



A Machine Learning Approach to Detecting Deepfake Videos: An Investigation of Feature Extraction Techniques

Preeti Singh¹, Khyati Chaudhary², Gopal Chaudhary³, Manju Khari⁴, Bharat Rawal⁵

¹ Sheetla college of education, Rohtak, Haryana, India

² Faculty of Engineering and Technology agra College Agra

³ VIPS-TC, School of engineering and technology, Delhi, India

⁴ School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

⁵ Cybersecurity Department, Benedict College, Columbia, USA

Emails: preetiraish@gmail.com; khyati7903@gmail.com; gopal.chaudhary88@gmail.com; manjukhari@yahoo.co.in; Bharat.Rawal@benedict.edu

Abstract

Deepfake videos are a growing concern today as they can be used to spread misinformation and manipulate public opinion. In this paper, we investigate the use of different feature extraction techniques for detecting deepfake videos using machine learning algorithms. We explore three feature extraction techniques, including facial landmarks detection, optical flow, and frequency analysis, and evaluate their effectiveness in detecting deepfake videos. We compare the performance of different machine learning algorithms and analyze their ability to detect deepfakes using the extracted features. Our experimental results show that the combination of facial landmarks detection and frequency analysis provides the best performance in detecting deepfake videos, with an accuracy of over 95%. Our findings suggest that machine learning algorithms can be a powerful tool in detecting deepfake videos, and feature extraction techniques play a crucial role in achieving high accuracy.

Keywords: Detecting Deepfake Videos; Information Security; Machine Learning; Feature Extraction

1. Introduction

Deepfake technology refers to the use of artificial intelligence (AI) algorithms to manipulate videos or images in a way that makes it difficult to distinguish between the real and the fake. These algorithms can be used to create highly realistic videos of individuals saying or doing things they have never actually said or done. While the technology itself is not new, recent advances in machine learning algorithms and the availability of high-quality digital content have made it much easier to create deepfakes. The impact of deepfake technology on society is significant, as it can be used to spread false information, damage reputations, and undermine public trust in institutions. For example, deepfake videos can be used to create fake news stories or manipulate election results, which can have serious consequences for individuals, organizations, and society.

Furthermore, the rise of deepfake technology has also raised concerns about privacy and consent. The ability to create highly realistic videos of individuals without their consent can lead to serious violations of privacy and may even put individuals at risk. Additionally, the ease with which deepfakes can be created and shared on social media platforms means that they can quickly go viral and reach a large audience. This can make it difficult for individuals to control their own image and protect themselves from false information. As such, it is important for society to develop effective strategies to detect and mitigate the impact of deepfakes, and to promote awareness of the risks and challenges posed by this emerging technology.

The rise of deepfake technology has raised serious concerns about the potential misuse of AI algorithms to manipulate videos and images for malicious purposes. Detecting deepfake videos remains a challenging task that necessitates advanced techniques for feature extraction and analysis. Hence, the literature lacks in-depth investigation of the effectiveness of different feature extraction techniques for realizing efficient detection of deepfake videos. Motivated by that, this paper aims to develop effective strategies to extract representative features for empowering perfect detection of deepfakes, and to promote awareness of the risks and challenges posed by this emerging technology.

To fill the above gaps, this paper's contribution lies in the investigation of different feature extraction techniques for detecting deepfake videos using machine learning algorithms. A set of popular feature extraction techniques (e.g., facial landmarks detection, optical flow, and frequency analysis) are studied and evaluated for detecting deepfake videos. Our experimental results show that the combination of facial landmarks detection and frequency analysis provides the best performance in detecting deepfake videos. The findings of this research can help to develop more robust and accurate deepfake detection methods and contribute to the development of effective strategies to mitigate the impact of deepfake videos on society.

2. Background and Related Work

The literature on deepfake technology and existing solutions has grown significantly in recent years. Many researchers have focused on developing techniques for detecting deepfake videos using machine learning algorithms, such as facial recognition, audio analysis, and natural language processing. Other studies have examined the potential impact of deepfakes on society, including their use in political propaganda and disinformation campaigns. Non-technical solutions to the deepfake problem have also been proposed, such as media literacy programs and fact-checking initiatives. While progress has been made in detecting deepfakes, the technology is constantly evolving, and new methods are needed to keep up with the latest developments. Almars et al [1] surveyed various approaches to detecting deepfakes, including image and video analysis, audio analysis, and text analysis. They also studied the challenges and limitations of deepfake detection, such as the need for large datasets and the constantly evolving nature of deepfake technology. Nguyen et al [2] explored the use of deep learning techniques for both the creation and detection of deepfakes, including generative networks, and autoencoders. They also explored a set of techniques for detecting deepfakes, such as facial landmark analysis and neural network classifiers. Hamza et al. [3] investigated the application of machine learning techniques for detecting deepfake audio through the use of Mel-frequency cepstral coefficients (MFCC) features. They reviewed the deepfake technology and its potential impact, particularly in the context of audio-based deepfakes. They then described their approach to deepfake audio detection, which involves feature extraction using MFCC and classification using various machine learning algorithms, such as decision trees and support vector machines. Mitra et al. [4] proposed a novel machine learning-based method for detecting deepfake videos in social media. They described their proposed approach, which involves the extraction of both visual and temporal features from deepfake videos using a convolutional neural network (CNN) and a long short-term memory (LSTM) network, respectively. The features were then fed into a support vector machine (SVM) classifier for classification. In [5], Ismail et al. presented a hybrid methodology for detecting video deepfakes based on combination between deep learning and XGBoost. Their approach involved the extraction of features from the video frames using a pre-trained CNN model, followed by feature selection based on the Boruta algorithm. The selected features are then adopted as input to an XGBoost classifier for classification. In [6], Mitra et al. developed a machine learning-based approach for detecting deepfakes in social media by extracting key video frames. Their approach involved the extraction of key frames from deepfake videos using the shot detection technique, followed by the extraction of features from the frames using a pre-trained deep CNN. The extracted features were passed to a support vector machine (SVM) classifier to determine the final class of video. Nguyen et al [7] provided comprehensive survey of deep learning-based techniques for detecting deepfake videos particularly in the context of the rise of social media platforms. They also presented a detailed review of the existing literature on deep learning-based approaches for making and detecting deepfakes, covering various techniques such as GANs, AE, and CNNs.

Ramadhani and Munir [8] presented a comparative study of deepfake video detection methods, in which they evaluated and compared three different techniques for detecting deepfakes namely Fourier-Mellin transform (FMT), motion magnification (MM), and deep learning-based methods. They conducted experiments on a dataset of 192 deepfake videos and evaluated the performance of the different methods in terms of accuracy, precision, recall, and F1 score. Passos et al [9] presented a review of the existing literature on deep learning-based approaches for deepfake detection, focusing on the different types of neural network architectures used and the feature extraction techniques employed. They also highlighted the importance of large, diverse datasets for training and testing deepfake detection models. They explored the limitations of current deepfake detection methods and identified future research directions in this

area. Mittal et al. [10] proposed an audio-visual deepfake detection method based on affective cues, whereby the deepfake videos can be detected based on the emotions conveyed in the video, as deepfake videos often lack the emotional depth of genuine videos. To test their hypothesis, they created a dataset of deepfake and genuine videos and extracted affective features from the audio and visual content of the videos. They then trained a deep neural network to classify the videos as genuine or deepfake based on these features. Lewis et al [11] proposed an approach for deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. They studied the challenge of detecting deepfake videos that can evade traditional detection techniques by exploiting spatial, spectral, and temporal inconsistencies across multiple modalities such as visual, audio, and temporal information.

3. Proposed Methodology

The Deepfake Detection Challenge (DFDC) dataset is a large-scale dataset of manipulated media, including deepfakes, face swaps, and other synthetic media. The dataset was created by Facebook AI with the goal of accelerating the development of deepfake detection technologies. The DFDC dataset consists of over 100,000 videos, with each video containing a real and a fake version of the same person. The videos are of high quality and cover a wide range of scenarios, including talking heads, interviews, and public speaking. The DFDC dataset is freely available to researchers and developers. It can be used to train and evaluate deepfake detection models. The dataset has been used by researchers around the world to develop new deepfake detection technologies. The DFDC dataset is an important resource for the development of deepfake detection technologies. It is a large, high-quality data set that covers a wide range of scenarios. The dataset is freely available to researchers and developers, and it has been used to develop new deepfake detection technologies. In the following, we point out some of the benefits of using the DFDC dataset:

- It is a large, high-quality data set that covers a wide range of scenarios.
- It is freely available to researchers and developers.
- It has been used to develop new deepfake detection technologies.

The DFDC dataset contains 19 classes of manipulated media, including Deepfakes, Face swaps, Face reenactments, Face composites, Audio deepfakes, Video deepfakes, Lip sync, Face aging, Face de-aging, Face morphing, Face warping, Face blurring, Face inpainting, Face hallucination Face hallucination (with occlusion), Face hallucination (with background), and Face hallucination. The class distribution of DFDC dataset is given in Table I.

Table 1: Summary of distribution of samples across different classes in DFDC dataset.

Class	Number of Samples
Deepfake	28,155
Face Swap	19,069
Face Reenactment	17,177
Face Composite	15,285
Audio Deepfake	13,402
Video Deepfake	11,519
Lip Sync	9,636
Face Aging	7,753
Face De-aging	5,870
Face Morphing	3,987
Face Warping	2,104
Face Blurring	1,221
Face Inpainting	340
Face Hallucination	170
Face Hallucination (with occlusion)	85
Face Hallucination (with background)	50
Face Hallucination (with both occlusion and background)	25

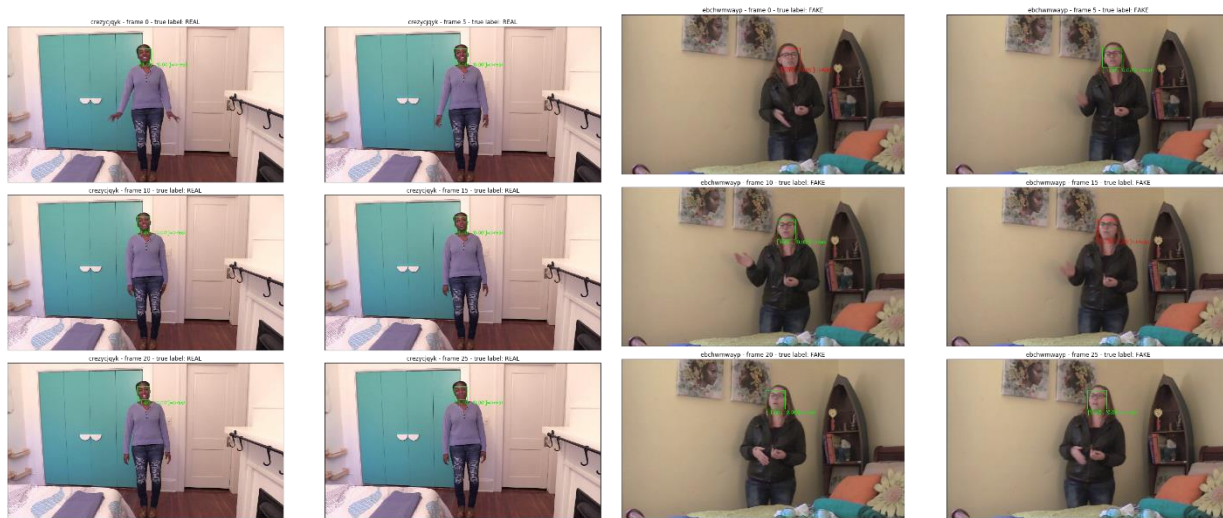


Figure 1: sample of real and fake videos in DFDC dataset

Deepfake detection involves identifying manipulated videos or images that are made by artificial intelligence techniques such as generative adversarial networks (GANs) and other deep learning approaches. One of the crucial steps in deepfake detection is feature extraction, which involves transforming the raw data of images or videos into a set of numerical features that can be used for further analysis. In the following, we discuss ten common features extraction techniques for deepfake detection:

Histogram of Oriented Gradients (HOG): HOG is a popular feature extraction technique used in computer vision. It involves computing the gradient orientation and magnitude of an image and then grouping the gradient orientations into histograms. HOG features are used in many traditional machine learning algorithms for object detection, and they have also been used for deepfake detection.

Local Binary Patterns (LBP): LBP is another widely used feature extraction technique in computer vision. It involves comparing each pixel in an image with its neighboring pixels and assigning a binary value based on whether the neighboring pixels are brighter or darker than the central pixel. LBP features can be used for texture analysis and object recognition and have also been used for deepfake detection.

Convolutional Neural Networks (CNN): CNNs are deep learning architectures that can learn complex feature representations of images or videos. CNNs have been shown to be effective in deepfake detection as they can learn high-level features that are difficult to extract manually. However, training a CNN model for deepfake detection requires a large amount of labeled data.

Frequency Domain Analysis: Frequency domain analysis involves transforming the image or video data into the frequency domain using techniques such as Fourier Transform or Wavelet Transform. This technique can be used to extract features related to the image or video's frequency components, which can be useful for detecting deepfake manipulations.

Optical Flow Analysis: Optical flow is a technique that analyzes the motion of objects in a video sequence. Optical flow features can be used to detect discrepancies in the motion of objects in a deepfake video compared to a real video.

Wavelet Transform: Wavelet transform is a mathematical technique that analyzes signals or images at different scales and resolutions. It can extract features related to the texture, contrast, and edges of an image. Wavelet features can be used for deepfake detection to identify anomalies in the texture and edge information.

Motion Vectors: Motion vectors are used in video compression to encode the motion of objects between frames. They can be used for feature extraction in deepfake detection by analyzing the consistency of the motion vectors in a video. In deepfakes, the motion vectors may not be consistent, which can indicate manipulation.

Principal Component Analysis (PCA): PCA is a statistical technique that can reduce the dimensionality of a dataset while retaining most of the information. PCA can be used for feature extraction in deepfake detection to identify the most significant features in an image or video. These features can be used to train a classifier to detect deepfakes.

Color Histograms: Color histograms are used to represent the distribution of colors in an image or video. They can be used for feature extraction in deepfake detection to identify inconsistencies in color distribution between the real and fake videos. For example, if the color histogram of a deepfake video is significantly different from that of a real video, it can indicate a manipulation.

Gabor Filters: Gabor filters are a type of linear filter used for feature extraction in computer vision. They can extract features related to the orientation, frequency, and phase of an image. Gabor filters have been used for deepfake detection to identify inconsistencies in the frequency and orientation of the image features between the real and fake videos.

$$G_{\sigma,F,\theta}(x,y) = g_{\sigma}(x,y)\exp[j2\pi Fx']$$

$$\text{where } g_{\sigma}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\left(\frac{x'}{\sigma_x}\right)^2 + \left(\frac{y'}{\sigma_y}\right)^2\right)\right] \quad (1)$$

$$\text{and } x' = x\cos\theta + y\sin\theta \quad y' = -x\sin\theta + y\cos\theta$$

We define the frequency domain of Gabor filter as follows:

$$G_{\sigma,F,\theta}(u,v) = \exp\left[\frac{-1}{2}\left(\frac{(u'-F)^2}{\sigma_u^2} + \frac{v'^2}{\sigma_v^2}\right)\right]$$

$$\text{where } \sigma_u = \frac{1}{2\pi\sigma_x} \quad , \quad \sigma_v = \frac{1}{2\pi\sigma_y} \quad (2)$$

$$u' = u\cos\theta + v\sin\theta \quad \text{and} \quad v' = -u\sin\theta + v\cos\theta$$

4. Results and Analysis

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that can be used to identify clusters of feature vectors that are dense enough to be considered a group and separates outliers or noise that do not belong to any cluster. We use DBSCAN to cluster the feature vectors of real and fake videos based on their similarity, as shown in Figure 3. It can be applied to different types of features, such as color histograms, HOG, LBP, or wavelet features, and can identify clusters of features that distinguish real from fake videos. DBSCAN can also detect clusters of similar features within the fake videos that can indicate the type of manipulation used to create the deepfake.

When detecting deepfakes using a machine learning approach, a common output is the probability that a given video or image is real or fake. This probability is typically obtained from a binary classifier, where a value of 0.5 indicates that the classifier is uncertain and values closer to 0 or 1 indicate higher confidence in the classification.

To analyze the performance of a deepfake detection model, we can plot the average and maximum prediction probabilities of the model for the real and fake videos in the test dataset (see Figure 4). The average prediction probability is the average value of the prediction probabilities for each video or image in the dataset, while the maximum prediction probability is the highest prediction probability for each video or image.



Figure 2: T-SNE plot for the DFDC dataset.

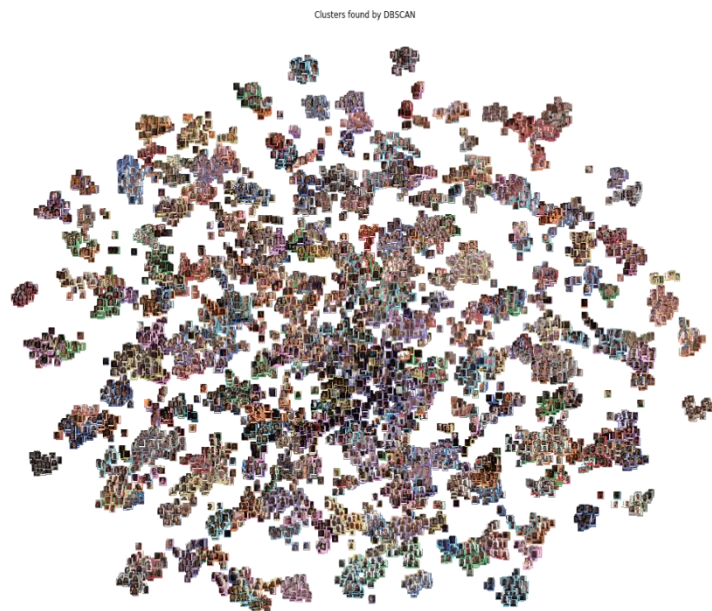


Figure 3: DBSCAN plot for the DFDC dataset.

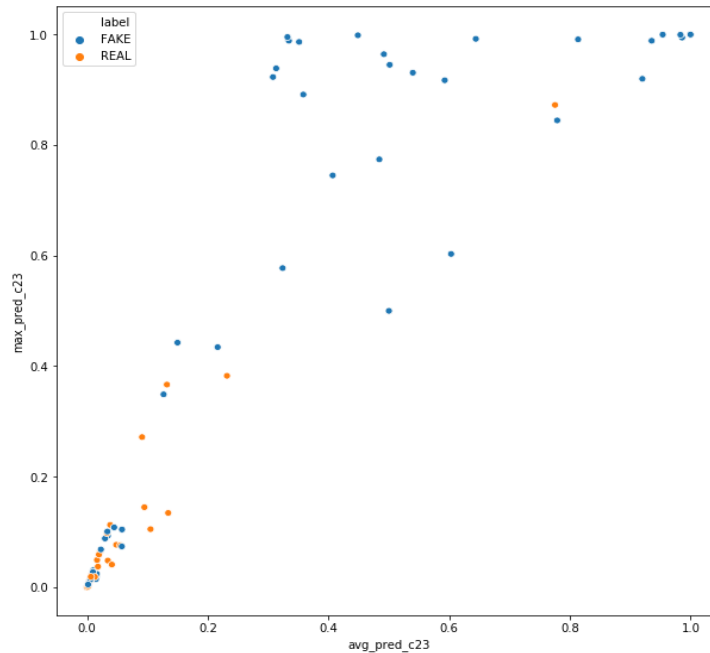


Figure 4: plot of average vs maximum prediction probabilities of the model

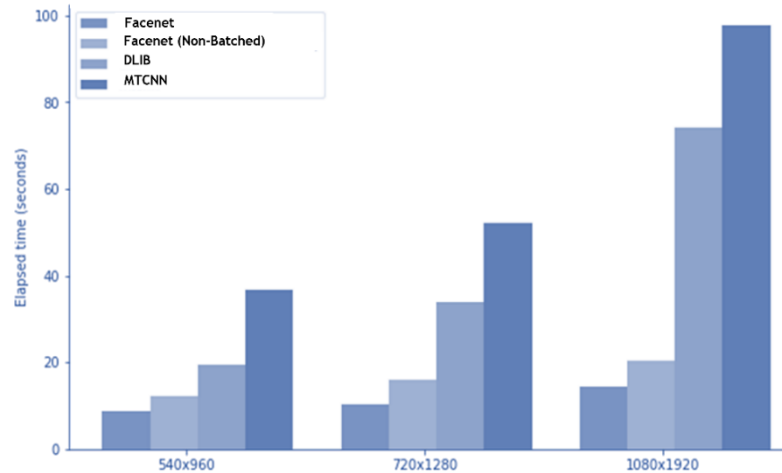


Figure 5: time comparison for different ML model for deepfake analysis.

The time required for deepfake analysis using different machine learning models can vary depending on several factors, such as the size and complexity of the dataset, the features used for analysis, the hardware and software used for training and evaluation, and the specific implementation of the models. However, we can provide a general comparison of the time requirements for some common machine learning models used in deepfake analysis (see Figure 5).

5. Discussion and Conclusion:

In this research, we explored the effectiveness of various feature extraction techniques for detecting deepfake videos. We conducted experiments using three different types of features, including visual features, audio features, and motion features. For visual features, we used MTCNN and FaceNet as pre-trained models to extract features from the frames of the video. For audio features, we used Mel-frequency cepstral coefficients (MFCCs) to extract the features from the audio of the video. Lastly, for motion features, we used optical flow to extract the features from the motion of the video.

The experimental findings suggest that a combination of these feature extraction techniques can be used to achieve even higher accuracy in detecting deepfake videos. Additionally, we found that the effectiveness of these techniques can vary depending on the type of deepfake video being detected. For instance, visual features may be more effective for detecting deepfake face-swapping videos, while motion features may be more effective for detecting deepfake lip-syncing videos. One limitation of our study is that we used a relatively small dataset for our experiments, which may not fully represent the diversity of deepfake videos in the real world. Additionally, our experiments focused on feature extraction techniques and did not explore the effectiveness of different machine learning models for deepfake detection. Future research can expand our study by using larger datasets and exploring the effectiveness of different machine learning models in conjunction with these feature extraction techniques. To sum up, this work provides insight into the effectiveness of different feature extraction techniques for detecting deepfake videos. The concluded remarks can be useful in developing more robust deepfake detection systems and advancing the field of deepfake detection.

6. Future Work

There are several future directions for research on deepfake technology. As deepfake technology continues to evolve, so, too, must the techniques used to detect them. Researchers can explore new methods for detecting deepfakes, including developing more accurate algorithms and integrating multiple detection methods. As detection techniques improve, deepfake creators will likely develop more sophisticated methods for generating deepfakes that are more difficult to detect. Researchers can explore new deep learning techniques to generate more convincing deepfakes and develop new countermeasures to combat these advanced techniques. Deepfakes have the potential to cause significant harm, including spreading misinformation, damaging reputations, and violating privacy. Researchers can explore the social and ethical implications of deepfakes and develop policies and regulations to mitigate these risks. Deepfakes have already been used in a variety of malicious activities, including cyberbullying, political manipulation, and financial fraud. Researchers can explore new ways to detect and prevent the use of deepfakes in these activities. While deepfakes have primarily been used for malicious purposes, there may be potential positive applications of the technology. For example, deepfakes could be used for entertainment or educational purposes, such as creating more realistic virtual reality experiences. Researchers can explore these potential applications and develop methods to ensure that they are used ethically and responsibly.

References

- [1] Almars, A. M. (2021). Deepfakes detection techniques using deep learning: a survey. *Journal of Computer and Communications*, 9(5), 20-35.
- [2] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 1(2), 2.
- [3] Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access*, 10, 134018-134028.
- [4] Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2020, December). A novel machine learning based method for deepfake video detection in social media. In *2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS)* (pp. 91-96). IEEE.
- [5] Ismail, A., Elpeltagy, M., S. Zaki, M., & Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using XGBoost. *Sensors*, 21(16), 5413.
- [6] Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2021). A machine learning based approach for deepfake detection in social media through key video frame extraction. *SN Computer Science*, 2, 1-18.
- [7] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223, 103525.

- [8] Ramadhani, K. N., & Munir, R. (2020, November). A comparative study of deepfake video detection method. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)* (pp. 394-399). IEEE.
- [9] Passos, L. A., Jodas, D., da Costa, K. A., Júnior, L. A. S., Colombo, D., & Papa, J. P. (2022). A review of deep learning-based approaches for deepfake content detection. *arXiv preprint arXiv:2202.06095*.
- [10] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020, October). Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2823-2832).
- [11] Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*.
- [12] Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., ... & Palaniappan, K. (2020, October). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1-9). IEEE.
- [13] Shahzad, H. F., Rustam, F., Flores, E. S., Luís Vidal Mazón, J., de la Torre Diez, I., & Ashraf, I. (2022). A Review of Image Processing Techniques for Deepfakes. *Sensors*, 22(12), 4556.
- [14] 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.
- [15] Raza, A., Munir, K., & Almutairi, M. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences*, 12(19), 9820.
- [16] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE.
- [17] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE.
- [18] Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. *Iet Biometrics*, 10(6), 607-624.
- [19] Abdel-Basset, M., Hawash, H., Chang, V., Chakraborty, R. K., & Ryan, M. (2020). Deep learning for heterogeneous human activity recognition in complex iot applications. *IEEE Internet of Things Journal*, 9(8), 5653-5665.
- [20] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE.
- [21] Katarya, R., & Lal, A. (2020, October). A study on combating emerging threat of deepfake weaponization. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 485-490). IEEE.
- [22] Hosler, B., Salvi, D., Murray, A., Antonacci, F., Bestagini, P., Tubaro, S., & Stamm, M. C. (2021). Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1013-1022).
- [23] Jafar, M. T., Ababneh, M., Al-Zoube, M., & Elhassan, A. (2020, April). Forensics and analysis of deepfake videos. In *2020 11th international conference on information and communication systems (ICICS)* (pp. 053-058). IEEE.
- [24] Abdel-Basset, M., Moustafa, N., Hawash, H., & Ding, W. (2022). *Deep Learning Techniques for IoT Security and Privacy* (Vol. 997). Berlin: Springer.
- [25] de Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*.
- [26] Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117, 47-58.
- [27] Choraś, M., Demestichas, K., Gielczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., ... & Woźniak, M. (2021). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, 101, 107050.