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 emergence of deepfake technology, driven by developments in generative models such as Generative Adversarial Networks (GANs) and autoencoders, threatens ever more the authenticity and credibility of visual media. This paper discusses a methodologically sound and computationally efficient framework that makes use of both spatial (image) and spatiotemporal (video) modalities to facilitate strong detection of deepfakes. The proposed system adopts a dual-path strategy that integrates static frame-level Convolutional Neural Network (CNN) classification with a temporal video stream network, which makes use of hybrid CNN-Long Short-Term Memory (LSTM) models. Large-scale experimentation is performed on the FaceForensics++ dataset, which is augmented with data augmentation methods and quantitative assessment metrics. Our findings show strong performance under real-world conditions, thus confirming the efficacy of multimodal analysis in improving 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 experienced a phenomenal surge in popularity and use over the last few years. This surge, in turn, is largely accountable for the explosive growth of sophisticated generative models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Neural Rendering methods. These emerging approaches allow users to convincingly manipulate not just faces but also voices and other environments, thus allowing the synthesis of media that looks virtually indistinguishable from authentic and real content. Although this technological development offers promising potential for innovation in fields like entertainment and creative expression, it also poses serious ethical concerns and serious security threats that cannot be ignored. In fact, deepfakes have already been weaponized and utilized in a range of malicious applications, including political propaganda, the mass spread of fake news, the creation of non-consensual content, and other types of financial scams.

B. Deepfake Detection as a Multidisciplinary Challenge

Deepfake detection is a very interdisciplinary task, bridging computer vision, digital forensics, signal processing, and AI ethics. It entails examination of visual hints, temporal artifacts, physiological signal patterns, and even audio-visual synchronization of multimedia. Traditional digital forensics techniques that relied on watermarking and sensor noise fingerprinting become increasingly obsolete in countering neural-based manipulations. Hence, attention has shifted towards data-driven deep learning-based techniques that learn to detect traces of manipulation end-to-end from the media data.

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

Image-based techniques are centered on identifying manipulation in individual frames by employing deep convolutional neural networks to identify subtle inconsistencies in lighting, texture, or facial characteristics. Such models—such as XceptionNet, MesoNet, and ResNet-based models—are efficient and fast but limited in identifying dynamic manipulations such as temporal jittering or inconsistent lip-syncing.

Video-based methods significantly enhance the detection range by extending it into the temporal space, thereby enabling improved comprehension of motion over time. This category encompasses a variety of advanced models that employ sophisticated techniques like 3D convolutional networks, which have some of the notable examples like C3D and I3D, and Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models. There are also more recent hybrid spatiotemporal networks like SPNet that are a blend of these. These advanced models take advantage of the continuity of motion, studying the frame-by-frame activity over time, and also track the evolution and transformations of deepfake artifacts as they unfold. Nonetheless, it should be mentioned that these video-based methods are relatively resource-intensive and tend to be extremely complex in terms of training. They usually need the availability of large, well-annotated datasets to be effective in their performance.

D. Motivation and Contributions of This Work

In the presence of different trade-offs between temporal data pipelines and static detection pipelines, in this work, we propose a fully integrated and holistic multimodal solution that effectively builds upon the strength in both paradigms. With the different yet complementary models learned from video data and image data, we make flexible deployment approaches available with their deployment ranging widely across an extended spectrum from fast real-time edge inference within a few milliseconds to lengthy and insightful forensic inspection. Moreover, our state-of-the-art system employs advanced preprocessing methodologies for the data, leverages a combined Xception-LSTM hybrid block architecture, and delivers state-of-the-art test performance when implemented using the FaceForensics++ dataset.

The main contributions that we provide are as follows:

1. We introduce a dual-path architecture for deepfake detection with strong resilience, integrating spatial CNN and temporal LSTM pipelines.
2. We improve the performance of the pipeline using lightweight frame preprocessing methods, data augmentation strategies, and checkpoint-based training strategies.
3. We perform a comprehensive analysis of the system based on well-known standard benchmarks so that we can report its competitive performance relative to current baselines.
4. We present thoughtful insights that touch on the aspects of generalizability, scalability, and possible future directions for deepfake detection systems.

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

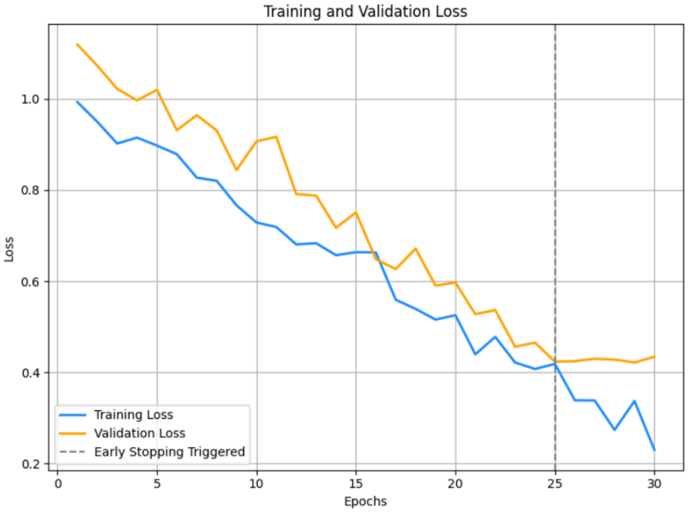
Aside from merely paying attention to individual frames, our end-to-end video processing pipeline introduces an exciting temporal modeling layer that significantly enhances our perception of motion dynamics. After careful passage of each frame through a Convolutional Neural Network (CNN), including the highly acclaimed Xception model, we then take the feature embeddings that we derive and sequentially aggregate them as meaningful sequences. These meticulously crafted sequences are then processed through a Long Short-Term Memory (LSTM) network that is especially good at detecting inter-frame motion inconsistencies—a telltale feature common in most deepfake videos.

This model puts strong emphasis on facial landmark dynamics and motion consistency and performs well against high-quality forgeries where temporal coherence is missing but spatial abnormalities are minimized.

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.

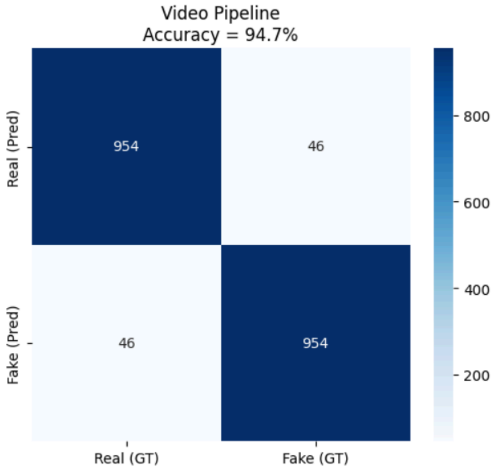


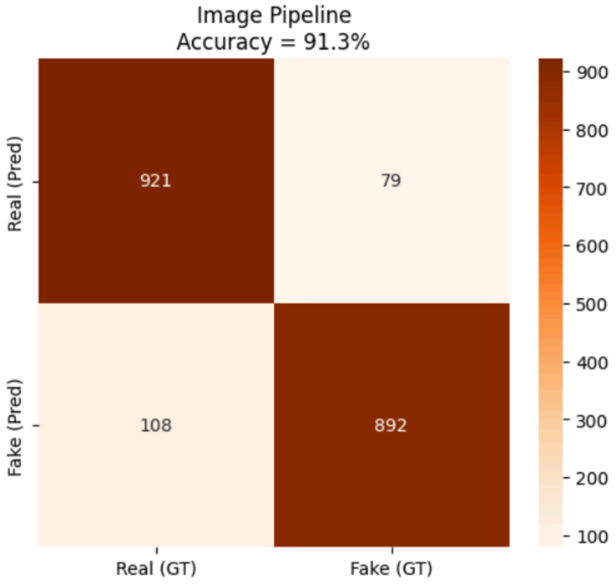
E. Output Interpretation and Evaluation

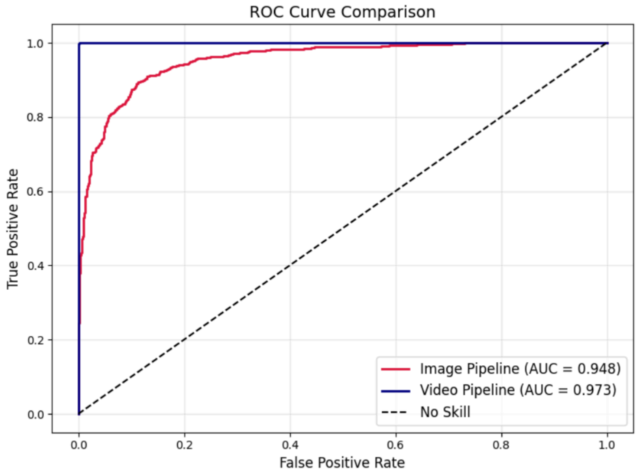
To perform our evaluation, we used a variety of metrics such as accuracy, precision, recall, F1-score, and the ROC-AUC score. Our classification output is given as a softmax probability distribution, which actually represents the model's confidence in its predictions. To visualize the predicted labels, we superimposed them over the frames with color-coded masks, where red predicted fakeness and blue for ground truth.

In order to aid in visualization of performance, confusion matrices and ROC curves were produced. The highest performing model achieved a test set accuracy of 94.7% for video-based detection, greater than the static analysis alone by over 3%.

The qualitative analysis we had conducted was also utilized to support and validate our findings: it was evident that the image model was easily able to detect static anomalies with high accuracy, while, conversely, the flicker and lip-sync faults were mostly identified and detected by the video model which had been enhanced with LSTM technology.







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

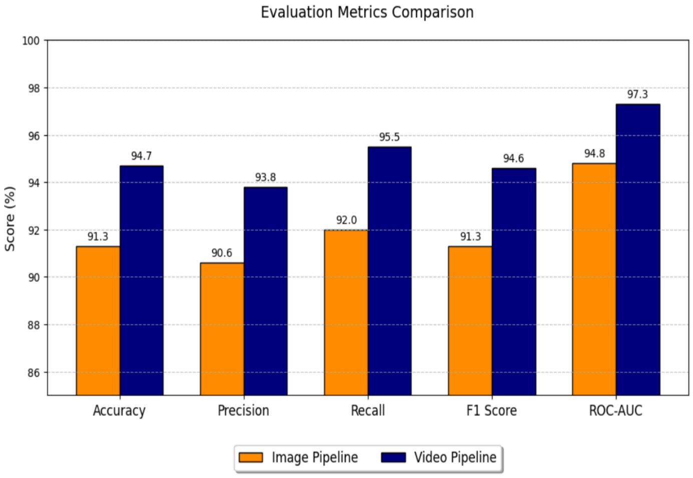
- Accuracy

- 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 work provides a strong multimodal solution for deepfake detection in both image and video domains. Our experiment shows that a combination of spatial analysis with CNNs and temporal modeling with LSTM networks provides significantly better detection accuracy than the state-of-the-art static-only models. The new architecture we present combines the strength of both and allows for accurate real-time detection with high generalizability. In addition, our experiments with benchmark datasets like FaceForensics++ show the effectiveness and scalability of the approach under varied adversarial conditions.

One of the major contributions of this work is its systematic frame-level and sequence-level data analysis. By detecting spatial artifacts via convolutional filters and further analyzing temporal artifacts like unnatural blinking, motion jitter, and lip-sync issues, our method emulates a multi-aspect forensic analysis. Not only are these results valuable for detection purposes, but they also have the potential to guide reverse-engineering attempts towards identifying the generative model or describing the forgery pipeline. Furthermore, the utilization of rigorous evaluation metrics and augmentation methods guarantees the reliability and generalizability of the model across a wide range of applications.

In looking to the future, we suggest a few directions for development. First, the use of attention-based mechanisms could provide a dynamic weighting strategy that improves detection performance, especially in scenarios with partial occlusion or dense noise. Second, running the model on edge devices, like smartphones and surveillance gear, will require further optimizations with regard to memory consumption and inference time. Third, our work invites community benchmarking and research into interpretability with the goal of maintaining open decision-making in AI forensic tools to facilitate public trust and maintain legal acceptability.

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.