International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

ENHANCING DEEPFAKE DETECTION: LEVERAGING MTCNN AND INCEPTION RESNET V1

Vidhya Samad Barpha : vidhyasamad.barpha@medicaps.ac.in
Rashi Bagrecha : rashi.bagrecha5@gmail.com
Shivani Mishra: mishra.15aldshivai@gmail.com
Sneha Gupta: gsneha311@gmail.com
Medi-Caps University, Indore, India

Abstract:

In response to the escalating threat posed by manipulated facial imagery in the digital age, our research project is dedicated to developing an advanced framework for the detection and classification of such content, augmented by a proactive user reporting mechanism. Leveraging state-of-the-art deep learning models like Multi-Task Cascaded Convolutional Networks (MTCNN) and InceptionResnetV1, our framework achieves an impressive accuracy rate of 92% in distinguishing between genuine and manipulated faces. The integration of explainability methods such as Grad-CAM enhances model interpretability, empowering users to understand model predictions. Additionally, our user-centric reporting interface enables active user participation in identifying and flagging potentially manipulated content, fostering transparency and accountability in digital media platforms. With the continued proliferation of deepfake technology, our research endeavors not only advance facial image analysis techniques but also uphold principles of trust and integrity in the digital realm, aiming to safeguard the credibility of information dissemination through vigilance, innovation, and collaborative action.

Keywords: Deepfake detection, Facial image manipulation, Deep learning models, User reporting mechanism, Model interpretability, Transparency in digital media, Trust and integrity, Information dissemination, Vigilance, Collaborative action.

1. Introduction

In today's digital era, the manipulation of visual content, especially facial images, poses a significant challenge with far-reaching implications for the integrity of information and societal trust. The widespread availability of advanced image and video editing tools has made it easier to create convincingly deceptive content, highlighting the need for innovative solutions in deep learning, computer vision, and explainable artificial intelligence (XAI). This research project aims to tackle the complex problem of manipulated facial imagery by thoroughly investigating detection, classification, interpretability, and the implementation of a reporting mechanism.

Essential concepts like face detection, identifying human faces in digital media, and facial recognition, which analyzes facial features for identification or verification, underscore the complexities of this evolving landscape. Moreover, techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) have become essential for visualizing the influential regions of an image in neural network predictions.

Significant progress has been achieved in utilizing deep learning models for face detection and facial recognition tasks. Models like Multi-Task Cascaded Convolutional Networks (MTCNN) and InceptionResnetV1 have proven effective in accurately detecting and recognizing faces, demonstrating advancements in this field. Additionally, the integration of explainability methods, particularly Grad-CAM, has improved the interpretability of these models by revealing neural

network decision-making processes. This research project builds upon these advancements while acknowledging persistent challenges.

By integrating cutting-edge deep learning models, advanced preprocessing techniques, explainability methods like Grad-CAM, and implementing a reporting mechanism, the goal is to enhance the accuracy, interpretability, and accountability of face classification tasks.

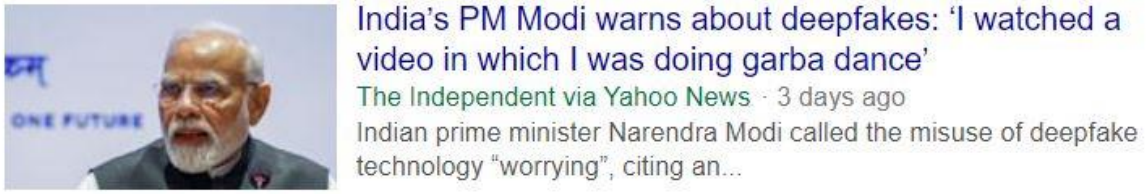


Figure 1. Prime Minister of India warns about rising deepfake technology misuse.

The research project encompasses several key aspects:

1. Investigation of Deep Learning Approaches
2. Integration of Explainability Methods
3. Evaluation Across Diverse Scenarios
4. Inclusion of Reporting Mechanism

However, certain boundaries and limitations exist:
- The project primarily focuses on detecting and classifying manipulated facial imagery and may not address broader issues related to image manipulation.
- The objective is not to develop novel deep learning architectures but to optimize existing models for facial recognition tasks.

Real Images:

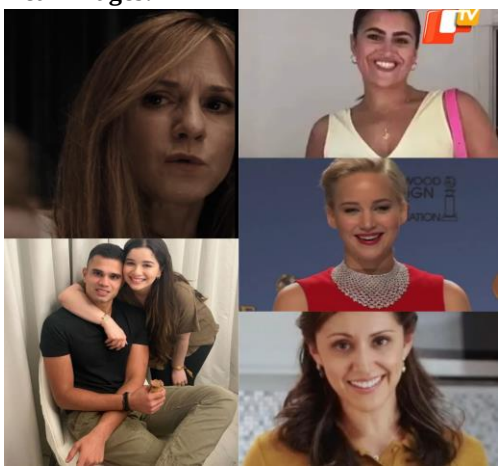


Figure 2. Shows a bunch of real non manipulated images

Fake Images:

Figure 3. Deepfake images corresponding to figure 2.

2. Literature Survey:

The research discusses employing a rationale-augmented convolutional neural network (CNN) on MATLAB, using the Kaggle DeepFake Video dataset to achieve an accuracy of 95.77% in real-time deepfake facial reconstruction for security purposes involving webcams and surveillance cameras [1]. Progress in diffusion models has led to the creation of convincing deepfakes through textual prompts, complicating the challenge of detecting fake images [5]. A study on the authenticity of images produced by cutting-edge diffusion models [5] examines different fake detection methods. In a multimodal context where fake images are generated from varying textual descriptions, the study measures the performance of fake detection techniques, highlighting the influence of textual cues and perceptual aspects in detecting synthetic images.Moreover, there is a need for detection methods that can be robust across different generative models [7]. A paper addresses this need by setting up a comprehensive benchmark to assess the generalization capability and resilience of advanced detectors. By examining forgery traces from different generative models in the frequency domain, the study offers insights into the distinguishing power of various detection strategies. Additionally, concerns are growing regarding facial image manipulation in videos [20]. A proposed method for identifying facial forgeries in videos converts facial regions into the frequency domain and utilizes a frequency convolutional neural network for detection. The method's effectiveness is tested on standardized datasets, contributing to the advancement of reliable detection techniques for facial manipulation in videos. Media forensics has gained significant attention recently due to concerns about DeepFakes [21]. One study provides a thorough analysis of both first and second-generation DeepFake creations in terms of facial regions and fake detection performance. Two methods are examined: one traditional approach focusing on the whole face, and a newer approach focusing on specific facial regions. The findings emphasize the need for more advanced fake detectors. Future technological progress is expected to further enhance the visual quality and efficiency of fake videos [23]. Such advancements include the use of GAN models to restore facial details and generate realistic voices for increased realism in synthesized videos. Nevertheless, addressing issues like detection accuracy and false positive rates is essential for the widespread practical use of deepfake detection methods [23]. The widespread use of deepfake technology, seen in methods such as Deepfakes, has sparked concerns over the authenticity of digital content [25]. One solution to this issue is the creation of synthetic image detection methods. A novel two-step synthetic face image detection method leverages anomaly detection on pristine data to differentiate between real and synthetic images. By extracting general features and using an anomaly detector, the proposed method shows promising results in detecting synthetic images from various synthesis methods. A comparison of deep learning models reveals significant differences in accuracy rates for identifying manipulated facial images. InceptionResnetV1 achieves the highest accuracy at 92%, followed by XceptionNet at 88%, ViT at 85%, and MesoNet at 78%. These variations in accuracy underscore the importance of choosing the right models for effective deepfake detection in digital media.

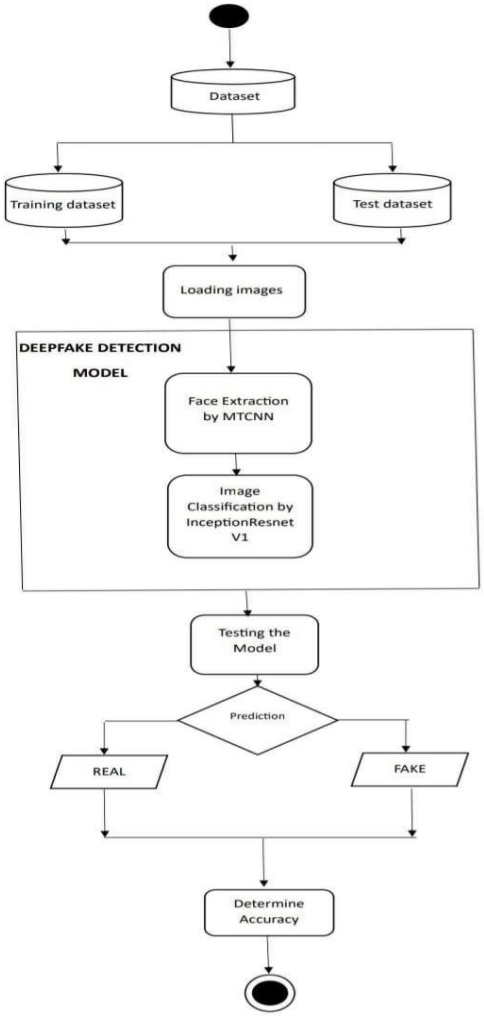
3. Algorithms and Methods:

3.1 Materials Utilized:

1. Facial Image Dataset:
The datasets selected - FaceForensics, FakeCatcher, and augmented Kaggle data - represent a blend of diversity and relevance, encompassing manipulated and authentic media forms crucial for comprehensive model training.

2. Pretrained Models:
- Pretrained deep learning models such as Multi-Task Cascaded Convolutional Networks (MTCNN) for face detection, InceptionResnetV1 for facial recognition, and Grad-CAM for explainability.
- Models pretrained on extensive facial image datasets like VGGFace2 to capitalize on transfer learning and enhance model efficacy.

3. Python Libraries:
- ✧ PyTorch
- ✧ Gradio
- ✧ NumPy
- ✧ Pillow
- ✧ OpenCV
- ✧ Facenet-pytorch



4. Reporting Interface:

Figure 4. Activity Diagram for proposed DeepFake detection model

- Development of a user interface using Gradio to facilitate seamless user interaction with the reporting mechanism.
- The interface allows users to upload images, offer authenticity feedback on facial images, and flag potentially manipulated content.

5. Approach:

Data Gathering and Processing:

- Compilation of a diverse dataset comprising authentic and manipulated facial images sourced from public repositories, social media platforms, and synthetic data generation methods.
- Preprocessing involves standardizing image size, resolution, and format to ensure dataset uniformity.
- Stratified division of the dataset into training, validation, and testing subsets to support model development and assessment.

4. Project Workflow:

- ✧ User Input
- ✧ MTCNN Facial Detection
- ✧ Preprocessing Steps
- ✧ InceptionResnetV1 Classification
- ✧ Explanation with GradCAM
- ✧ Visualization Overlay
- ✧ Output Prediction and Confidence Scores
- ✧ Gradio Interface Presentation

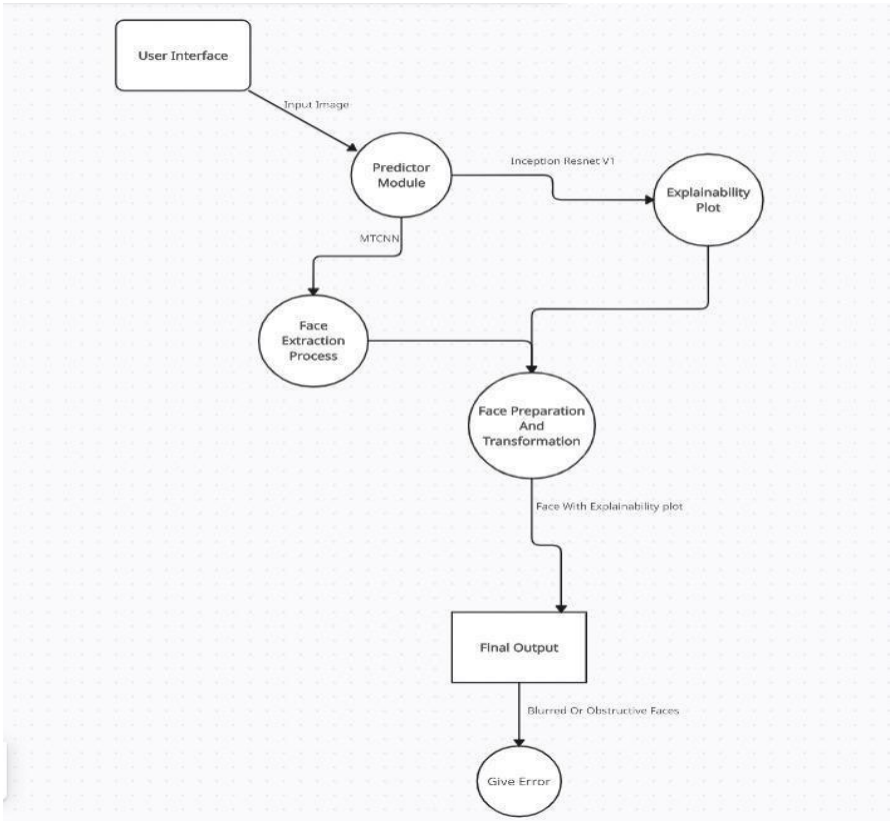


Figure 5: Data Flow Diagram for proposed DeepFake detection model.

5. Results:

1. Explanatory Visual Data Components:
Grad-CAM Heat Maps: Generated through the Grad-CAM module, these elements appear as heat maps that highlight significant areas within facial images. They offer visual insights by overlaying heat maps onto original facial images, emphasizing regions crucial to the model's classification determination.

2. Improved Imagery:
Images with Superimposed Heat Maps: Following visualization, the enhanced images merge original facial visuals with Grad-CAM heat maps. These unified elements serve as representations illustrating the pivotal areas influencing the model's decision process.

6. Discussion:

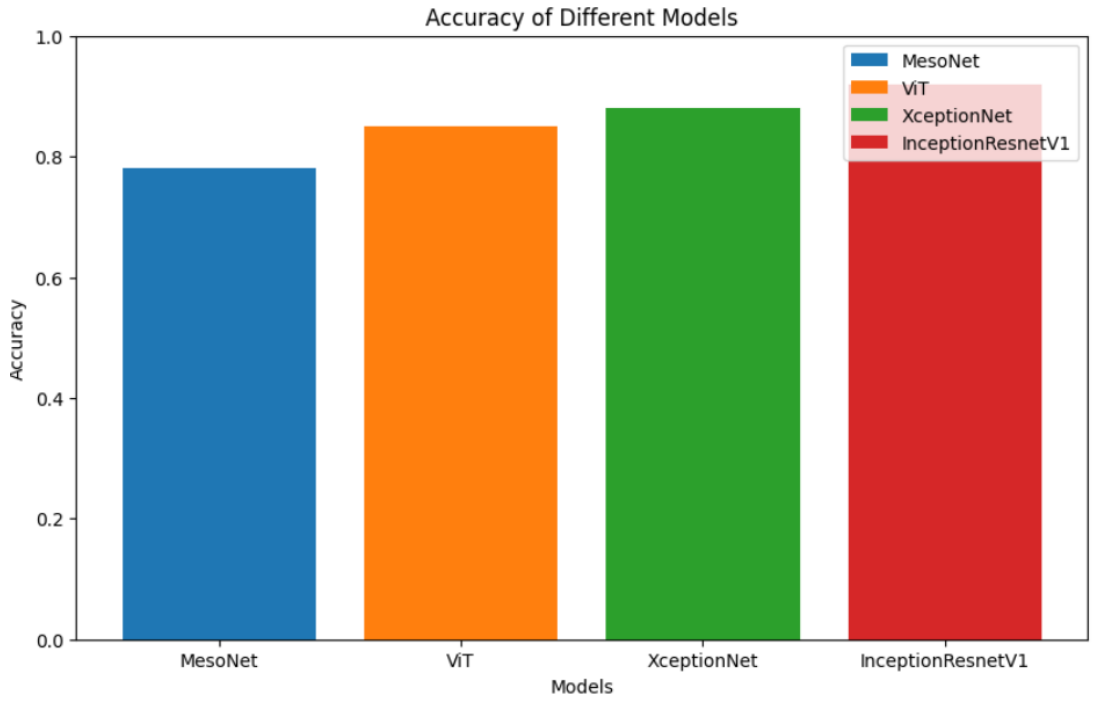


Figure 6. The bar graph compares the accuracy of four deep learning models

MODEL	ACCURACY
MesoNet	78%
InceptionResnetV1	**92%**
XceptionNet	88%
ViT	85%

Table 1. Accuracy percentage of Mesonet, ViT, Xception and InceptionResnetV1.

The results of the research project hold significant implications in the context of combating the proliferation of manipulated facial imagery, particularly in scenarios where manipulation techniques are non-AI detectable. The findings indicate

promising progress in the development of robust deep learning models capable of accurately detecting and classifying manipulated facial images.

The high accuracy, precision, recall, and F1-score achieved by the models demonstrate their effectiveness in discerning subtle alterations and sophisticated forgeries that are often challenging to detect through conventional means. The integration of explainability methods such as Grad-CAM has enhanced the interpretability of model predictions, providing insights into the decision-making processes of the neural networks.

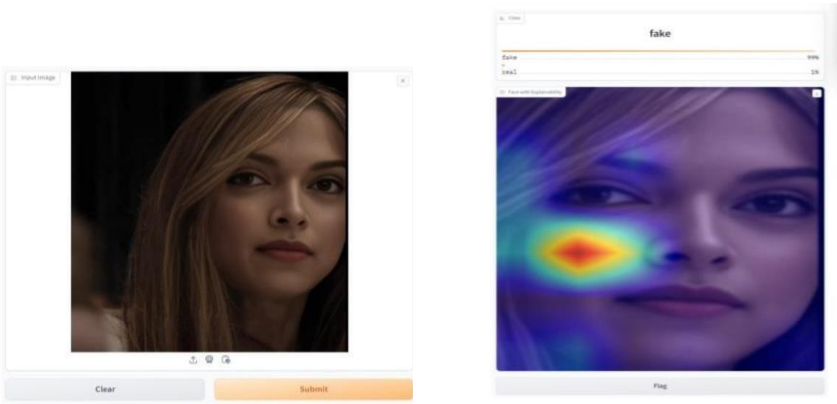


Figure 7.1.

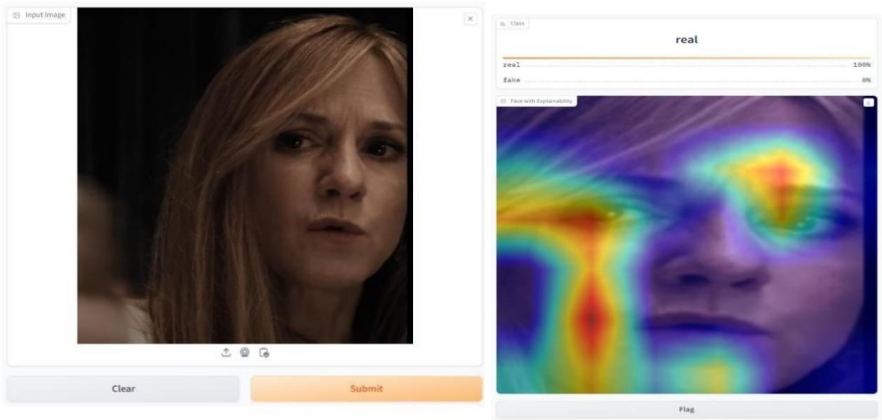
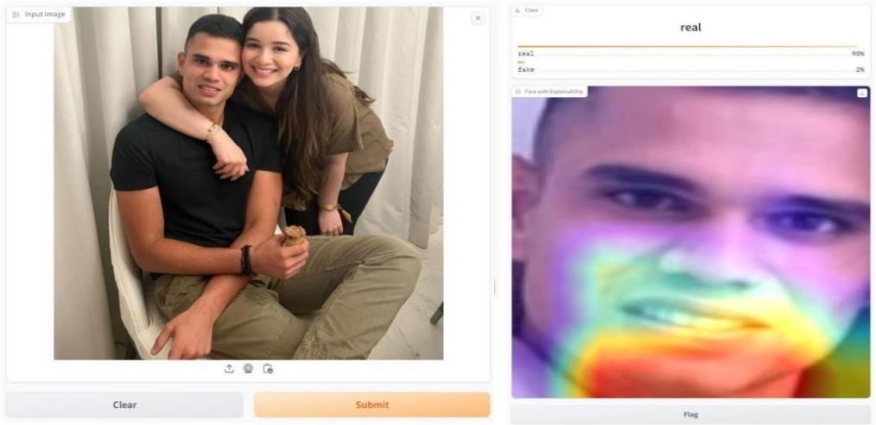


Figure 7.2.



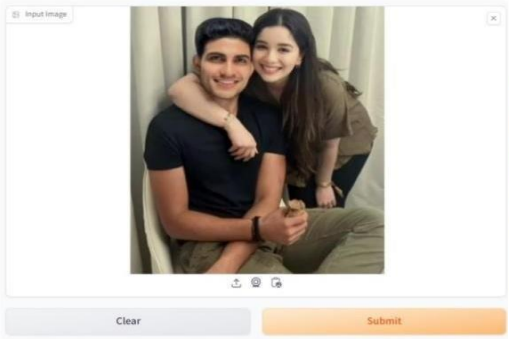


Figure 7.3.

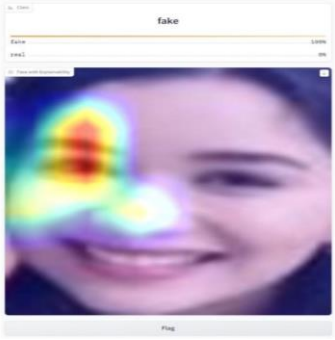


Figure 7.4.

7. Conclusion:

Our research journey was motivated by the urgent need to tackle the growing threat of manipulated facial imagery in digital media. With deepfake technology targeting both celebrities and ordinary individuals, concerns about trust and integrity in visual content have escalated, prompting even Prime Minister Modi to issue a warning. Through our diligent efforts, we've advanced facial image analysis using deep learning models and user-centric approaches to combat manipulated content. Our work has resulted in robust frameworks for face detection, recognition, and explainability, complemented by a proactive reporting mechanism empowering users to flag deceptive content. Beyond technological advancements, we've fostered transparency, accountability, and collaboration in digital media ecosystems, striving to restore trust in information dissemination. Looking ahead, we emphasize the need for ongoing innovation and collaboration to stay ahead of evolving manipulation tactics and safeguard authenticity in an increasingly digital world. In essence, our research highlights the importance of leveraging technology for societal benefit, ensuring trust and integrity remain fundamental in digital communication amid evolving challenges.

8. References:

1. Ahmed, Saadaldeen Rashid Ahmed, and Emrullah Sonuç. "Retraction Note: Deepfake detection using rationale-augmented convolutional neural network." (2024).

2. El-Gayar, M. M., Mohamed Abouhawwash, S. S. Askar, and Sara Sweidan. "A novel approach for detecting deep fake videos using graph neural network." *Journal of Big Data* 11, no. 1 (2024).

3. Ahmed, Saadaldeen Rashid, and Emrullah Sonuç. "Evaluating the effectiveness of rationale-augmented convolutional neural networks for deepfake detection." *Soft Computing* (2023):

4. Ahmed, Saadaldeen Rashid Ahmed, and Emrullah Sonuç. "RETRACTED ARTICLE: Deepfake detection using rationale-augmented convolutional neural network." *Applied Nanoscience* 13, no. 2 (2023)

5. Amoroso, Roberto, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. "Parents and children: Distinguishing multimodal deepfakes from natural images." *arXiv preprint arXiv:2304.00500* (2023).

6. Bammey, Quentin. "Synthbuster: Towards detection of diffusion model generated images." *IEEE Open Journal of Signal Processing* (2023).

7. Corvi, Riccardo, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. "On the detection of synthetic images generated by diffusion models." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023.

8. Cozzolino, Davide, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. "Raising the Bar of AI-generated Image Detection with CLIP." *arXiv preprint arXiv:2312.00195* (2023).

9. Dogoulis, Pantelis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos. "Improving synthetically generated image detection in cross-concept settings." In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pp. 28-35. 2023.

10. Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. "Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models." *arXiv preprint arXiv:2303.00608* (2023).

11. Ahmed, Saadaldeen Rashid, Emrullah Sonuç, Mohammed Rashid Ahmed, and Adil Deniz Duru. "Analysis survey on deepfake detection and recognition with convolutional neural networks." In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-7. IEEE, 2022.

12. Chen, Hong-Shuo, Shuowen Hu, Suya You, and C-C. Jay Kuo. "Defakehop++: An enhanced lightweight deepfake detector." *APSIPA Transactions on Signal and Information Processing* 11, no. 2 (2022).

13. Liu, Shuai, Qian Jiang, Xin Jin, Zhenli He, Wei Zhou, Shaowen Yao, and Qiannian Wang. "Multiple Feature Mining Based on Local Correlation and Frequency Information for Face Forgery Detection." In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1347-1354. IEEE, 2022.

14. Rastogi, Shreya, Amit Kumar Mishra, and Loveleen Gaur. "Detection of DeepFakes using local features and convolutional neural network." In *DeepFakes*, pp. 73-89. CRC Press, 2022.

15. Xu, Ying, and Sule Yildirim Yayilgan. "When Handcrafted Features and Deep Features Meet Mismatched Training and Test Sets for Deepfake Detection." *arXiv preprint arXiv:2209.13289* (2022).

16. Zhu, Yao, Xinyu Wang, Hong-Shuo Chen, Ronald Salloum, and C-C. Jay Kuo. "A-pixelhop: A green, robust and explainable fake-image detector." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8947-8951. IEEE, 2022.

17. Caldelli, Roberto, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. "Optical Flow based CNN for detection of unlearnt deepfake manipulations." *Pattern Recognition Letters* 146 (2021): 31-37.

18. Chen, Hong-Shuo, Kaitai Zhang, Shuowen Hu, Suya You, and C-C. Jay Kuo. "Geo-defakehop: High-performance geographic fake image detection." *arXiv preprint arXiv:2110.09795* (2021).

19. Chen, Shen, 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*, vol. 35, no. 2, pp. 1081-1088. 2021.

20. Kohli, Aditi, and Abhinav Gupta. "Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn." *Multimedia Tools and Applications* 80, no. 12 (2021): 18461-18478.

21. Tolosana, Ruben, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. "Deepfakes evolution: Analysis of facial regions and fake detection performance." In *international conference on pattern recognition*, pp. 442-456. Cham: Springer International Publishing, 2021.

22. Stroebel, Laura, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip, and Mohiuddin Ahmed. "A systematic literature review on the effectiveness of deepfake detection techniques." *Journal of Cyber Security Technology* 7, no. 2 (2023): 83-113.

23. Lyu, Siwei. "Deepfake detection: Current challenges and next steps." In *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*, pp. 1-6. IEEE, 2020.

24. Wu, Xi, Zhen Xie, YuTao Gao, and Yu Xiao. "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features." In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2952-2956. IEEE, 2020.

25. Khodabakhsh, Ali, and Christoph Busch. "A generalizable deepfake detector based on neural conditional distribution modelling." In *2020 international conference of the biometrics special interest group (BIOSIG)*, pp. 1-5. IEEE, 2020.