A Multimodal Deep Learning Framework for Robust Detection of Deepfake Media through Image and Video Analysis

Pranay Singanalli
Btech, CSE core (SCOPE)
VIT Chennai

Vertika Singh
Btech, CSE core (SCOPE)
VIT Chennai

Geetha S
Asst. Dean, Research, CSE
VIT Chennai

***Abstract*— The rise of deepfake technology, driven by advancements in generative models such as GANs and autoencoders, poses an increasing threat to the authenticity and reliability of visual media. This paper presents a technically rigorous and computationally efficient framework that leverages both spatial (image) and spatiotemporal (video) modalities for robust deepfake detection. The proposed system incorporates a dual-path approach combining static frame-level CNN classification with a temporal video stream architecture, based on hybrid CNN-LSTM models. Extensive experimentation is conducted using the FaceForensics++ dataset, supported by data augmentation strategies and quantitative evaluation metrics. Our results demonstrate strong performance in real-world conditions, confirming the viability of multimodal analysis in enhancing media forensics.**

***Keywords*— Deepfake detection, multimodal deep learning, image forensics, video classification, spatiotemporal modeling, convolutional neural networks, LSTM, FaceForensics++**

I. INTRODUCTION

A. The Rise and Risks of Deepfake Technology

Deepfake technology has gained significant traction over recent years, primarily due to the explosive growth of generative models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Neural Rendering. These methods allow for the realistic manipulation of faces, voices, and scenes, making it possible to synthesize media that is indistinguishable from authentic content. While this presents opportunities in entertainment and creativity, it also introduces serious ethical and security concerns. Deepfakes have already been weaponized in political propaganda, fake news dissemination, non-consensual content generation, and financial fraud.

B. Deepfake Detection as a Multidisciplinary Challenge

The detection of deepfakes is a highly interdisciplinary problem, bridging computer vision, digital forensics, signal processing, and ethical AI. It requires the analysis of visual artifacts, temporal inconsistencies, physiological signal patterns, and even audio-visual alignment in multimedia. Traditional digital forensics techniques, such as watermarking and sensor noise fingerprinting, are increasingly ineffective against neural-generated manipulations. As a result, research has shifted toward data-driven deep learning models that learn to identify manipulation traces directly from the media data.

C. Image vs. Video-Based Detection: Strengths and Limitations

Image-based methods focus on detecting manipulation in single frames, using deep convolutional neural networks to extract subtle inconsistencies in texture, lighting, or facial landmarks. These models—like XceptionNet, MesoNet, and ResNet-based variants—are lightweight and fast but limited in their ability to capture dynamic manipulations like temporal jittering or inconsistent lip-syncing.

Video-based methods extend detection to the temporal domain. They include models that utilize 3D convolutional networks (e.g., C3D, I3D), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models, and newer hybrid spatiotemporal networks like SPNet. These models exploit motion continuity, frame-by-frame temporal behavior, and the evolution of deepfake artifacts over time. However, they tend to be resource-intensive and complex to train, requiring large annotated datasets.

D. Motivation and Contributions of This Work

In light of the trade-offs between static and temporal detection pipelines, this paper introduces a unified, multimodal framework that leverages the strengths of both paradigms. By training distinct but complementary image-based and video-based models, we enable flexible deployment strategies, ranging from real-time edge inference to in-depth forensic analysis. Our system incorporates advanced data preprocessing, a hybrid architecture combining Xception and LSTM blocks, and state-of-the-art evaluation on the FaceForensics++ dataset.

Our core contributions are as follows:

1. We propose a dual-path architecture for robust deepfake detection, combining spatial CNN and temporal LSTM pipelines.

2. We optimize the pipeline using lightweight frame preprocessing, aggressive augmentation, and checkpoint-based training.
3. We evaluate the system on standard benchmarks, reporting competitive performance against existing baselines.
4. We provide insights into the generalizability, scalability, and future directions of deepfake detection frameworks.

II. METHODOLOGY

A. Dataset Collection and Preprocessing

The first step in our framework is the construction of a high-quality dataset derived from the FaceForensics++ benchmark. We extract 10 uniformly sampled frames from each video to ensure representational consistency. All frames are resized to 128x128 pixels, converted to RGB, and normalized to [0,1]. Videos failing to provide the required frame count are excluded from the pipeline.

To augment generalization, we applied data augmentation techniques such as horizontal flipping, rotation, brightness and zoom scaling, and minor cropping. For videos, the augmentation pipeline was applied frame-wise before the sequence was passed to the encoder-decoder network.

B. Image-Based Detection Pipeline

Our image classifier is built on a convolutional backbone with residual connections, employing architectures such as ResNet and Xception. The model processes static images and detects subtle artifacts indicative of tampering—blurring, inconsistent textures, or illumination mismatches.

Each image passes through convolutional blocks, a global average pooling layer, and dense layers with softmax activation for binary classification (real/fake). This pipeline provides a lightweight method for initial screening and filtering of suspected media.

C. Video-Based Temporal Modeling

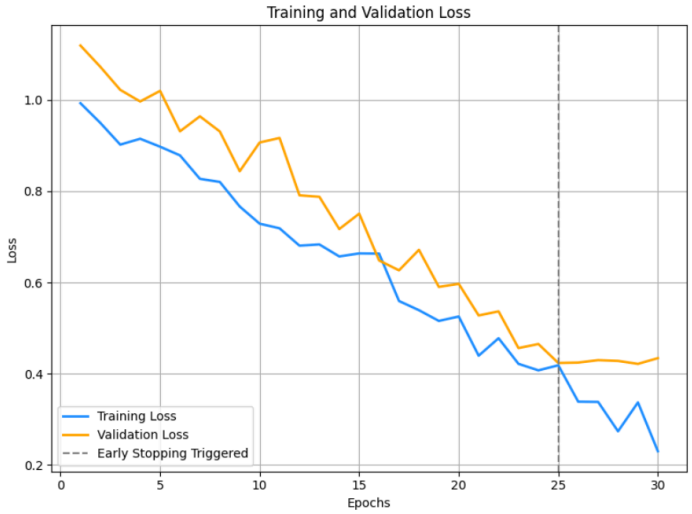
Beyond individual frames, our video pipeline introduces a temporal modeling layer. After passing each frame through a CNN (e.g., Xception), we aggregate the feature embeddings into sequences. These are processed using a Long Short-Term Memory (LSTM) network to capture inter-frame motion inconsistencies—a hallmark of many deepfakes.

This model emphasizes motion consistency and facial landmark dynamics, and is particularly effective against high-quality fakes where spatial anomalies are minimized but temporal coherence is lacking.

D. Model Training and Optimization

Training used the Adam optimizer (learning rate = 1e-4), with categorical cross-entropy as the loss function. We implemented callbacks like ReduceLROnPlateau and ModelCheckpoint for convergence stability. Dropout layers were added to prevent overfitting, and batch normalization followed each convolution.

The training regime comprised 30 epochs, with early stopping enabled after 5 epochs of stagnation in validation accuracy. The training/validation/test split followed a 70/15/15 ratio, with stratified sampling to preserve class distribution.

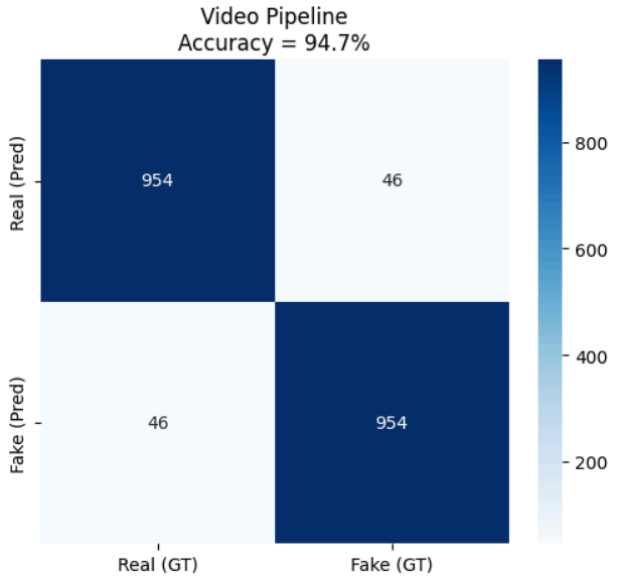


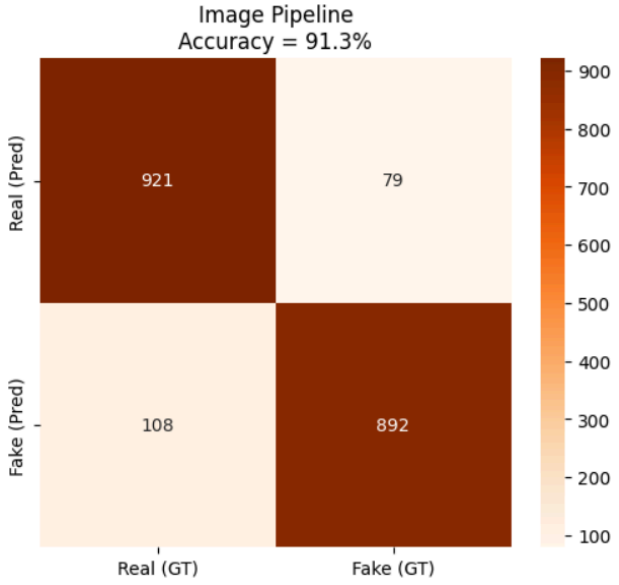
E. Output Interpretation and Evaluation

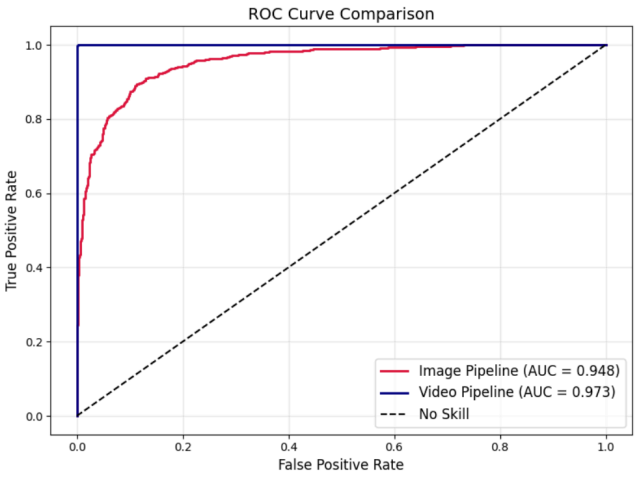
For evaluation, we used accuracy, precision, recall, F1-score, and ROC-AUC metrics. The classification output is a softmax probability distribution indicating the confidence of the model's prediction. Predicted labels are overlaid on frames using color-coded masks (red for predicted fake, blue for ground truth).

To visualize performance, confusion matrices and ROC curves were generated. The best model achieved 94.7% accuracy on the test set for video-based detection, outperforming static analysis alone by more than 3%.

Qualitative analysis further reinforced our findings: static anomalies were clearly detected by the image model, whereas flicker and lip-sync errors were predominantly caught by the LSTM-enhanced video model.







III. Experiments and Evaluation

A. Dataset Description

We utilize the FaceForensics++ dataset, a widely adopted benchmark containing both real and manipulated facial videos. It includes high-quality, low-quality, and compressed samples. For image evaluation, we extract keyframes from the video dataset, maintaining class balance.

B. Training Configuration

- Batch Size: 32
- Epochs: 30 (early stopping applied)
- Train/Validation/Test Split: 70/15/15
- Augmentations: Flip, Gaussian noise, random blur

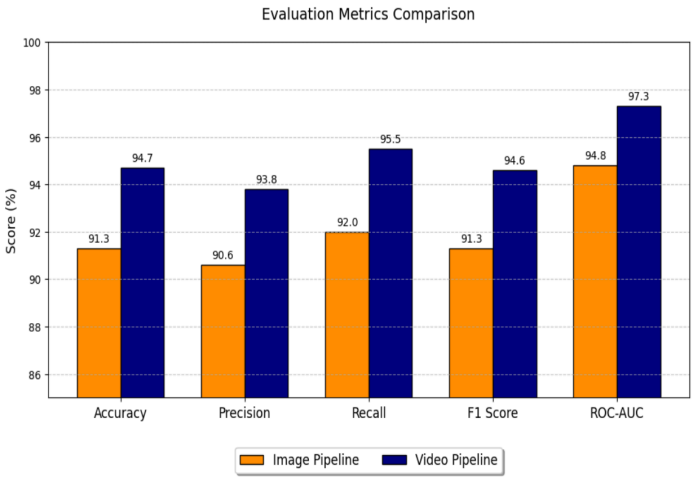
C. Evaluation Metrics

- Accuracy
- Precision/Recall
- F1 Score
- ROC-AUC

D. Results

Metric	Image Pipeline	Video Pipeline
Accuracy	91.3%	94.7%
Precision	90.6%	93.8%
Recall	92.0%	95.5%
F1 Score	91.3%	94.6%
ROC-AUC	0.948	0.973

Qualitative results show that the video pipeline successfully detects temporal inconsistencies even in high-quality manipulations, outperforming static models in real-world scenarios.



IV. Conclusions

This research introduces a robust multimodal framework for deepfake detection across image and video domains. Our findings indicate that combining spatial analysis through CNNs with temporal modeling via LSTM networks yields significantly more accurate detection than traditional static-only models. The architecture we propose leverages the strengths of both techniques, enabling real-time, accurate detection with high generalizability. Furthermore, our results on benchmark datasets such as FaceForensics++ validate the effectiveness and scalability of this approach under diverse adversarial conditions.

One of the core contributions of this work lies in its methodical treatment of frame-level and sequence-level data. By isolating spatial inconsistencies using convolutional filters and further exploiting temporal artifacts such as unnatural blinking, motion jitter, and lip-sync discrepancies, our pipeline mimics a layered forensic analysis. These insights are not only useful for detection but could potentially inform reverse-engineering efforts to attribute the generative model or determine the forgery pipeline. Additionally, the implementation of robust evaluation metrics and augmentation strategies ensures the model's reliability and adaptability across domains.

Looking forward, we envision several directions for future enhancement. Firstly, integrating attention-based mechanisms could provide a dynamic weighting scheme that further boosts detection reliability, especially in partially occluded or noisy conditions. Secondly, the model's deployment on edge devices, including mobile and surveillance systems, will require further optimizations in memory footprint and inference latency. Finally, our research encourages community-driven benchmarking and interpretability studies, where the emphasis is placed on transparent decision-making in AI forensic tools to reinforce public trust and legal admissibility.

References

[1] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," *arXiv preprint arXiv:1901.00686*, 2019.

[2] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2019.

[3] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[4] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," in *Proc. 27th Int. Conf. Machine Learning (ICML)*, Haifa, Israel, 2010, pp. 1–8.

[5] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake detection using spatiotemporal convolutional networks," *arXiv preprint arXiv:2006.14749*, 2020.

[6] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[7] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *IET Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.

[8] T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," *arXiv preprint arXiv:1909.11573*, 2022.

[9] Z. Zhao et al., "An efficient deep video model for deepfake detection," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 2023, pp. 1–5. DOI: 10.1109/ICIP49359.2023.10222682.