Deepfake Detection Using Deep Learning: A Neural Network-Based Approach to Identifying AI- Generated Media

Krishna Shrimali , Chirag Yadav
Department of Artificial Intelligence and Machine Learning, Universal college of Engineering Mumbai, India

Krishna.shrimali@universal.edu.in Chirag.yadav@universal.edu.in

***Abstract*— Over the last few decades, rapid progress in AI, machine learning, and deep learning has resulted in new techniques and various tools for manipulating multimedia. Though the technology has been mostly used in legitimate applications such as for entertainment and education, etc., malicious users have also exploited them for unlawful or nefarious purposes. For example, high-quality and realistic fake videos, images, or audios have been created to spread misinformation and propaganda, foment political discord and hate, or even harass and blackmail people. The manipulated, high-quality and realistic videos have become known recently as Deepfake. Various approaches have since been described in the literature to deal with the problems raised by Deepfake. Given the ease with which deep fake videos/images may be generated and shared, the lack of an effective deep fake detection system creates a serious problem for the world. However, there have been various attempts to address this issue, and deep learning-related solutions outperform traditional approaches.[5-6]**

***Keywords*— Deepfake detection, video or image manipulation, deep learning , fake detection.**

I. INTRODUCTION

In a narrow definition, deepfakes (stemming from “deep learning” and “fake”) are created by techniques that can superimpose face images of a target person onto a video of a source person to make a video of the target person doing or saying things the source person does.[1] One of the major global concerns of modern society regards the development and rapid dissemination of fake information through fast-content consumption platforms, such as TikTok, Twitter, Facebook, and Instagram . Such content may vary from text-based messages to, most recently, image and video automatic manipulation using a family of machine learning (ML)-based approaches called deep learning.[3].

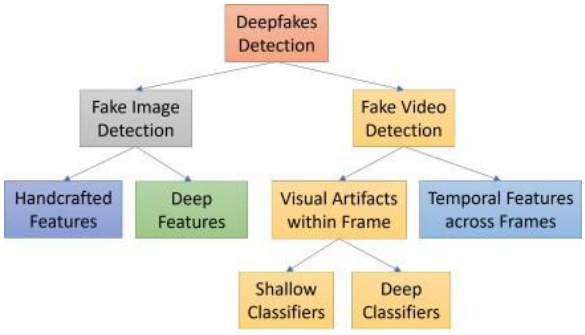

Figure 1: Categories of reviewed papers relevant to deepfake detection methods where we divide papers into two major groups, i.e., fake image detection and face video detection. [1]

The proliferation of manipulated media, particularly deepfakes, poses a significant threat to information integrity and societal trust. Deepfakes, synthetic videos or images convincingly altered to depict someone saying or doing something they never actually did, leverage powerful deep learning techniques to achieve an unprecedented level of realism. This ease of creation, coupled with the potential for widespread dissemination through social media, makes deepfakes a potent tool for misinformation campaigns, political manipulation, and even identity theft. Consequently, the development of robust and reliable deepfake detection methods is crucial. This research explores the application of deep learning models, trained on a comprehensive dataset of both real and manipulated media, to effectively distinguish authentic content from deepfakes and mitigate their harmful impact. Many software apps/tools are available through which deep fake images are created without a programming knowledge and technical side background information. Usually the profile pictures from the social media are taken and fake images or videos are developed with a help of the expert. Security enhancement in the detection of face swap and the accuracy are very low.[7]

TABLE I Overview of deepfake detection method

SL.No	Authors	Methodology	Techniques	Key Features	Databases Used
1	Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. [20]	Eye blinking	Long ter m recurrent CNN	Use LRCN to understand the temporal patterns of eye blinking.	Consist of 49 interview & presentation videos, & their correspondin g generated deepfakes.
2	Afchar, Darius et al. [18]	MesoNet	CNN	Two dee p networks Mes0-4 & Mesoinception-4 are introduced to examine deep false videos at the mesoscopic level of analysis.	Two databases: Deep fake one constituted from online videos & Face Forensics one created by face2face approach
3	Sabir, Ekraam et al. [19]	Spatio-temporal features wit h RCN	RCN	RCN, whic h combines the convolutional network DenseNet with the gated recurrent uni t cells, is used	FaceForensics++ dataset, including 1000 videos

				to investigate temporal differences between frames.	
4	Chintha, Akash et al. [9]	Spatio-temporal features wit h LSTM	convolutional bidirectional recurrent LSTM network	For face feature extraction, an XceptionNet CNN is employed, an d audio embeddings are generated by stacking numerous	FaceForensics++, Celeb- DF & ASVSpoof 2019 logical Access audio dataset
				convolution modules.	
5	Fernandes, Steven et al. [24]	Using attribution based confidence (ABC) metric	ResNet50 model pretrained on VGGFace2	Without access to training data, deep fake videos are detected using the ABC measure.	VidTIMIT, COHFACE

II. LITERATURE REVIEW

Deepfake detection has been an area of extensive research, with various deep learning-based techniques proposed to counter the increasing sophistication of synthetic media. Afchar et al. (2018) introduced **MesoNet**, a lightweight CNN-based architecture designed specifically for detecting deepfake videos[10]. Their approach focuses on capturing low-level image artifacts that are often present in manipulated content. Similarly, Nguyen et al. (2022) conducted a comprehensive survey on deepfake generation and detection techniques, highlighting the strengths and weaknesses of various deep learning approaches, including CNNs, autoencoders, and GAN-based methods.[11]

Another important work by Heidari et al. (2023) presents a systematic and comprehensive review of deepfake detection methods, emphasizing the challenges of dataset biases and adversarial robustness. The study discusses the effectiveness of transformer- based architectures like **Vision Transformers (ViTs)** and their potential in improving classification performance.[9] Rana et al. (2022) provided a systematic literature review on deepfake detection, emphasizing the need for adversarial training and real-time detection techniques to counter evolving threats.[6]

More recent studies, such as Suratkar & Kazi (2023), explored the application of **transfer learning** in deepfake detection, leveraging pre-trained architectures like **XceptionNet and EfficientNet** to enhance detection accuracy. Passos et al.[8] (2024) reviewed deep learning-based approaches for deepfake content detection and highlighted the importance of **multimodal detection techniques**, integrating audio-visual and physiological cues such as blinking and lip movement inconsistencies.[3]

Despite these advancements, challenges remain, including the need for more **generalizable models** that can detect deepfakes across different datasets and manipulation techniques. The integration of **explainable AI (XAI) methods** and federated learning approaches is a promising research direction for improving transparency and privacy-preserving deepfake detection (Heidari et al., 2023).[9]

TABLE II Literature Survey

Paper Title	Year	Advantages	Disadvantages
Deepfake Detection: A systematic Literature.	2022	A systematic literature review provides a thorough overview of existing research on deepfake detection, summarizing and synthesizing findings from multiple studies.	The quality of the studies reviewed can vary significantly. Some studies might have methodological weaknesses, small sample sizes.
How deep learning fake video and how to detect it	2021	The paper would provide a detailed exploration of deep learning techniques used to create fake videos, including generative adversarial networks (GANs) and other neural network architectures.	Deep learning models for both creating and detecting fake videos can be highly complex and computationally intensive.

Fighting deepfake by exposing the convolutional traces on images.	2020	Using convolutional traces as a detection mechanism could be a novel and promising approach.	The convolutional traces left by deepfake generation might be subtle and vary across different generative models.
Unmasking deepfake with simple features.	2019	Using simple features can make the detection method more efficient and less computationally demanding.	Simple features may not capture the subtle and sophisticated artifacts introduced by advanced deepfake generation techniques
Two stream neural network for tempered face detection.	2017	A two-stream neural network can process information from two different perspectives or modalities simultaneously	Two-stream neural networks are inherently more complex than single-stream models

A. Proposed System

The proposed deepfake detection system utilizes a deep learning-based approach to effectively classify real and manipulated images/videos. The system is designed to handle various deepfake generation techniques by leveraging convolutional neural networks (CNNs) and transformer-based architectures such as Vision Transformers (ViTs) and EfficientNet for feature extraction (Heidari et al., 2023).[9] By analyzing spatial and temporal inconsistencies in deepfake videos, the model improves classification accuracy and robustness against adversarial attacks (Afchar et al., 2018).[10]

The proposed framework consists of several key modules: (1) **Preprocessing**, where input images or frames are resized, normalized, and enhanced using contrast adjustment techniques; (2) **Feature Extraction**, where deep learning models such as XceptionNet and ResNet extract meaningful features from facial regions (Suratkar & Kazi, 2023); (3) **Classification**, where the extracted features are passed through a deep neural network (DNN) or Long Short-Term Memory (LSTM) network for final classification into real or fake categories (Nguyen et al., 2022).[11]

Additionally, the system incorporates explainable AI (XAI) techniques to provide interpretability and enhance trust in deepfake detection results (Rana et al., 2022).[6] A real-time detection component is integrated using optimized models deployed on edge devices, ensuring efficient processing and usability in digital forensics and social media moderation (Passos et al., 2024).[3] Future improvements may include integrating multimodal detection, leveraging physiological cues such as blinking patterns and heartbeat signals to enhance detection accuracy (Heidari et al., 2023).[9]

III. METHODOLOGY

The methodology for deepfake detection using deep learning consists of several key phases: data acquisition, preprocessing, model selection and training, and evaluation. The dataset is obtained from publicly available sources such as Kaggle, comprising both real and deepfake images or videos. Preprocessing techniques, including frame extraction (for video-based detection), resizing, normalization, and data augmentation, are applied to improve model robustness and prevent overfitting (Rossler et al., 2019). For feature extraction, deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or hybrid models (e.g., CNN-LSTMs) are employed to capture both spatial and temporal inconsistencies in manipulated media. Transfer learning with pre-trained networks like Xception, EfficientNet, and ResNet-50 is used to enhance model performance (Chollet, 2017). The model is trained using labeled datasets with loss functions such as binary cross-entropy and optimization algorithms like Adam or SGD. Evaluation metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are employed to measure the model's effectiveness in classifying deepfake content (Afchar et al., 2018). The final model is then deployed as a web-based or standalone application for real-time deepfake detection. [12-13]

A. System Architecture

The system architecture for deepfake detection using deep learning comprises multiple stages, including data preprocessing, feature extraction, model training, and classification. Initially, real and fake images/videos are collected from publicly available datasets such as Kaggle. These data samples undergo preprocessing techniques, including resizing, normalization, and augmentation, to enhance model generalization. Next, deep learning models, such as Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), extract spatial and temporal features crucial for distinguishing authentic and manipulated media. The extracted features are then passed through a classification layer, typically a fully connected neural network with a softmax or sigmoid activation function, to categorize the input as real or fake. To improve accuracy, transfer learning techniques with pre-trained models (e.g., Xception, EfficientNet) are often employed, as they have demonstrated superior performance in image forensics tasks (Nguyen et al., 2019). The final classification results are evaluated using performance metrics such as accuracy, precision, recall, and F1-score to assess model effectiveness (Afchar et al., 2018). [10-11]

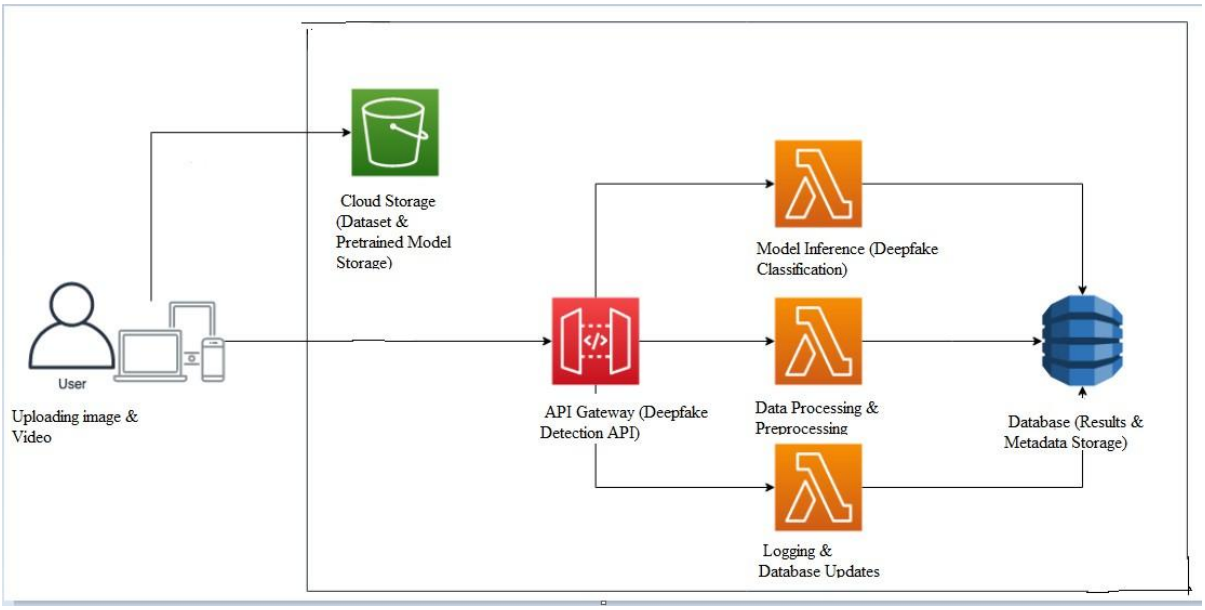

Figure 1 : System Architecture of deepfake detection

I. RESULTS & DISCUSSION

In this research, we have developed a deepfake detection system utilizing machine learning techniques with a user-

friendly interface. The project is built using Streamlit, a Python library that enables an interactive web application for users to upload and analyze images and videos. To maintain a clean and reproducible development environment, we employed a virtual environment. The model selected for deepfake classification is EfficientNetB5, a state-of-the-art convolutional neural network known for its efficiency and accuracy. The dataset used for training the model was sourced from Kaggle, consisting of both real and fake images and videos. These datasets were preprocessed using OpenCV and MediaPipe to detect faces, ensuring that only facial features were passed through the model for classification.

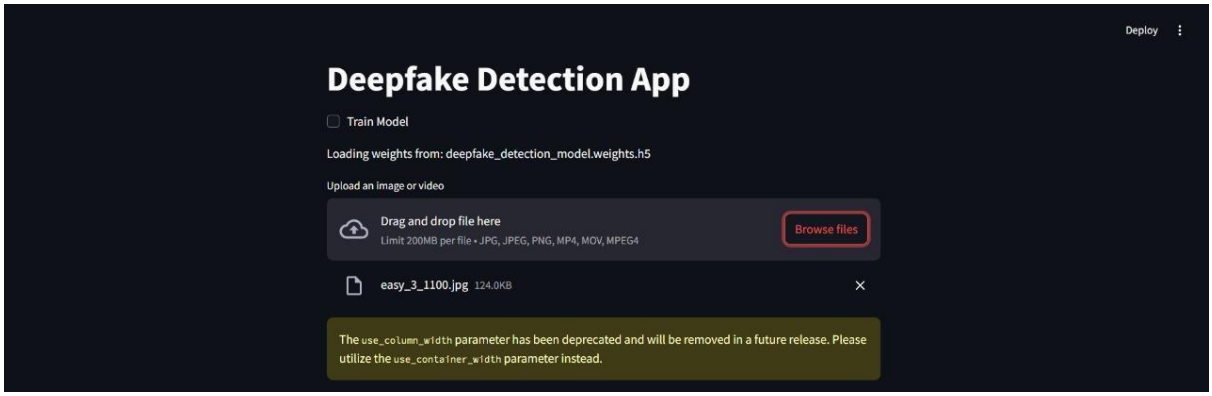

Figure 2 : User interface of the project

The model training process involved splitting the dataset into training and validation sets, with data augmentation techniques applied to improve generalization. We incorporated L2 regularization, batch normalization, and dropout layers to enhance the model's robustness. The Adam optimizer with a reduced learning rate and gradient clipping was used for better convergence. Additionally, class weighting was applied to balance the dataset, and early stopping with a learning rate reduction strategy was implemented to prevent overfitting. The trained model was integrated into the Streamlit interface, allowing users to upload images or videos for classification. The predictions are made by detecting facial regions, preprocessing them, and passing them through the EfficientNetB5 model, which outputs a probability score indicating whether the input is real or fake. This system provides an efficient and accessible solution for deepfake detection, which can be further improved with larger datasets and advanced training strategies.

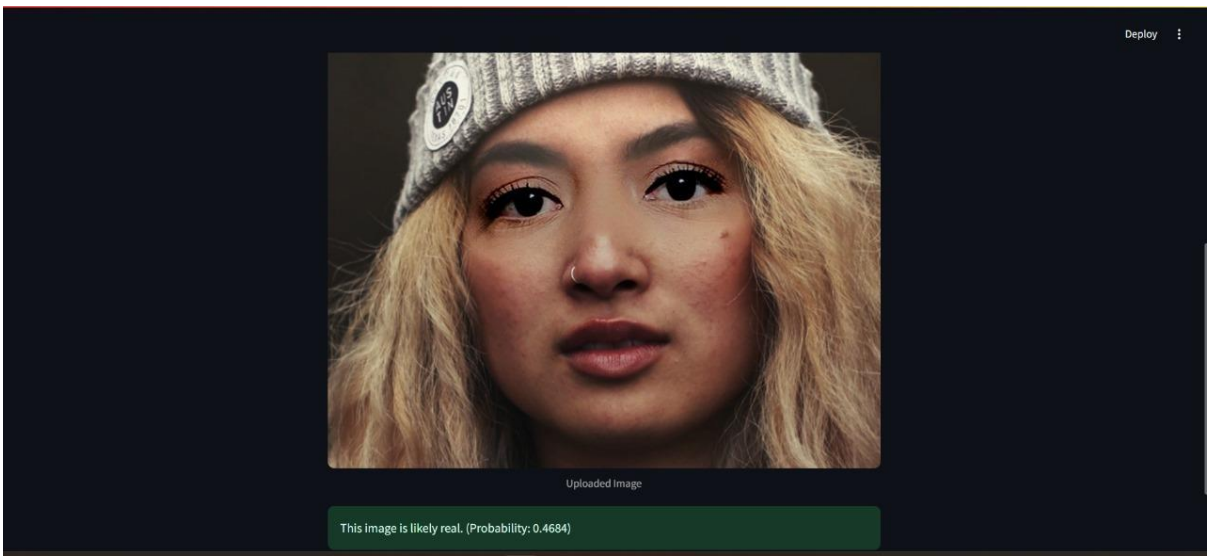

Figure 3 : Real Image Detection

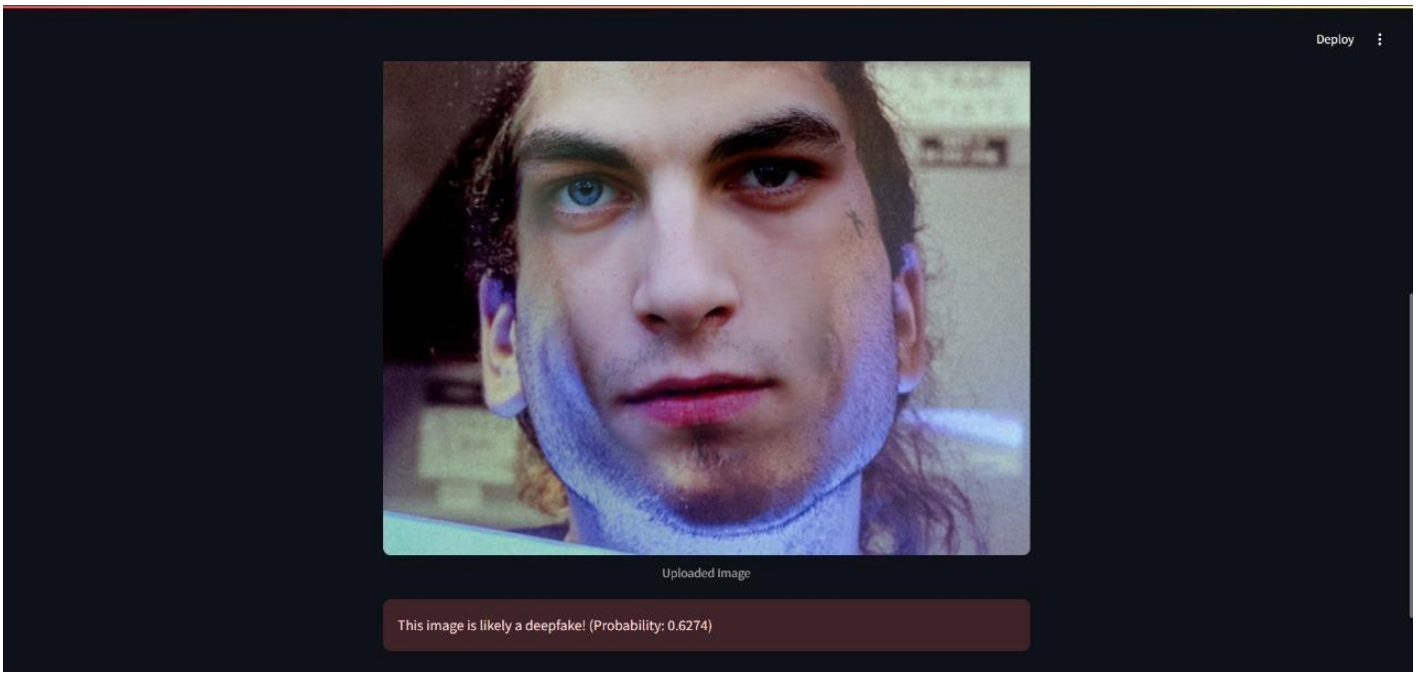

Figure 3 : Fake Image Detection

II. CONCLUSION

Deepfake detection using deep learning has become a crucial area of research due to the increasing sophistication of synthetic media and its potential misuse. This study highlights the effectiveness of various deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid approaches such as CNN-LSTMs, in identifying manipulated images and videos (Afchar et al., 2018)[10]. Transfer learning with pre-trained architectures like Xception and EfficientNet has significantly improved detection accuracy, demonstrating the importance of feature extraction from both spatial and temporal domains (Chollet, 2017)[12]. Despite recent advancements, challenges such as dataset bias, adversarial attacks, and the generalization of detection models remain open research problems (Nguyen et al., 2022) [11]. Future research should focus on developing more robust and explainable AI models, integrating multimodal detection techniques, and enhancing real-time deepfake detection systems to mitigate potential threats (Heidari et al., 2023). [9]

FUTURE SCOPE

As deepfake generation techniques continue to evolve, the need for more robust and adaptive detection methods is critical. Future research should focus on improving the generalization capabilities of deepfake detection models to ensure effectiveness across different datasets and real-world scenarios (Nguyen et al., 2022)[11]. The integration of multimodal detection approaches, including audio-visual analysis and physiological cues, can enhance model reliability by identifying inconsistencies beyond visual artifacts (Heidari et al., 2023).[9] Additionally, explainable AI (XAI) techniques should be explored to provide interpretability in deepfake detection, helping researchers and law enforcement agencies better understand model decisions (Rana et al., 2022).[6]

Furthermore, adversarial robustness remains a key challenge, as deepfake generation models are increasingly using techniques to bypass detection systems. Future studies should investigate adversarial training strategies and continual learning methods to adapt detection models against emerging threats (Afchar et al., 2018)[10]. Real-time deepfake detection in social media and digital forensics applications is another important area for development, requiring efficient deployment of deep learning models on edge devices and cloud-based platforms (Suratkar & Kazi, 2023)[8]. Advancements in federated learning and privacy-preserving AI techniques can also enhance deepfake detection while maintaining user privacy and data security (Passos et al., 2024).[3]

REFERENCES

1) Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *arXiv preprint arXiv:1909.11573v5*. https://arxiv.org/abs/1909.11573

2) Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). DeepFake detection based on discrepancies between faces and their context. *arXiv preprint arXiv:2008.12262v1*. https://arxiv.org/abs/2008.12262

3) Passos, L. A., Jodas, D., Costa, K. A. P., Souza Júnior, L. A., Rodrigues, D., Del Ser, J., Camacho, D., & Papa, J. P. (2024). A review of deep learning-based approaches for deepfake content detection. *arXiv preprint arXiv:2202.06095v3*. https://arxiv.org/abs/2202.06095

4) Abdul Kareem, H. S. (Year). Detection of deepfake in face images using deep learning. *[Journal Name]*, *Volume*(Issue), Page Numbers. https://doi.org/10.31185/wjcm.92

5) Mary, A., & Edison, A. (2023). Deepfake detection using deep learning techniques: A literature review. *2023 International Conference on Control, Communication and Computing (ICCC)*. https://doi.org/10.1109/ICCC57789.2023.10164881

6) Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *Volume*(Issue), Page Numbers. https://doi.org/10.1109/ACCESS.2022.3154404.

7) Suganthi, S. T., Ayoobkhan, M. U. A., Kumar, K. V., Bacanin, N., Venkatachalam, K., Hubálovský, Š., & Trojovský, P. (Year). Deep learning model for deepfake face recognition and detection. *[Journal/Conference Name]*, *Volume*(Issue), Page Numbers. https://doi.org/[DOI]

8) Suratkar, S., & Kazi, F. (2023). Deepfake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, *48*(9), 9727–9737. https://doi.org/10.1007/s13369-022-07321-3

9) Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (Year 2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. *[Journal Name]*, *Volume*(Issue), Page Numbers. https://doi.org/10.1002/widm.1520

10) Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network.
Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1-7.

11) Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2307-2311.

12) Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251-1258.

13) Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1-11.