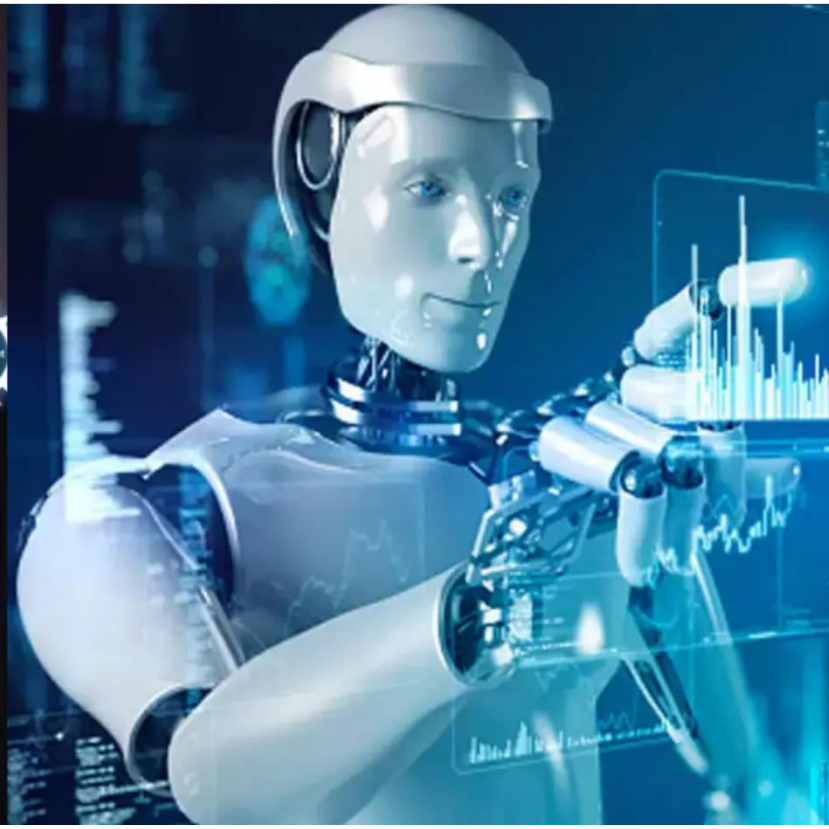
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

Deepfake Video and Image Detection using Deep Learning

Kanumuri Veera Sai Sindhu[1] Kummari Srihari[2] Nambula Veerendra[3] Kumbha Venkat Kyathi[4]

Mittapalli Pardhasaradhi[5]

Students, Department of Computer Science Engineering (Artificial Intelligence & Machine Learning)

Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[1234]

Associate Professor, Department of Computer Science Engineering (Artificial Intelligence & Machine Learning) [5]

Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[5]

ABSTRACT: The rise of deep learning-based tools has made it easier to create highly convincing face-swapped images and videos, often leaving little to no visible evidence of tampering. These AI-generated manipulations, commonly referred to as Deepfakes (DF), pose serious concerns regarding misinformation and digital fraud. Although generating deepfakes is a relatively simple process, accurately detecting them remains a major challenge due to their increasingly sophisticated nature [1].

This study presents an advanced deepfake detection framework capable of analysing both videos and images. For video-based detection, a Convolutional Neural Network (CNN) is used to extract frame-level features, while a Recurrent Neural Network (RNN) identifies inconsistencies across frames to detect temporal anomalies introduced by deepfake generation methods [2]. To enhance image-based detection, we integrate a MobileNet-based CNN model, which efficiently extracts features and classifies images while maintaining computational efficiency [3].

Our system is tested on a large dataset of manipulated images and videos, demonstrating strong performance in differentiating real content from AI-generated forgeries. By combining MobileNet for static image detection and a CNN-RNN pipeline for video analysis, our approach enhances the accuracy and reliability of deepfake identification [4].

KEYWORDS: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), MobileNet, AI-generated Content, Image and Video Forgery Detection, Deepfake Detection.

I. INTRODUCTION

The widespread availability of advanced smartphone cameras and high-speed internet has led to an exponential rise in digital media creation and sharing. With the increasing power of computational resources, deep learning techniques have progressed to a level where they can generate highly realistic synthetic content, which was once deemed unattainable.

One of the most pressing concerns arising from these advancements is deepfake (DF) technology, which employs Generative Adversarial Networks (GANs) to manipulate both videos and images seamlessly. While GANs manipulate videos by swapping the face of a target individual with that of a source, they ensure facial features and expressions align realistically. This involves breaking the video into frames, altering each frame, and then reconstructing the sequence to maintain fluidity. However, due to computational constraints, these models produce fixed-size face images that require resizing, often introducing distortions. Our approach detects such inconsistencies by examining frame-level artifacts and comparing the generated facial areas with their surrounding regions.

To address this challenge, we propose an AI-powered deepfake detection system that effectively identifies manipulated media in both videos and images. Our video detection framework leverages a ResNext-based Convolutional Neural Network (CNN) to extract spatial features from individual frames. These extracted features are then processed through

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to analyse temporal inconsistencies introduced during deepfake video generation.

For image-based detection, we employ a MobileNet-based CNN model, which enables efficient feature extraction while maintaining computational efficiency. Unlike traditional deepfake detection approaches, our method effectively captures subtle visual distortions and resolution inconsistencies caused by GAN-based transformations.

Deepfake creation tools often operate under computational and time constraints, producing face images at a limited resolution that require Affine transformations to align with the target face. This process introduces detectable artifacts, especially inconsistencies between the altered face region and its surrounding areas. Our system identifies these artifacts to accurately distinguish between authentic and manipulated media.

We evaluate the performance of our proposed method using large-scale deepfake datasets consisting of both images and videos. The combined ResNext-CNN + RNN-LSTM architecture for video analysis and MobileNet-CNN for image detection provides a highly efficient, accurate, and scalable solution for detecting deepfake content.

II. LITERATURE SURVEY

The increasing prevalence of deepfake content has raised concerns regarding its misuse in spreading misinformation, necessitating the development of reliable detection techniques. Various approaches have been proposed for deepfake detection, focusing on both video-based and image-based analysis.

One category of methods examines motion inconsistencies, detecting unnatural facial deformations and irregular head movements caused by improper motion synthesis. However, such methods struggle when deepfake models accurately replicate natural head motions. To enhance motion-based detection, we employ LSTM networks that process sequential frame data, allowing for better temporal feature extraction and anomaly detection [5].

Another widely explored technique detects audio-visual synchronization errors, where discrepancies between lip movements and speech patterns indicate manipulation. While effective, this method is less reliable in manually synchronized videos or silent clips. Our approach integrates ResNext CNN for feature extraction, ensuring that fine-grained visual inconsistencies are detected independently of audio cues [6].

Frame-level temporal analysis identifies artifacts in lighting, texture, and color variations across consecutive frames. However, advanced deepfake generation methods increasingly minimize such inconsistencies. To address this, we preprocess video frames by detecting and cropping faces, ensuring a uniform dataset where variations in extracted features are more pronounced. By feeding sequential frame representations into an LSTM network, we enhance the model's ability to track subtle alterations over time [7].

Another effective approach utilizes physiological cues like blood flow and skin texture variations, which are difficult for deepfake algorithms to replicate. However, such methods are sensitive to lighting conditions and resolution. To mitigate these challenges, our dataset preprocessing ensures uniform video frame extraction, optimizing detection accuracy across different lighting environments [8].

For image-based detection, frequency domain analysis has been employed to reveal deepfake-induced distortions, though it is less effective against high-quality fakes. Our approach leverages MobileNet with fine-tuned layers, enhancing feature extraction capabilities for improved classification [9]. Additionally, we employ image augmentation techniques such as rotation, width shifting, height shifting, and horizontal flipping, improving model robustness against various manipulations [10].

Local texture analysis detects inconsistencies in image fine details but may fail when excessive post-processing is applied. We refine this technique by using MobileNet's pretrained feature extractor, fine-tuning its layers to identify subtle deepfake traces. Furthermore, metadata analysis, though useful for flagging inconsistencies, can be bypassed by

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

altering metadata. Our model circumvents this limitation by relying primarily on pixel-level and high-level feature extraction methods [4].

By integrating ResNext CNN for spatial feature extraction, LSTM for sequential processing in videos, and MobileNet for image detection, our method builds upon existing approaches to improve deepfake detection accuracy across both video and image datasets.

III. PROPOSED SYSTEM

Numerous tools are available for generating deepfakes, but there is a noticeable scarcity of solutions for identifying them. Our deepfake detection method aims to make a substantial impact in curbing their proliferation online. We are in the process of creating a web-based platform that allows users to upload images or videos to determine whether they are authentic or manipulated. This initiative has the potential to evolve from a web platform into a browser extension capable of automatically detecting deepfakes. Prominent applications such as WhatsApp and Facebook could incorporate this system to scrutinize media before it is circulated among users. A primary goal of this project is to assess its effectiveness and acceptance, focusing on aspects like security, ease of use, precision, and dependability. Our approach is designed to identify various forms of deepfakes, including replacement, retrenchment, and interpersonal deepfakes. Figure 1 represents the simplified system architecture of the proposed system: -

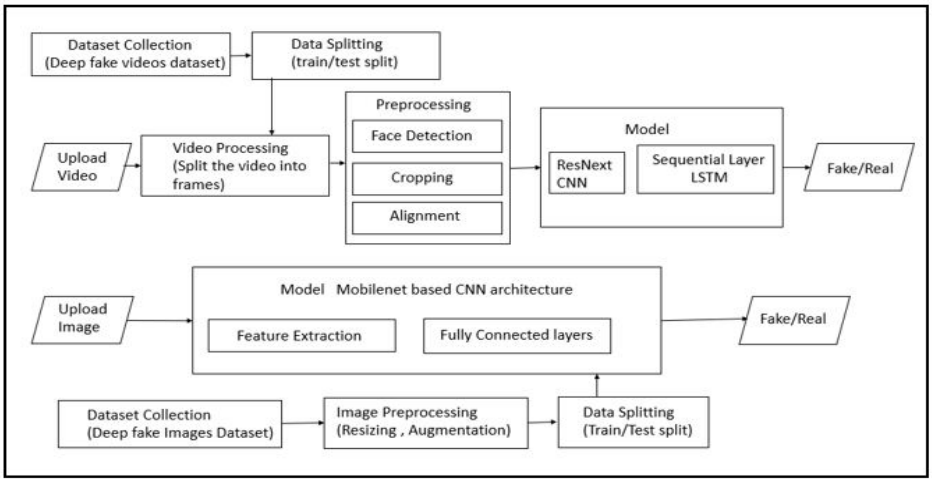


Fig. 1: System Architecture

A. Dataset Collection:
The dataset used for this project consists of both images and videos, structured to ensure a balanced and effective training process.

Video Detection: For video detection, we are using a mixed dataset containing an equal number of real and manipulated deepfake videos. The videos are collected from multiple sources, including YouTube, FaceForensics++ [4] and the Celeb-DF [5] dataset. The dataset is split into 70% for training and 30% for testing, ensuring a robust model training process.

Image Detection: For image detection, the dataset collected from multiple sources is divided into three sets: training, validation, and testing. Each set contains two directories, one for real images and another for fake images. The dataset comprises 256x256 pixel JPEG images of human faces, labelled as either real or fake.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

B. Preprocessing:

Video Detection:

For videos, preprocessing involves splitting the video into individual frames, detecting faces in each frame, and cropping the detected faces. To maintain uniformity, the mean number of frames per video is calculated, and all videos are processed to match this number. Frames without faces are discarded to ensure that only relevant data is used. Since processing an entire 10-second video at 30 frames per second would be computationally expensive, only the first 100 frames of each video are used for training.

Image Detection:

For images, preprocessing involves resizing all images to a uniform dimension of 256x256 pixels to maintain consistency across the dataset. Normalization is applied to scale pixel values between 0 and 1, improving model performance. Augmentation techniques such as rotation, flipping, and contrast adjustment are used to enhance the dataset and improve model generalization.

C. Video Classification Model:

a) Model Architecture:

The video classification model follows a two-stage pipeline: feature extraction using a ResNeXt-50 (32x4d) CNN and sequential processing using an LSTM layer. The Data Loader is responsible for preloading face-cropped video frames, splitting the dataset into training and testing sets, and feeding these frames into the model in mini-batches. Each video is broken down into frames, which are individually processed by the CNN before being analysed in sequence by the LSTM network. This structure enables the model to capture both spatial and temporal features, making it highly effective in detecting deepfake videos.

b) ResNeXt CNN for Feature Extraction

Instead of manually designing a classifier, the model uses ResNeXt-50 (32x4d) CNN as a feature extractor. This architecture improves upon traditional ResNet by employing grouped convolutions, which enhance feature representation. The CNN is fine-tuned with additional layers and an optimized learning rate to ensure stable convergence during training. The CNN extracts high-level features from each video frame, capturing texture, structural details, and inconsistencies that may indicate a deepfake. The last pooling layer of ResNeXt produces a 2048-dimensional feature vector, which is then passed into the LSTM network. This method ensures that the extracted features contain sufficient information for deepfake classification.

C) LSTM for Sequence Processing:

Once the ResNeXt CNN extracts frame-level features, they are processed by an LSTM (Long Short-Term Memory) network to analyse the temporal relationships between frames. The primary challenge in deepfake detection is identifying inconsistencies between frames, which is where LSTMs excel. The model employs 2048 LSTM units with a dropout rate of 0.4 to reduce overfitting. The LSTM processes the extracted feature sequences and compares frames at time t with frames at t-n (where n represents previous frames in the sequence).

This comparison helps detect unnatural transitions or temporal inconsistencies that are common in deepfake videos. The final classification layer is a 2-node neural network, which outputs probability scores indicating whether the video is real or fake.

D. Image Classification Model

a) Model architecture:

For detecting deepfake images, a MobileNet-based Convolutional Neural Network (CNN) is employed. This model leverages the efficiency of MobileNet, which is pre-trained on the ImageNet dataset and fine-tuned for deepfake detection. By learning from a dataset containing both real and manipulated images, the model effectively distinguishes between them with high accuracy. The CNN processes each input image, extracts relevant features, and classifies it as either real or fake based on distinctive characteristics.

b) CNN for Feature Extraction:

The architecture consists of multiple layers designed to extract crucial spatial features from input images. A MobileNet backbone is used as the primary feature extractor, capturing fine-grained details such as textures, edges, and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

irregularities that indicate potential deepfake manipulations. To reduce computational complexity while retaining essential features, a Global Average Pooling layer is applied, converting feature maps into a compact representation. Additionally, fully connected layers refine these extracted features, enabling the model to make more precise classifications. To enhance generalization and prevent overfitting, dropout layers are incorporated within the network, ensuring robustness when dealing with unseen data.

c) Fully Connected Layers and Classification:
Once feature extraction is completed, the model flattens the obtained feature maps and passes them through fully connected layers equipped with ReLU activation functions. These layers further refine the extracted features and enable efficient decision-making. The final output layer employs a sigmoid activation function, which assigns a probability score indicating whether an image is real or fake. The model is optimized using the Binary Cross-Entropy loss function, which is well-suited for binary classification tasks. Additionally, the Adam optimizer is used to enhance learning efficiency, ensuring faster convergence during training. To improve model performance and prevent overfitting, early stopping is applied, halting training when validation loss stops improving. Furthermore, model checkpointing is used to save the best-performing model, allowing for optimal inference results.

E. Prediction
Video Prediction:
For video classification, the uploaded video is split into individual frames. Face detection and cropping are performed on each frame to focus on key facial features. The processed frames are passed through the ResNeXt CNN for feature extraction.

The extracted feature vectors are fed into the LSTM network, which analyses temporal inconsistencies. The final classification decision is made based on the overall sequence, with the model outputting a probability score indicating whether the video is real or fake.

Image Prediction:
For image classification, the uploaded image undergoes preprocessing, including resizing and normalization. The image is then fed into the trained CNN model, which extracts features and determines whether it is real or fake based on its learned patterns.

IV. RESULT

The proposed model effectively classifies input images and videos as either real or deepfake, utilizing extracted feature representations for accurate detection. The output consists of a classification label along with a confidence score, indicating the probability of the image belonging to the predicted category. Figure 2 shows the result of video detection, while figure 3 shows the result of image detection.

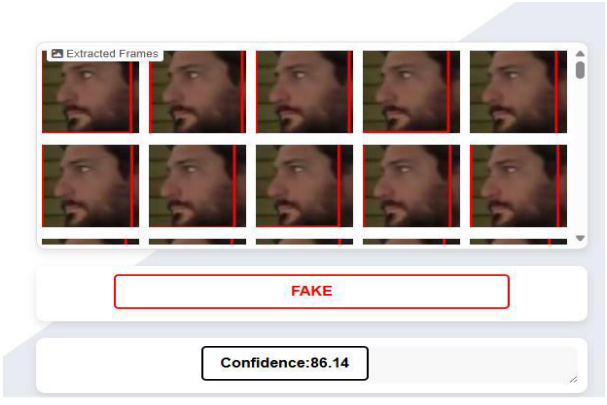


Figure 2: Video Detection

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

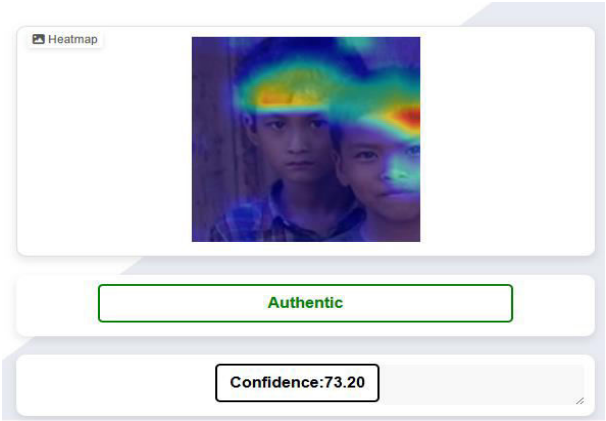


Figure 3: Image Detection

V. CONCLUSION

This research introduces a neural network-based framework for detecting manipulated content in both images and videos. The proposed approach effectively classifies media as real or deepfake while providing a confidence score to indicate prediction reliability. By integrating ResNeXt CNN for feature extraction and LSTM for sequential processing, the model captures both spatial and temporal inconsistencies that are indicative of deepfake manipulations. For image-based detection, the model analyses facial features and artifacts that commonly appear in synthetic media. In video-based detection, it processes individual frames and tracks temporal relationships between them to identify inconsistencies over time.

The use of deep learning techniques enables the model to detect subtle alterations that might be imperceptible to the human eye. Inspired by the mechanisms of Generative Adversarial Networks (GANs) and autoencoders, our approach effectively reverses the deepfake creation process by identifying manipulated patterns.

The system demonstrates high accuracy in distinguishing between real and synthetic media, making it a valuable tool for digital forensics, security applications, and misinformation control.

The effectiveness of this approach suggests its potential for real-world deployment in areas where detecting media authenticity is critical.

Future enhancements may involve training on larger datasets, improving generalization across diverse conditions, and optimizing computational efficiency for real-time applications. By refining the detection process, this method can contribute to safeguarding digital integrity in an era of increasing AI-generated media.

VI. LIMITATIONS

The proposed method effectively detects deepfake content in both images and videos; however, it does not consider audio-based manipulations, making it incapable of detecting audio deepfakes. In video detection, the model focuses solely on visual inconsistencies such as facial artifacts and unnatural expressions, while in image detection, it analyses spatial patterns and pixel-level anomalies without multimodal verification. Since deepfake technology often involves synchronized modifications of both visual and auditory components, the absence of an audio analysis module limits the model's ability to detect comprehensive manipulations. To enhance detection accuracy, future work aims to incorporate audio deepfake detection by integrating speech recognition, voice synthesis analysis, and lip-sync verification, ensuring a more robust approach to identifying manipulated media.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403146|

REFERENCES

[1] T. Korshunov and S. Marcel, "DeepFakes: A New Threat to Face Recognition? Assessment and Detection," in *arXiv preprint arXiv:1812.08685*, 2018.
[2] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing DeepFakes Using Inconsistent Head Poses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
[3] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018.
[4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 506-520, 2021.
[5] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.
[6] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.
[7] X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019, pp. 1006–1015.
[8] T. Qi, L. Zhen, S. Wang, X. Xu, and J. See, "DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms," in *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 1317–1326.
[9] M. Frank, A. Roesch, R. R. Schindler, and H. Seidler, "DeepFake Detection Using Frequency Domain Analysis," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.
[10] J. Dang, S. Liu, and Z. Sun, "On the Detection of Digital Face Manipulation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019, pp. 2702–2711.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details