Title :- Deepfake Detection Using Deep learning Models

1. Abstract :

The rapid advancement of synthetic image manipulation techniques has raised concerns regarding their widespread societal impact. Despite efforts to develop reliable methods for identifying deepfake videos, current approaches often lack universality, allowing deceptive content to spread unchecked on social platforms. Accurately detecting these misleading deepfakes is crucial in mitigating their harmful effects. In this study, we propose an innovative deep learning framework designed specifically for detecting fraudulent video clips.

While traditional deepfake detection methods mainly focus on facial manipulation and expression alteration, establishing a comprehensive basis for real-time analysis of various fake videos poses a significant challenge. To tackle this, we introduce a hybrid approach that integrates successive video frames and utilizes a ResNeXt-Swish-BiLSTM model—a convolutional BiLSTM-based residual network optimized for classification. This method enables the identification of subtle artifacts indicative of deepfake manipulation, enhancing the model's ability to distinguish between authentic and synthetic content.

To assess the effectiveness and robustness of our model, we conducted extensive experiments using two well-known datasets: the DeepFake Detection Challenge dataset (DFDC) and the Celeb-DF dataset, along with the Face Forensics deepfake collections (FF++). Our results demonstrate an impressive accuracy of 99.24% when evaluated on the DFDC dataset alone, showcasing the efficacy of our approach in detecting synthetic content. Moreover, even when tested on the combined FF++, Celeb-DF, and DFDC datasets, our model achieves a commendable accuracy rate of 96.36%. These results highlight the superior performance of our proposed method compared to existing techniques, offering promising prospects for combating the spread of deepfake content in digital media.

Keywords: deepfake detection, deep learning, forensic analysis, video manipulation, authenticity

2. Introduction :

In today's digital landscape, the manipulation of images and videos poses a significant threat to society. With the advancement of artificial intelligence (AI), particularly machine learning (ML) methods, discerning between altered and original media has become increasingly challenging. Various conventional techniques, including content alteration and computer-generated imagery tools like GIMP, Photoshop, and CANVA, are employed for this purpose. Deepfake technology, rooted in Deep Learning (DL), has emerged as a powerful contender in the realm of video content customization, facilitated by DL networks (DNN) and algorithms like Generative Adversarial Networks (GAN).

Facial swapping, a key process in deepfake creation, involves replacing a person's face in a video with another individual's image while preserving the original movements and expressions. The proliferation of techniques such as StyleGAN and StyleGAN2 has made it difficult for the human eye to distinguish between real and synthesized images. Platforms like YouTube, Instagram, and TikTok, along with apps like ZAO2 and FaceApp3, have facilitated the widespread use of GAN-based face-swapping methods, blurring the line between reality and fabrication.

As deepfake technology becomes increasingly sophisticated, its misuse has led to various societal issues, including the spread of fake news and propaganda through manipulated images and videos. Detection of deepfakes remains challenging due to advancements in DL methods and the ability to adapt to environmental changes. To address these challenges, we propose a robust deepfake detection method: ResNeXt-Swish-BiLSTM.

This innovative approach combines the Residual Network (ResNeXt) architecture with the Swish activation function and Bidirectional Long Short-Term Memory (BiLSTM) layers for feature extraction and classification. By leveraging the smoothness of the Swish function and the capabilities of BiLSTM in capturing meaningful attributes from input features, our model achieves a high-recall rate in detecting deepfakes.

We evaluated the proposed model using datasets like DFDC, FF++, and Celeb-DF, ensuring its suitability and generalizability across different scenarios. Our approach demonstrates resilience against common cyberattacks such as compression, noise, blurring, and variations in scale and orientation.

The remainder of this paper is organized as follows: Section 2 reviews existing deepfake detection techniques, Section 3 details our methodology, Section 4 presents experimental results and analyses, and Section 5 concludes the study.

3. Related Work:

In contemporary deepfake generation, three prominent types are recognized: Face Swap (FS), Lip Sync, and Puppeteer. FS deepfakes entail replacing the face of a target person with that of a source individual to create a fictitious video. Lip sync deepfakes synchronize facial movements with audio, simulating the illusion of speech. Puppeteer deepfakes maintain the facial expressions of the source person on the target face, enhancing the authenticity of impersonation. While current detection methods often target specific types of deepfakes, comprehensive techniques capable of addressing all variations are less explored.

Agarwal et al. [2] introduced a lip sync detection method utilizing phoneme and viseme discrepancies, employing manual and CNN-based viseme-phoneme mapping. Existing detection methods broadly fall into two categories: those based on manually crafted features [3] and those employing Deep Learning (DL) [4,5]. For instance, Yang et al. [5] utilized a Support Vector Machine (SVM) classifier trained on 68-D face landmarks, effective for high-quality videos but less so for low-quality ones.

Massod et al. [6] exploited texture-based eye and teeth characteristics for detecting FS and F2F deepfakes, utilizing attributes like eye color and nose shape. However, this method is limited to faces with clean teeth and open eyes.

DL-based approaches, such as Doke et al.'s [3] LSTM-CNN model, leverage sequence inconsistencies for categorization but face challenges in accuracy for top-level deepfakes. Xia et al. [7] proposed MesoNet for deepfake identification, while Nguyen et al. [8] introduced a bubble network for detecting various manipulations, albeit with limitations in testing across diverse datasets.

Abdul Qadir[1] et al proposed ResNet- Swish- BiLSTM, an optimized convolutional BiLSTM- based residual network for deep fake detection with its limitation in less accuracy over testing while integrating diverse datasets.

While these advancements show promise, they often focus on specific deepfake types and may lack generalizability. Table 1 provides a summary of recent developments in the deepfake detection domain.

4. Proposed Methodology

This section outlines the details of our proposed system. The ResNeXt-Swish-BiLSTM deepfake video identification technique. Our approach integrates the Swish activation function, which enhances learning behavior by minimizing the movement of negative values across the network. This smoothness feature optimizes the model's performance. To extract facial landmark characteristics from input videos, we utilized the FaceRecognition Python library. The Residual Blocks (RBs) are responsible for extracting key facial features to discern between authentic and fake videos. Additionally, BiLSTM layers are employed to focus on capturing the most relevant features for deepfake frame classification.

4.1 Data Acquisition

To ensure the efficiency of our model for real-time prediction, we gathered data from various available datasets such as FF++ and DFDC. We augmented these datasets by creating a new dataset comprising a mixture of collected data. This approach enhances the model's accuracy and real-time detection capabilities across different types of videos. To mitigate training bias, we balanced the dataset with 70% real and 30% fake videos. Notably, we excluded audio-altered videos from the DFDC dataset, as audio deepfakes were beyond the scope of this study. Subsequently, we selected 1500 real and 1500 fake videos from the DFDC dataset, 1000 real and 1000 fake videos from the

FF++ dataset, and 500 real and 500 fake videos from the Celeb-DF dataset. This aggregation resulted in a total dataset comprising 3000 real and 3000 fake videos.

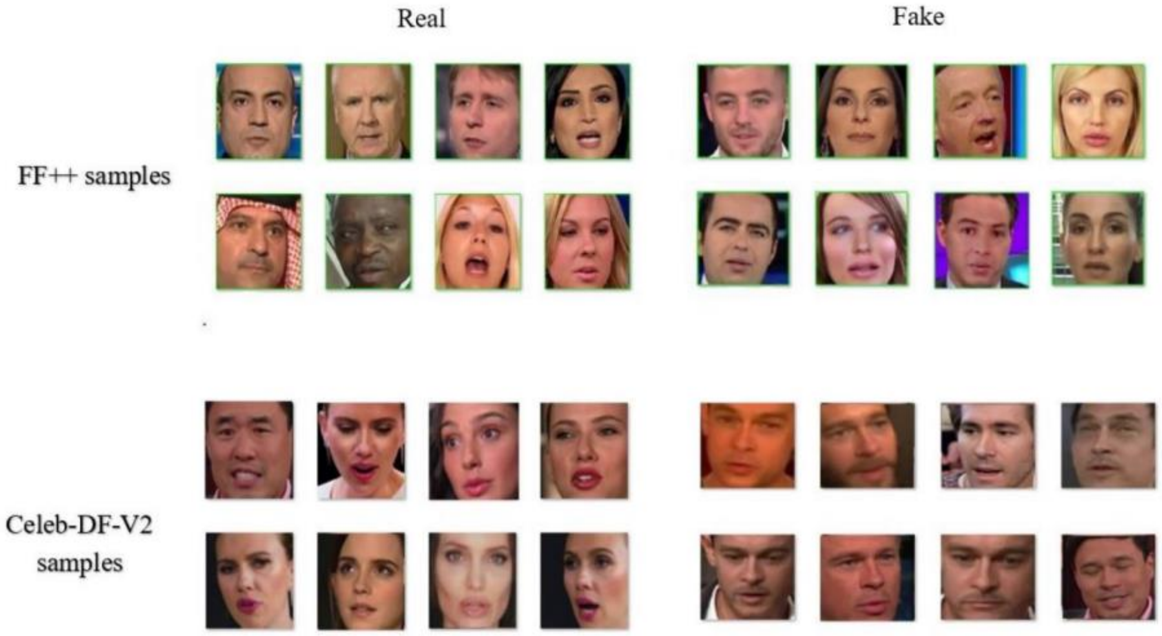

Fig 1. Samples from FF++ and Celeb-DF datasets

4.2. Pre-processing locating and cropping faces in videos

In this process, the system detects faces within extracted frames, with a focus on the facial area for visual modifications. We utilized the FaceRecognition module in Python to gather facial data. This method employs 2D and 3D facial landmarks to locate the facial region accurately. Seven points of interest (POI) were chosen for analysis, including the outward even edge (OE), outward left edge (LE), chin (C), frontal head (FH), outward proper cheek (PC), odd cheek (OC), and the center of the face (MF).ss

4.3 Standardization and segmentation

In our approach, we implement feature standardization during data loading by adjusting the extracted attribute using Equation (1):

$$z = \frac{X - \mu}{S}$$

Here, μ and S represent the mean and standard deviation of the feature columns, respectively, while X denotes the input face attribute vector. Our method accommodates both frame-based and segment-based operations. We've devised fragments with a frame duration of 100, incorporating a 25-frame overlap to facilitate smooth operation. To maintain computational efficiency, our projected resolution necessitates a frame rate of 25 frames per second.

4.4 Feature computation

Feature computation involves extracting relevant characteristics from video images after identifying faces. In our approach, we modified the pre-trained ResNext CNN network by incorporating the low-value elimination technique within the current framework to address this challenge. The rationale behind employing the Swish activation function is to enhance the model's ability to handle negative numbers throughout the neural network, thereby aiding in capturing complex underlying patterns in visual perception. Additionally, the inclusion of supplementary RNN (Recurrent Neural Network) layers and extra coatings assists in selecting a representative set of facial features that are subsequently used for classification.

4.5 Proposed model ResNeXt-swish-bilstm

The proposed model, named ResNext-Swish-BiLSTM, incorporates a combination of ResNext50 as the base CNN architecture, Swish activation function, and Bidirectional LSTM (BiLSTM) for processing spatiotemporal data. This model is designed to effectively handle the complexities of video data and classify it into authentic or deepfake categories.

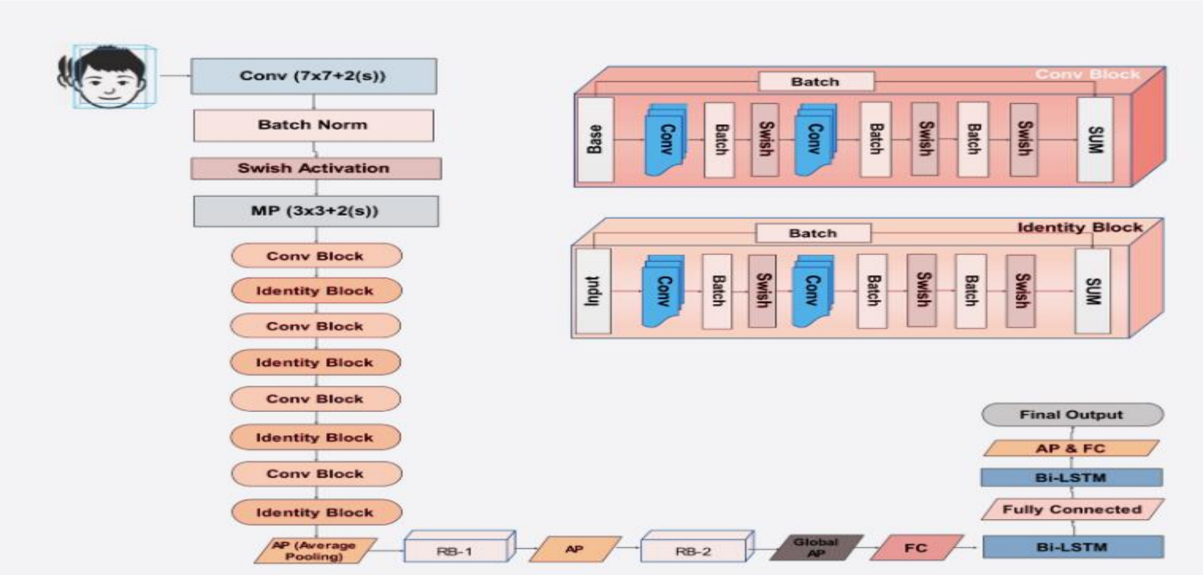

Fig 3. The ResNeXt - Swish - BiLSTM Architectural Description

Fig 3. shows how the ResNext50 architecture, pre-trained on a large dataset, serves as the backbone of the model. It is followed by an Adaptive Average Pooling layer to reduce the spatial dimensions of the feature maps. Subsequently, the feature maps are reshaped and fed into a Bidirectional LSTM (BiLSTM) layer. This BiLSTM layer processes the temporal information in both forward and backward directions, enabling the model to capture sequential dependencies effectively.

To enhance the model's learning capability and generalization, dropout regularization is applied before the final linear layer. This helps prevent overfitting by randomly dropping a fraction of the connections during training. The linear layer computes the final classification scores, which are then passed through the Swish activation function to introduce non-linearity and capture complex patterns in the data.

Overall, the proposed model architecture effectively combines convolutional and recurrent neural network components to process spatiotemporal information and classify video frames as authentic or deepfake with high accuracy.

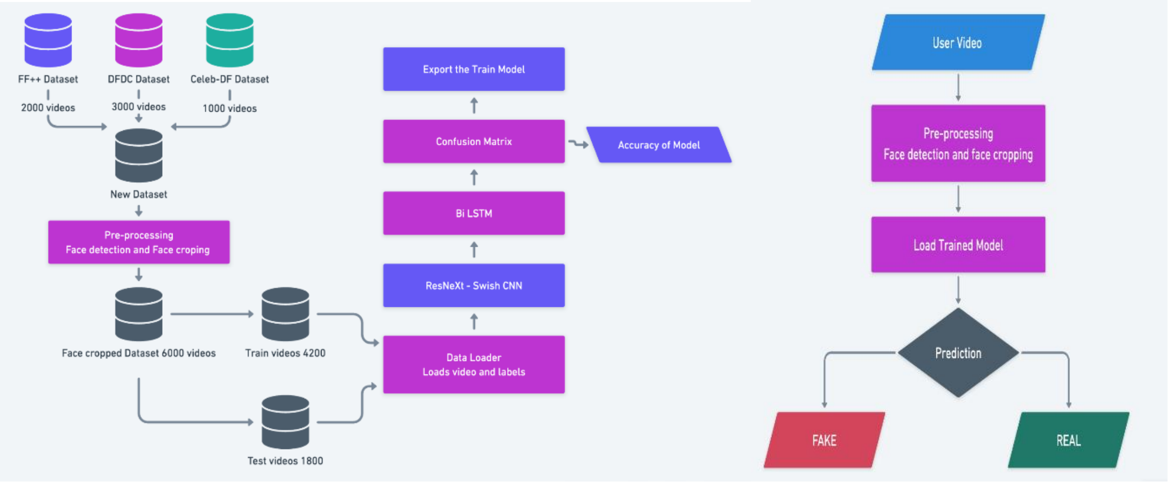

Fig 2. Proposed Work Flow

Fig 2. flowchart outlines the creation and maintenance of a deepfake detection model: video data (including datasets like FF++, DFDC, and Celeb-DF) is pre-processed (face detection and cropping), divided into training and testing sets, used to train deep learning models (like ResNext-Swish CNN and BILSTM), evaluated for accuracy, and finally exported for real-world use.

5. Result metric

The table compares the performance of different methods for detecting face simulations. Face simulations, also known as deepfakes, are artificial images or videos that have been manipulated to make it look like a real person is doing or saying something they never did. The table shows the title of the paper that described the method, the author(s) of the paper, the method used, the dataset the method was tested on, and the accuracy of the method.

Table 1. Comparative analysis with existing methods from various papers

Study	**Method**	**Dataset**	**Performance (AC)**
E.D. Cannas et al. [12]	Group of CNN	FF++(c23)	84%
Tarasiou et al. [13]	A lightweight architecture	DFDC	78.76%
Nirkin et al. [14]	FACE X-RAY	Celeb-DF	81.58%
Ciftci et al. [10]	Bio Identification	Celeb-DF	90.50%
Sabir et al. [11]	CNN + GRU + STN	FF++, DF	96.90%
		FF++, F2F	94.6%
Abdul Qadir et al. [1]	ResNet-Swish-BiLSTM	FF++ and DFDC	78.33%
		FF++ and F2F	98.08%
	Proposed Method	FF++, DFDC and Celeb-DF	**96.36%**
		FF++ and DFDC	**95.66%**
		DFDC	**99.23%**
		FF++	**97.82%**

Table 1. Shows compares the results from our paper(Proposed Method) to other papers that we have referred from. The methods are all based on deep learning, a type of machine learning that uses artificial neural networks to learn patterns in data. The datasets used to train and test the methods include FF++, DFDC, and Celeb-DF. The accuracy of the method is measured by the Area Under the Curve (AUC), which is a measure of how well a classification model can distinguish between positive and negative examples.

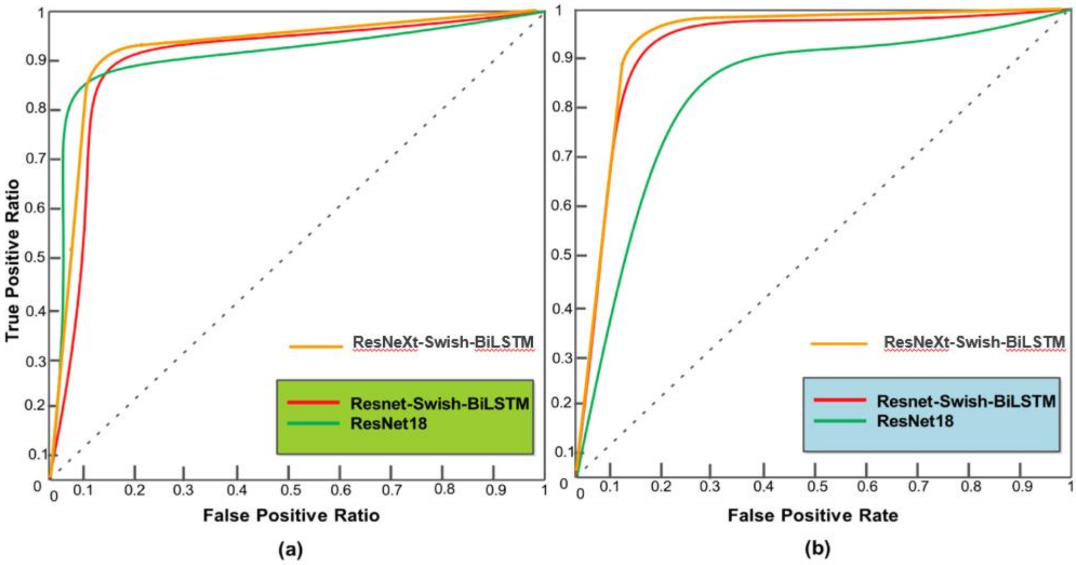

Fig 4. AOC and ROC Curve Comparison a. FF++ Dataset b. DFDC dataset

Fig 5. Tells how above AOC curve results show that our model is working better than other models on the datasets FF++ and DFDC with an area above 0.96 for both.

Overall, the methods achieve high accuracy on the datasets they are tested on. The highest accuracy, 99.24%, is achieved by the method proposed in the paper "Recurrent convolutional strategies for face manipulation detection in videos" by Sabir et al. This method uses a combination of convolutional neural networks (CNNs), gated recurrent units (GRUs), and spatial transformer networks (STNs).

6. Conclusion

This study has provided a comprehensive strategy to identify deepfakes based on the fusion of our unique facial features. Unlike many other systems, the proposed model is easy to use, understandable, efficient, and resilient at the same time. We introduced a novel ResNeXt-Swish-BiLSTM deepfake detection model. The utility of the proposed visual tampering detection technique was extensively tested on the deepfake data source FF++, DFDC and Celeb-DF. We evaluated the proposed method across deepfake collection to demonstrate its generalizability for unusual scenarios. We found that the suggested modelling technique can distinguish between the modified and the unmodified digital footage with a high-recall rate and recognizes various visual modifications. The FF++and DFDC and Celeb-DF combined dataset, which show the highest Accuracy value of 0.9636 has been used to evaluate the proposed technique in detail, respectively. Therefore, after thoroughly examining the ResNeXt-Swish-BiLSTM at the statistical and digital media levels, we can conclude that our work in the field of advanced digital investigation, such as criminal forensics.

6. Future Work

There is need for further development in deepfake detection methods to address the increasing sophistication of forgeries. Promising future directions include leveraging Generative Adversarial Networks (GANs) to generate more realistic training data, exploring multimodal analysis of video and audio content, developing interpretable AI models for transparent decision-making, continuously adapting to novel manipulation techniques, and achieving real-time detection capabilities to mitigate the spread of misinformation.

7. References

1. Qadir, A., Mahum, R., El-Meligy, M. A., Ragab, A. E., AlSalman, A., & Awais, M. (2024). An efficient deepfake video detection using robust deep learning. Heliyon, 10(5).

2. Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020). Detecting deep-fake videos from phoneme-viseme mismatches. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 660-661).

3. Kolagati, S., Priyadharshini, T., & Rajam, V. M. A. (2022). Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model. International Journal of Information Management Data Insights, 2(1), 100054.

4. Y. Doke, P. Dongare, V. Marathe, M. Gaikwad and M. Gaikwad, "Deep Fake Video Detection Using Deep Learning." Journal homepage: www.ijrpr.com ISSN, vol. 2582, pp. 7421.

5. Yang, X., Li, Y., & Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8261-8265). IEEE.

6. Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence, 53(4), 3974-4026.

7. Xia, Z., Qiao, T., Xu, M., Wu, X., Han, L., & Chen, Y. (2022). Deepfake video detection based on MesoNet with preprocessing module. Symmetry, 14(5), 939.

8. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019, May). Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2307-2311). IEEE.

9. Suratkar, S., & Kazi, F. (2023). Deep fake video detection using transfer learning approach. Arabian Journal for Science and Engineering, 48(8), 9727-9737.

10. Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE transactions on pattern analysis and machine intelligence.

11. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), 3(1), 80-87.

12. Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021, January). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012-5019). IEEE.

13. Zhang, W., Zhao, C., & Li, Y. (2020). A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. *Entropy*, *22*(2), 249.

14. Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2021). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6111-6121.

15. https://www.kaggle.com/datasets/jaswanthravi/deepfakedata/data