

Deepfake Detection using rPPG Signals

B. Tech. Project Report

Submitted by

Sumit Sunil Girnar 112003045

Swapnil Santosh Gite 112003046

Om Prakash Gurav 112003047

Under the guidance of

Dr. P. R. Deshmukh

COEP Technological University, Pune



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
COEP TECHNOLOGICAL UNIVERSITY, PUNE-5**

May 2024

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

CERTIFICATE

Certified that this project titled, "Deepfake Detection using rPPG Signals"
has been successfully completed by

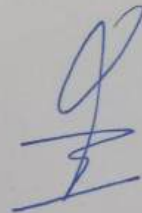
Sumit Sunil Girnar	112003045
Swapnil Santosh Gite	112003046
Om Prakash Gurav	112003047

and is approved for the partial fulfillment of the requirements for the degree
of "B.Tech. Computer Engineering".



SIGNATURE

Dr. P. R. Deshmukh
Project Guide
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.



SIGNATURE

Dr. P. K. Deshmukh
Head
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

ORIGINALITY REPORT

13%
SIMILARITY INDEX

9%
INTERNET SOURCES

10%
PUBLICATIONS

0%
STUDENT PAPERS

PRIMARY SOURCES

1 www.mdpi.com
Internet Source

2 arxiv.org
Internet Source

3 Yuezheng Xu, Ru Zhang, Cheng Yang, Yana Zhang, Zhen Yang, Jianyi Liu. "New Advances in Remote Heart Rate Estimation and Its Application to DeepFake Detection", 2021 International Conference on Culture-oriented Science & Technology (ICCST), 2021
Publication

4 Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Niessner. "FaceForensics++: Learning to Detect Manipulated Facial Images", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019
Publication

5 www.researchsquare.com
Internet Source

2%

1%

1%

1%

1%

PT

Abstract

The advancement of generative techniques for producing fake portrait videos poses a significant societal challenge, particularly with the emergence of realistic deepfakes used for political propaganda, imitating celebrities, and manipulating identities. This project focuses on identifying fake videos, especially those created using video forging techniques like Face2Face, FaceSwap, NeuralTexture, and DeepFakes . It aims to classify videos into five categories: real videos and four types of deepfake manipulations. By analyzing minute shifts in facial skin tone resulting from blood circulation through remote visual photoplethysmography (rPPG), prior studies suggest that the typical heartbeat patterns observed in real-life videos are altered or completely disrupted in deep fake videos. These disruptions in facial color changes can be exploited, making the rPPG signal a robust biological indicator for detecting deepfakes. The proposed approach entails utilizing a spatial-temporal PPG map to detect heartbeat signals across various facial areas, augmenting the model with neural network methodologies such as spatial and temporal attention modules, as well as transformers.

Chapter 1

Introduction

The rise in advanced and accessible technology has led to an increase in deepfake videos across social media platforms, hence a significant challenge in modern society. Deepfakes involve digitally manipulating images or videos to replace a person's identity with another's which leads to further spreading misleading information or creating fake news. Notably, deepfake technology has been misused to produce false videos of prominent public figures like Barack Obama and Joe Biden, worsening concerns about the spread of misinformation. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have emerged as popular tools for creating deepfakes [2], leveraging large datasets to create realistic images and videos. Various methods exist for manipulating faces in videos. These methods include computer graphics-based techniques like Face2Face and FaceSwap, as well as learning-based approaches such as DeepFakes and NeuralTextures. Deep learning methodologies such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) have been proposed to address this challenge.

Face2Face is a technique that allows for the mapping of facial expressions from one video onto another video, all while maintaining the identity of the person in the target video. Similarly, FaceSwap involves transferring the

facial region from one video to another. DeepFakes is a way of changing faces in videos using GAN. It's like creating a mask for someone's face and putting it on another person's face in a video. NeuralTextures (NT) utilizes learned neural textures of the target individual to reenact facial expressions through deferred neural rendering.

Some Deepfake detection methods make use of biological features in detection such as eye blinking, eyebrow, ear, eye movement, mouth movement, and heartbeat detection[8]. [5] study focuses on utilizing rPPG signals, which were previously utilized for heart rate estimation, for deepfake detection.

Chapter 2

Literature Review

2.1 What is rPPG (Remote Photoplethysmography)

rPPG, also known as remote photoplethysmography, is a technique for assessing blood flow utilizing light. It operates by discerning alterations in how light is absorbed or dispersed by blood circulation in a specific area of the body. These changes also affect the face, causing subtle variations in skin tone as the pulse causes the skin to appear slightly lighter and darker over time. Although these changes are not visible to the naked eye, they can be detected by analyzing how the intensity of light reflected from specific regions of the skin (ROIs) changes over time [9].

Analyzing these periodic pulsations allows for the extraction of physiological indicators such as heart rate, respiratory rate, and heart rate variability from the recorded skin tone information in the video [13]. Heartbeats lead to periodic changes in blood flow, which in turn cause periodic variations in capillary volume. These variations are typically represented by the blood volume pulse (BVP) and reflect changes in intravascular blood flow and hemoglobin content.

The variations in facial skin color are correlated to the heart rate due to changes in blood flow, altering light absorption and scattering, thereby

reflecting subtle changes in skin color that correspond to the pulse rate. This property is used to extract heartbeat from the facial colour variation in the video. The figure 2.1.a shows the variation of one of the ROIs of a face in the video across the frames.

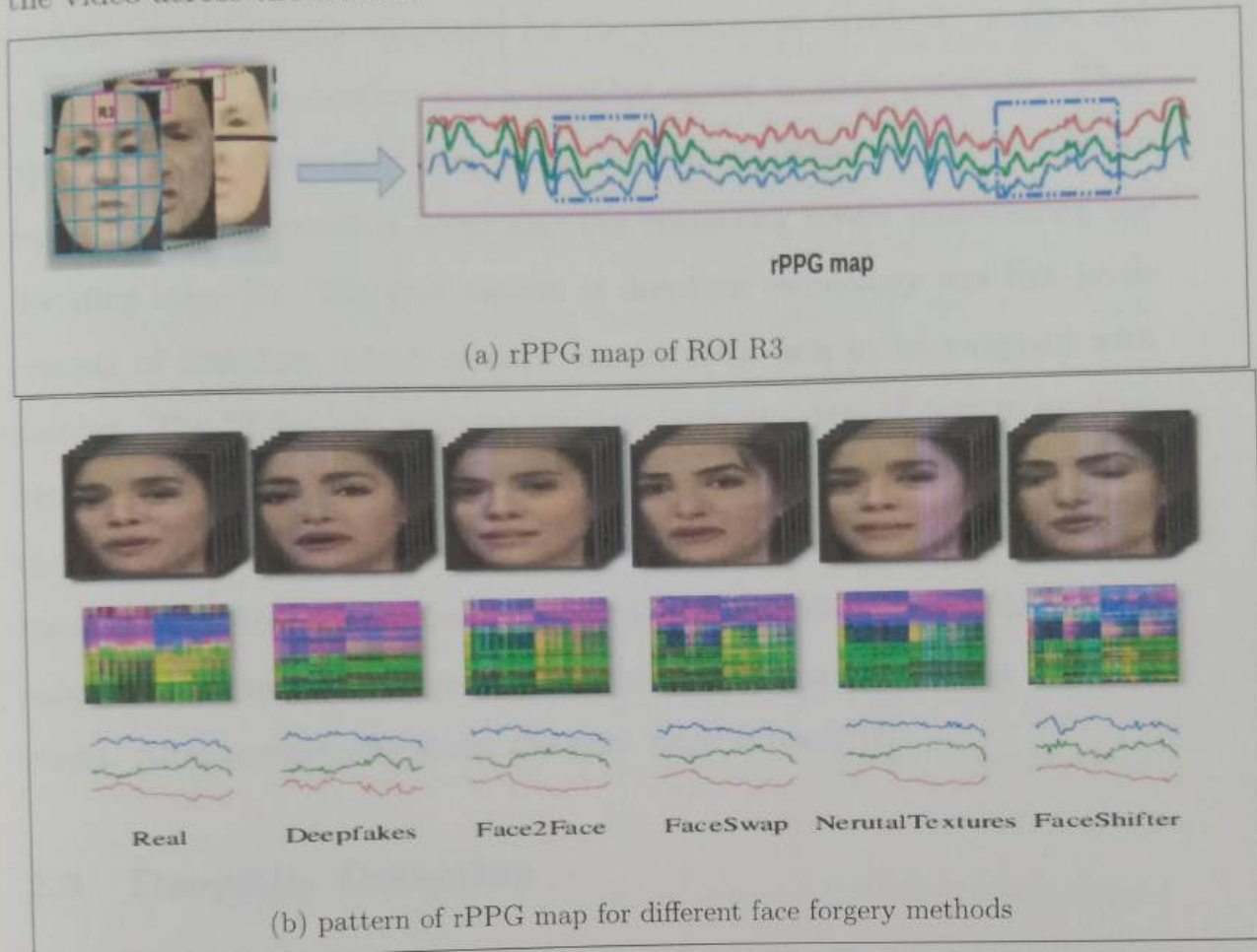


Figure 2.1: Demonstration of rPPG map

Since different face forgery methods affect different ROIs of faces, it has been observed that each of the corresponding PPG maps exhibits unique rhythmic patterns. These patterns are utilized to detect the source of deepfakes, i.e., the method or model used to create them, and subsequently classify videos as real or fake. Figure 2.1(b) shows pattern of the rPPG map for different face forgery methods.

2.2 Deepfake Creation

Deepfake techniques use advanced algorithms to create realistic fake content, including videos, images, text, and voices. Generative Adversarial Networks (GANs) are frequently employed for the creation of synthetic images and videos. Deep autoencoders are a popular model in deep networks. They consist of two symmetrical Deep Belief Networks (DBNs), with several layers representing the encoding stage and the remaining layers representing the decoding stage [7]. The first success of deepfake technology was the development of FakeApp, which allowed one person's face to be swapped with another. The "FakeApp" software requires large amounts of data to produce better results. The data is passed to the system to train the model, which then inserts the face into the target video. Creating fake videos in FakeApp involves extracting all the images from the source video into a folder, cropping and aligning them correctly, and then processing them using a trained model. After merging the faces, the final video is ready [7].

2.3 Deepfake Detection

Early deepfake detection methods relied on hand-crafted features such as color histograms, which were manually designed to capture specific video content aspects. These features were then fed into machine learning algorithms like SVMs or random forests for classification. However, with the emergence of deep learning, Convolutional Neural Networks (CNNs) became the preferred choice for deepfake detection. Initial CNN models like Mesonet and CapsuleNet were developed to detect crucial facial features in both original and deepfake images.

The hypothesis that deepfake videos exhibit spatial and temporal consis-

tencies drove the design of detection models. Various models were created to calculate local spatial inconsistencies, such as Multi-scale Patch Similarity [3], which measures pixel-wise differences between different areas. For temporal analysis, Recurrent Neural Networks (RNNs) were employed to explore temporal dependencies and patterns in video frames. Subsequently, the Attention mechanism was developed to enhance RNN models by assigning weights to each frame based on its relevance to deepfake detection.

Another approach used for deepfake detection is by analyzing biological features as discussed in [8]. Biological signals like heart rate, eye movements, eye blinking, and detection of ear and mouth movements can be utilized for deepfake detection.

Deepfakes lack natural eye blinking because they are generated by AI algorithms that often do not include realistic blinking patterns. The DeepVision [6] algorithm analyzes the human eye blinking pattern to detect deepfakes, focusing on detecting anomalies such as rapid and repeated eye blinking within a short period. They devised a methodology that integrates Convolutional Neural Networks (CNNs) and Long-term Recurrent Convolutional Neural Networks (LRCNs) to classify the state of eyes as open or closed, taking into account the historical patterns of eye behavior. By aligning faces and focusing on eye regions, the model effectively detects eye blinking in videos, showing promising results in identifying deepfake videos compared to standard datasets.

The creation of face-swap and lip-sync deepfakes often overlooks the human ear, which provides both static biometric signals and dynamic cues from jaw movement. While face-swaps may accurately depict a co-opted identity, the ears typically belong to the original person, and in lip-sync deepfakes, ear dynamics are not synchronized with mouth and jaw movements. [1] describes

a forensic technique leveraging these static and dynamic aural properties for deepfake detection.

The detection of heartbeat signals using photoplethysmography (rPPG) technology has long been employed in biomedicine for heart rate monitoring. This technology captures slight variations in skin color from video recordings, which are disrupted by facial pixel modifications in deepfake videos. Previous studies have demonstrated that analyzing the rPPG signal can effectively detect deepfake videos, as deepfakes struggle to maintain consistent heartbeat signals, providing a reliable indicator for forgery detection.

In their study,[4] introduced a method that not only differentiates between deepfake and authentic videos but also pinpoints the generative technique employed to produce the deepfake. Current CNN-based techniques focus on learning the differences within the generator, which harbor valuable insights that can be revealed through disentanglement alongside biological signals. These methods extract PPG characteristics from both genuine and synthetic videos, employing an advanced classification network to discern the generative model employed in each video. This methodology is grounded on the idea that spatio-temporal patterns in biological signals can be interpreted as representations of residuals.

[11] adopts a Multi-scale Spatial-Temporal PPG map for the detection of heartbeat signals across multiple facial regions. The approach suggests a two-stage network that includes a Mask-Guided Local Attention module (MLA) to identify specific local patterns in PPG maps, along with a temporal transformer. This temporal transformer helps in enhancing interactions among features from adjacent PPG maps across extended distances. The rPPG maps are further utilized for binary classification (distinguishing real videos from fake ones) and for multi-class classification of videos (identifying

the source model used to create deep fakes, such as FaceSwap, NeuralTextures, etc.). Figure 2.2 demonstrates the generalized model for the binary classification of videos (real or fake) using the rPPG map.

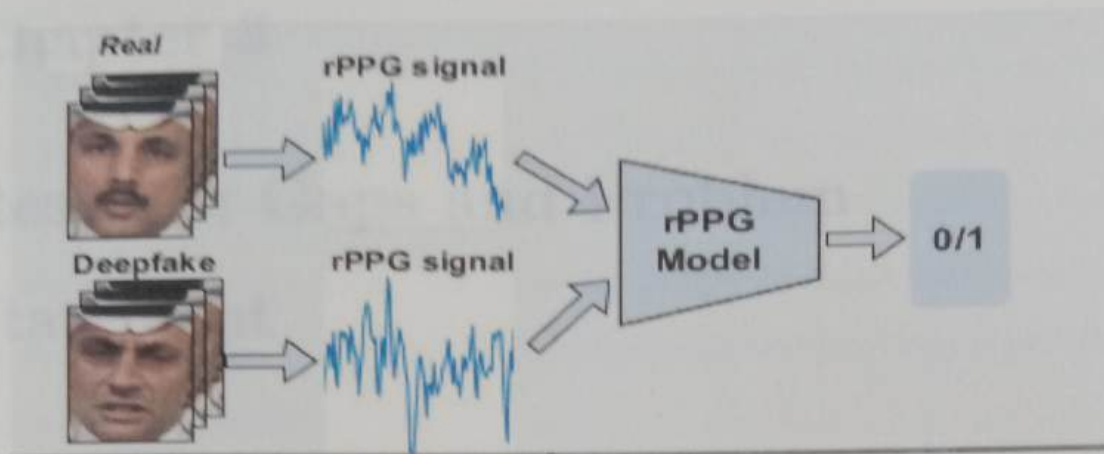


Figure 2.2: Binary classification of videos

Chapter 3

Research Gaps and Problem Statement

3.1 Research Gap

In the realm of digital media forensics, the emergence and proliferation of deepfake technology have introduced a significant challenge to the authenticity and integrity of visual content. While substantial efforts have been directed toward detecting deepfake images, there exists a notable research gap in effectively utilizing Remote Photoplethysmography (rPPG) signals for detecting deepfake videos. Current methodologies primarily focus on analyzing uncompressed video data, disregarding the prevalent use of compressed videos on social media platforms. Compression algorithms employed by these platforms, such as spatial and temporal compression, alongside advanced video codecs like H.264 or H.265, significantly alter the visual characteristics of videos, complicating traditional detection techniques. Prior research in deepfake detection has often relied on intricate features such as edge detection, eye blinking patterns, and color imbalances, which pose challenges when videos undergo compression.

3.2 Problem Statement

The research aims to develop a model for the binary classification of videos as real or fake and further identification of the deepfake generation method used (e.g., FaceSwap, NeuralTexture) by analyzing the remote photoplethysmography (rPPG) signal of a video. Additionally, the model will be trained using moderately compressed videos to evaluate its effectiveness in detecting deepfakes in compressed video formats. This research aims to advance the state-of-the-art in deepfake detection and enhance trust and reliability in digital media content.

Chapter 4

Proposed Methodology/ Solution

4.1 Dataset

We assessed the effectiveness of our approach using the FaceForensics++ [10] dataset, comprising 1000 original video sequences modified with four manipulation methods: DeepFakes, Face2Face, FaceShifter, and NeuralTextures. The dataset encompasses videos of three quality levels: raw (C0), lightly compressed (C23), and low quality (C40). For our study, we utilized 1000 videos for each manipulation method, totaling 5000 videos. Specifically, we selected videos of quality C23 to evaluate our model's performance on compressed video data. Figure 4.1 shows the sample frames of the FaceForensics++ dataset. The first column displays the original, unaltered frames, while the subsequent columns show examples from DeepFakes (DT), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT), respectively.

4.2 Data Preprocessing

The initial 300 frames of the video were extracted and processed. For each frame, facial regions were extracted, and 81 facial landmarks were computed using the DLIB library. These landmarks were used to identify key facial features, enabling the subsequent removal of eyes and background components



Figure 4.1: FaceForensics++ dataset [12]

based on the landmark information.

Subsequently, face alignment was conducted for each detected face. This process involved rotating the facial images to achieve alignment based on the angles of the eyes, ensuring consistent orientation and positioning across all facial images. Following alignment, each frame was resized to a fixed dimension. The preprocessing stages are illustrated in Figure 4.2.

Each 300 preprocessed frame is stacked together to form an aligned facial video. A motion magnification algorithm is then applied to align facial video to form a motion-magnified video. This algorithm acts as a visual motion microscope, amplifying subtle motions within the video sequence enabling the visualization of minute deformations that would otherwise remain imperceptible.

The initial 300 frames of the motion-magnified video are again extracted and processed. Each frame is divided into N (25) rectangular blocks, each block acts as a region of interest (ROI). Average pooling is independently applied to each block across each color channel. This process yields the formation of rPPG (Remote photoplethysmography) map. Figure 4.3 illustrates the preprocessing of motion-magnified video. Each row of the map corresponds to the temporal variation of one block across the RGB channel.

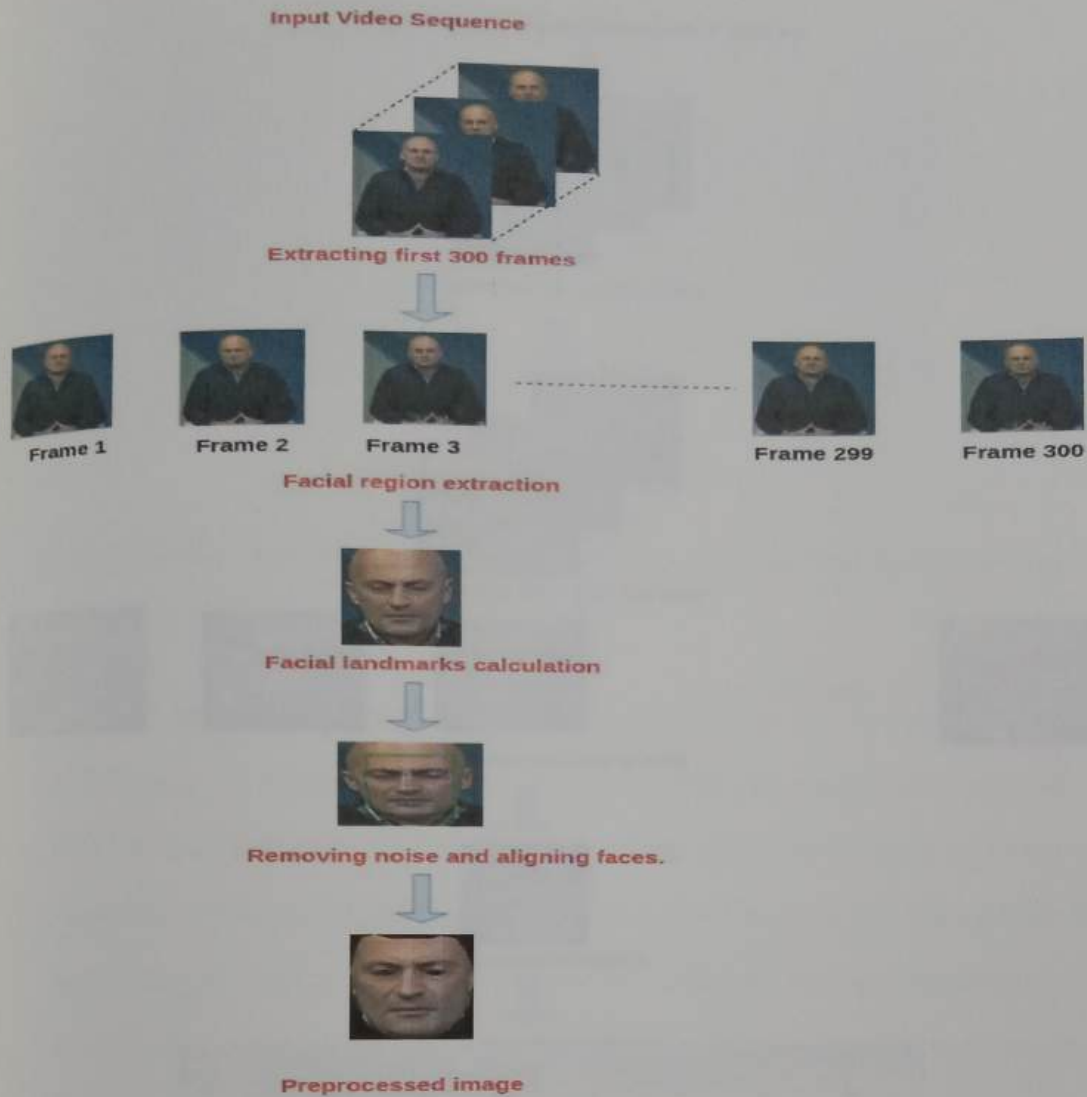


Figure 4.2: Preprocessing of an original video

The final size of the rPPG map is $2^{N-1} \times T \times C$, where N is the count of ROIs, C is the number of channels, and T is the number of frames.

4.3 Spatial and Temporal attention

Two types of attention weights were designed: spatial (s) and temporal (t).

1. Spatial Attention:

The same face can appear different under various conditions, such as lighting and other environmental factors. Therefore, it is necessary to adjust which

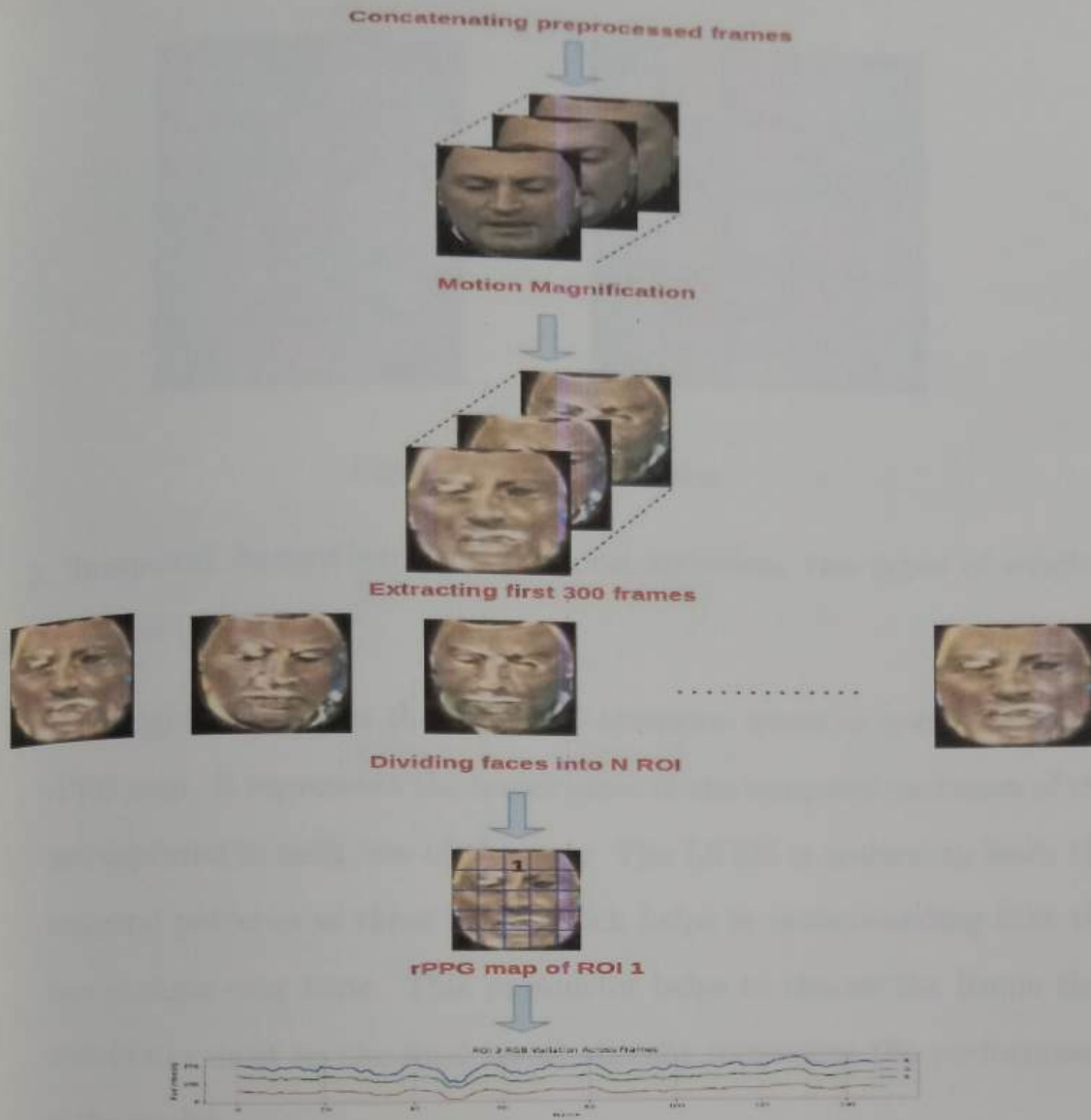


Figure 4.3: Preprocessing of motion magnified video

regions of interest (ROIs) to focus on. This process optimizes the model by selecting only those ROIs that contribute the most to the final result. The spatial attention weights highlight different blocks (ROIs) to adapt to these variations. Our proposal involves training a spatial attention network to produce spatial attention. This network includes a convolutional layer with 64 kernels of size 15×1 , followed by a batch normalization layer and a max-pooling layer.

Figure 4.4 shows how different ROIs are highlighted.

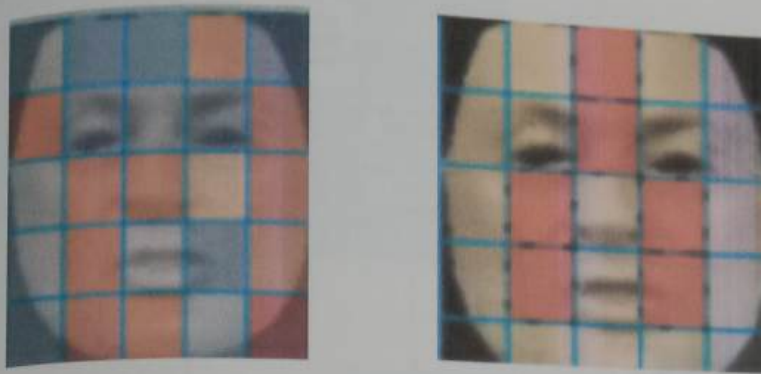


Figure 4.4: Spatial Attention

2. Temporal Attention: For temporal attention, two types of weights are calculated ($t1$, $t2$).

- **t1:** Weight $t1$ signifies the temporal attention given to each row of the rPPG map. It represents the importance of the temporal variation of the face captured in each row of the map. The LSTM is trained to learn the temporal patterns in these rows, which helps in understanding how the face changes over time. This parameter helps to choose the frame that contributes most to the final result thereby increasing the performance of the model.
- **t2:** A pre-trained model, MesoNet architecture is used for this purpose. MesoNet was initially trained to assess the authenticity of an image, determining how likely it is to be fake. In our case, each frame of the video is passed through the MesoNet classifier, which calculates a score indicating the level of fakeness. Frames with higher fakeness scores are given more weight in the final decision-making process. The MesoNet network comprises four convolutional layers followed by two fully-connected layers.

The final temporal weight is calculated as $t = t1 + t2$.

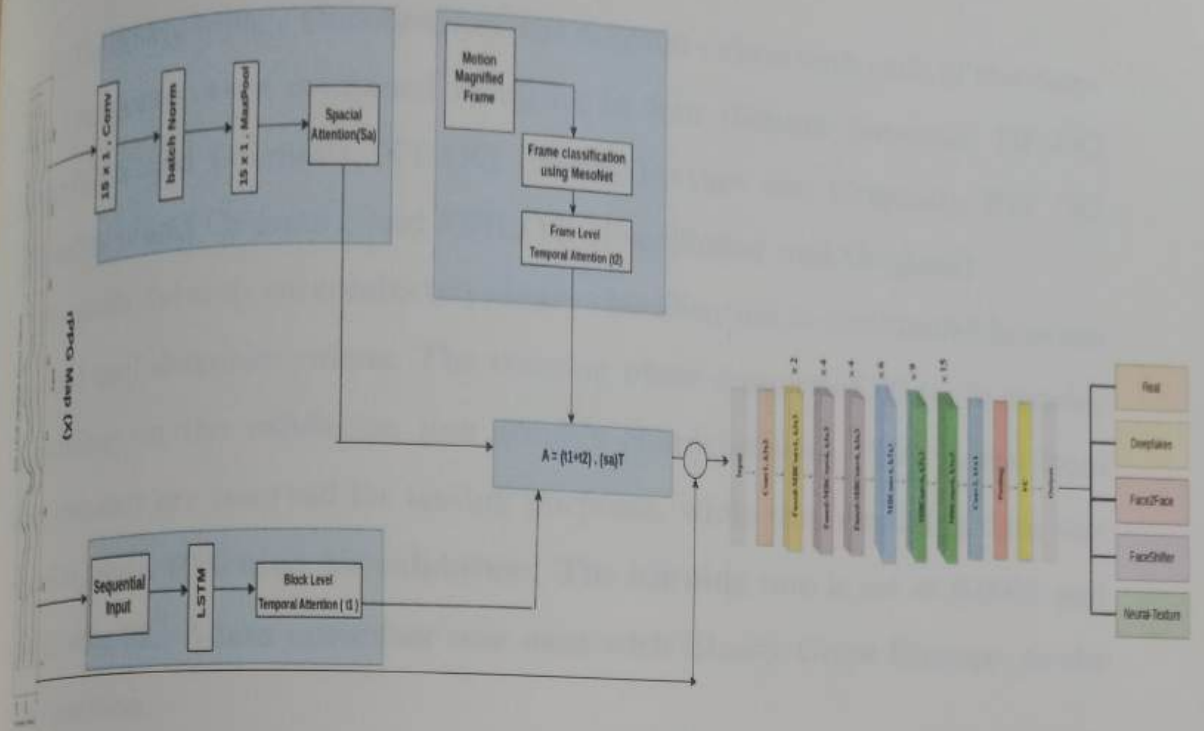


Figure 4.5: Architecture of Model

4.4 Implementation Detail

The preprocessing of the video leads to the creation of the rPPG map (X), which is then utilized to generate the spatial attention (Sa) and temporal attention weights ($t1$ and $t2$) through the LSTM and pre-trained MesoNet architecture, respectively. This resultant weight matrix will be

$$A = Sa \times (t1 + t2)^T$$

This weight matrix (A) is element-wise multiplied with the rPPG map (X) to create the attention rPPG map, which is subsequently fed into the classifier model. The EfficientNetV2 model is employed as the classifier. The training parameters of the EfficientNetV2, spatial attention network, and LSTM are jointly trained using categorical cross-entropy loss with the Adam optimizer. The parameters of the MesoNet model were trained independently. Detailed architecture is shown in Figure 4.5.

To facilitate binary classification, the original videos with each of the deepfake categories were combined resulting in four distinct datasets: DF_OG (Deepfakes and Original), NT_OG (NeuralTexture and Original), F2F_OG (Face2Face and Original), and FSH_OG (FaceShifter and Original).

For each dataset, we conducted binary classification to distinguish between original and deepfake videos. The training phase comprised 50 to 70 epochs, depending on the validation loss graph's characteristics. 200 videos from each dataset are reserved for testing purposes, while the remaining data was divided, with 10% used for validation. The learning rate is set at 0.0001 and batch size 32. Adam optimizer was used with Binary Cross Entropy as the loss function.

For the multi-class classification of videos, the dataset was created by combining videos from each category, including Original, DeepFakes, Face2Face, FaceShifter, and NeuralTexture, resulting in the ALL_OG dataset. Each video was labeled with a corresponding class: Original (0), DeepFakes (1), Face2Face (2), FaceShifter (3), and NeuralTexture (4). The model was trained for 50 epochs and batch size 32 using the Adam optimizer with a learning rate of 0.0001 and categorical crossentropy as the loss function.

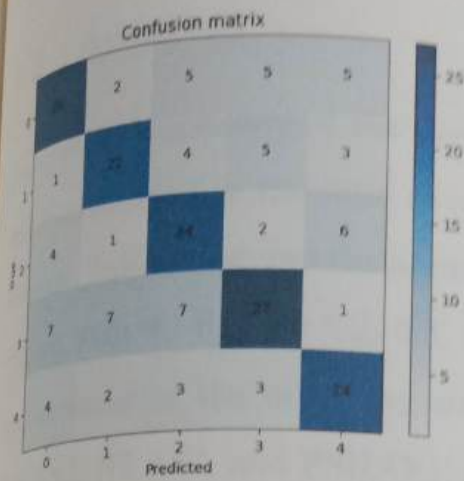
Chapter 5

Results and Discussion

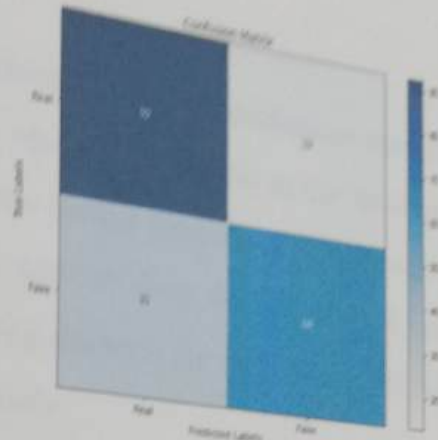
The evaluation metrics used include the confusion matrix for detailed classification performance analysis and accuracy for overall model performance assessment.

5.1 Confusion Matrix

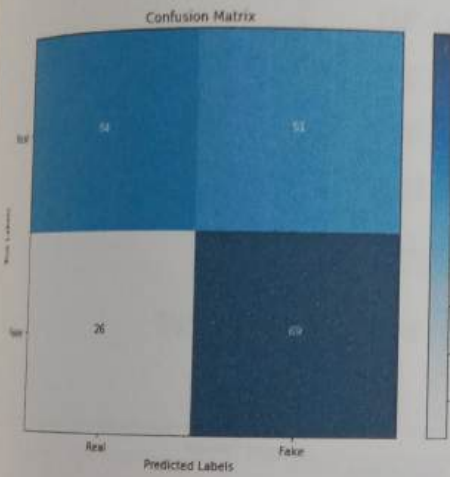
A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It consists of four primary components that are utilized to determine the classifier's evaluation metrics. The results of a classification algorithm of all the datasets are presented in a confusion matrix shown in Figure 5.1 .



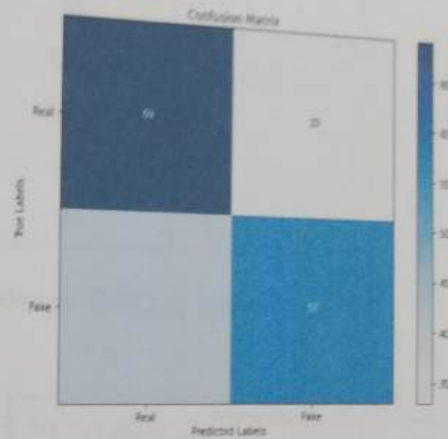
(a) ALL_OG



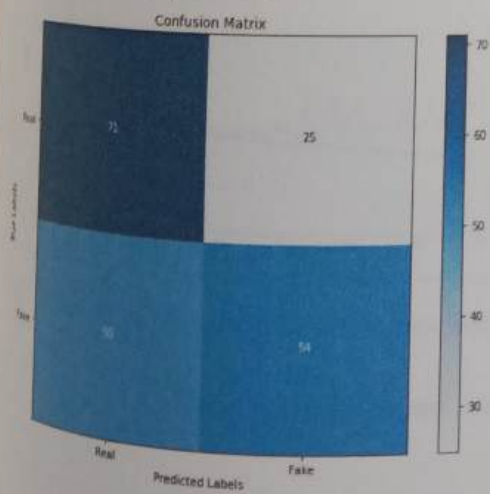
(b) DF_OG



(c) F2F_OG



(d) FSH_OG



(e) NT_OG

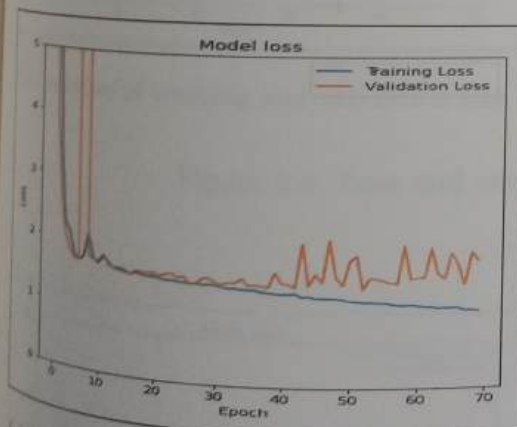
Figure 5.1: Confusion Matrix

5.2 Accuracy

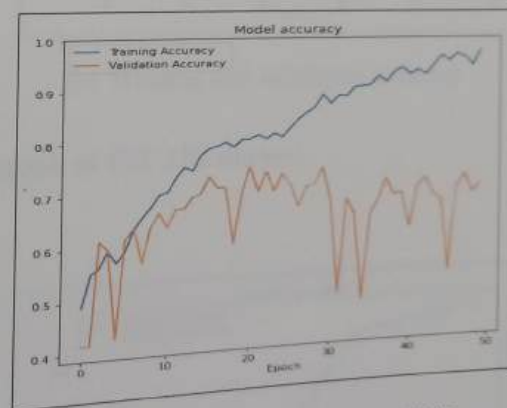
Accuracy is a measure of the overall correctness of the classification system and is calculated as the ratio of correctly classified instances to the total instances. Table 5.1 shows the accuracy of the testing data along with datasets trained with it. Figures 5.2, 5.3, 5.4, 5.5, and 5.6 show the variation of loss and accuracy of the validation and training datasets for ALL_OG, DF_OG, F2F_OG, NT_OG, and FSH_OG, respectively.

Dataset Name	Accuracy (%)
DF_OG	.78
NT_OG	62.5
F2F_OG	70.99
FSH_OG	63
ALL_OG	64

Table 5.1: Accuracy of Testing Data on Different Datasets

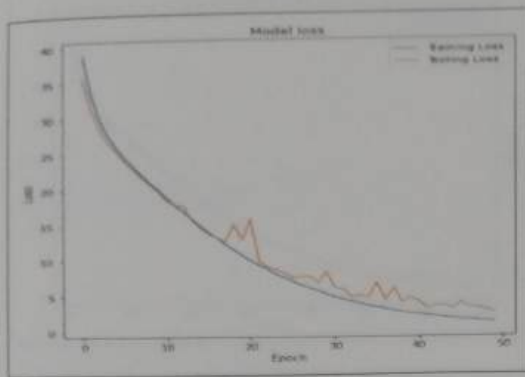


(a) Variation of training and validation loss

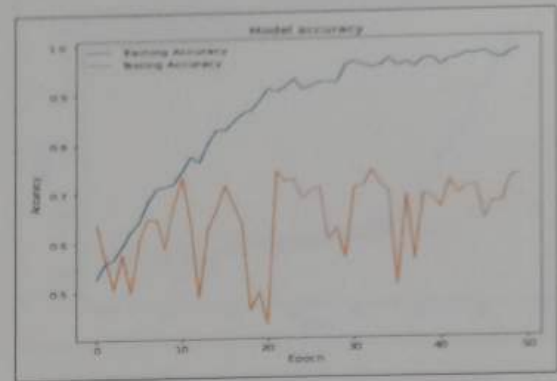


(b) Training and validation accuracy

Figure 5.2: Loss and accuracy graph of ALL_OG dataset

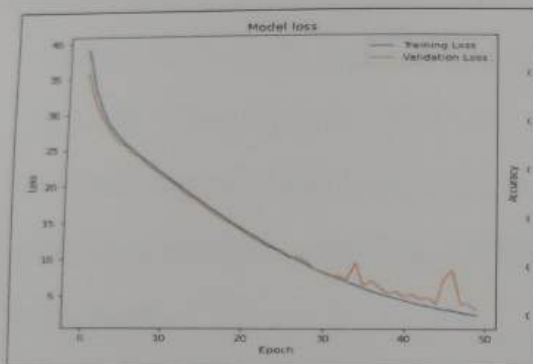


(a) Variation of training and validation loss

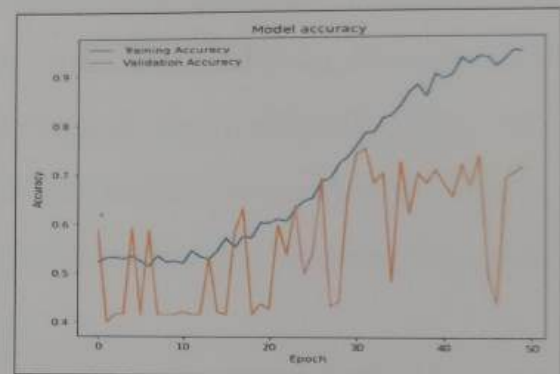


(b) Training and validation accuracy

Figure 5.3: Loss and accuracy graph of DF_OG dataset

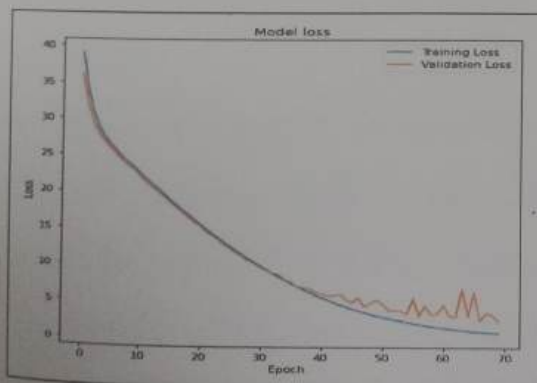


(a) Variation of training and validation loss

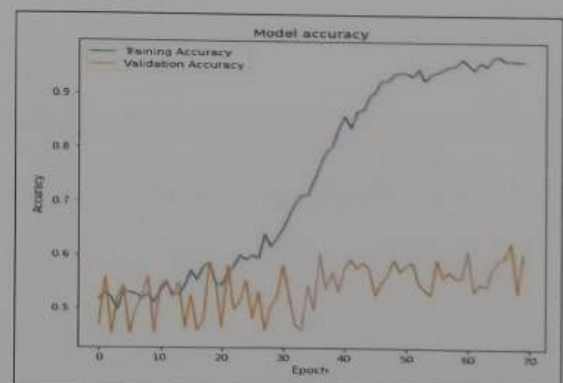


(b) Training and validation accuracy

Figure 5.4: Loss and accuracy graph of F2F_OG dataset

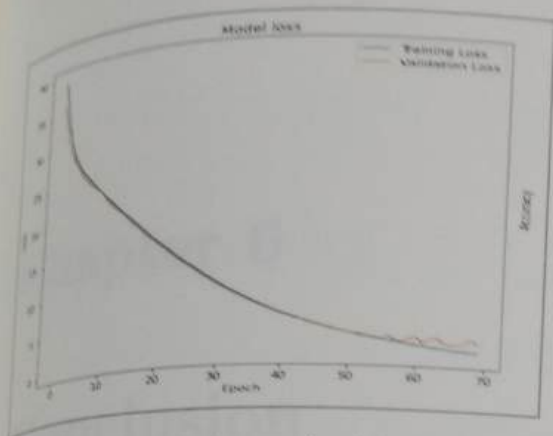


(a) Variation of training and validation loss

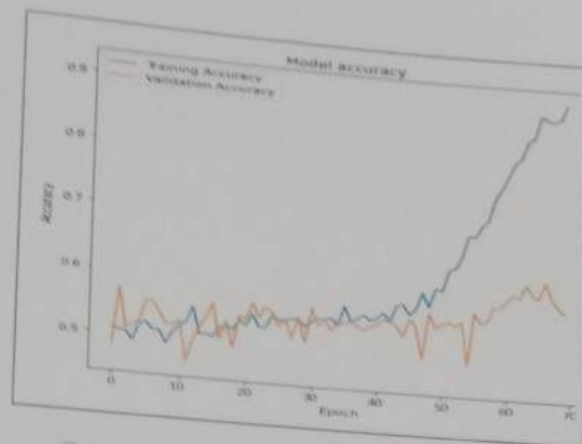


(b) Training and validation accuracy

Figure 5.5: Loss and accuracy graph of NT_OG dataset



(a) Variation of training and validation loss



(b) Training and validation accuracy

Figure 5.6: Loss and accuracy graph of FSH.OG dataset

Chapter 6

Conclusion

In conclusion, our project has effectively employed a combination of RPPG maps, spatial and temporal attention mechanisms, and EfficientNetV2 to classify videos into real or manipulated categories, covering four distinct deepfake manipulation methods. By harnessing the power of rPPG, our methodology has demonstrated a remarkable ability to accurately classify compressed videos, marking a significant advancement in deepfake detection technology.

We achieved an accuracy of 78% for DF_OG, 62.5% for NT_OG, 71% for F2F_OG, and 63% for FSH_OG, demonstrating the effectiveness of our approach for binary classification. Additionally, we attained a 64% accuracy in source classification, indicating our efficiency in identifying the source of deepfakes.

To further improve the accuracy of detecting moderately compressed deepfake videos, integrating our approach with complementary deepfake detection methods can be beneficial. Additionally, training the model with datasets comprising videos generated by state-of-the-art deepfake creation models like DeepFaceLab and StyleGAN could enhance its robustness. This strategy aims to leverage diverse deepfake generation techniques, improving the model's ability to generalize and detect a wider range of deepfake variations.

Bibliography

- [1] Shruti Agarwal and Hany Farid. Detecting deep-fake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 981–989, 2021.
- [2] Giuseppe Boccignone, Sathya Bursic, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Sabrina Patania. Deepfakes have no heart: A simple rppg-based method to reveal fake videos. In *International Conference on Image Analysis and Processing*, pages 186–195. Springer, 2022.
- [3] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1081–1088, 2021.
- [4] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2020.
- [5] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020.

- [6] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.
- [7] Bahar Uddin Mahmud and Afsana Sharmin. Deep insights of deepfake technology: A review. *arXiv preprint arXiv:2105.00192*, 2021.
- [8] Kundan Patil, Shrushti Kale, Jaivanti Dhokey, and Abhishek Gulhane. Deepfake detection using biological features: a survey. *arXiv preprint arXiv:2301.05819*, 2023.
- [9] Shobha Phansalkar, Judy Edworthy, Elizabeth Hellier, Diane L Seger, Angela Schedlbauer, Anthony J Avery, and David W Bates. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. *Journal of the American Medical Informatics Association*, 17(5):493–501, 2010.
- [10] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.
- [11] Jiahui Wu, Yu Zhu, Xiaoben Jiang, Yatong Liu, and Jiajun Lin. Local attention and long-distance interaction of rppg for deepfake detection. *The Visual Computer*, 40(2):1083–1094, 2024.
- [12] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. pages 379–389, 01 2022.
- [13] Yuezheng Xu, Ru Zhang, Cheng Yang, Yana Zhang, Zhen Yang, and Jianyi Liu. New advances in remote heart rate estimation and its ap-

plication to deepfake detection. In *2021 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 387–392. IEEE, 2021.