# Deepfake Detection
# Ai – Manipulated or Human

Samitha S.[1], Dr. Sreejith Vignesh B P[2]
*[1]Junior Researcher, Department of Information Technology,*
*Sri Krishna Adithya College of Arts and Science*
*[2]Associate Professor & Head, Department of Information Technology*
*Sri Krishna Adithya College of Arts and Science*

***Abstract*—The unprecedented development of generative AI makes possible, with unprecedented ease and accuracy, deepfakes that are very difficult to distinguish from human-generated content. This exposes significant risks in domains such as security, politics, journalism, and digital trust. This paper proposes an AI-driven deepfake detection framework that classifies media into manipulated or human-authentic categories with high reliability. This framework combines multimodal analysis that inspects visual artifacts, audio inconsistencies, facial dynamics, and temporal patterns with deep neural network architectures trained on large-scale datasets of real and synthetic media. By embedding feature extraction techniques coupled with anomaly detection models, the proposed approach identifies subtle irregularities created along the generation pipeline of deepfakes. Experimental results demonstrate strong performance across various manipulation types and real-world scenarios, thus underlining robustness and generalization capability. This research contributes to a growing demand for automated, scalable, and trustworthy deepfake detection instruments that will help preserve the integrity of digital content in an age when synthetic media is becoming increasingly practical.**

***Index Terms*—Deep Fake Detection, Synthetic Media, Manipulated Media, AI Generated Media, Facial Forgery Analysis, Multimedia Forensics, Convolutional Neural Networks, Deep Neural Networks, Feature Extraction, Anomaly Detection, Audio Visual Inconsistencies, Temporal Artifact Analysis, Media Integrity, Digital Authenticity, Fake vs Real Classification, Machine Learning, Computer Vision, Generative Adversarial Networks, Anti Spoofing Techniques**

## I. INTRODUCTION

Impacted how digital content is generated, disseminated, and distributed. The most emblematic consequence of this technology is, therefore, deepfake media, which entails highly realistic digital images, videos, and audio generated through deep learning algorithm. The advent of highly advanced generative AI tools has radically s. Even though such technology is extremely innovative, enabling a great deal of creative possibilities, it also brings about a huge set of problems, including not being able to recognize whether a digital image is generated by a human or is a deepfake.

Because of the fast-evolving nature of deepfakes, it is becoming increasingly challenging for current detection techniques to keep up. In view of this, there is an immediate requirement for intelligent and automated solutions that are able to detect minute signs and discrepancies indicative of deepfakes. Deepfake Detection AI is one solution that is able to capitalize on machine learning and computer vision technology, as well as audio-visual analysis, and help distinguish whether content is Deepfaked or authentic and human.

This work examines the design, methodology, and efficacy of an AI-based deepfake detector, its significance in fighting synthetic media dangers, and its emphasis on promoting trust in the face of an AI-driven world gaming systems into a platform of equal empowerment and fun in gaming environments.

## II. PROBLEM MOTIVATION WITH REAL - WORLD STATISTICS

1. Explosive Growth of Deepfake Content

The number of deepfake videos online has been doubling every six months, according to multiple cybersecurity analyses.

By 2023, more than 95% of all deepfakes online were pornographic, and 90% targeted women, creating massive privacy and harassment risks.

2. Financial Fraud Is Rising Fast

In 2024, a multinational company in Hong Kong lost $25 million after scammers used AI-generated deepfake video and voice of the CFO to authorize a fraudulent transfer.

Deepfake-enabled identity fraud increased by 300% between 2020 and 2023, according to global fraud-prevention firms.

3. Political Manipulation and Misinformation

During recent elections worldwide, deepfake political videos spread millions of views within hours, influencing public opinion before fact-checkers could respond.

A study found that 1 in 3 people cannot reliably distinguish a deepfake video from a real one.

4. Threat to National Security

Intelligence agencies warn that deepfake technology can be used for:

- impersonating military officials
- fabricating diplomatic statements
- manipulating public sentiment during crises

5. Psychological and Social Impact

Exposure to deepfakes reduces trust in all media even authentic content.

This leads to the "liar's dividend," where real wrongdoers claim genuine evidence is fake.

### A. Why Chess Matters:

It is essential to note that fake detection is important because sometimes AI-produced video, photo, and audio clips are so convincing that it becomes difficult to identify what is real. It is hard to differentiate AI-manipulated content from human-created content. Deepfakes may lead to spreading misinformation, financial scams, reputation damages, and shaping public opinion. However, detecting deepfakes can prevent the loss of trust and integrity within the digital space.

### B. Real-World Impact:

Real-world deepfake detection is emerging increasingly nowadays, as synthetic media grows more and more sophisticated, pervasive, and ubiquitously undeniable. Manipulated videos and AI-generated content can easily circulate today on social platforms, influencing public perception and eroding trust in information. Deepfake detection systems help counter the effects of manipulated media by establishing a reliable method that distinguishes between media that is genuinely created and that which has been artificially generated. This fortifies digital trust and helps protect one's identity from misuse, impersonation, and harm to one's reputation.

Beyond personal safety, the technology contributes to stability in society. Detection tools are key for governments and security agencies to uncover political statements that have been faked, military footage manipulated, or synthetic propaganda perpetrated to disturb public order. In democratic contexts, deepfake detection supports election integrity by helping journalists, fact-checkers, and citizens distinguish real political communication from manipulated content designed to mislead.

The corporate world also has its share. Deepfake-based fraud, such as a call by somebody mimicking the CEO, can authorize financial transfers. Detection AI helps organizations guard communication channels and defend against high-stakes social engineering attacks. At the same time, law enforcement agencies and digital forensic teams use such tools in order to validate evidence so that tampered media does not compromise an investigation or a judicial process.

Deepfake detection on social media ultimately supports healthier online ecosystems by enabling the early identification of harmful or misleading synthetic content. It fosters content moderation and decreases misinformation. More generally, such technology would encourage public awareness and media literacy in helping people understand the digital content they encounter more critically.

The AI for deepfake detection reinforces trust, security, and accountability at every layer of society; hence, it constitutes an indispensable technology in a world where the differences between real and artificial media are increasingly obscured

## III. LITERATURE REVIEW & REVIEW OF RECENT RELATED STUDIES

Deepfake detection methods have progressed at a fast pace to counter the growing sophistication of synthetic media produced by deep learning models like GAN and autoencoders. In the early stages of deepfake detection research, the methods primarily targeted image analysis for each frame of the image by means of convolutional neural network architectures to differentiate between real and doctored faces based on pixel-level discrepancies and compression artifacts. Xception and Efficient-Net architectures gained widespread recognition due to their excellent performance on standard benchmark sets and their capability to learn subtle texture and boundary discrepancies produced by face swapping and reenactment techniques.

When the quality of generated deepfakes increased, spatial forgery alone became increasingly hard to detect, hence the emergence of temporal and physiological methods. Temporal methods, whether through recurrent neural networks or 3D convolutional networks, inspect video sequences based on motion patterns, lip syncing, or gaze movements expected in a forged video, as these are hard to mimic. Another group is based on signals similar to those in biological systems. These methods believe that imitation AI has difficulties mimicking biological rates in terms of heartbeat, micro-expressions, or blink rates in forged videos. The other line of research involves frequency and compression domain analysis. The techniques convert images into frequency space (such as DCT or FFT) or directly analyze the JPEG artifact to extract characteristic spectra from generated images. The frequency-based countermeasures will generalize better to unknown manipulation methods because the countermeasures rely on statistical regularities rather than spatial information. Also, social media recompression tools distort low-level textures, and deepfake images will likely pass through a social media recompression step.

Benchmarks and datasets have played a crucial role. Large-scale datasets of genuine and fake video and image samples have made training and comparison of approaches feasible. These datasets usually contain variants of manipulations (face swaps, reenactment, expression manipulations), varying levels of compressions, and different individuals and illumination conditions. They demonstrate that models that are strong within their training set usually fail when evaluated on new datasets or new manipulation attacks, illustrating the problem of generalization in the field of deepfake detection.

More recently, the transformer architecture and hybrid CNN-transformer networks have received increased attention. Vision transformers (ViTs) and convolutional vision transformers (CvTs) make use of the self-attention technique. This technique is particularly useful in capturing the global information of the frames and can be beneficial in recognizing the overall inconsistencies in the facial structures. Research validates that attention, combined with the CNN technique of extracting features, is more efficient in dealing with the robustness of visually varying data. Multimodal learning extends this vision-based work by also considering the analysis of the audio. It examines the relationships between lip motion, speech, and audio modality characteristics inorder to detect inconsistencies between the audio and vision streams. The assumption here is that deepfake tools are likely processing the vision independently or mismatches the vision and audio, making it possible to notice the inconsistencies or anomalies. The task of multimodal learning is performed by late fusion, early fusion, or learning an embedding space.

Another theme present throughout the literature is the idea of robustness and applicability. There is now considerable research evaluating detector performance for cross-dataset generalization, adversarial examples, and simulation of real-world perturbations such as noise, compression, scaled imagery, and occlusion. Another area being explored is the field of explanation methods including saliency map visualization and attention visualization for identifying the region or frames responsible for the detection made by the detector. This is particularly relevant for applications involving forensics where it is necessary for the evidence to be explicable.

Illmar Raag TITLE: Parent Involvement
Recent work on deepfake detection (2023–2024) has focused on hybrid models, surveys, and in-depth analysis of existing datasets and benchmarks.
In Neural Computing and Applications, a 2024 paper presents a deepfake detection model that integrates convolutional neural networks and convolutional vision transformers. This technique utilizes the strength of CNNs in local feature extraction and vision transformers to handle the global features to address the reality that deepfakes are very realistic and pervasive across social media platforms. According to the authors, the hybrid architectures comprising both CNNs and CvT outperform the purely CNN-based architectures in terms of robustness to both the local artifacts and global inconsistencies within the manipulated facial images.

Survey studies have also contributed much to the integration of the area. A survey by IEEE explores recent deepfake detection techniques for images and videos, classifying them based on the type of model used (CNN-based models, transformer-based models, recurrent models, frequency-domain models, and multimodal models), the modality of the target (image, video, audiovisual), and the modality of the manipulation type. It also looks at the history of deepfake datasets, classifying them into 'generations' depending on size, diversity, and realism, and important topics including generalization, domain transfer, and adversarial attacks against detectors.
Another recent survey conducted by the Electronics journal gives a general view of the methods for facial forgery detection proposed by different researchers between 2019 and 2023. Some of the techniques addressed in this survey include conventional methods based on CNN detectors, physiological signal techniques, frequency-based detection methods, and novel models involving transformers, and how they have been compared among different popular datasets and sets of facial images. Of significance to our study, this survey recognizes some challenges associated with facsimile detection, which include data imbalances, dynamic generative models, data biases, and the detection of poor and highly compressed deepfakes found in some social media sites.
In light of these more modern studies, what is clear about these relevant "Deepfake Detection AI – Manipulated or Human" projects is that first, there is an increasing approach to the use of deep architectures that combine both CNN and transformers. Secondly, there is an emphasis on more than just achieving accurate results on one dataset. Thirdly, there is an understanding through the survey articles on deepfake detection that there is not a single "perfect" AI for the task and that perhaps what is needed is not one stop in the development of deep AI for image detection but a fluid constant evolution of techniques and methods that keep up with other generative AI advancements. How your project fits into this debate and focus group will help to contextualize and locate your project within relevant and universal patterns of deep AI detection.

## IV. DATASET DESCRIPTION

Deepfake Detection - AI-Manipulated or Human
A deep fake detection project utilizes a dataset that holds real or human-generated data and AI-manipulated data. The project aims to develop models that can classify the authenticity of the media and that which is synthetically generated or modified. Below is a structured description of the dataset for documentation in reports, research articles, or project documentation.

Dataset Description
The data for the topic "Deepfake Detection: AI-Manipulated or Human" contains a set of videos, images, or audio samples that can be classified basically into two main categories:

1. Real (Human-Generated
This class consists of authentic media taken directly from real people with no digital alterations.
It comprises:

- Video captured using cameras or smartphones originally
- True facial expressions, movements, and speech
- Lighting, shadows, and background details
- Original audio without music and sounds

This is used as ground truth to train the system on recognizing how actual human-created content will look and will sound.

2. Deepfake AI-Manip

This class covers media that has been digitally modified or completely created through AI technologies such as:

- Face Swapping
- Face reenactment
- Lip sync manipulation
- Voice Cloning
- GAN synthesized faces

Deepfakes samples commonly display:

- Small discrepancies in facial features
- Irregular blinking or unusual eye movement
- Texture mismatches
- Inconsistencies in

Audio LIP synchronization errors All of these samples assist in teaching the model about the minutc of manipulation

## V. PROBLEM STATEMENT

### 5.1 Deepfake Detection – AI-Manipulated or Human

Artificial intelligence is developing quickly, which makes it increasingly easy to produce fake videos, photos, and audio that are extremely realistic so called "deepfakes". This kind of content manipulated by AI is hardly distinguishable from "authentic" human-created content. Therefore, as AI-generated content becomes more sophisticated, the methods of authenticating its origin become ineffective.

The persistent problem arises from the difficulty in conclusively determining whether digital content is genuine or artificially created. Such ambiguity gives way to misinformation, identity theft, reputational abuse, political manipulation, and a consequence of a lack of trust in digital communications. As long as effective detection schemes are not in place, deepfakes will easily proliferate social networks with the end effect of influencing public views even before detection.

Thus, the problem solved in this study is the need for efficient automatic deep fake detection methods that are able to distinguish AI-processed media from human-generated ones. In particular, it is a problem of enhancing digital security and trust in image/authentic media through improved digital techniques.

If you want, I can also develop a shorter version, a technical version, and/or problem definition of project for your project report.

## VI. EXISTING SYSTEM

Deepfake detection systems are built to find out if digital pictures or videos have been changed using intelligence. These systems use ideas from Computer Vision and Deep Learning where strong models are trained to tell real from face data. The main aim is to spot inconsistencies that humans can't see, like strange textures lighting that doesn't match or messed up facial features.

- One common way to make these systems is to use Convolutional Neural Networks (CNNs) which're great at looking at image data.
  - These models learn from datasets and can spot tiny problems in deepfake pictures.

- For videos models like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks are used to look at changes over time.
  - They help spot issues like blinking, unnatural facial expressions or bad lip syncing.
  - Some detection methods also look at frequency and artifact analysis.
  - These methods check hidden patterns in images like inconsistencies added when making or compressing images.

- Biological signal-based detection is another approach.
  - It looks at human signs like eye movement, head motion and even tiny skin color changes that show heartbeat.
  - These are often not copied well in deepfake content.
  - Many real-w`orld tools have been made to use these detection techniques.

For example, Microsoft Video Authenticator and Deepware Scanner are used to check media authenticity.

- These systems are often trained using benchmark datasets like FaceForensics++ and DFDC which have both real and fake media, for training and evaluation.

- Deepfake detection systems still have challenges.
- As deepfake technology gets better it gets harder to spot changes
- Many models struggle to work across different types of deepfakes.
- Adversarial techniques can sometimes trick detection systems.
- So ongoing research is focused on making accuracy making real-time detection methods and combining many approaches to make more strong and reliable systems.

## VII. PROPOSED SYSTEM

Deepfake detection systems help find videos and images made using smart computers. These systems are important because it's getting harder to tell real from fake with computers getting smarter. We need a system to stop fake news protect people's online identity and make sure we can trust what we see online.

- The system starts by collecting real and fake images and videos from the internet.
- It then prepares the data by breaking videos into pictures making images the same size and adjusting the colors.
- The system focuses on people's faces to make sure it's looking at the things to tell if it's real or fake.

The system uses computer models to look at faces and find clues that they are fake. These clues include lighting, strange expressions and tiny mistakes in the picture. The model is trained on lots of labeled data so it can learn what makes something fake.

- The model learns by looking at lots of examples of real content.
- We test it to make sure it works well and can tell the difference between real and fake.
- We keep updating it so it can catch ways of making deepfakes.

When it's ready the system can be used on websites or phones to check if something is fake, in time. It can help on media, news sites and cybersecurity tools so people can quickly check if something is real or not. This helps stop media from spreading and keeps the internet safer.

## VIII. RESEARCH DESIGN METHODOLOGY

### a) Research Objective

The main goal of this research is to build a system that can accurately detect facial images and videos. This system uses learning techniques to identify fake faces.

### b) Dataset Collection

A large dataset of images. Videos are collected from public sources. This dataset includes both fake faces, with different lighting conditions, facial expressions and backgrounds.

Examples of datasets used are FaceForensics++ and DeepFake Detection Challenge dataset.

### c) Data Preprocessing

The data is prepared for use. Video data is broken down into frames. Images are resized to be the size.

- The faces in the images are. Extracted.
- The data is also. Noise is removed.
- The dataset is made diverse by flipping, rotating and cropping the images.

### d) Feature Extraction

- Deep learning models are used to extract features from the images and videos.
- These features include inconsistencies and color mismatches.
- In videos unnatural motion patterns are also analyzed.

### e) Model Selection and Architecture

- Different deep learning models are considered for use.
- The chosen model classifies inputs as real or fake.
- Pre-trained models may be used to improve efficiency.

### f) Model Training

- The model is trained using labeled data.
- The dataset is divided into training, validation and testing sets.
- Algorithms such as Adam are used to minimize errors.

### g) Performance Evaluation

- The system is evaluated using metrics such as accuracy and precision.

- A confusion matrix is used to understand how well the system classifies inputs.
- Cross-validation ensures the model is reliable.

h) Implementation Tools and Technologies

The system is built using tools such as Python and TensorFlow.

Libraries, like OpenCV are used for image processing.

i) System Deployment

The final model is deployed as a real-time detection system.

Users can upload images or videos. Get instant results.

j) Limitations and Future Work

- The system may ` incorporated for accuracy.

## IX. CONCLUSION

In the end it is clear that detecting deepfakes is an important area of research. This is because artificial intelligence is getting better and better and people are using media in bad ways. This study showed a system that uses learning to find fake pictures and videos of faces. It does this by looking for things that do not look right and patterns that are hard to see. The system we talked about showed that some models, like Convolutional Neural Networks can tell what is real and what is a deepfake. This works when these models are trained on different kinds of data that have been prepared well. If we take the time to get the features from the data make the models work better and use good tests then the system will work well and be accurate.

Deepfake technology is getting better too which makes it harder to detect. So, we need to keep updating our systems use sets of data and make new models that combine different approaches. We can also try to detect deepfakes in time and look at both pictures and sound to get better results. So deepfake detection systems are very important. They help stop information protect people's identities and make sure we can trust what we see and hear online. This is a deal in today's world, where technology is such a big part of our lives and deepfake detection plays a crucial role, in it and that is why deepfake detection is so important.

## REFERENCE

[1] Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1–11.

[2] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) dataset," arXiv preprint arXiv:2006.11475, 2020.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A facial video forgery detection network," in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), 2018, pp. 1–7.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1251–1258.

[5] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 1831–1839.

[6] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of capsule networks to detect fake images and videos," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2019, pp. 2622–2626.

[7] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.