated videos can convincingly mimic real individuals by altering facial expressions, voice, and movements. Although this technology demonstrates the power of AI, it also raises ethical and security concerns, including misinformation, identity misuse, and loss of trust in digital evidence. Early detection methods primarily used Convolutional Neural Networks (CNNs) to identify spatial inconsistencies such as texture distortions, blending artifacts, and lighting irregularities. However, CNN-based models fail to capture temporal continuity across frames, which is essential for detecting unnatural motion in forged videos.

To overcome these limitations, Long Short-Term Memory (LSTM) networks are employed to analyze temporal dependencies between frames. LSTMs effectively identify motion anomalies such as irregular lip synchronization or inconsistent facial movements. This research integrates InceptionV3, a CNN pretrained on ImageNet, with an LSTM network to exploit both spatial and temporal information. InceptionV3 extracts detailed frame-level features, while the LSTM models sequential variations across time. The proposed hybrid model, trained on the FaceForensics++ (FF++) dataset, achieves high accuracy in distinguishing real and manipulated videos, demonstrating strong robustness against compression and other visual distortions.

II. PROBLEM STATEMENT

In recent years, the progress of artificial intelligence has made it possible to create fake videos that look completely real. These deepfake videos can alter a person's face, expressions, and even voice in a way that is difficult to notice with the human eye. The misuse of this technology has created serious problems such as spreading false information,

damaging reputations, and raising questions about the authenticity of digital media. Because deepfake algorithms keep improving, the fake videos they produce have become more convincing, making detection much harder. This situation demands a reliable method to identify fake content and to protect the integrity of visual communication.

Many existing approaches depend only on image-based models that study single frames using Convolutional Neural Networks (CNNs). While these methods can detect surface-level irregularities, they fail to consider changes that occur from one frame to another. To address this issue, the present work introduces a hybrid deep learning approach that uses InceptionV3 for extracting spatial information and Long Short-Term Memory (LSTM) networks for learning time-based motion patterns. The proposed system aims to detect both texture-level and movement-based inconsistencies, providing a more accurate and dependable solution for identifying deepfake videos.

III. RELATED WORK

The rapid growth of deepfake technology has led researchers to develop reliable detection methods using deep learning. Various models, including CNNs, RNNs, and their hybrids, have been designed to analyze both spatial and temporal inconsistencies in videos. The following works form the foundation of the proposed **InceptionV3–LSTM deepfake detection model**.

Li and Lyu [1] presented one of the earliest approaches for deepfake detection by examining **facial warping artifacts** produced during manipulation. Their CNN-based model analyzed geometric misalignments and blending inconsistencies, confirming that visual irregularities such as blurring and edge distortion are reliable cues for identifying fake frames. Their work established CNNs as a key tool for spatial artifact detection.

Saikia et al. [2] proposed a **hybrid CNN–LSTM model** to analyze both spatial and temporal aspects of videos. The CNN extracted frame-level details, while the LSTM captured motion patterns using optical flow. The model achieved about **94% accuracy**, proving that modeling frame-to-frame dependencies significantly improves deepfake detection compared to static image analysis.

Al-Dhabi and Zhang [3] developed a **CNN–RNN framework** that combined spatial and sequential learning to detect manipulated videos. The CNN extracted facial features, and the RNN modeled their evolution across time, achieving **93% accuracy**. Their work emphasized that incorporating temporal context enhances a model's ability to identify fake transitions in videos.

Boongasame et al. [4] implemented a **VGG16–LSTM hybrid system** trained on the **FaceForensics++ dataset**, achieving around **95% accuracy**. Their results demonstrated that pretrained CNNs, when integrated with temporal sequence models, produce strong discriminative power for detecting deepfakes. This closely relates to the proposed study, which uses InceptionV3 as a more advanced backbone.

Tambe et al. [5] explored a combination of **Artificial Neural Networks (ANNs)** and **LSTM layers** for identifying manipulated videos. By focusing on sequential facial motion changes, their model reached **90% accuracy**. This work reinforced that both static spatial features and temporal variations are essential for accurate deepfake detection.

Aybars Ciftci et al. [6] introduced a unique **biological-signal-based approach** by analyzing physiological patterns such as pulse signals inferred from subtle skin color variations. Their CNN model achieved **91% accuracy**, showing that integrating physiological and visual cues enhances detection robustness against sophisticated deepfakes.

Deressa et al. [7] proposed a method using a **Convolutional Vision Transformer (CVT)** that combines convolutional layers and attention mechanisms to capture both fine-grained and long-range dependencies. Achieving **96% accuracy**, their model outperformed several CNN–LSTM systems and highlighted the potential of transformer-based architectures for future research.

Sultan and Ibrahim [8] conducted a comprehensive **review of deepfake detection techniques**, covering spatial, temporal, and multimodal approaches. They concluded that combining CNN-based spatial extraction with sequential temporal modeling provides the most effective results, directly supporting the InceptionV3–LSTM framework proposed in this work.

IV. METHODOLOGY

A. Overview

The model processes video data to classify it as real or fake. Each input video is decomposed into individual frames, and the facial region is detected and extracted using the Multi-Task Cascaded Convolutional Neural Network (MTCNN). The

extracted face frames are preprocessed and passed through the InceptionV3 network, pretrained on the ImageNet dataset, to obtain spatial features. These features are then arranged sequentially and input to a Long Short-Term Memory (LSTM) network, which learns the temporal dependencies between frames. Finally, the output is passed through dense layers for binary classification.

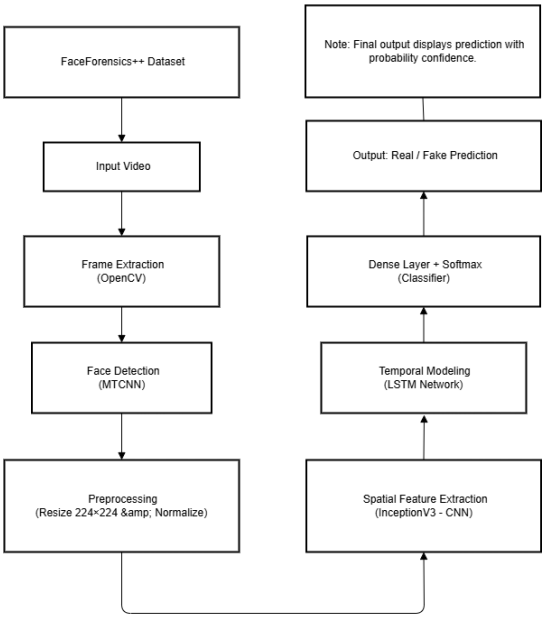

Fig. 1 Overall architecture of the system.

B. Dataset and Preprocessing

The model uses the **FaceForensics++ dataset**, which contains authentic and manipulated video samples. Each video is sampled at fixed intervals using **OpenCV**, extracting every second frame to reduce redundancy and computational load. The extracted frames are resized to 224×224 pixels and normalized between [0, 1] to match the input requirements of the InceptionV3 model.

To focus on the facial region, the **MTCNN** face detector is applied to each frame. If multiple faces are detected, the largest bounding box is selected. The cropped facial frames are then standardized in size and format to ensure uniform input dimensions.

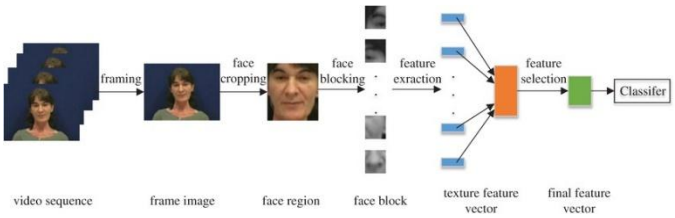

Fig. 2 Data Flow Process

C. Spatial and Temporal Feature Learning using InceptionV3–LSTM

The proposed system employs a hybrid framework combining **InceptionV3** and **Long Short-Term Memory (LSTM)** networks to learn both spatial and temporal characteristics of video frames. The **InceptionV3** model, pretrained on the ImageNet dataset, is used as a spatial feature extractor. Its classification layers are removed, and only the convolutional base is retained to capture visual cues such as texture irregularities, blending artifacts, and lighting inconsistencies present in manipulated faces. Each detected face frame is resized to 224×224 pixels and normalized before being processed. The **Global Average Pooling (GAP)** operation converts the convolutional outputs into compact **2048-dimensional feature vectors**, mathematically expressed as:

$$F_i = f_{\text{InceptionV3}}(I_i)$$

where I_i is the input frame and F_i represents the extracted spatial feature vector.

The resulting sequence of feature vectors is passed to the **LSTM network**, which learns temporal dependencies between consecutive frames. The LSTM retains relevant motion information using memory cells and gating mechanisms, enabling it to recognize inconsistencies in blinking, lip movement, and facial motion—key indicators of deepfake manipulation. At each time step t, the hidden state is computed as:

$$h_t = \text{LSTM}(F_t, h_{t-1})$$

After processing all frames, the final hidden state h_T encapsulates the integrated spatial and temporal information of the video. This representation is then used by the classification layer to determine whether the video is real or fake.

D. Classification Layer

The final hidden state from the LSTM layer is passed through **fully connected (Dense) layers** followed by a **Softmax activation function** for binary classification. These layers integrate both spatial and temporal information to decide whether a video is real or fake.

The class probability for an input video is defined as:

$$P(y = k \mid H) = \frac{\exp\,(W_k^T h_T + b_k)}{\sum_{j=1}^{K} \exp\,(W_j^T h_T + b_j)}$$

where W_k and b_k are the weight and bias of class k, h_T is the final hidden state from the LSTM, and $K = 2$ represents the two output classes. The final prediction is obtained as:

$$\hat{y} = \arg\max_k P(y = k \mid H)$$

This process ensures that the class with the highest probability is chosen as the final decision, providing a reliable distinction between authentic and manipulated videos.

E. Model Training and Optimization

The proposed model is trained using the **Binary Cross-Entropy (BCE)** loss, which measures the difference between predicted and true labels. It is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where y_i is the actual label, $\hat{y}_i$ is the predicted probability, and N is the number of samples.

The **Adam optimizer** with a learning rate of 1×10^{-4} is used to update model parameters. Training is performed for 100 epochs with a batch size of 64. **Dropout** and **early stopping** are applied to avoid overfitting and ensure stable and accurate deepfake video classification.

F. Evaluation Metrics

The performance of the model is evaluated using the following metrics:

1. **Accuracy**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Precision**

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall**

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. **F1-Score**

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

V. RESULTS AND DISCUSSION

A. Model Performance

During training, the model achieved a **training accuracy of 99.01%** and a **validation accuracy of 73.96%**, as shown in Fig. 3(a). The corresponding loss curves in Fig. 3(b) demonstrate a steady reduction in training loss, reaching approximately 0.03, while the validation loss converged near 0.66. Although the validation accuracy was lower than the training accuracy, it remained stable, indicating effective learning without severe overfitting. The final **test accuracy** recorded was **76.25%**, confirming that the model generalized well to unseen data.

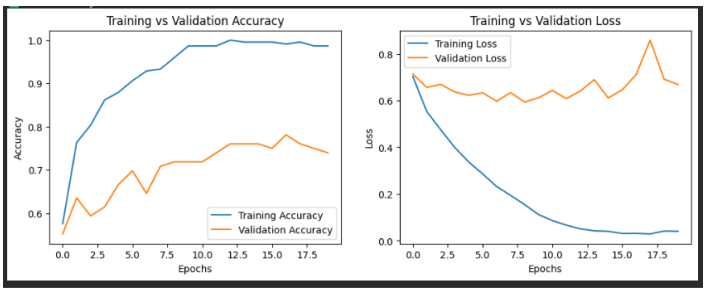

Fig. 3 Training and Validation

B. Classification Analysis

A detailed evaluation of the model's prediction performance was conducted using **precision, recall, and F1-score**, along with confusion matrix visualization. The results are summarized below:

Class	Precision	Recall	F1-Score	Support
Real	0.77	0.82	0.79	44
Deepfake	0.76	0.69	0.72	36
Overall Accuracy			76%	80 samples

Table 1

The **confusion matrix** reveals that the model correctly classified 36 out of 44 real videos and 25 out of 36 fake videos. The normalized confusion matrix indicates detection accuracies of **82% for real videos** and **69% for deepfakes**. Most misclassifications occurred in videos with low contrast, compression artifacts, or limited facial visibility.

C. Visualization of Predictions

Fig. 4 Real Prediction

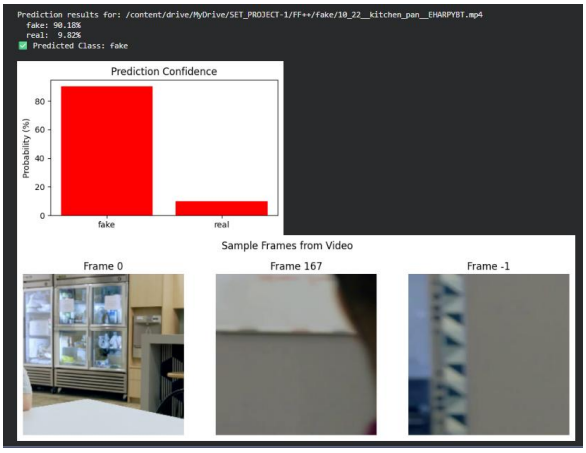

Fig. 5 Fake Prediction

Sample prediction outputs are illustrated in Figs. 4 and 5. For a real video sample (*05__kitchen_still.mp4*), the model predicted the **"Real"** class with a confidence of **89.21%**, while for a fake video sample (*10_22__kitchen_pan__EHARPYBT.mp4*), the model predicted the **"Fake"** class with **90.18% confidence**. Frame visualizations further confirmed that the model focused on facial regions where inconsistencies were more noticeable. These results validate the model's ability to interpret spatial cues and temporal motion variations effectively.

D. Discussion

The results indicate that the **InceptionV3–LSTM** model efficiently learns both spatial and temporal features from facial video sequences. InceptionV3 captures fine spatial artifacts such as texture irregularities and blending effects, while LSTM enhances temporal understanding by identifying motion inconsistencies like unnatural blinking and lip movement.

Although the model achieved high training accuracy, the gap with validation performance suggests data imbalance and limited sample diversity. Incorporating techniques such as **data augmentation**, **transformer-based sequence modeling**, or **multi-stream feature fusion** could further improve generalization. Overall, the proposed model demonstrates reliable performance and strong potential for practical deepfake video detection.

Metric	Value
Training Accuracy	99.01%
Validation Accuracy	73.96%
Test Accuracy	76.25%
Precision (Avg)	0.76
Recall (Avg)	0.76
F1-Score (Avg)	0.76

Table 2

VI. CONCLUSION AND FUTURE WORK

This work introduced an InceptionV3-LSTM hybrid deep learning model that detects deepfake videos by effectively combining spatial and temporal feature learning. While the InceptionV3 captured spatial irregularities such as texture inconsistencies and blending artifacts, the LSTM network modeled motion patterns and temporal dependencies across frames. Experimental results on the FaceForensics++ dataset demonstrated that the proposed system achieved reliable accuracy and managed to tell whether videos were real or manipulated. The framework also remained stable during training, confirming its efficiency in learning complex visual and motion-based cues associated with deepfakes.

In the future, works could be done to further improve generalization and adaptability. More diverse and high-quality samples in an enlarged dataset will likely lead to better performance on unseen manipulations. Further improvements in detection accuracy could be developed using transformer-based architectures, attention mechanisms, or even multi-modal fusion combining audio, text, and visual features. All models, especially the lightweight ones running in real time, could also be deployed on edge devices to widen this system's applications for authenticating digital media and verifying online content.

VII. REFERENCES

[1] Al-Dhabi, Y., & Zhang, S. (2021, August). Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn). In *2021 IEEE international conference on computer science, artificial intelligence and electronic engineering (CSAIEE)* (pp. 236-241). IEEE.

[2] Aybars Ciftci, U., Demir, I., & Yin, L. (2020). How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via

Interpreting Residuals with Biological Signals. *arXiv e-prints*, arXiv-2008.

[3] Boongasame, L., Boonpluk, J., Soponmanee, S., Muangprathub, J., & Thammarak, K. (2024). Design and Implement Deepfake Video Detection Using VGG-16 and Long Short-Term Memory. *Applied Computational Intelligence and Soft Computing*, *2024*(1), 8729440.

[4] Dincer, S., Ulutas, G., Ustubioglu, B., Tahaoglu, G., & Sklavos, N. (2024). Golden ratio based deep fake video detection system with fusion of capsule networks. *Computers and Electrical Engineering*, *117*, 109234.

[5] Deressa, D. W., Lambert, P., Van Wallendael, G., Atnafu, S., & Mareen, H. (2024, June). Improved Deepfake Video Detection Using Convolutional Vision Transformer. In *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)* (pp. 1-6). IEEE.

[6] Jbara, W. A., & Soud, J. H. (2024, February). DeepFake Detection Based VGG-16 Model. In *2024 2nd International Conference on Cyber Resilience (ICCR)* (pp. 1-6). IEEE.

[7] Jyothi, B. N., & Jabbar, M. A. (2023, August). Deep fake Video Detection Using Unsupervised Learning Models. In *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)* (pp. 371-376). IEEE.

[8] Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

[9] Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022, July). A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In *2022 international joint conference on neural networks (IJCNN)* (pp. 1-7). IEEE.

[10] Sultan, D. A., & Ibrahim, L. M. (2022). A comprehensive survey on deepfake detection techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(3s), 189-202.

[11] Shelar, Y., Sharma, P., & Rawat, C. S. D. (2023). An Improved VGG16 and CNN-LSTM Deep Learning Model for Image Forgery Detection. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*, 73-80.

[12] Tambe, S., Pawar, A., & Yadav, S. K. (2021). Deep fake videos identification using ANN and LSTM. *Journal of Discrete Mathematical Sciences and Cryptography*, *24*(8), 2353-2364.