Deep-fake Detection using rPPG signals

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—Deepfake videos pose significant risks to digital security and misinformation by generating highly realistic synthetic content. Although deep fakes have found applications in entertainment and visual effects, they also pose serious risks, including misinformation, identity theft, and political manipulation. High-profile cases, such as the fake video of Sachin Tendulkar endorsing a political party and deepfake speeches attributed to former U.S. presidents, have demonstrated the growing challenge of distinguishing real content from synthetic fabrications.Traditional detection methods often struggle with advanced deep-fake techniques that convincingly mimic facial expressions and movements.r, we propose a novel deep-fake detection approach utilizing remote photoplethysmography (rPPG) signals to analyze subtle physiological cues such as heart rate variability, which are often missing or inconsistent in deep-fakes. Using a data set of 6,000 real and deep-fake videos, we extract rPPG features and train a deep learning model to distinguish between authentic and manipulated videos. Our approach demonstrates the effectiveness of physiological signal analysis in deep-fake detection.**

***Index Terms*—Deepfake, Remote photoplethysmography(rPPG), Heart rate variability, Deep learning model, Manipulated videos,**

1 Introduction

Deepfake technology, powered by generative adversarial networks (GANs) and other AI models, has made it increasingly difficult to differentiate real videos from synthetic ones. Although traditional deep-fakefake detection methods focus on inconsistencies in facial expressions, lighting, and artifacts, these approaches can struggle with highly refined deep-fakes.

One promising avenue for detection is remote photoplethysmography (rPPG), a noncontact method that extracts heart rate signals from subtle skin color variations caused by blood flow. Since deep-fake videos lack natural physiological signals or exhibit unrealistic pulse patterns, rPPG-based detection can serve as a robust biometric defense.

In this study, we propose a deepfake detection model using rPPG signals, combined with deep learning techniques, to improve accuracy. We evaluate our method on a dataset of 6,000 videos and compare it with existing approaches. The results highlight the potential of physiological-based deepfake detection for real-world applications in media forensics and security.

2 Literature Review

2.1 Deepfake Detection Techniques

Deepfake videos are primarily generated using Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). These methods synthesize realistic human faces and expressions by training on extensive datasets. Common face manipulation techniques include FaceSwap, Face2Face, DeepFakes, NeuralTextures, and FaceShifter, each employing distinct methodologies to replace or modify facial features.

2.1.1 FaceSwap and Face2Face: Enable real-time facial reenactment, making them popular in entertainment and social media.

2.1.2 NeuralTextures: utilizes learned textures to enhance facial expressions, making detection more difficult.

2.1.3 DeepFakes: rely on autoencoders to swap faces in a convincing manner, often seen in viral videos and manipulated political content.

2.1.4 FaceShifter: Itis an advanced deepfake generation model that improves identity preservation and facial blending by using Adaptive Feature Fusion (AFF) to produce more seamless face-swapped videos, making detection even more challenging.

2.2 Deepfake Detection Approaches

Traditional deepfake detection techniques have relied on pixel-level inconsistencies, unnatural eye blinking patterns, and frame-by-frame inconsistencies. Machine learning-based classifiers, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been employed to differentiate real and manipulated content. Recent advancements incorporate transformers and attention-based architectures, allowing models to better analyze temporal coherence in deepfake videos. These techniques improve upon basic CNN models by assigning weights to key regions of interest, highlighting areas where deepfake artifacts are likely to occur

2.3 Biological Signal-Based Detection

Emerging research suggests that physiological signals, such as heartbeat rhythms and facial micro-expressions, provide promising avenues for deepfake detection.

2.3.1 Remote Photoplethysmography: Remote Photoplethysmography (rPPG) has been widely used in biomedical applications for heart rate estimation and is now being explored for deepfake detection.

2.3.2 DeepRhythm: The DeepRhythm approach specifically introduces attention-guided extraction of rPPG maps, making it possible to detect irregular or missing heartbeat patterns in deepfake videos.

2.3.3 Other Biological signals: Other studies have explored eye movement and mouth synchronization inconsistencies as additional biological signals for detection.

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

2.4 Challenges in Detecting Deepfakes in Compressed Videos

One major challenge in deepfake detection is the impact of video compression algorithms (e.g., H.264, H.265). Compression removes fine details, making many visual detection techniques ineffective. Existing detection models often fail in real-world applications where social media platforms automatically compress videos.

- Many detection methods trained on high-resolution datasets struggle when applied to compressed formats.
- Physiological signals like rPPG remain more resilient to compression, making them an attractive alternative
- Researchers are now exploring hybrid models that combine spatial-temporal attention with physiological data to improve detection accuracy across different video qualities.

Addressing these challenges, our approach focuses on detecting physiological discrepancies in both high-quality and compressed videos, offering a robust and generalizable solution.

3 Proposed Methodology

Our method is designed to detect DeepFake videos by analyzing heartbeat rhythms from facial regions using remote photoplethysmography (rPPG). The methodology follows a systematic flow:

3.1 Dataset Preparation

Datasets available are

- FaceForensics++ (FF++): Contains DeepFake, Face2Face, FaceSwap, NeuralTextures and FaceShifter videos.
- Out of them, we used Real and Deepfake videos for training our model, and evaluated it on Real and Face2Face videos

3.2 Data Preprocessing and Face Segmentation

3.2.1 Face Detection and Landmark Extraction: • Face Detection: Detect the face in each frame using MTCNN (Multi-task Cascaded Convolutional Networks).

- Facial Landmarks: Identify 81 key landmarks on the face using Dlib.
- Region of Interest (ROI) Selection:

– Remove eyes and background as they introduce noise.

– Focus on forehead, cheeks, and under-eye areas, where heartbeat signals are strongest

3.2.2 Face Tracking and Stabilization:

- If multiple faces are detected, retain the one that is closest to the previously detected face.
- Frames without detected faces are discarded (if more than 50 frames are lost, the video is skipped)

3.3 Motion-Magnified Spatial-Temporal Representation (MMSTR)

3.3.1 Motion Magnification for Heartbeat Enhancement:

- Why? The heartbeat signal in facial regions is subtle and may not be directly visible.
- How? Apply the Eulerian Video Magnification (EVM) technique to amplify small color changes caused by blood flow.
- Output: A motion-magnified face video where color changes due to the heartbeat are enhanced.

3.3.2 Generating the MMST Map: Divide the face into N non-overlapping blocks (ROIs) (e.g., 5x5 grid = 25 blocks). Extract the average RGB colour intensity per block over time. Construct an MMST (Motion-Magnified Spatial-Temporal) Map, where:

- Rows = different facial regions (N blocks)
- Columns = time (frames)
- Values = RGB intensity variations (representing heartbeat patterns)

3.4 Dual-Spatial-Temporal Attentional Network (DualST AttenNet)

DeepRhythm utilizes a dual-attention mechanism to focus on meaningful areas while ignoring noise

3.4.1 Spatial Attention (Where to Focus?): Some facial regions provide stronger heartbeat signals than others. The model applies a Dual-Spatial Attention Mechanism:

- Prior Spatial Attention (Fixed Weights): Focuses on predefined robust regions (e.g., under the eyes).
- Adaptive Spatial Attention (Learned Weights): Adjusts dynamically based on video conditions (e.g., lighting changes).

3.4.2 Temporal Attention (When to Focus?): Some frames contain more distinctive DeepFake artifacts than others. Two types of temporal attention are applied:

- • Block-Level Temporal Attention: Uses LSTM to analyze variations in facial regions over time.
- Frame-Level Temporal Attention: Uses MesoNet (a CNN-based model) to assign importance scores to frames.
- The final weight matrix A = (t * s) X ensures that the model gives higher importance to frames and regions with strong heartbeat signals

3.5 Deep Neural Network for DeepFake Classification

- The weighted MMST map (with spatial and temporal attention applied) is passed to a deep neural network for classification.
- ResNet18 is used as the final classifier.
- Adam Optimizer is used
- The network is trained to output 1 for fake videos and 0 for real videos.

3.6 Model Training

3.7 L2 Regularization:

- Helps in mitigating overfitting by adding a penalty for large weights, preventing the model from relying too heavily on specific features.
- Encourages better generalization by ensuring that the model does not memorize the training data but instead learns meaningful patterns.

3.7.1 Early Stopping:

- • Continuously monitors the validation loss during training and stops the process when no further improvement is observed.
- • Prevents the model from over-training on the training data, reducing the risk of poor performance on new, unseen data.
- • Ensures that the model retains optimal performance without excessive training, leading to better real-world applicability

3.7.2 Combined Effect::

- Helps in reducing variance by balancing model complexity and performance, preventing overfitting to training data.
- Enhances generalization by ensuring that the learned features remain relevant across different datasets.
- Stabilizes the overall training process, making it more efficient and reducing the risk of unnecessary computations.

4 Results and Discussion

4.1 Accuracy

Accuracy is a measure of the overall correctness of the classification system and is calculated as the ratio of correctly classified instances to the total instances. Accuracy = (TP+TN) / (TP+FN+FP+TN). The model was trained for 500 epochs with a batch size of 32, using the Adam optimizer and binary cross-entropy loss function. To improve generalization and reduce overfitting, L2 regularization was applied to penalize large weight values, and Early Stopping was implemented to halt training when the validation loss stopped improving.

4.1.1 Training and Validation Performance: By Epoch 95, the model achieved:

- Training Accuracy: 63.53
- Training Loss: 0.9619
- Validation Accuracy: 52.78
- Validation Loss: 0.9498

The incorporation of L2 regularization helped in reducing model variance by discouraging overly complex weight distributions, while Early Stopping ensured that training was halted before overfitting could occur. The accuracy and loss values fluctuated in the early stages but eventually stabilized. The Accuracy vs. Epochs and Loss vs. Epochs graphs, included in this section, provide a clear visualization of the learning behavior.

4.1.2 Testing Performance on Deepfake: To evaluate real-world generalization, the model was tested on a separate dataset consisting of 20 % deepfake data, yielding:

- Test Accuracy: 61.00
- Test Loss: 0.9789

These results suggest that while the model learned useful patterns from the training data, there is still room for improvement in generalization. The moderate test accuracy could be attributed to class imbalance, feature complexity, and potential overfitting to the training dataset. Further enhancements, such as fine-tuning hyperparameters, adding data augmentation techniques, or exploring alternative architectures, could lead to better performance.

4.1.3 Face2Face (F2F) Testing Results: In addition to deepfake detection, the model was tested on the Face2Face dataset, producing the following results:

- Face2Face Test Accuracy: 61.88
- Face2Face Test Loss: 0.9943
- DeepRhythm Confidence Score: 58.53

The performance on Face2Face videos suggests that the model retains its ability to detect manipulated content across different deepfake types, although the accuracy remains moderate.

5 Conclusion

In this project, we implemented a deepfake detection model using a ResNet18- based architecture, incorporating L2 regularization and Early Stopping to improve generalization and prevent overfitting. Our approach focused on extracting spatial and temporal features from video frames to classify them as real or fake. The dataset used for training consisted of deepfake and genuine videos, with preprocessing techniques applied to enhance feature extraction. Achieved the following performance:

- Training Accuracy: 63.53
- Validation Accuracy: 52.78
- Test Accuracy: 61.00

Identified challenges related to generalization and robustness, requiring further improvements.

Planned enhancements include:

- Hyperparameter tuning (learning rate, batch size)
- Applying advanced augmentation techniques
- Training on additional datasets like DFDC (Deepfake Detection Challenge)
- Evaluating performance on different video compression levels

Future work will focus on optimizing network architecture, inference speed, and computational efficiency to support real-time deployment. Additionally, we may explore the feasibility of making the model real-time, enhancing its practical applicability for deepfake detection in live scenarios. Optimizing the network architecture, inference speed, and computational efficiency will be key considerations for real-time deployment.

References

[1] DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms.

[2] Munawar et al., "Forged Video Detection Using Deep Learning: A Systematic Literature Review," Applied Computational Intelligence and Soft Computing, 2023.
[3] F. Mittelbach and M. Goossens, *The LaTeXCompanion*, 2nd ed. Boston, MA, USA: Pearson, 2004.
[4] G. Grätzer, *More Math Into LaTeX*, New York, NY, USA: Springer, 2007.
[5] M. Letourneau and J. W. Sharp, *AMS-StyleGuide-online.pdf,* American Mathematical Society, Providence, RI, USA, [Online]. Available: http://www.ams.org/arc/styleguide/index.html
[6] H. Sira-Ramirez, "On the sliding mode control of nonlinear systems," *Syst. Control Lett.*, vol. 19, pp. 303–312, 1992.
[7] A. Levant, "Exact differentiation of signals with unbounded higher derivatives," in *Proc. 45th IEEE Conf. Decis. Control*, San Diego, CA, USA, 2006, pp. 5585–5590. DOI: 10.1109/CDC.2006.377165.
[8] M. Fliess, C. Join, and H. Sira-Ramirez, "Non-linear estimation is easy," *Int. J. Model., Ident. Control*, vol. 4, no. 1, pp. 12–27, 2008.
[9] R. Ortega, A. Astolfi, G. Bastin, and H. Rodriguez, "Stabilization of food-chain systems using a port-controlled Hamiltonian description," in *Proc. Amer. Control Conf.*, Chicago, IL, USA, 2000, pp. 2245–2249.