iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** IV **Month of publication:** April 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.61211

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue IV Apr 2024- Available at www.ijraset.com*

# Deepfake Detection System

Pragati Patil[1], Vaishali Shirsath[2], Harsh Churi[3], Shravani Gavali[4], Jayesh Khandare[5]
[1, 2]*Professor,* [3, 4, 2]*Students, Department of Information Technology, Vidyavardhini's College of Engineering and Technology Vasai, India*

***Abstract:*** ***The emergence of deepfake technology presents a profound challenge to the integrity and trustworthiness of multimedia content online. To address this issue, this study proposes a novel deepfake detection system that integrates a hybrid Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) architecture. The proposed system employs a multi-faceted approach to training by utilizing several diverse datasets encompassing real and synthetic videos. This comprehensive training strategy ensures that the model learns robust features representative of both authentic and manipulated content across various contexts and scenarios. By incorporating a range of datasets, including those specifically curated to represent different deepfake generation techniques and quality levels, the model gains a more comprehensive understanding of the intricate nuances present in manipulated videos. During the training phase, the CNN component extracts high-level spatial features from individual frames of the input videos, effectively capturing visual patterns indicative of deepfake manipulation. Subsequently, the LSTM network models the temporal dynamics inherent in video sequences, enabling the detection of subtle inconsistencies over time. Additionally, the RNN component facilitates the capture of contextual dependencies across sequential frames, further enhancing the model's discriminative capabilities. Extensive experimentation is conducted to evaluate the proposed approach, encompassing benchmark datasets as well as additional datasets curated to represent a wide range of deepfake manipulation scenarios. The results demonstrate the superior performance of the hybrid RNN-LSTM architecture compared to state-of-the-art deepfake detection techniques, particularly in terms of accuracy, robustness, and generalization across diverse datasets. In conclusion, the proposed deepfake detection framework offers a powerful and reliable solution to mitigate the proliferation of fake multimedia content online. By leveraging multiple datasets during training, the model achieves heightened sensitivity to the subtle cues indicative of deepfake manipulation, thereby safeguarding the integrity and credibility of digital media platforms.***

***Keywords: Computer vision, Res-Next Convolution Neural Network, Long short-term memory (LSTM), Face Recognition, GAN (Generative Adversarial Network), PyTorch***

## I. INTRODUCTION

Deepfake detection has emerged as a critical imperative in the digital age, as the proliferation of synthetic media, particularly deep fakes, poses unprecedented challenges to the veracity and integrity of online content. These modifications, involving complex techniques like face-switching and voice replication, blur the boundaries between reality and fabrication, making it increasingly challenging for individuals to distinguish authentic from altered content. As a result, the stakes have never been higher for maintaining trust, combating misinformation, and safeguarding individuals and institutions from potential harm. The term "deepfake" itself encapsulates the profound impact of deep learning techniques, notably generative adversarial networks and variational autoencoders , in synthesizing hyper-realistic images and videos. Coined in 2017 following a notorious incident involving the manipulation of celebrity faces in pornographic material, deepfake technology has since undergone rapid evolution and refinement. This evolution owes much to the availability of expansive face datasets like CelebA and FFHQ, coupled with advancements in deep learning frameworks such as TensorFlow and PyTorch. While deepfake technology has found benign applications across various domains, including entertainment, education, art, and journalism, its darker potentialities cannot be overlooked. Malicious actors exploit deepfake technology to propagate misinformation, sow discord, and perpetrate acts of defamation, blackmail, and cybercrime. The ability to impersonate public figures, manipulate public opinion, and fabricate evidence imparts the significant need for enhanced detection and mitigation techniques. In this study, we undertake an extensive examination of the latest techniques for identifying deepfake content, covering both conventional machine learning models and modern deep learning approaches. We interrogate the challenges and limitations inherent in current methodologies, including the scarcity of large-scale and diverse datasets, the delicate balance between accuracy and efficiency, and the vulnerability to adversarial attacks from deepfake generators. Moreover, we chart a course for future research endeavours, highlighting avenues for enhancing the resilience and efficacy in deepfake detecting frameworks in an ever-developing perspective of synthetic media manipulation. Through this holistic exploration, we endeavour to equip stakeholders with the insights and tools necessary to confront the formidable challenges posed by deepfake technology and safeguard the integrity of online information ecosystems.

## II. RELATED WORK

1) The emergence of Deepfakes has democratized the creation of convincing face-swapped videos, necessitating automated detection methods due to their potential for misuse. The DeepFake Detection Challenge (DFDC) Dataset, comprising footage from willing participants, addresses the need for scalable Deepfake detection training data. A benchmark competition using this dataset facilitates unbiased evaluation of Deepfake detection models, encouraging advancements in the field. The competition incentives experts to dedicate resources to model training, providing insights into the current state of Deepfake detection.
2) Fake news dissemination, facilitated by technologies like DeepFake, poses global concerns, with notable use cases in cinema and internet memes. Major tech companies and research initiatives, like the DeepFake Detection Challenge, aim to combat this issue, but despite advancements in detection methods, face manipulation techniques continue to evolve, necessitating updated approaches like leveraging datasets such as Celeb-DF and convolutional neural network models for detection and classification.
3) This study addresses the pressing need for reliable deepfake detection methods by employing cost-sensitive deep learning techniques. Four pre-trained CNN models are utilized, achieving high accuracy's, notably 98% with XceptionNet on CelebDf-V2 dataset. Key frame extraction optimizes video processing efficiency, while a cost-sensitive neural network approach addresses dataset imbalance, showcasing the adaptability and effectiveness of the proposed methodology.
4) Deepfakes, originating from deep learning, involve face swapping, lip-sync, and puppet-master techniques, posing security threats including political manipulation and privacy violations. While deepfakes offer creative potentials in entertainment and visual effects, they also enable malicious activities such as scams and non-consensual pornography. Detection methods, primarily based on deep learning, are actively researched, spurred by initiatives like DARPA's Media Forensics and Facebook's Deepfake Detection Challenge. Surveys on deepfake creation and detection strategies highlight reenactment, replacement, and detection approaches, underlining the escalating research trend in deepfake technology.
5) Deepfakes, comprising over 15,000 media online, pose threats to democracy and privacy, with most targeting politicians and celebrities for defamation. Countermeasures include ML and DL techniques like CNN and RNN, but most focus on spatial features, lacking temporal analysis. A hybrid deep learning approach incorporating temporal feature analysis, particularly optical flow, is investigated to accurately detect deepfakes, evaluated on various performance metrics.
6) A new deep convolutional Transformer model is proposed for deepfake detection, integrating convolutional pooling and re-attention to analyse local and global image features. Keyframes, retaining high-resolution frame information, are extracted to mitigate compression-induced information loss. Extensive experiments demonstrate superior performance over baselines in both within- and cross-dataset evaluations, highlighting the importance of global features and keyframe utilization for robust detection.
7) In order to edit photographs or videos to represent events or scenarios that did not occur, deep learning algorithms are used to create synthetic media known as "deepfakes," as this study explains. It examines the evolution and effects of deepfake technology in relation to a larger social and cultural framework using the Social Construction of Technology (SCOT) framework. The study examines the moral and societal ramifications of deepfake technology, including its propensity to propagate false information, erode public confidence in the media, and violate people's right to privacy.It looks at the legislative attempts to address the exploitation of deepfakes and safeguard people's rights and interests, as well as the regulatory and policy reactions to the widespread usage of deepfake technology.
8) The article discusses the rising alarm regarding deepfake video growth, which seriously jeopardizes digital media trust, security, and privacy. The ID-Reveal method, which uses identity-aware deep learning algorithms to detect deepfake films while protecting the participants' identities, is presented in this study. It talks about using identity-aware features—like facial landmarks and signature expressions—to discriminate between real and fake content according to the identities of the people shown. The ID-Reveal model's architecture, which includes identity-aware feature extraction modules and classification layers for deepfake detection, is described in the study.

## III.PROBLEM STATEMENT

The rapid evolution and widespread availability of deepfake technology have precipitated a pressing need for robust and effective detection mechanisms. Synthetic media manipulation, exemplified by deep fakes, poses profound challenges to the veracity and trustworthiness of online content. As manipulated videos, audio, and images become increasingly sophisticated, individuals and organizations face heightened risks of misinformation, defamation, and other forms of exploitation. Existing detection methods, while promising, often strive to keep tempo with the ever-developing nature of deepfake generation techniques and the scale of the problem.

Moreover, the scarcity of large-scale and diverse datasets hampers the development of resilient detection models capable of generalizing to unseen manipulation scenarios. Addressing these challenges requires innovative approaches that harness the power of both traditional machine learning and modern deep learning techniques. By tackling these pressing issues head-on, researchers can contribute to safeguarding the integrity and credibility of digital media platforms and protecting individuals and institutions from the harmful consequences of deepfake manipulation. Furthermore, the proliferation of deepfake technology has outpaced the development of effective countermeasures, exacerbating the urgency of the problem. Malicious actors exploit deepfake technology for a myriad of nefarious purposes, including political manipulation, financial fraud, and social engineering attacks. The potential ramifications extend beyond individual harm to societal destabilization and erosion of trust in online information sources. In the absence of robust detection mechanisms, the dissemination of fabricated content threatens to undermine democratic processes, exacerbate social divisions, and erode public confidence in media integrity. Thus, the imperative to develop and deploy advanced deepfake detection solutions has never been more pressing. By addressing the multifaceted challenges posed by deepfake manipulation, researchers can play a pivotal role in safeguarding the digital landscape and preserving the integrity of online discourse for future generations.

## IV. PROPOSED SYSTEM

### A. *Proposed Method*

The methodology for developing a DeepFake detection system utilizing a combination of Long Short-Term Memory, Recurrent Neural Networks (LSTM RNN) and Convolutional Neural Networks (CNN) involves a two-step process. Firstly, the CNN component is utilized to extract relevant features and patterns from video frames, enabling a deep understanding of spatial relationships within the frames. The CNN processes each frame independently, generating a rich representation of visual features. Afterward, the LSTM RNN is utilized to scrutinize the temporal connections and successive trends detected within the extracted features. This LSTM RNN captures the temporal progression of features throughout frames, taking into account the surrounding context of previous and subsequent frames. By combining the strengths of CNN for spatial analysis and LSTM RNN for temporal analysis, this methodology enhances the model's ability to discern authentic and manipulated sequences, effectively detecting DeepFake videos. The model learns to recognize patterns in both spatial and temporal dimensions, significantly improving accuracy and robustness in DeepFake detection.

### B. *Collecting Relevant Data*

A crucial first step in creating a deep fake detection system is data collection. To provide personalized suggestions, the algorithm needs information about users and books. Here are some methods followed for gathering pertinent data. The composition of the dataset was altered by preprocessing the DFDC dataset, which include audio alerted videos (audio deep fake is outside the purview of this piece). From the revised DFDC dataset, 1500 actual and 1500 false videos were then chosen. Furthermore, the FaceForensic++ (FF) dataset provided 1000 actual and 1000 fraudulent videos. Four cutting-edge techniques, Face2Face, FaceSwap, DeepFakes, and NeuralTextures, were used to make the modifications in the dataset, and 500 actual and 500 fake clips were taken from the Celeb-DF dataset. A thorough dataset with 3000 real videos and 3000 fraudulent videos totaling 6000 videos overall was produced by this selection process.

TABLE I
DATASET SOURCES

| Name | No. of Videos |
|---|---|
| FaceForensic++ [9] | 2000 |
| DFDC Dataset [1] | 3000 |
| Celeb Df [10] | 1000 |

### C. *Data Pre-processing*

Videos are pre-processed in this first stage so as to remove extraneous noise and concentrate only on identifying and trimming the relevant face region. First, every video is divided into its component frames. Next, each frame undergoes face identification, and the recognised face is clipped. After that, each processed input is used to create a new video by compiling the chopped frames. Frames without facial features are ignored in this technique.The first step in the video preprocessing process is to divide the video into frames. Face detection is applied to each frame, and the detected face is clipped out of the frame.

Then, a new video is created using the cropped frames that result, highlighting only the features on the face. Every video in the dataset goes through this sequential process once more, producing a collection of processed videos that only have the face content. The average frame count of every video is used to establish a threshold value, which guarantees consistency in the amount of frames in every video.

This decision is influenced by the need to preserve consistency as well as computational limitations, particularly in situations where the Graphics Processing Unit (GPU) has limited power. In order to ensure computational feasibility, 150 frames was selected as the threshold value.

Only the first 150 frames of each movie are stored as frames in the new dataset. Sequential frame order is maintained when demonstrating the use of Long Short-Term Memory (LSTM), focusing on the first 150 frames instead than a random selection. The resulting films are stored at 30 frames per second (fps) and 112 × 112 resolution, which corresponds to the experimental GPU's computing capability.

Figure 1 : Pre-Processed Dataset

*D. Dataset Split*

The training dataset, which comprises 4,080 movies, accounts for 70% of the total, while the testing dataset, consisting of 1,920 videos, comprises 32% of the videos. A balanced distribution is maintained by both the training and testing sets, with each segment including 50% real and 50% fake videos.

*E. Model Training*

In combination an existing ResNext CNN framework for frame-level feature extraction is used to mix a convolutional neural network (CNN) and a recurrent neural network (RNN). In order to differentiate between real and false films, an LSTM network's training phase uses these extracted properties. Labels associated with videos are imported into the framework and integrated for training by using the Data Loader on the training segment of the videos.

In particular, feature extraction makes use of the resnext5032x4d model, which is optimised for improved performance on deeper neural architectures. The current ResNext concept is used rather than writing the code from scratch. The process of fine-tuning include choosing the best learning rate and incorporating extra layers that are required to guarantee that the gradient descent of the model converges effectively. The sequential LSTM is fed with 2048-dimensional feature vectors that are obtained after ResNext's final pooling layers.

A single LSTM layer with 2048 latent dimensions, 2048 hidden layers, and a dropout probability of 0.4 receives these feature vectors in 2048 dimensions. By comparing frames at 't' seconds with those at 't-n' seconds—where 'n' denotes any specific number of frames before to 't'—this LSTM algorithm allows for the temporal evaluation of videos. The framework comes with an activation function for Leaky ReLU.

The average correlation rate between input and output is taught to the model through a linear layer consisting of 2048 input features and 2 output features. An image size of H x W is guaranteed by an adaptive average pooling layer with a single output parameter. A consecutive Layer is used to process frames in a consecutive manner. Batch training is made easier with a batch size of 4, and the Softmax layer offers information on the model's confidence throughout the prediction stage.

*F. Flowchart*

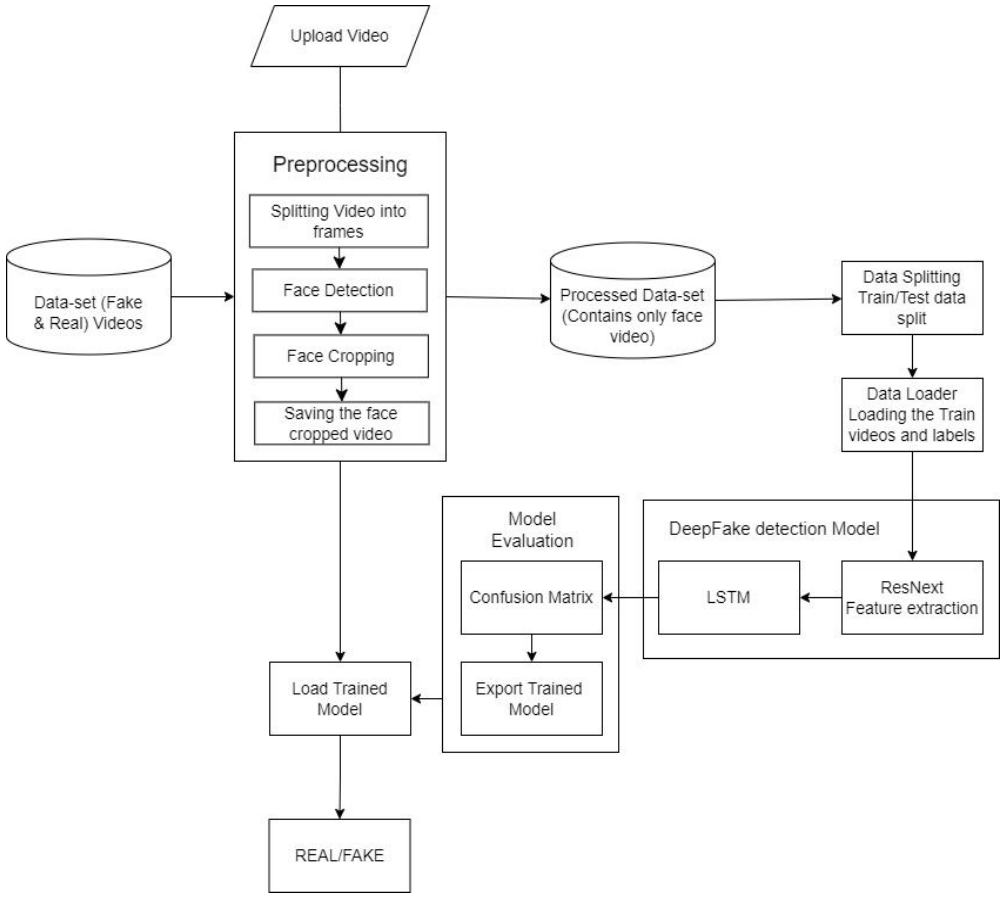



Figure 2 : Flowchart of Deepfake Detection

## V. RESULTS

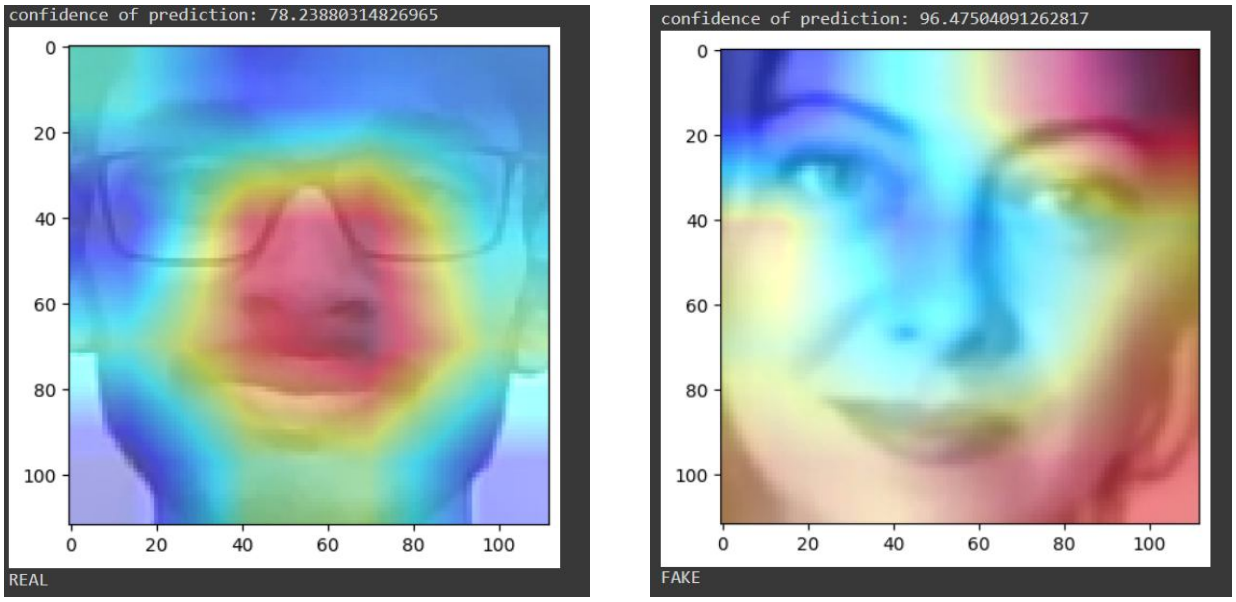



Figure 3 : Working in Google Colab

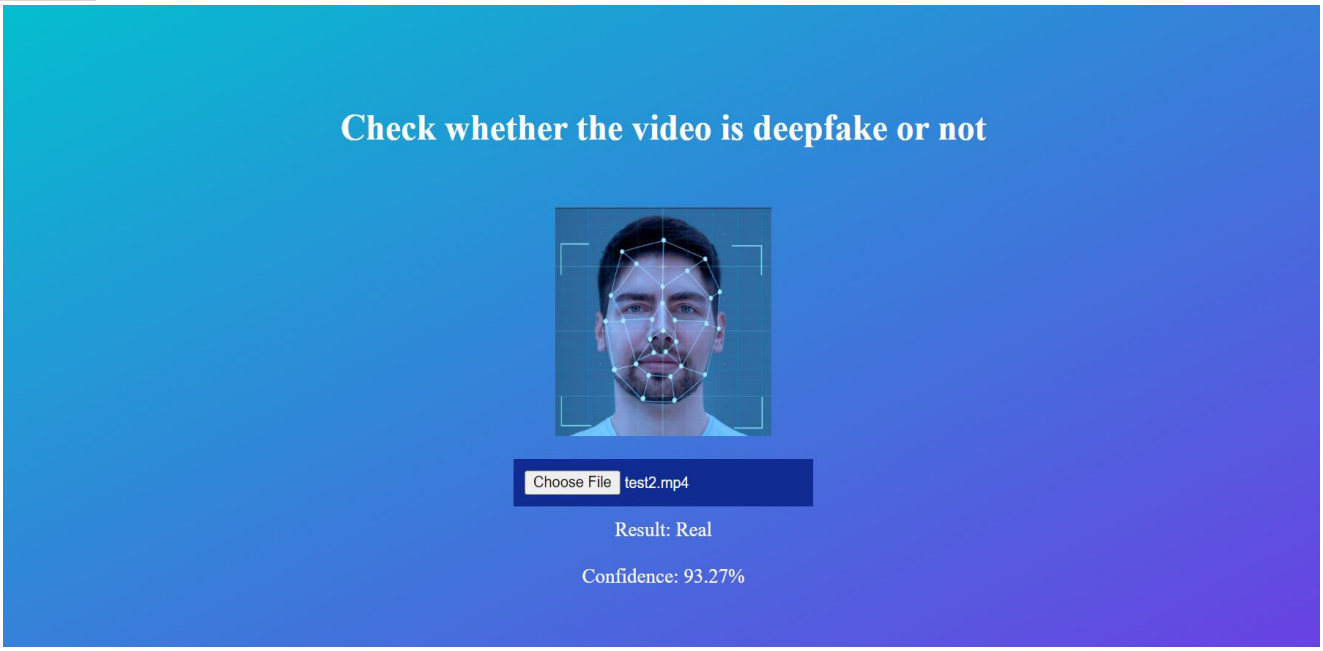



Figure 4 : UI for Deepfake Detection

Deepfake detection outcomes that utilised RNN, CNN, and LSTM models exhibit notable efficacy in accurately discerning manipulated media. RNNs excel in capturing temporal dependencies, making them adept at analyzing sequential data such as video frames for subtle manipulation patterns. CNNs, renowned for their spatial feature extraction capabilities, effectively detect anomalies in pixel-level details, enabling robust identification of forged content based on visual cues. LSTMs, specialized in retaining long-range dependencies, enhance contextual understanding across multiple time steps, further bolstering the model's ability to differentiate authentic behavior from manipulated sequences. With assessment criteria like accuracy, precision, recall, and F1-score acting as benchmarks for assessing detection efficacy, leveraging these architectures singly or in combination provides competitive performance in deepfake detection. Achieving the best detection results requires fine-tuning hyperparameters and optimising model topologies.

## VI. FUTURE SCOPE

The future of deepfake detection holds significant promise with advancements in AI and machine learning techniques alongside increased collaboration across sectors. Research into more sophisticated models, including transformer-based architectures, could lead to even more accurate detection systems by leveraging contextual understanding and generative capabilities. Integrating multiple modalities such as audio, text, and metadata analysis alongside visual cues could enhance detection resilience. Improving the interpretability of models will be crucial for building trust among users. Real-time detection systems capable of rapidly flagging manipulated content will be essential for timely intervention. Generating diverse datasets and addressing ethical concerns will further improve detection capabilities and ensure responsible deployment. Regulatory frameworks and industry standards will play a vital role in deterring malicious use and fostering a safer online environment. Overall, the future of deepfake detection lies in advancing technology, fostering transparency, and promoting collaboration to productively combat new challenges of deepfake manipulation.

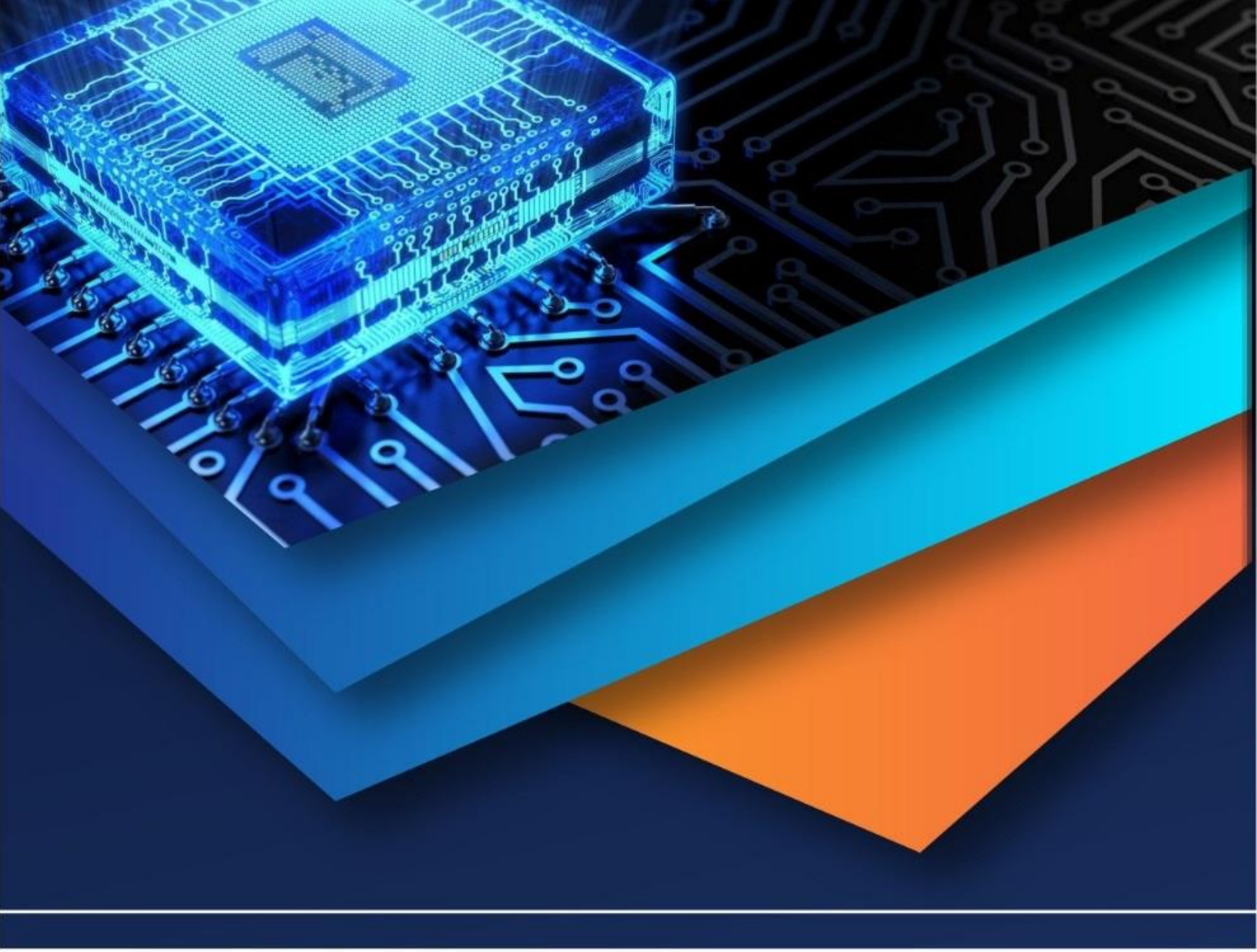
## VII. CONCLUSIONS

In summary, the development of deepfake technology poses a serious threat to the reliability and integrity of digital media ecosystems. The proliferation of synthetic media manipulation, exemplified by deep fakes, threatens to undermine public discourse, erode trust in online information sources, and facilitate the spread of misinformation and disinformation. Despite these formidable obstacles, the scientific community has achieved notable progress in crafting detection methods that utilize both conventional machine learning and advanced deep learning techniques. However, ongoing efforts are required to address the persistent challenges of generalization to unseen manipulation techniques, resilience to adversarial attacks, and scalability to the vast and diverse landscape of online content.

By fostering interdisciplinary collaborations, investing in large-scale datasets, and advancing innovative detection methodologies, researchers can mitigate the risks posed by deepfake manipulation and safeguard the integrity of digital media platforms. Ultimately, the endeavor to counteract deepfake technology is not solely a technical hurdle but a moral obligation—one that requires united efforts to uphold the values of honesty, openness, and reliance in the digital era. Besides technological progress, cooperation among academia, industry, and policymakers is essential to address the various challenges presented by deepfake manipulation in a thorough manner. By fostering dialogue and sharing resources, stakeholders can develop holistic strategies to combat the spread of synthetic media manipulation effectively. Moreover, public awareness campaigns and digital literacy initiatives play a pivotal role in empowering individuals to critically evaluate online content and discern between authentic and manipulated media. As we traverse the intricate terrain of digital data, it is crucial that we stay watchful, flexible, and dedicated to preserving the credibility of our online conversations. Together, through collective action and unwavering dedication, we can confront the threats posed by deepfake technology and preserve the trustworthiness of our digital world for generations to come.

## REFERENCES

[1] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.
[2] Gustavo Cunha Lacerda and Raimundo Claudio da Silva Vasconcelos. A machine learning approach for deepfake detection. 2022
[3] Faysal Mahmud, Yusha Abdullah, Minhajul Islam, and Tahsin Aziz. Unmasking deepfake faces from videos using an explainable cost-sensitive deep learning approach. 2023.
[4] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223:103525, October 2022.
[5] Pallabi Saikia, Dhwani Dholaria, Priyanka Yadav, Vaidehi Patel, and Mohendra Roy. A hybrid cnn-lstm model for video deepfake detection by leveraging optical flow features. 2022.
[6] Tianyi Wang, Harry Cheng, Kam Pui Chow, and Liqiang Nie. Deep convolutional pooling transformer for deepfake detection. ACM Transactions on Multimedia Computing, Communications, and Applications, 19(6):1–20, May 2023.
[7] A. O. Kwok and S. G. Koh, "Deepfake: a social construction of technology perspective," Current Issues in Tourism, vol. 24, no. 13, pp. 1798–1802, 2021.
[8] D. Cozzolino, A. R¨ossler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15 108–15 117.
[9] Rossler, Andreas & Cozzolino, Davide & Verdoliva, Luisa & Riess, Christian & Thies, Justus & Niessner, Matthias. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. 1-11. 10.1109/ICCV.2019.00009.
[10] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY