Deep Learning Framework for Robust Deep Fake Image Detection: A Review

Ankita Bhandarkar
Center of Excellence (Digital Technology)
Department of Computer Science and Engineering
G H Raisoni College of Engineering, Nagpur, India
ankita.bhandarkar@raisoni.net

Prashant Khobragade
Department of Computer Science & Engineering,
G H Raisoni College of Engineering, Nagpur, India
prashant.khobragade@raisoni.net

Raju Pawar
Department of Computer Science & Engineering,
G H Raisoni College of Engineering, Nagpur, India
raju.pawar@raisoni.net

Prasad Lokulwar
Center of Excellence (Digital Technology)
Department of Computer Science and Engineering
G H Raisoni College of Engineering, Nagpur, India
prasad.lokulwar @raisoni.net

Pranay Saraf
Department of DIC
G H Raisoni College of Engineering, Nagpur, India
pranay.saraf@raisoni.net

***Abstract-* Deepfake technology, which uses deep learning to change pictures and videos, makes it difficult to identify and tell the real and fake images. People are worried about false information, data breaches, and bad uses of deepfake generation methods that are getting better and better very quickly. As the name suggests, this study looks at the current deep learning methods for finding deepfake pictures, focusing on their pros and cons. It looks at how well CNN method, autoencoders method, RNN method and attention-based models can discover little visual blemishes and blunders that were included amid the deepfake creation handle. The study also talks about how data addition, transfer learning, and ensemble learning can help improve the efficiency of identification. Additionally, it talks about how deepfake technology is changing and the need for flexible models that can successfully fight more complex fakes. Problems like small datasets, heavy processing needs, and the chance of hostile attacks are talked about, along with possible answers such as combining blockchain technology and real-time detection systems. This review gives an in-depth look at the latest progress and gaps in deepfake detection. The goal is to help make tools that are stronger, more scalable, and more reliable to fight the growing threat of deepfake picture manipulation.**

Keywords: CNN, GNN, Deep fake images, Artificial Intelligence (AI), Deep Learning

I. INTRODUCTION

Deepfakes, a term derived from "deep learning" and "fake," are artificial intelligence (AI) advancements that create highly realistic synthetic images and videos. These images, initially developed for entertainment and creative arts, have evolved into tools with potentially malicious applications, such as political misinformation, identity fraud, and reputational damage. Detecting deepfake images is challenging due to their increasing sophistication, where subtle manipulations can deceive even the most trained human eyes.[1] Traditional forensic techniques, which rely on visual inconsistencies, are becoming less effective as deepfake technology improves. Due to its capacity to memorize and discover complex designs in information, particularly pictures, deep learning has become a potential way to find deepfakes. Convolutional Neural Networks (CNNs) are a powerful way to find deepfake pictures because they instantly pull out relevant features from data without the need for feature building. This article talks about how to make and use a deep learning system based on CNN architecture that is perfect for finding deepfake pictures. The suggested framework aims to offer a strong and dependable answer to this growing issue, as part of the larger effort to protect the purity of digital media [2].

The fast growth of AIM (Artificial Intelligence and Machine) learning has changed many areas, including how images and videos are processed. As a strong tool for making very accurate fake media, deep learning has become known as "deep fakes." Deep learning methods were used to make these fakes, which are a big problem because they spread false information and pose security and privacy risks [3].

979-8-3315-1795-3/24/$31.00©2024 IEEE

Fig. 1. Deepfake Technology and Its Impact

The expansion of deep fake innovation has raised critical concerns in segments like news coverage, legislative issues, and social media, where the realness of visual substance is fundamental. In spite of the developing mindfulness of the suggestions of deep fakes, existing discovery strategies regularly battle to keep pace with the modernity of control strategies [4]. Conventional approaches, which regularly depend on handcrafted highlights and machine learning classifiers, are progressively lacking in tending to the challenges postured by deep fakes. There's a squeezing require for strong and versatile arrangements that use the capabilities of present day deep-learning strategies to progress discovery precision and productivity [5]. This inquire about centers on planning and executing a deep learning system pointed at improving the location of deep fake pictures. By utilizing progressed neural arrange structures, especially convolutional neural systems (CNNs), the think about points to form a framework that not as it were recognizes controlled pictures with tall precision but moreover adjusts to the advancing scene of deep fake era procedures. The targets of this investigate are to analyze the current state of deep fake discovery strategies, create a comprehensive deep learning system, assess its execution against built up benchmarks and compare it with existing strategies [6]. Deepfake innovation, a combination of deep learning and fake, employments artificial intelligence and machine learning to make practical manufactured pictures, recordings, or sound by controlling or manufacturing individuals' appearance and conduct. Its center is Generative Ill-disposed Systems (GANs), which comprise of two neural systems that produce fake substance and endeavor to recognize between genuine and fake information. At first created for genuine purposes, deepfakes have ended up a effective device with possibly hurtful applications, raising ethical, lawful, and social concerns. They have been utilized to manufacture political addresses, fake news, and celebrity pantomimes, undermining open believe in computerized media, and posturing a genuine risk to political frameworks, notorieties, and privacy [7]. The effect of deepfake innovation expands past person hurt, challenging the establishment of media judgment and data dispersal. In cybersecurity, deepfakes have been utilized for advanced phishing assaults, extortion, and personality burglary, exposing vulnerabilities in verification frameworks. As deepfake innovation advances, it gets to be progressively troublesome to distinguish these creations, requiring the improvement of strong location components. This presents a mechanical breakthrough and societal challenge, requiring collaboration between AI specialists, policymakers, and cybersecurity experts [8].

TABLE I. SUMMARY OF DIFFENT DEP LEARNING MODEL WITH KEY STRENGTH

Model	Data Requirements	Model Flexibility	Key Strengths
Convolutional Neural Network (CNN)	Moderate	Moderate	Efficient in image feature extraction
Generative Adversarial Network (GAN)	Very High	High	Can generate and detect highly realistic fake images
Recurrent Neural	High	Low	Captures temporal
Hybrid Model (CNN + RNN)	Very High	Very High	Combines spatial and temporal strengths effectively

II. BACKGROUNG WORK

In study [15] Digital picture forgery is a serious problem, and knowing the identification and categorization of altered digital photographs is critical. Copy-move forgeries are among the most popular due to their simplicity of execution. To counteract this, passive image forensic approaches have emerged, with deep learning-based systems being considered cutting-edge for image processing and counterfeit detection. However, current deep learning algorithms are time and resource-costly.

The Deepfake images developed with deep generative models pose serious hazards to web platforms. Researchers have been working on detecting these photos, but the approaches are not yet suitable for use due to two recent changes [16]. First, lightweight approaches for customizing big generative models can produce a high number of modified generators, expanding the danger surface. Existing protections do not apply well to these user-customized models. New machine learning algorithms and ensemble modelling are addressed as ways to increase generalization performance versus these models.

The research study of [17] Deep Learning (DL) is a powerful technique for handling various problems, including data analytics and disease diagnosis. In any case, it has raised security, equity, and national security concerns. Deepfake, a prevalent DL-based application, makes fake pictures and recordings that are troublesome for people to recognize. This paper examines the issues, openings, and prospects of Deepfake innovation and proposes mechanized strategies to identify and assess dangers.

This study investigate [18] explores the flexibility of deep neural organize (DNN) models for facial acknowledgment. It assesses the impact of DNNs on assault vulnerabilities, finds singularities by analyzing distorted channel reaction conduct in covered up layers, and alters the handling pipeline.

Deepfake images undermine the genuineness and astuteness of mixed media fabric within the advanced period. To solve this, DeepSight could be a novel arrangement that mixes outfit learning approaches and deep learning models. It utilizes various classifiers prepared on diverse highlight representations recovered from crude picture information to decide in case the picture has been modified. The highlight vectors are categorized utilizing Arbitrary Woodland, KNearest Neighbors, and XGBoost with hyper-parameter optimization. DeepSight outperforms cutting-edge calculations in test appraisals, with DenseNet and XGBoost yielding the finest precision of 97.5%.

TABLE II. SUMMARY OF RELATED WORK IN DEEP FAKE IMAGE ANALYSIS

Methods	**Approach**	**Finding**	**Limitation**
Convolutional Neural Networks (CNN) [19]	Feature extraction through convolutional layers	Achieved high accuracy in detecting manipulated regions	Requires large datasets and is computationally intensive
Recurrent Neural Networks (RNN) [20]	Sequential analysis of temporal inconsistencies	Effective in detecting temporal artifacts in videos	Less effective for still images; struggles with subtle changes
Autoencoders [21]	Image reconstruction and anomaly detection	Capable of identifying inconsistencies by reconstruction loss	Prone to false positives, especially with high-quality fakes
Capsule Networks [22]	Captures spatial relationships of image features	Better performance on complex manipulations	Limited scalability and less tested on diverse datasets
Ensemble Learning [23]	Combines multiple models for better accuracy	Enhances detection by leveraging strengths of individual models	Requires more computational resources and complex tuning
Frequency Domain Analysis [24]	Detects discrepancies in frequency patterns	Identifies subtle artifacts not visible in spatial domain	Less effective against advanced, high-resolution fakes
Optical Flow Analysis [24]	Analyzes motion inconsistencies across frames	Effective for detecting deep fake videos	Not suitable for static image analysis
Deep Learning with Transfer Learning [25]	Utilizes pre-trained models for fake detection	Offers high accuracy with limited training data	Performance drops with entirely new fake techniques

III. DEEP LEARNING TECHNIQUES

These neural network based models can adapt to evolving manipulation techniques without human intervention, providing the computational power and flexibility needed to keep pace with advancements [10].

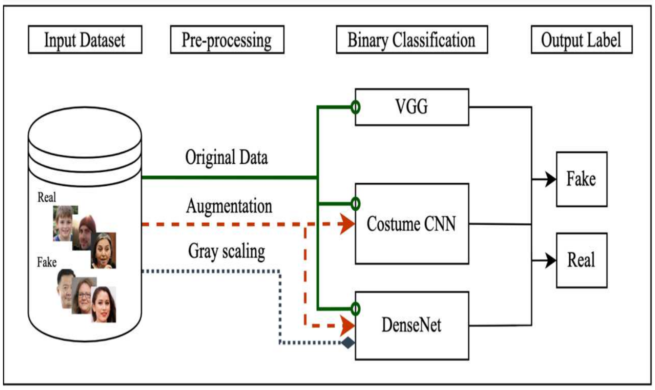

Fig. 2: Representation of methodolgy for deep fake detection

CNN: Computer vision devices, especially CNNs, are profoundly compelling in recognizing deepfakes by naturally capturing spatial highlights from pictures. These instruments utilize convolutional layers to prepare input pictures, extricating surfaces, edges, and facial points of interest, which are frequently modified in controlled media. Their capacity to handle high-dimensional information with negligible preprocessing makes them a pivotal component in any deepfake location framework [11].

GANs: GANs, comprising of a generator and a discriminator, are utilized to make manufactured information and recognize between genuine and created information. This antagonistic prepare produces practical pictures and empowers the advancement of vigorous location models. Preparing a profound learning show with GAN-based information improves discovery systems' viability in recognizing controls and recognizing true substance from fakes [12], [13],[14].

TABLE III. COMPARATIVE ANALYSIS OF DEEP LEARNING TECHNIQUES FOR DEEPFAKE DETECTION

Aspects	CMM Model	GAN Model	RNN Model	Autoencoders
Primary Use	Image-based feature extraction [26]	Data generation and discrimination	Sequential data analysis (video frames)	Anomaly detection (image reconstruction)
Processing Speed	Fast for images [27]	Medium (depends on training complexity)	Slower due to sequential data processing	Fast for image reconstruction
Data Requirement [28]	Large labeled datasets	Requires synthetic and real data	High (video sequences)	Moderate (training on anomalies)
Strengths	High spatial feature detection [29]	Creates robust fake/real distinctions	Captures temporal inconsistencies	Effective at identifying subtle manipulations
Weaknesses [30]	Less effective for temporal changes	Computationally intensive	Less accurate for single-frame detection	Prone to false positives in high-quality fakes
Typical Application	Image-based deepfake detection	Training detection models using fakes	Deepfake video detection	Detecting anomalies in images/videos
Robustness to Manipulations	High for image-level manipulations	Medium-high, depends on training	High for video-based manipulations	Medium, affected by quality of input

IV. DATASET AVAILABLE

A. Celeb-DF

Celeb-DF is a large-scale deepfake dataset with 6,229 videos, featuring realistic manipulations across diverse ages, ethnicities, and genders. It offers challenging examples for developing advanced detection models.

B. DFDC (Deepfake Detection Challenge)

DFDC comprises over 100,000 videos with various deepfake techniques, making it ideal for training and evaluating detection models. It's widely referenced, offering diverse manipulation complexities for robust model development.

C. FaceForensics

FaceForensics contains over 500,000 frames from 1,004 videos, focusing on frontal faces. It offers two compression levels, making it valuable for studying image/video forgeries and benchmarking detection models.

D. FakeAVCeleb

FakeAVCeleb is an audio-visual deepfake dataset containing manipulated videos and corresponding synthesized audio. It covers lip-syncing, voice cloning, and face swapping, aiding comprehensive research on synchronized deepfake detection.

E. WildDeepfake

WildDeepfake features 7,314 face sequences from 707 real-world deepfake videos. Collected from the internet, it provides varied conditions, making it ideal for developing models against real-world deepfakes.

TABLE IV. DETAILS OF DEEP FAKE DATSET AVAILABLE

Name of Dataset	Records	Class	Variant	Type
Celeb-DF [31]	590 original videos + 5,639 DeepFake videos	Videos	DeepFake vs. Original	Face Manipulation
DFDC (Deepfake Detection Challenge) [32]	Over 100,000 videos	Videos	DeepFake vs. Real	Face Manipulation
FaceForensics [33]	Over 500,000 frames from 1,004 videos	Images and Videos	Manipulated vs. Original	Face Manipulation
FakeAVCeleb [34]	Deepfake videos with synthesized cloned audios	Videos and Audio	DeepFake vs. Original	Audio-Visual Manipulation

WildDeepfake	7,314 face image sequences from 707 deepfake moving videos	Videos	DeepFake vs. Real	Real-world Deepfake
DeeperForensics-1.0	60,000 videos (48,475 source, 11,000 manipulated)	Videos	Manipulated vs. Original	Face Forgery
WaveFake	Over 100,000 generated audio clips	Audio	DeepFake vs. Real	Audio Manipulation
KoDF (Korean DeepFake Detection Dataset)	Numerous deepfake videos in Korean language	Videos	DeepFake vs. Original	Face Manipulation

V. Limitations and Challenges in Current Deepfake Detection

A. High-Quality Deepfake Generation

Because deepfake generation methods are improving so quickly, fakes are becoming more lifelike and high-quality, which makes them harder to spot. Modern deepfake algorithms can make face emotions that are very subtle, keep lips moving in sync, and even make skin patterns that look very real. A lot of the time, these high-quality changes get past recognition models, especially ones that were learned on older or less complex datasets. As deepfake technology changes, tracking methods have a hard time keeping up.

B. Computational Complexity and Resource Demands

For training and reasoning, deepfake detecting models, especially those that use deep learning, need a lot of computer power. To get correct results, you often need GPUs with a lot of power, a lot of memory, and a lot of training time. This makes it very hard to identify things in real time, especially video material. Also, it's not possible to use models that use a lot of resources on consumer electronics or low-power systems. This makes it harder to use and expand deepfake detection methods in real life.

C.Dataset Limitations and Biases

How well deepfake recognition models work depends a lot on the quality and variety of the datasets that are used to train them. A lot of the statistics that are already out there are limited in size, variety, and the number of people from different racial, age, and gender groups that they include. This makes models skewed or less good at finding deepfakes involving groups that aren't well represented. Also, some datasets were made using old deepfake techniques, which makes them less useful for training models to find newer, more complex changes. To make recognition work better, we need to get rid of these flaws and add more diverse and up-to-date cases to our datasets.

VI. Conclusion

Deepfake picture recognition has become an important area of study because deepfake creation methods are getting smarter and can be used in bad ways to share false information and break digital security. Additionally, this review shows how well deep learning systems like CNNs, RNNs, autoencoders, and attention-based models can spot deepfake pictures by using their skill to find small visual errors. Even though there has been a lot of progress, it is still hard to make sure that these models are strong and flexible enough to handle new deepfake methods and hostile attacks. Adding advanced techniques like data reinforcement, transfer learning, and ensemble learning to recognition methods has shown promise in making them more accurate. But for creating models that are more adaptable, we need samples that are bigger, more varied, and of higher quality. Because deepfake technologies change all the time, real-time recognition is still a difficult but necessary goal.

References

[1] Antora, K. F., et al. (2024). Aiblocknet - Framework for Authenticity Validation Using Blockchain & Machine Learning for Fake Image Detection. https://doi.org/10.2139/ssrn.4938931

[2] Balasubramanian, S. B., et al. (2022). Deep fake detection using cascaded deep sparse auto-encoder. PeerJ Computer Science, 8, e1040. https://doi.org/10.7717/peerj-cs.1040

[3] Bethu, S., et al. (2023). Framework for Deep Fake Motion Detection using Deep Learning. ICACECS 2023. https://doi.org/10.2991/978-94-6463-314-6_18

[4] Castillo Camacho, I., & Wang, K. (2021). Deep-Learning-Based Methods for Image Forensics. Journal of Imaging, 7(4), 69. https://doi.org/10.3390/jimaging7040069

[5] Ghai, A., Kumar, P., & Gupta, S. (2024). Deep-learning-based image forgery detection framework. Information Technology & People, 37(2), 966–997. https://doi.org/10.1108/ITP-10-2020-0699

[6] Heidari, A., et al. (2024). Deepfake detection using deep learning methods: A review. WIREs Data Mining, 14(2). https://doi.org/10.1002/widm.1520

[7] Hsu, C.-C., et al. (2020). Deep Fake Image Detection Based on Pairwise Learning. Applied Sciences, 10(1), 370. https://doi.org/10.3390/app10010370

[8] Vaishnavi, K. D. V. N., et al. (2023). Deep learning for robust deep fake detection. World J. Adv. Res. Rev., 21(3), 2283–2289. https://doi.org/10.30574/wjarr.2024.21.3.0889

[9] Mcuba, M., et al. (2023). Deepfake Audio Detection for Digital Investigation. Procedia Computer Science, 219, 211–219. https://doi.org/10.1016/j.procs.2023.01.283

[10] Nguyen, T. T., et al. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision Image Understanding, 223. https://doi.org/10.1016/j.cviu.2022.103525

[11] Pashine, S., et al. (2021). Deep Fake Detection: Survey of Facial Manipulation Solutions. arXiv preprint. https://doi.org/10.48550/ARXIV.2106.12605

[12] Passos, L. A., et al. (2024). Review of deep learning-based approaches for deepfake content detection. Expert Systems, 41(8). https://doi.org/10.1111/exsy.13570

[13] Patel, Y., et al. (2023). Deepfake Generation and Detection: Case Study. IEEE Access, 11, 143296–143323. https://doi.org/10.1109/ACCESS.2023.3342107

[14] Pu, J., et al. (2020). NoiseScope: Detecting Deepfake Images. ACM, 913–927. https://doi.org/10.1145/3427228.3427285

[15] Abbas, M. N., et al. (2021). Lightweight Deep Learning Model for Image Forgery Detection. IEEE SAMI. https://doi.org/10.1109/SAMI50585.2021.9378690

[16] Abdullah, S. M., et al. (2024). Advances in Deepfake Image Detection. arXiv preprint. https://doi.org/10.48550/ARXIV.2404.16212

[17] Hadke, R. T., & Khobragade, P. (2015). Approach for class imbalance using oversampling. Int. J. Innov. Res. Comput. Commun. Eng., 3(11), 11451-11455.

[18] Goswami, G., et al. (2019). Detecting & Mitigating Adversarial Perturbations. Int. J. Comput. Vis., 127(6–7), 719–742.

[19] N. J. Zade, N. P. Lanke, B. S. Madan, P. Ghutke, and P. Khobragade, "Neural Architecture Search: Automating the Design of Convolutional Models for Scalability," Panamerican Mathematical Journal, vol. 34, no. 4, pp. 178–193, 2024. https://doi.org/10.52783/pmj.v34.i4.1877

[20] M. Bende, M. Khandelwal, D. Borgaonkar and P. Khobragade, "VISMA: A Machine Learning Approach to Image Manipulation," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-5, doi: 10.1109/ISCON57294.2023.10112168.

[21] K. Agnihotri, P. Chilbule, S. Prashant, P. Jain and P. Khobragade, "Generating Image Description Using Machine Learning Algorithms," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/ICETET-SIP58143.2023.10151472.

[22] Ankit Duddalwar, Prashant Khobragade; A statistical approach for hospital management system using machine learning. AIP Conf. Proc. 6 August 2024; 3139 (1): 100007. https://doi.org/10.1063/5.0224460

[23] Golchha, Rishi, Prashant Khobragade, and Ashish Talekar. "Design of an Efficient Model for Health Status Prediction Using LSTM, Transformer, and Bayesian Neural Networks." 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET). IEEE, 2024.

[24] R. N. Wadibhasme, A. U. Chaudhari, P. Khobragade, H. D. Mehta, R. Agrawal and C. Dhule, "Detection And Prevention of Malicious Activities In Vulnerable Network Security Using Deep Learning," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, 2024, pp. 1-6, doi: 10.1109/ICICET59348.2024.10616289.

[25] P. K. Pande, P. Khobragade, S. N. Ajani and V. P. Uplanchiwar, "Early Detection and Prediction of Heart Disease with Machine Learning Techniques," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, 2024, pp. 1-6, doi: 10.1109/ICICET59348.2024.10616294.

[26] R. Golcha, P. Khobragade and A. Talekar, "Multimodal Deep Learning for Advanced Health Monitoring A Comprehensive Approach for Enhanced Precision and Early Disease Detection," 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 2024, pp. 1-6, doi: 10.1109/ICITIIT61487.2024.10580622.

[27] Ajani, S. N. ., Khobragade, P. ., Dhone, M. ., Ganguly, B. ., Shelke, N. ., & Parati, N. . (2023). Advancements in Computing: Emerging Trends in Computational Science with Next-Generation Computing. International Journal of Intelligent Systems and Applications in Engineering, 12(7s), 546–559

[28] S. N. Atkar, R. Agrawal, C. Dhule, N. C. Morris, P. Saraf and K. Kalbande, "Speech Emotion Recognition using Dialogue Emotion Decoder and CNN Classifier," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 94-99, doi: 10.1109/ICAAIC56838.2023.10141417.

[29] Saraf, P.D., Bartere, M.M., Lokulwar, P.P. (2022). Survey on Edge, Fog Assisted IoT Framework Using Intelligent Learning Techniques. In: Gunjan, V.K., Zurada, J.M. (eds) Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Lecture Notes in Networks and Systems, vol 237. Springer, Singapore. https://doi.org/10.1007/978-981-16-6407-6_17

[30] M. Jaipurkar, N. Ragit, C. Tambuskar and P. D. Saraf, "IPL Data Analysis and Visualization Using Microsoft Power BI Tool," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10306783.

[31] P. Keshattiwar, P. Lokulwar and P. Saraf, "Data Defender's Shield in Safeguarding Information through Advanced Encryption and Access Management in Cloud-Based Applications," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, 2024, pp. 1-6, doi: 10.1109/ICICET59348.2024.10616375.

[32] C. Dhule, R. Agrawal, S. Dorle and B. Vidhale, "Study of Design of IoT based Digital Board for Real Time Data Delivery on National Highway," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 195-198, doi: 10.1109/ICICT50816.2021.9358560.

[33] C. Dhule and T. Nagrare, "Computer Vision Based Human-Computer Interaction Using Color Detection Techniques," 2014 Fourth International Conference on Communication Systems and Network Technologies, Bhopal, India, 2014, pp. 934-938, doi: 10.1109/CSNT.2014.192.

[34] A. A. Deshmukh, S. D. B. Sonar, R. V. Ingole, R. Agrawal, C. Dhule and N. C. Morris, "Satellite Image Segmentation for Forest Fire Risk Detection using Gaussian Mixture Models," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 806-811, doi: 10.1109/ICAAIC56838.2023.10140399.