STM JOURNALS

Journal of
Image Processing & Pattern Recognition Progress

ISSN: 2394-1995
Volume 11, Issue 3, 2024
September–December
DOI (Journal): 10.37591/JoIPPRP

https://journals.stmjournals.com/joipprp

Review JoIPPRP

An Analysis of Multimodal Fusion in Deepfake Detection for Video Samples

Devang Mulye[1], Parakh Pawar[2,*], Aditya Yadav[3], S. Poornima[4]

Abstract

In today's rapidly evolving digital landscape, deepfake technology stands as both a marvel and a threat to privacy and security. Deepfakes, hyper-realistic synthetic media created using artificial intelligence (AI), can deceive and manipulate on an unprecedented scale, from political propaganda to compromising videos of public figures. This research navigates deepfake detection, focusing on two advanced methodologies: the vision transformers (ViT) image classifier and the Meso4 method. The ViT model utilizes convolutional neural networks (CNNs) for image analysis, while the Meso4 method examines images at a mesoscopic level. These methodologies are evaluated for their effectiveness in differentiating genuine content from altered media. Motivated by the threats posed by deepfakes to individuals and society, this project evaluates the ViT model using the Open Forensics dataset and the Meso4 model using a combination of self-generated content and the Face2Face dataset. The ViT model achieved 99.9% accuracy within its dataset and 85% accuracy for subtly deceptive images. Conversely, the Meso4 model showed a 60% accuracy and biases, particularly towards female subjects, likely due to its training on adult content. To address privacy and security concerns, five additional features are proposed: eye movement error detection, lip sync inconsistency, facial expression analysis, facial texture inconsistency, and background inconsistency. These features enhance detection accuracy and reliability and are integrated into a customized model, augmenting the ViT model's strengths and addressing its limitations. The proposed model improves accuracy and computation speed, making it practical for real-world applications. This research highlights the urgency of advancing robust and unbiased deepfake detection methods to safeguard privacy and security in an era where truth can be easily manipulated.

Keywords: Deepfake detection, Meso4, ViT, eye movement, skin texture, facial expression, lip sync

INTRODUCTION

Deepfakes, crafted with advanced AI, wreak havoc on digital content, fueling misinformation, privacy violations, and cybersecurity threats [1, 2]. To address these concerns, this research delves into image and video analysis [3]. We compared two pre-trained models: Meso4 [4], which is a heatmap model offering meticulous distortion detection but lacking accuracy, and ViT [5], which excels in accuracy but struggles with video compatibility. Recognizing these limitations, we propose a hybrid approach that leverages ViT image processing for individual video frames while addressing computational concerns [6]. Furthermore, [7] we introduce a novel model that analyzes five key facial features: eye movement, skin texture, facial expressions, background inconsistencies, and lip sync [8]. Focusing solely on the face, this model

***Author for Correspondence**
Parakh Pawar
E-mail: parakhpawar623@gmail.com

[1-3]Student, Department of Information Technology, SIES Graduate School of Technology, Navi Mumbai, Maharashtra, India
[4]Associate Professor, Department of Information Technology, SIES Graduate School of Technology, Navi Mumbai, Maharashtra, India

Received Date: June 27, 2024
Accepted Date: August 05, 2024
Published Date: September 12, 2024

Citation: Devang Mulye, Parakh Pawar, Aditya Yadav, S. Poornima. An Analysis of Multimodal Fusion in Deepfake Detection for Video Samples. Journal of Image Processing & Pattern Recognition Progress. 2024; 11(3): 19–27p.

extracts and compares these features across frames to enhance the detection accuracy and can be combined with ViT for further image classification improvements. This study addresses existing model limitations and enhances deepfake detection by exploring key facial features [9]. Analysis of various fusion methods for the recognition of different facial parts and features in earlier works motivated confidence in building this application for deepfake detection using multimodal fusion [10–15].

Challenges in deepfake detection persist, necessitating ongoing research efforts. We outline future directions, including enhancing the robustness of the model against adversarial attacks, improving the interpretability of detection results, and exploring novel detection techniques beyond facial analysis. Collaborative efforts and open-source initiatives play a pivotal role in advancing research on deepfake detection. We highlight the importance of knowledge sharing and collaborative platforms in accelerating innovation and fostering community-driven solutions to combat evolving threats posed by deepfake technology. It promotes responsible AI use by contributing to robust detection tools and public awareness, empowering us to navigate the digital landscape with trust [16] and discernment in the face of evolving deepfake technology [17].

Overall, this study contributes to the development of robust deepfake detection tools and promotes public awareness of the ethical implications of synthetic media manipulation. By empowering individuals to navigate the digital landscape with trust and discernment, we can mitigate the risks posed by the proliferation of deepfake technology and uphold the integrity of digital content.

DETECTION MODELS

The Meso4 and ViT models are discussed here, which are used to analyze the deepfake outcomes. The Meso4 deepfake detection network (Figure 1) [4] employs a cascading series of convolutional layers, each followed by batch normalization to enhance the training stability and accuracy. Feature extraction progresses through these layers, culminating in fully connected layers and dropout for robust classification. Max pooling reduces dimensionality, whereas the final sigmoid activation outputs a probability score for the input image to be a deepfake [18]. This concise architecture leverages the key elements for effective deepfake detection.

In this study, the Meso4 model incorporates an innovative enhancement through the integration of a threshold value derived from a heatmap, offering an enriched approach to image analysis and data visualization [19]. The workflow of the model begins with the loading of an image, followed by the conversion of the image into a NumPy array presented in matrix form for heatmap generation [20]. The heatmap plays a pivotal role in determining the threshold value and is subsequently utilized to calculate the confidence level. This threshold value is integral to the decision-making process employed in the Meso4 model.

After initializing the model and loading the pre-trained weights, the image was preprocessed for subsequent predictions. The model classifies an image as either genuine or manipulated. The fusion of the heatmap with the model contributes to the discernment of the discriminative regions. It facilitates the identification of crucial areas in an image, thereby improving the resistance of the model to adversarial attacks [21].

Our investigation identified the Vision Transformer (ViT) as a promising architecture for deepfake detection [22]. Notably, a ViT-based image classifier achieved over 99% accuracy on its training data (70,000 real/fake images) but suffered from limited generalizability and high computational demands [5]. The proposed machine-learning workflow addresses these challenges. The pipeline (Figure 2) utilizes the PIL library for efficient data loading, followed by label mapping and dataset splitting for robust evaluation. A pre-trained ViT model was integrated [23], with the input image size adjusted for optimal performance. Image transformations enhance model robustness and collate functions streamline batched data preparation. Leveraging the pre-trained weights facilitates faster convergence during training. The model was trained and evaluated on a validation set and its practical application was

demonstrated through real-time image predictions. The results are presented as interpretable scores and labels to empower users with informed decision-making.

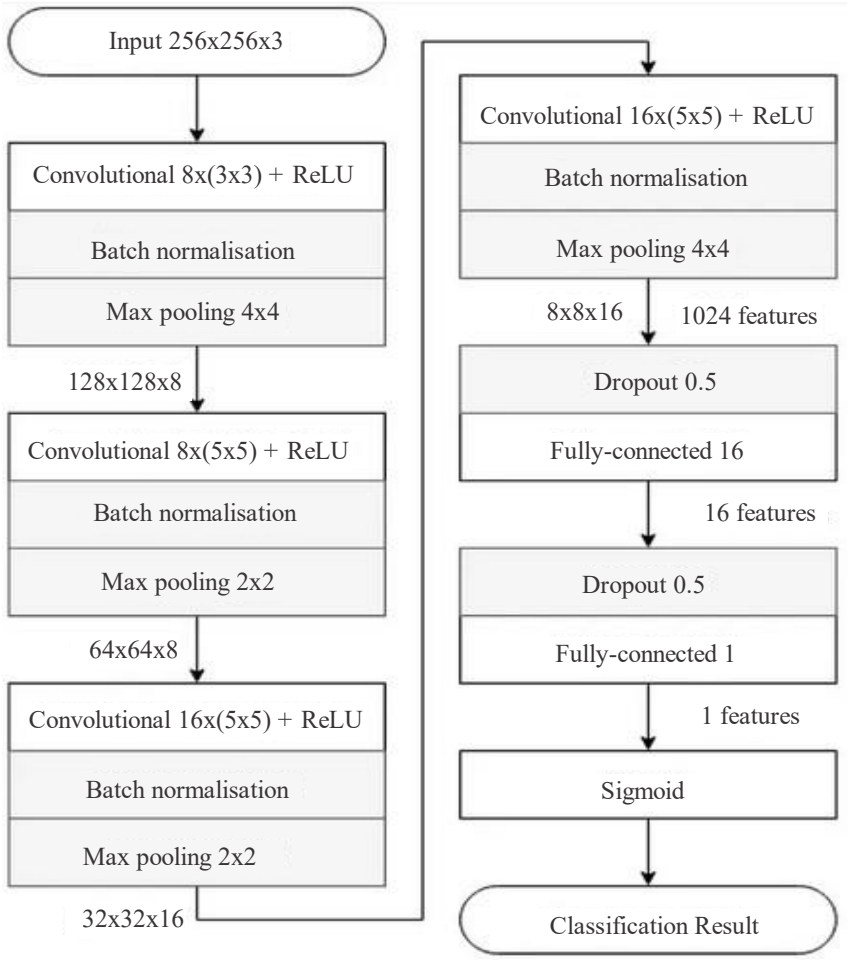

Figure 1. Architecture of Meso4 Model [24].

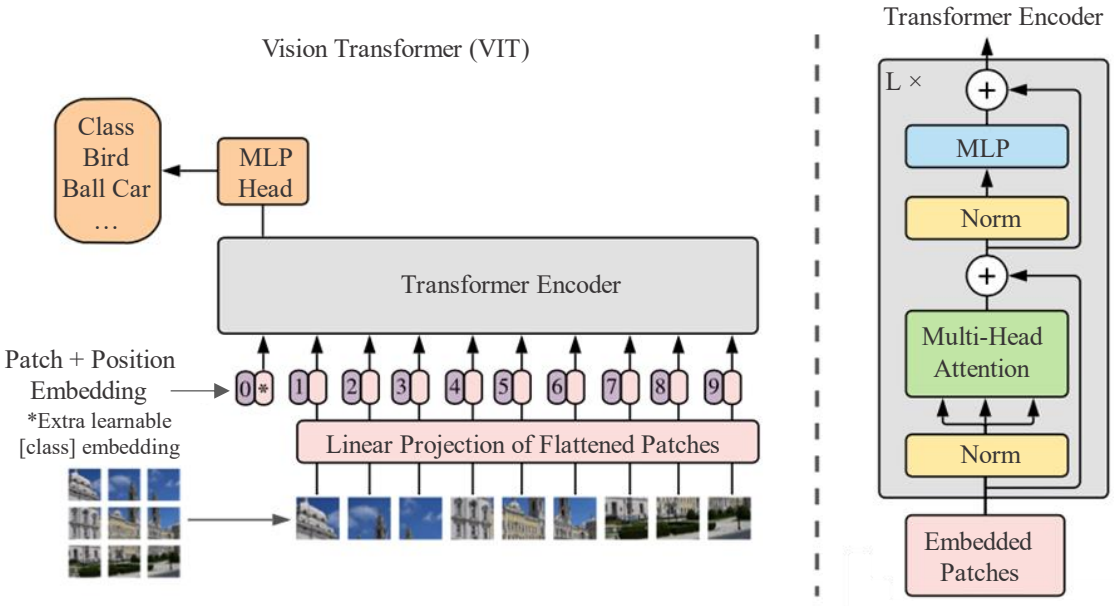

Figure 2. Architecture of ViT Model [25].

DATASET DETAILS

This paper presents a comparative analysis of Meso4 and ViT image classifiers, followed by enhancements of the ViT model. We employed meticulous dataset selection tailored to the requirements of each model.

- *Meso4:*2,043 images (960 fake, 1,081 real) categorized by difficulty level for training, testing, and validation.
- *ViT:* More extensive dataset (190,305 images):
- *Testing:* 5,492 fake, 5,413 real.
- *Training:* 70,000 fake, 70,000 real.
- *Validation:* 19,600 fake, 19,800 real.

For additional feature testing, we used the EfficientNetB4 dataset (Kaggle):

- *Training:* 3462 fake images, 700 real images, 3462 fake videos, and 700 real videos.
- *Validation:* 678 fake images, 136 real images, 678 fake videos, and 136 real videos.
- *Testing:* 790 image directories and 790 video files (all labeled CSV files).

These diverse datasets enabled robust training and evaluation of the image and video classification capabilities of both models. The performance metrics were meticulously computed and compared, as summarized in Table 1. This approach ensures a comprehensive and fair comparison by considering the specific training data and testing conditions of each model.

While the results of both Meso4 and ViT can be displayed using text output, we added a heatmap function to better visualize the output and obtain a greater understanding of the working of the models. Meso4 (Figures 3 and 4) shows the heatmap and the affected areas, whereas ViT (Figures 5 and 6) displays the output as a numerical value.

Table 1. Comparative study of findings from Meso and ViT models.

Metric	**Meso4**	**ViT**
Model size	Compact	Large
Complexity	Moderate	Easy
Training time	Fast	Medium — Slow
Inference speed	Faster	Slower
Accuracy	60% (fluctuates as per dataset size)	>90% (reduces to >45 on unknown datasets)
Robustness	Tasks-specific	Generalizes well
Resource requirements	Less	Higher
Transfer learning	Good	Strong
Availability	Pre-trained models available	Pre-trained models available

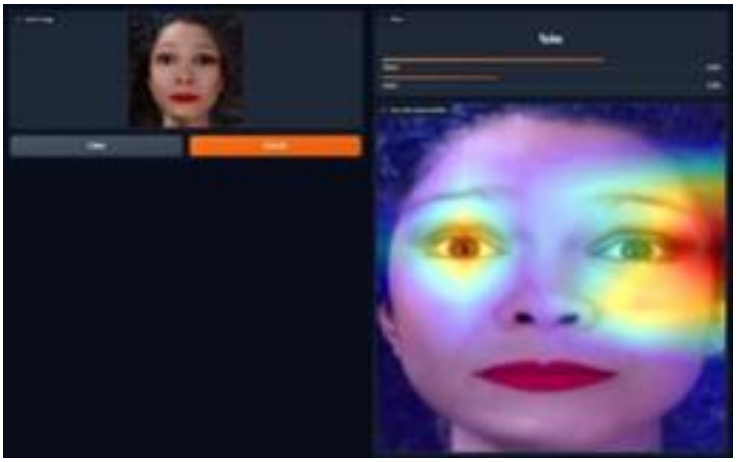

Figure 3. Meso4 outputs for fake image.

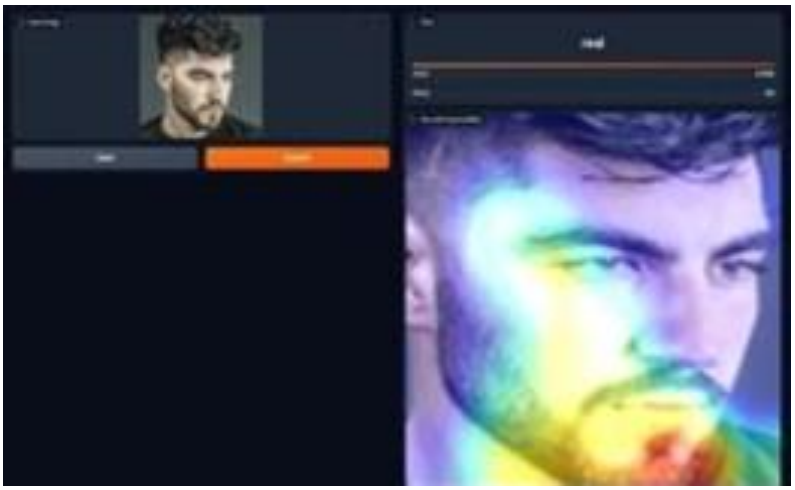

Figure 4. Meso4 outputs for real image.

Figure 5. ViT output for the real image.

The accuracy of the Meso4 model started at 0.65 at the first epoch, and increased to 0.99 by the tenth epoch, indicating that the model was learning and improving its predictions with each epoch. However, the accuracy curve plateaus, indicating that the model was no longer significantly improved. After several observations, it was found that the accuracy decreased because of saturation and overfitting. This situation motivated us to further extend our investigation to deepfake detection.

PROPOSED MULTIMODAL FUSION

While meso4 offered video compatibility but lower accuracy than ViT, ViT excelled in accuracy (in and outside the training data) yet lacked video capabilities. To address this, we propose a custom model (Figure 7) that leverages ViT image processing for individual video frames but with significant

computational limitations. To enhance deepfake detection research [7], suggest carrying with it a novel model analysis and fusion of five key features [26, 10–15, 27] of a human face is proposed.

1. *Eye movement error detection:* Identifies inconsistencies in eye movement patterns.
2. *Skin texture detection:* Analyses skin irregularities potentially indicative of manipulation.
3. *Facial expression detection:* Examines facial movements for unnatural expressions.
4. *Background inconsistencies detection:* Compares background elements for coherence across frames.
5. *Lip sync inconsistency:* The alignment of speech with lip movements is measured for possible mismatches.

Figure 6. ViT output for the real image.

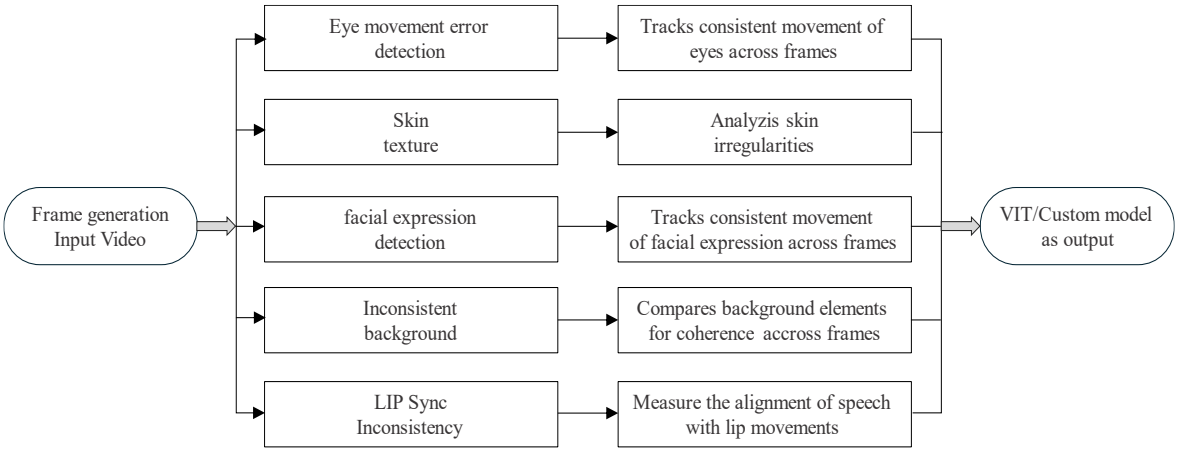

Figure 7. Proposed multimodal fusion for deepfake detection.


```
Frame frame_0.jpg is real
Frame frame_1.jpg is real
Frame frame_10.jpg is real
Frame frame_100.jpg is real
Frame frame_101.jpg is real
Frame frame_102.jpg is real
Frame frame_103.jpg is real
Frame frame_104.jpg is real
Frame frame_105.jpg is real
Frame frame_106.jpg is real
Frame frame_107.jpg is real
Frame frame_108.jpg is real
Frame frame_109.jpg is real
Frame frame_11.jpg is real
Frame frame_110.jpg is real
Frame frame_111.jpg is real
Frame frame_112.jpg is real
Frame frame_113.jpg is real
Frame frame_114.jpg is real
Frame frame_115.jpg is real
Frame frame_116.jpg is real
Frame frame_117.jpg is real
Frame frame_118.jpg is real
Frame frame_119.jpg is real
Frame frame_12.jpg is real
...
Frame frame_97.jpg is real
Frame frame_98.jpg is real
Frame frame_99.jpg is real
The video is FAKE with 13.59% manipulation
```

Figure 8. Sample output of tested video.

```
Video Path: E:\Test_1\Video_Dataset\train\fake_video\00014.mp4
Video Size: (640, 480)
Video Frame Rate: 30.0
Video Frame Count: 1385
Video Duration: 46.166666666666664 seconds
```

Figure 9. Sample output of tested video.

This model focuses solely on the face within the video, extracting and comparing these features across frames for a comprehensive analysis and enhanced detection accuracy [28]. It can also be combined with ViT to enhance the image classification capabilities at the cost of computation power and time. The proposed features are tested on various videos to analyze the efficiency of these outcomes for identifying deepfakes. A few sample outputs from the proposed methodology are shown in Figures 8 and 9. Still, work on and analyze the features of videos with different time durations. More investigations are yet to be verified and observed in detail to confirm the conclusions of this study.

CONCLUSION

This research spearheads a multi-pronged attack against deepfakes, recognizing the limitations of the existing models. We forge a hybrid approach, harnessing Meso4's real-time efficiency and interpretability for quick responses while exploiting ViT's superior accuracy and generalization for diverse tasks. In addition, we introduce a custom facial feature model that dissects five key areas: eye

movement, skin texture, facial expressions, background coherence, and lip sync. This model meticulously analyzes facial regions across video frames, offering enhanced detection accuracy and seamless integration with ViT for further image classification improvements.

Crucially, ethical considerations underpin our work. We advocate responsible use, regulatory compliance, and content management practices to ensure responsible technological advancements. The adaptability and scalability of the system empower it to combat real-time threats in data streams and large content batches, making it valuable for various scenarios. Its seamless integration with social media platforms allows for automatic identification and removal of deepfakes, thereby protecting the credibility of online information. Furthermore, this research translates into the environmental realm, where the system's capabilities can be harnessed for environmental monitoring, ensuring the authenticity of multimedia content, and identifying manipulated media in surveillance systems.

Looking ahead, the continuous refinement and adaptation of the detection system are paramount. Collaboration and knowledge sharing with the community will be crucial for staying ahead of evolving manipulation techniques and navigating the ever-shifting deepfake landscape. This research serves as a testament to the power of collective effort in the fight against deepfakes, paving the way for a more trustworthy and responsible digital future.

REFERENCES

1. Taeb M, Chi H. Comparison of deepfake detection techniques through deep learning. J Cybersecur Priv. 2022;2:89–106. DOI: 10.3390/jcp2010007.
2. Ramachandran S, Nadimpalli AV, Rattani A. An experimental evaluation on deepfake detection using deep face recognition. 2021 International Carnahan Conference on Security Technology (ICCST), Hatfield, United Kingdom, 2021, pp. 1–6. DOI: 10.1109/ICCST49569.2021.9717407.
3. Lyu S. Deepfake detection: Current challenges and next steps. IEEE Int Conf Multimed Expo Workshops (ICMEW). IEEE Publications; 2020. p. 1–6. DOI: 10.1109/ICMEW46912.2020.9105991.
4. Khichi M. Deepfake or real image prediction using Mesonet [dissertation]. Delhi Technological University. 2021. Available from: http://dspace.dtu.ac.in:8080/jspui/handle/repository/18952.
5. Liang J, Wang D, Ling X. Image classification for soybean and weeds based on ViT. J Phys Conf Ser. 2021;2002:012068. DOI: 10.1088/1742-6596/2002/1/012068.
6. Wodajo D, Atnafu S. Deepfake video detection using convolutional vision transformer. [Preprint]. ArXiv:2102.11126. 2021.
7. Pan D, Sun L, Wang R, Zhang X, Sinnott RO. Deepfake detection through deep learning. 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Leicester, UK, 2020. pp. 134–43. DOI: 10.1109/BDCAT50828.2020.00001.
8. Aghasanli A, Kangin D, Angelov P. Interpretable-through-prototypes deepfake detection for diffusion models. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2023, pp. 467–74. DOI: 10.1109/ICCVW60793.2023.00053.
9. Turan SG. Deepfake and digital citizenship: A long-term protection method for children and youth. In: Deep Fakes, Fake News, and Misinformation in Online Teaching and Learning Technologies. IGI Global; 2021. p. 124–42. DOI: 10.4018/978-1-7998-6474-5.ch006.
10. Poornima S, Subramanian S. Unconstrained iris authentication through fusion of RGB channel information. Int J Pattern Recognit Artif Intell. 2014;28:1456010. DOI: 10.1142/S0218001414560102.
11. Poornima S, Subramanian S. An efficient feature level fusion for a multimodal biometric system using correlation filter. Int J Appl Eng Res. 2014;9:8975–8.
12. Poornima S, Thaya IM. Comparison analysis of feature level fusion in multimodal biometrics system. Int J Appl Eng Res. 2014;9:8908-8912.
13. Poornima, S., Fathima Nadheen, M. Feature level fusion in multimodal biometric authentication system. Int J Comput Appl. 2013;69(18):36-40.

14. Poornima S. Performance study of fusion in multimodal biometric verification using ear and iris features. In: Proceedings of the International Conference on Research Trends in Computer Technologies (ICRTCT); Feb 2013. ICRTCT, 4:133-136.
15. Poornima S, Rajavelu C, Subramanian S. Comparison and a neural network approach for iris localization. Procedia Comput Sci. 2010;2:127–32. DOI: 10.1016/j.procs.2010.11.016.
16. George AS, George AH. Deepfakes: The evolution of hyper-realistic media manipulation. Partners Univ Innov Res. 2023;1:58–74.
17. Westerlund M. The emergence of deepfake technology: A review. Technol Innov Manag Rev. 2019;9:39–52. DOI: 10.22215/timreview/1282.
18. Shahzad HF, Rustam F, Flores ES, Luís Vidal Mazón J, de la Torre Diez I, Ashraf I. A review of image processing techniques for deepfakes. Sensors. 2022;22:4556. DOI: 10.3390/s22124556.
19. Wang Z, Guo Y, Zuo W. Deepfake forensics via an adversarial game. IEEE Trans Image Process. 2022;31:3541–52. DOI: 10.1109/TIP.2022.3172845.
20. Silva SH, Bethany M, Votto AM, Scarff IH, Beebe N, Najafirad P. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. Forensic Sci Int Synergy. 2022;4:100217. DOI: 10.1016/j.fsisyn.2022.100217.
21. Pashine S, Mandiya S, Gupta P, Sheikh R. Deep fake detection: Survey of facial manipulation detection solutions. [Preprint] ArXiv:2106.12605. 2021.
22. Joseph Z, Nyirenda C. Deepfake detection using a two-stream capsule network. 2021 IST-Africa Conference (IST-Africa), South Africa, South Africa, 2021, pp. 1–8.
23. Heo YJ, Yeo WH, Kim BG. DeepFake detection algorithm based on improved vision transformer. Appl Intell. 2023;53:7512–27. DOI:10.1007/s10489-022-03867-9.
24. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: A compact facial video forgery detection network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China. 2018, pp. 1–7. DOI: 10.1109/WIFS.2018.8630761.
25. Boesch G. (2023). Vision Transformers (ViT) in Image Recognition - 2024 Guide. [online] viso.ai. Available from: https://viso.ai/deep-learning/vision-transformer-vit/
26. Ahmed OB, Asghar AH, Bahwerth FS, Assaggaf HM, Bamaga MA. [RETRACTED ARTICLE]: The prevalence of aminoglycoside-resistant genes in Gram-negative bacteria in tertiary hospitals. Appl Nanosci. 2023;13:1093–3. DOI: 10.1007/s13204-021-01887-4.
27. Das S, Seferbekov S, Datta A, Islam MS, Amin MR. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 3769–78. DOI: 10.1109/ICCVW54120.2021.00421.
28. Coccomini DA, Messina N, Gennaro C, Falchi F. Combining EfficientNet and vision transformers for video deepfake detection. In: Sclaroff S, Distante C, Leo M, Farinella GM, Tombari F, editors. Image Analysis and Processing – ICIAP 2022. Cham (Germany): Springer; 2022. p. 239–51. DOI: 10.1007/978-3-031-06433-3_19.

