AN OPTIMIZED ROBUST SOLUTION TO DETECT DEEPFAKE THROUGH ADVANCED COLOR ANALYSIS AND PROFILING

MANOJ KUMAR MISHRA[1], VIJAY PAL SINGH[2]
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SUNRISE UNIVERSITY ALWAR (RAJ.) INDIA
Mkmishra12@gmail.com

ABSTRACT

With the rise of deepfake technology, the ability to create realistic synthetic videos and images has become a significant concern for various applications, including security and trustworthiness. This research paper focuses on the development and application of the Color Characteristics method to compare and identify deepfake content within videos and images.

The proposed methodology involves extracting and analyzing color features from frames of videos and images, employing advanced image processing techniques. By utilizing color characteristics as distinctive markers, the research aims to distinguish between authentic and deepfake content. The study explores the efficiency of this method in detecting manipulated frames and determining the extent of alterations.

Comparative analysis involves assessing two sets of videos or images, one authentic and the other manipulated through deepfake techniques. The evaluation is based on the degree of divergence in color characteristics, providing insights into the identification of deepfake content and the specific frames where alterations occur. Results from this research contribute to the ongoing efforts in developing robust methods for detecting deepfake content, ensuring the integrity and reliability of visual media in various domains.

1. INTRODUCTION ABOUT DEEP FAKE

Deepfake technology, a portmanteau of "deep learning" and "fake," refers to the use of artificial intelligence (AI) and machine learning algorithms to create highly realistic but fabricated images, videos, or audio recordings. The term gained prominence in the late 2010s as advancements in AI, particularly in deep learning techniques, made it possible to generate convincing synthetic media.[1,2]

Origin and Development:

Early Stages: Deepfake technology emerged from academic research in computer vision, natural language processing, and deep learning. Researchers explored techniques for generating and manipulating multimedia content using neural networks.[3]

Advancements: With the proliferation of deep learning frameworks and increased computational power, deepfake algorithms became more sophisticated and accessible to a wider audience. Generative adversarial networks (GANs) and autoencoders played significant roles in improving the realism of generated media.[4]

Open Source Tools: The release of open-source deepfake tools and libraries, such as TensorFlow and PyTorch, democratized the creation of deepfake content, leading to widespread experimentation and dissemination of the technology.

Current Stage:

Proliferation: Deepfake technology has become increasingly prevalent across various online platforms, social media, and entertainment industries. The ease of access to tools and tutorials has enabled individuals with minimal technical expertise to create deepfake content.[5]

Applications: Deepfake technology finds applications in various domains, including entertainment, advertising, political satire, and cybersecurity. While some applications are benign or entertainment-focused, others raise ethical and security concerns.[6]

Advantages:

1. Creative Expression: Deepfake technology allows for innovative and creative storytelling, enabling filmmakers, artists, and content creators to explore new narrative possibilities.[7]

2. Special Effects: In the entertainment industry, deepfake technology can be used to generate lifelike visual effects and enhance the production value of films, TV shows, and video games.[8]

3. Training Data Augmentation: Deepfake algorithms can generate synthetic data to augment training datasets for machine learning models, facilitating more robust and diverse model training.[9]

Drawbacks:

1. Misinformation and Manipulation: The ease of creating convincing deepfake content poses significant risks for misinformation, propaganda, and social engineering. Deepfakes can be used to spread false narratives, manipulate public opinion, and incite discord.[10

2. Privacy Concerns: Deepfake technology raises privacy concerns as it can be used to create non-consensual pornography ("deepfake porn") or to impersonate individuals without their consent.[11]

3. Identity Theft: The ability to convincingly mimic someone's appearance and voice raises concerns about identity theft and impersonation, potentially leading to fraud or blackmail.[12]

4. Erosion of Trust: The proliferation of deepfake content undermines trust in media and digital content, making it increasingly challenging to discern between authentic and manipulated content.[13]

Deepfake images and videos are digital media that have been manipulated using artificial intelligence (AI) techniques to produce highly realistic but fabricated content. The term "deepfake" is derived from "deep learning" and "fake." These manipulated media can be created using various AI methods, including deep learning algorithms such as generative adversarial networks (GANs) and autoencoders.

In deepfake videos, the faces or actions of individuals are often replaced or superimposed onto existing video footage, making it appear as though they are saying or doing something they did not actually do. Similarly, in deepfake images, faces can be swapped or modified to create false scenarios or misrepresent individuals.

The proliferation of deepfake technology has raised concerns about its potential misuse, including spreading misinformation, creating fake news, manipulating public opinion, and even impersonating individuals for malicious purposes. As a result, there is growing interest in developing detection methods and tools to identify and mitigate the spread of deepfake content.

Despite the risks associated with deepfake technology, it also has legitimate applications, such as in the entertainment industry for special effects and visual storytelling. However, the ethical and societal implications of deepfakes continue to be a subject of debate and research.

Detecting deepfake images and videos involves analyzing various aspects of the media to identify signs of manipulation. Here are some common techniques and approaches used to determine whether an image or video is a deepfake:

1. Forensic Analysis: Forensic analysis involves examining the digital properties of the media, such as metadata, compression artifacts, and inconsistencies in pixel patterns. Changes made during the editing process may leave traces that can be detected through careful examination.[18]

2. Facial and Body Movement Analysis: Deepfake videos often exhibit unnatural facial expressions, lip movements, and body gestures. Analyzing the movement of facial features, such as blinking, eye movements, and lip-syncing, can help identify inconsistencies that indicate manipulation.[19]

3. Audio Analysis: In videos with accompanying audio, discrepancies between the visual and auditory components may suggest manipulation. Lip-syncing errors or mismatched speech patterns can be indicators of a deepfake.

4. Artifact Detection: Deepfake generation processes may introduce artifacts or distortions that are not present in authentic media. These artifacts can include strange reflections, unnatural lighting, or inconsistencies in shadows and reflections.

5. Pattern Recognition: Machine learning algorithms can be trained to recognize patterns specific to deepfake images and videos. These algorithms analyze large datasets of both authentic and manipulated media to learn distinguishing features and identify anomalies.[20]

6. Comparison with Source Material: Comparing the suspected deepfake with known authentic footage or images of the same subject can reveal discrepancies or alterations. This approach relies on having access to reliable reference material for comparison.

7. Deep Learning Techniques: Advanced deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can be trained to detect deepfakes by learning patterns and features indicative of manipulation. These models can analyze the visual and auditory components of media to identify anomalies.[21]

8. Consistency Checks: Analyzing the consistency of elements within the image or video, such as lighting, shadows, reflections, and perspective, can reveal inconsistencies that are unlikely to occur in genuine media.

2. LITERATURE REVIEW

Historical Background and Case Study

2.1 Case Study of Deepfake Pics and Videos

2.1.1 The Birth: Emergence of Deepfakes

Deepfakes, a term that has become synonymous with digitally manipulated videos, owe their existence to an anonymous Reddit user known simply as u/deepfakes. This groundbreaking technology found its initial foothold on Reddit, a vast platform with numerous subreddits dedicated to various topics. Among them was a community created by u/deepfakes in November 2017, aptly named r/deepfakes. It was within this community that the very first face-swaps using the Deepfakes algorithm made their way into the public domain.

The Deepfakes algorithm, the software responsible for these transformative videos, was also first unveiled on this subreddit. Regrettably, Reddit chose to remove r/deepfakes after a change in site regulations that prohibited what they termed "involuntary pornography." I personally had the opportunity to witness the birth of Deepfakes on Reddit, as I was an active Reddit user during that period. I observed parts of this transformative moment unfold before my eyes, and it greatly contributed to my decision to undertake this thesis.

In an attempt to capture this pivotal moment—the public emergence of Deepfakes—I sought to examine the reasons behind u/deepfakes' decision to make this technology accessible to the public. Considering the removal of the subreddit, reconstructing this history was no easy feat. Yet, through diligent exploration of web archives, I managed to uncover snapshots of r/deepfakes pages on archive.is. These snapshots provide insight into the moment when u/deepfakes shared the code that would mark the inception of Deepfakes:

"Animators will have a machine learning-based tool to create natural character animations. People will train models to detect fake images, and others will train models to create undetectable fakes. Face swapping is nothing compared to creating realistic 3D avatars and placing them in virtual reality." [14]

It is worth noting that the author downplayed his own role in the proliferation of potentially harmful machine learning applications, arguing that numerous similar technologies were being explored within the industry. For instance, the technology referred to as Face2Face, which shares common applications with Deepfakes, had been researched and documented a year before the advent of Deepfakes. [15]

While we delve deeper into this phenomenon, it's essential to establish a shared glossary for the most frequently used Deepfakes-related terms, in order to prevent any potential confusion in the upcoming sections. This glossary will serve as a valuable reference:

- **Deepfakes:** Videos that employ face-swapping techniques facilitated by the Deepfakes technology.
- **Deepfakes Technology:** The ensemble of tools and technologies essential for creating Deepfakes.
- **r/deepfakes:** The initial Reddit community, or subreddit, dedicated to Deepfakes technology.
- **u/deepfakes:** The Reddit user who shared the original source code that forms the basis of Deepfakes technology.
- **Deepfakes Phenomenon:** The comprehensive range of socio-technical practices that surround Deepfakes technology and the videos it produces.

2.1.2 The Videos: A Taxonomy of Deepfakes in the Wild

As an integral aspect of my empirical research, I embarked on an expedition across the web to uncover instances of Deepfakes. This quest led me to discern a recurring pattern in the characteristics of these videos, culminating in a simple taxonomy to classify my findings. This taxonomy, though not intended as an exhaustive catalog of all potential Deepfakes, serves as a valuable tool for highlighting the most significant variations in the utilization of this technology. It will prove instrumental in our comprehensive analysis of this phenomenon.[16]

The taxonomy I propose encompasses the following categories, presented without any particular order:

- The Technology Demonstration Deepfake
- The Satirical Deepfake

- The Meme Deepfake

- The Pornographic Deepfake

- The Deceptive Deepfake

2.1.3 Technology Demonstration Deepfakes serve as illustrative examples designed to showcase the inner workings of the technology. These often feature side-by-side comparisons of the original video and the manipulated Deepfake version. An illustrative instance can be found in a video created by YouTube user derpfakes, which features actor Alec Baldwin performing a parody of Donald Trump on Saturday Night Live (SNL), contrasted with the same video where Trump's face has been seamlessly replaced by Baldwin's. Some of these videos are deliberately created to highlight the altered segments from the outset, explicitly demonstrating the technology's transformative effect on the original footage. Others might incorporate an authenticity puzzle [17], prompting the audience to distinguish the genuine segments from the fabricated ones, often before the puzzle is resolved and the manipulation is unveiled in the video's conclusion.

This categorization provides an initial glimpse into the diverse applications of Deepfakes, setting the stage for more comprehensive exploration and analysis.

3. PROPOSED ALGORITHM FOR DEEPFAKE DETECTION:

1. For the Original Video:

i. Take the original video.

ii. Divide the video into individual frames (pictures).

iii. Take one frame at a time.

iv. Calculate the hash value for this frame (e.g., H11).

v. Save this hash value in the database.

vi. Repeat the process for all frames in the video.

2. For the Second (Altered or Fake) Video:

vii. Take the second video.

viii. Divide this video into individual frames.

ix. Take one frame at a time.

x. Calculate the hash value for this frame (e.g., H21).

xi. Save this hash value in the database.

xii. Repeat the process for all frames in the second video.

3. Comparison of Hash Values:

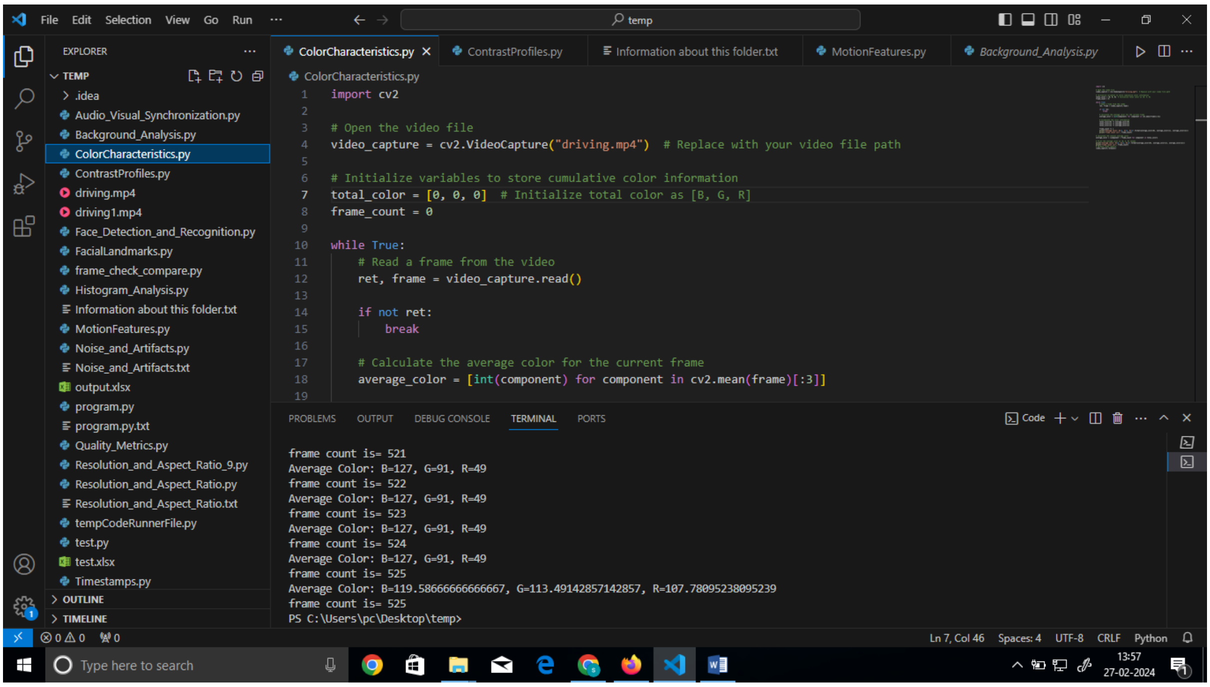
- Now, compare the hash values between the two videos.
- If (H11 = H21) for each corresponding frame, the frames are the same and not altered.
- If they are not equal, it means the frames are different.
- Continue this comparison for all hash values.

4. Final Evaluation:

- Count the number of matching frames (Count).
- If Count equals the total number of frames in either the first or second video (i or j), it indicates that the video is original.
- If Count is not equal to the total number of frames in either video, it suggests the video is altered.

4. IMPLEMENTATION, EXAMPLES, SCREENSHOTS AND RESULT ANALYSIS

For checking this algorithm we used the python language and take the video so that we can compare the color characters (R,G,B values) of videos.



```python
import cv2

# Open the video file
video_capture = cv2.VideoCapture("driving.mp4")  # Replace with your video file path

# Initialize variables to store cumulative color information
total_color = [0, 0, 0]  # Initialize total color as [B, G, R]
frame_count = 0

while True:
    # Read a frame from the video
    ret, frame = video_capture.read()

    if not ret:
        break

    # Calculate the average color for the current frame
    average_color = [int(component) for component in cv2.mean(frame)[:3]]

```

```
frame count is= 521
Average Color: B=127, G=91, R=49
frame count is= 522
Average Color: B=127, G=91, R=49
frame count is= 523
Average Color: B=127, G=91, R=49
frame count is= 524
Average Color: B=127, G=91, R=49
frame count is= 525
Average Color: B=119.58666666666667, G=113.49142857142857, R=107.78095238095239
frame count is= 525
PS C:\Users\pc\Desktop\temp>
```

Fig 4.1. Output screen of colour characteristics

Table 4.1 colour analysis of a video frame by frame

Frame no.	Colour values		
	Blue	Red	Green
1	119	125	121
2	119	125	121
3	119	125	121
4	119	125	121
5	119	125	121
6	119	125	121
7	119	125	121
8	119	125	121
9	119	125	121
10	119	125	121
11	119	125	121
12	119	125	121
13	119	125	121
14	119	125	121
15	119	125	121

16	119	125	121
17	119	125	121
18	119	125	121
19	119	125	121
20	119	125	121
21	119	125	121
22	119	125	121
23	119	125	121
24	119	125	121
25	119	125	121
26	119	125	121
27	119	125	121
28	119	125	121
29	119	125	121
30	119	125	121
31	119	125	121
32	119	125	121
33	119	125	121
34	119	125	121
35	119	125	121
36	119	125	121
37	119	125	121
38	119	125	121
39	119	125	121
40	119	125	121
41	119	125	121
42	119	125	121
43	119	125	121
44	119	125	121
45	119	125	121
46	119	125	121
47	119	125	121
48	119	125	121
49	119	125	121
50	119	125	121
51	119	125	121
52	119	125	121
53	119	125	121
54	119	125	121
55	119	125	121
56	119	125	121
57	119	125	121
58	119	125	121

59	119	125	121
60	119	125	121
61	119	125	121
62	119	125	121
63	119	125	121
64	119	125	121
65	119	125	121
66	119	125	121
67	119	125	121
68	119	125	121
69	119	125	121
70	119	125	121
71	119	125	121
72	119	125	121
73	119	125	121
74	119	125	121
75	119	125	121
76	119	125	121
77	119	125	121
78	119	125	121
79	119	125	121
80	119	125	121
81	119	125	121
82	119	125	121
83	119	125	121
84	119	125	121
85	119	125	121
86	119	125	121
87	119	125	121
88	119	125	121
89	119	125	121
90	119	125	121
91	119	125	121
92	119	125	121
93	119	125	121
94	119	125	121
95	119	125	121
96	119	125	121
97	119	124	121
98	119	124	121
99	119	124	121
100	119	124	120

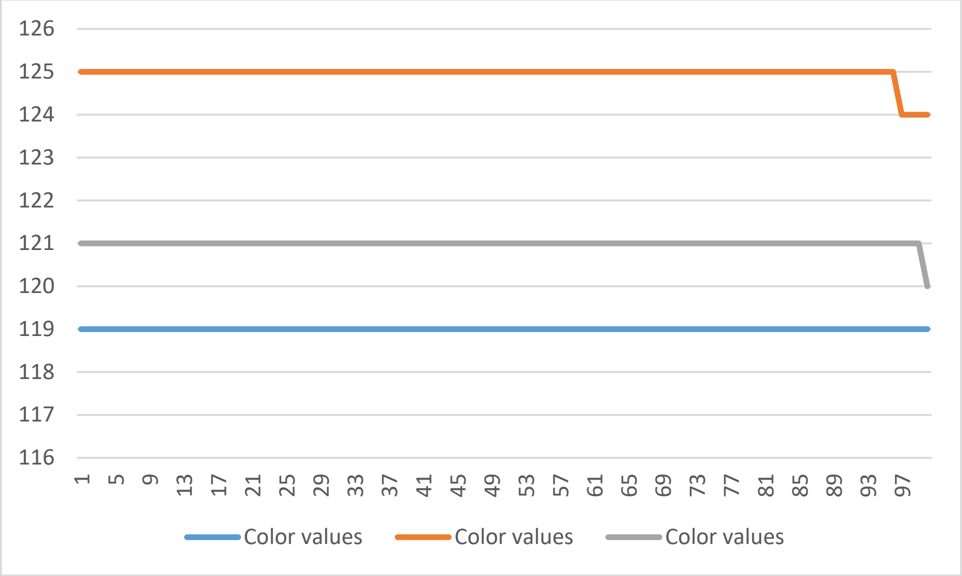


Fig. 4.2 RGB comparison frame by frame

5. CONCLUSION

By analyzing result of the algorithm, I can say that by applying this technology of proposed algorithm, we can protect the videos form the deepfake, and if we can protect our videos and pics from deepfake we can save of society for many kinds of conspiracy and rumors that can be occur with the due to deepfake video, deepfake pics and deepfake news.

1. The proposed algorithm produces the video with the hash value and if someone want to change in video than we can calculate the hash value of alter (deepfake) video and we can compare both values, by the process we can check the video is original or not, and we can secure our videos or images.

By the comparison of the previous and the proposed algorithm, we can check and say that the proposed algorithm is more secure and reliable from the previous one. This proposed algorithm is fulfilling all the goal (Checking the colour analysis) of video frame by frame. This is major difference between previous and the purposed algorithm.

REFERENCES

[1] Keane, S. (2018, September 5). Congress wrestles with ‘deepfake’ threat to Facebook. Retrieved 9 October 2018, from https://www.cnet.com/news/congress-wrestles-with- deepfake-threat-to-facebook/

[2] Knight, W. (2018, August 7). The Defense Department has produced the first tools for catching deepfakes. Retrieved 9 October 2018, from https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first- tools-for-catching-deepfakes/

[3] Chesney, R., & Citron, D. (2018, February 21). Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy? Retrieved 12 October 2018, from https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy

[4] Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., & Nießner, M. (2018). Headon: real-time reenactment of human portrait videos. ACM Trans. Graph., 37, 164:1-164:13.

[5] Karras, T., Laine, S., & Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks. CoRR, abs/1812.04948.

[6] Sharma, S. (2018, August 04). Celebrity Face Generation using GANs (Tensorflow Implementation). Retrieved from https://medium.com/coinmonks/celebrity-face-generation-using-gans-tensorflow-implementation-eaa2001eef86

[7] Seibold, C., Samek, W., Hilsmann, A., & Eisert, P. (2017). Detection of Face Morphing Attacks by Deep Learning. IWDW.

[8] Cozzolino, D., Poggi, G., & Verdoliva, L. (2017). Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection. IH&MMSec.

[9] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). FaceForensics: A Large-scale Video Dataset for Forgery Detection in HumanFaces. CoRR, abs/1803.09179.

[10] Rahmouni, N., Nozick, V., Yamagishi, J., & Echizen, I. (2017). Distinguishing computer graphics from natural images using convolution neural networks. 2017 IEEE Workshop on Information Forensics and Security (WIFS), 1-6.

[11] Raghavendra, R., Raja, K.B., Venkatesh, S., & Busch, C. (2017). Transferable Deep- CNN Features for Detecting Digital and Print-Scanned Morphed Face Images. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), 1822-1830.

[12] Guera, D., & Delp, E.J. (2018). DeepFake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video andSignal Based Surveillance (AVSS), 1-6.

[13] Li, Y., & Lyu, S. (2018). Exposing DeepFake Videos by Detecting Face Warping Artifacts. CoRR, abs/1811.00656.

[14] Brickell EF, editor. Advances in cryptologyCRYPTO‟92: 12th Annual international cryptology conference, santabarbara, California, USA. Proceedings. Springer; 2003.

[15] Neeraj Guhagarkar, Sanjana Desai, Swanand Vaishyampayan, Ashwini Save” DEEPFAKE DETECTION TECHNIQUES: A REVIEW”, 9th National Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021), Volume 1, Issue 4 (2021, ISSN(Online): 2581-7280.

[16] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image syn- thesis from semantic layouts via conditional imle. In Proceed- ings of the IEEE/CVF International Conference on Computer Vision, pages 4220–4229, 2019.

[17] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, Cuong M. Nguyen “Deep Learning for Deepfakes Creation and Detection: A Survey”, arXiv:1909.11573v4, 6 Feb 2022.

[18] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–7. IEEE, 2018.

[19] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 666–667, 2020.

[20] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 3(1):80–87, 2019.

[21] Abdul Qader M. Almars “Deepfake detection techniques using deep learning. A survey”, Journal of Computer and Communications, Vol.9 No.5, May 2021, DOI: 10.4236/jcc.2021.95003 , ISSN Print: 2327-5219, ISSN Online: 2327-522