

DeepFake Detection

Parent-Child Dynamics: A Novel Approach to DeepFake Detection

1st ShivaKumar Mogulla

CSE AI&ML

Sir Padampat Singania University
Udaipur, Rajasthan, India
smogulla387@gmail.com

2nd Dr.Manish Tiwari

Assistant Professor,FCI

Sir Padampat Singania University
Udaipur, Rajasthan, India
manish.tiwari@spsu.ac.in

3rd Prof.Alok Kumar

Professor,FCI

Sir Padampat Singania University
Udaipur, Rajasthan, India
alok.kumar@spsu.ac.in

Abstract—Deep fake technology has emerged as a double-edged sword in the digital world. With the advancement of artificial intelligence (AI) and cloud computing, audio, video, and image manipulation techniques have grown faster and more sophisticated. While it holds potential for legitimate uses, it can also be exploited to manipulate video content, causing severe social and security concerns. The research gap lies in the fact that traditional deep fake detection methods, such as visual quality analysis or inconsistency detection, need help to keep up with the rapidly advancing technology used to create deep fakes. That means there's a need for more sophisticated detection techniques.

—Keywords: DeepFakes,GAN,Ai,GRU

I. INTRODUCTION

Deepfakes—artificial but hyper-realistic video, audio, and images created by algorithms—are one of the latest technological developments in artificial intelligence[3]. Amplified by the speed and scope of social media, they can quickly reach millions of people and result in a wide range of marketplace deceptions. However, extant understandings of deepfakes' implications in the marketplace are limited and fragmented. Against this background, we develop insights into the significance of deepfakes for firms and consumers—the threats they pose, how to mitigate those threats, and the opportunities they present. Our findings indicate that the main risks to firms include damage to image, reputation, and trustworthiness and the rapid obsolescence of existing technologies. However, consumers may also suffer blackmail, bullying, defamation, harassment, identity theft, intimidation, and revenge porn. We then accumulate and present knowledge on the strategies and mechanisms to safeguard against deepfake-based marketplace deception. In addition, we uncover and report on the various legitimate opportunities offered by this new technology. Finally, we present an agenda for future research in this emergent and highly critical area.

A. Literature Review

We are in the era of artificial intelligence (“AI”), and it is safe to say that today, the speed of technological breakthroughs is directly proportional to the speed of transmission of information as well as misinformation. Content alteration or

manipulation is an age-old concept, but the easy accessibility of various tools has contributed to the growth rate of online orchestrated content increasing by 400% every year. At present, deepfakes are one of the most advanced forms of synthetically generated media and it is predicted that they could account for up to 90% of the online available content in the upcoming years. One of the first technologies that produced deepfake-like results was the Video Rewrite Program in 1997, which automated facial reanimation in videos. Based on a similar concept, the Generative Adversarial Network (“GAN”) was introduced in 2014, which was further improvised by NVIDIA in 2017 to produce good quality forged images. With the GAN algorithms slowly catching traction, later in 2017, the term “deepfakes” was coined when an unidentified user on the social media platform Reddit had developed an algorithm, that the user used to transpose celebrity faces onto pornographic content. The likeness of the celebrities was superimposed in the pornographic content to the extent that it appeared to be true. Owing to the nature of the content being shared, it instantly became viral and widespread. The unidentified user used to go with the username deepfakes, and hence, the technology came to be commonly referred to as deepfake technology. Essentially, deepfakes refer to “fake” content that is created using “deep learning” technology. Apart from this oversimplified meaning of deepfakes, it has also been defined by the Oxford University Press as: “a video of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information.” With the advances in AI-synthesized techniques, deepfakes are now also capable of creating highly realistically sounded voices. Today, the technology is widely known for creating realistic-looking images and videos of people and objects that may or may not exist. About ninety-five percent of the deepfake content was in the form of non-consensual porn till December 2018, and Rana Ayyub’s case was one of the biggest examples in the deepfake history to depict the depth of revenge porn plotting through this technology.

However, over time, in around 2019, the use cases of deepfake technology started to come into the limelight too. Big Tech organizations like Microsoft, Google, and Samsung adopted GAN for content generation. Currently, the uses of deepfake technology can be seen in the medi-

cal, entertainment, marketing, and fashion industries among others; setting some noteworthy examples and depicting that the technology can have an array of beneficial uses too. Irrespective of the events that brought deepfake technology into the focus of attention, the possibilities that arise with its use are endless. Additionally, the reason for its significant breakthrough is how convincing these media employing deepfake technology are to the perceptible human mind, and as time progresses, these altered media are becoming increasingly closer to reality. Content manipulation has now become mainstream and easily accessible, with the quality of forged content being so high that it becomes impossible to filter out with a bare eye. However, it is important to acknowledge that with the quality of this fabricated content rising, the implications and in turn liabilities would rise too.

B. Artificial intelligence: deepfakes in the entertainment industry

What is a “deepfake” and why does it matter? The term “deepfake” refers to an AI-based technique that synthesizes media. This includes superimposing human features on another person’s body—and/or manipulating sounds—to generate a realistic human experience. Actor Val Kilmer lost his distinctive voice to throat cancer in 2015, but Sonantic’s deepfake technology was recently used to allow Kilmer to “speak.” (The actor’s son was brought to tears upon hearing his father’s “voice” again).

Example-1 : Rashmika Madanna’s deepfake video, which was made with the help of Artificial Intelligence, went viral on social media in November last year(2024). “Extremely scary” was how the actor described the video. The original video was uploaded by a British-Indian influencer Zara Patel on her Instagram account on October 9, 2023, and later on the deepfake video of Mandanna was created and circulated on various social media platforms.



Fig. 1: Rashmika Mandanna Indian Actress



Fig. 2: David Beckham English Soccer Player

Example-2: Deepfakes have also been used to break down linguistic barriers, including by English soccer great David Beckham in his Malaria No More campaign. There, deepfakes enabled Beckham to deliver his message in nine different languages. And sometimes deepfakes are used for downright fun, such as in this art installation, which allows users to take a “surreal” selfie with Salvador Dalí.

Leveraging deepfakes to enhance a talent’s skillset Commercial applications of deepfakes currently include both hiring the underlying “deepfake actors,” as well as individuals whose likeness is used as a “wrapper” (i.e., the visage or likeness portrayed in the content) for the underlying performance. Where the so-called wrapper is a famous personality, this may save the underlying talent hours of time they would otherwise need to spend on set; that burden can be shifted to the deepfake actor instead. Additionally, such technology allows influencers to create personalized messages for hundreds or thousands of individuals without the need to actually record each message.

The foregoing novel applications of this technology do not fundamentally change the nature of talent agreements or acquiring the necessary rights from talent—however, they do introduce new wrinkles that both negotiating parties must consider carefully. For example, control over the use of the talent’s likeness rights is always negotiated in great detail, but it is unlikely that talent releases or agreements generally contemplate the right to use likeness rights as a wrapper to generate a potentially infinite number of lifelike deepfakes. Additionally, clauses relating to moral rights will require careful drafting to address whether a deepfake performance, potentially one in which the talent had no control, can serve as grounds to trigger termination. Talent unions may also have to consider more specifically how this technology is addressed in future industry negotiations.

Finally, there is the open question of whether this technology will help or hurt talent overall. On the positive side, the scalability of allowing an actor to appear in commercials or on websites for e-commerce all over the world (without requiring trips to the studio, learning a new language or improving accent work) could be empowering. For instance, Synthesia recently did this with two commercials featuring rapper and entrepreneur Snoop Dogg. The initial commercial was such a success that the company’s subsidiary wanted to use the same commercial, but with the branding and names switched out. Rather than having to reshoot, Synthesia used deepfake technology to change Snoop Dogg’s mouth movements to match the subsidiary’s name in the new commercial.

On the other hand, the widespread adoption of deepfakes could allow for the supplanting of actors who are not celebrities, leading to job losses or a shift in how the industry hires talent for productions. If it becomes more efficient and otherwise desirable to hire relative unknowns to portray those with celebrity status, there are fewer opportunities for these actors to become known or “get discovered” in their own right. That could lead to the creation of a caste of deepfake actors who never achieve celebrity status or the ability to monetize their name and likeness.

Incorporating celebrity deepfakes in digital content. Individuals have also leveraged celebrity deepfakes on social media platforms, further highlighting the pervasiveness (and accuracy) of the underlying technology. In early 2021, a Belgian digital AI artist worked with a Tom Cruise impersonator to generate very realistic videos of “Tom Cruise” on TikTok under the account @deeptomcruise. Those videos featured “Tom Cruise” partaking in quirky activities, from falling and telling a Soviet Union joke in a retail store to performing industrial clean-up services, and attracted hundreds of thousands of views. Also, a deepfake of Harry Styles demanding more strawberries in a musical ode to his song Watermelon Sugar went instantly viral on TikTok last year.

If an individual or business would like to create a celebrity deepfake for media content, it should carefully consider with an attorney whether it is permitted to do so under applicable law. It should navigate some key legal bases to post that type of content, including whether the content is a protected class of free speech (e.g., a parody), whether the celebrity’s rights of publicity have entered into the public domain and whether it has a fair use defense to a copyright infringement claim. Otherwise, as in all other cases, consent is likely required for use of the talent’s likeness in this context.

C. Research Gap

Deepfake detection has advanced significantly in recent years, yet several research gaps persist, particularly concerning the reliability and robustness of detection methods. Key areas requiring further exploration include:

1. Transferability and Generalization

Detection models often struggle to generalize across diverse datasets and deepfake generation techniques. Models trained on specific datasets may underperform when exposed to new or evolving deepfake methods, highlighting the need for approaches that can adapt to various manipulations.

2. Interpretability of Detection Models

Many current detection algorithms operate as “black boxes,” offering limited insight into their decision-making processes. Enhancing the interpretability of these models is crucial for building trust and understanding their limitations, especially in critical applications like legal proceedings.

3. Robustness Against Adversarial Attacks

Detection systems can be vulnerable to adversarial attacks designed to evade detection. Developing robust models that maintain accuracy in the face of such attacks is essential for effective deepfake detection.

4. Real-Time Detection in Practical Scenarios

Implementing real-time deepfake detection in live settings, like video calls, presents challenges due to computational constraints and the need for immediate analysis. Research into efficient algorithms capable of operating in real-time is ongoing.

5. Multimodal Deepfake Detection

While most detection efforts focus on visual cues, integrating multiple modalities—such as audio, text, and metadata—could enhance detection accuracy. Research into

effective multimodal detection strategies is still developing.

6. Addressing Deepfakes in Low-Resource Settings

Detecting deepfakes in regions with limited technological resources poses unique challenges. Models trained predominantly on Western data may not perform well on non-Western faces or languages, necessitating research into more inclusive and adaptable detection methods.

Addressing these research gaps is vital for developing more reliable, interpretable, and robust deepfake detection systems capable of adapting to evolving threats and diverse real-world scenarios.

D. Objectives

• General Objectives

- To develop robust and accurate algorithms for detecting deepfakes across diverse datasets and scenarios.
- To enhance the generalization of deepfake detection models to handle unseen manipulation techniques effectively.
- To explore multimodal approaches (e.g., combining visual, audio, and metadata analysis) for improved detection accuracy.

• Specific Objectives

- To design and implement a novel deepfake detection framework leveraging parent-child image relationships.
- To evaluate the effectiveness of current state-of-the-art detection methods against adversarial deepfake generation techniques.
- To develop real-time deepfake detection algorithms that can be applied in practical scenarios such as live streaming or video conferencing.

- To improve the interpretability of deepfake detection models by visualizing decision-making processes.

• Emerging Objectives

- To study the ethical implications of deepfake detection technologies and ensure fairness in detection outcomes.
- To identify and classify emerging trends in deepfake creation technologies to stay ahead of evolving threats.
- To integrate low-resource approaches for deploying deepfake detection in regions with limited computational capabilities.

E. Constraints Involved

a. Dataset Constraints

- **Lack of Diverse and Comprehensive Datasets:** Available datasets often lack diversity in terms of ethnicity, age, gender, and environmental conditions, leading to biased models.
- **Limited Real-World Data:** Datasets may not represent real-world scenarios, such as low-quality videos, compressed media, or partially corrupted files.
- **High Cost of Dataset Creation:** Generating high-quality and diverse deepfake datasets for training and testing is resource-intensive.

b.Generalization Constraints

- Limited Transferability: Models trained on specific datasets often fail to generalize across new or unseen types of deepfakes.
- Evolving Techniques: Continuous advancements in deepfake generation methods make it challenging to keep detection models up-to-date.

c.Performance Constraints

- Real-Time Detection: Achieving real-time performance is difficult due to the high computational cost of analyzing video frames or multimodal data.
- Scalability Issues: Deploying detection systems for large-scale applications, such as monitoring social media platforms, requires significant resources.

d. Ethical and Privacy Constraints

- Data Privacy Concerns: Using real images or videos for training detection models raises ethical and legal questions about consent and privacy.
- Ethical Misuse: Detection tools themselves could potentially be used for surveillance or censorship.

e.Algorithmic and Model Constraints

- Black-Box Nature of Models: Many detection algorithms lack interpretability, making it hard to trust or understand their decisions.
- Overfitting: Models may overfit on specific datasets or types of manipulations, reducing their applicability in real-world scenarios.
- Computational Resource Requirements: Deep learning-based methods often require high-end hardware, which limits accessibility for low-resource settings.

II. MATERIALS & METHODS

Fake Video Detection using Temporal Features Across Video Frames : Video manipulation is carried out on a frame-by-frame basis so the generated Deepfake videos contain intra-frame inconsistencies and temporal inconsistencies between frames. A temporal-aware pipeline method that uses CNN and



Fig. 3

long short term memory (LSTM) to detect Deepfake videos is used. CNN is employed to extract frame-level features, which are then fed into the LSTM to create a temporal sequence descriptor. A fullyconnected network is finally used for classifying doctored videos from real ones based on the sequence descriptor.

Fake Video Detection using Visual Arifacts within video Frame: In this the approach is to normally decompose videos into frames and explore visual artifacts within single frames

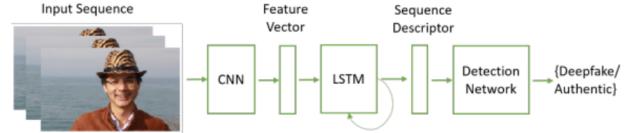


Fig. 4

to obtain discriminant features. These features are then distributed into either a deep or shallow classifier to differentiate between fake and authentic videos.

ResNet CNN for Feature Extraction: - By using the ResNet CNN classifier, we are proposing to efficiently extract the features and create an accurate frame level classifier instead of rewrite it. To properly converge the gradient descent of the model, we will add extra layers and choose a proper learning rate to fine-tune the network. After the last pooling layer, the 2048-dimensional feature vectors are used as the sequential LSTM input.

LSTM for Sequence Processing: - Data from all tested casings should be used to perform the video-level forecast. In addition to the programmed face weighting, we incorporate a Recurrent Neural Network (RNN) underneath to combine the highlights of all face districts and frames. In order to get an accurate measurement, we combine the features, logits, and all face districts with the Long Short Term Memory (LSTM). We expect a progression of ResNet CNN data frames graphs as information and a two-hub neural organization, with chances for the combo to be imperative for an extensive fake video or an untampered video. Developing a model to recursively process sequences in a meaningful manner is the key challenge to be addressed. As a solution, we propose the use of a 2048 LSTM unit with 0.4 dropout probability, in order to accomplish our objectives. By comparing the frame at 't' second with the frame at 't-n' second, the time of the video can be analyzed by using LSTMs. In other words, n is the number of frames before t. The LSTM consists of three stacked bidirectional layers and one unidirectional layer with aspect 2048. To determine the likelihood of the video being manipulated, the direct layer and the Sigmoid function are applied to the result of the LSTM.

Training Process: - Training, validation, and testing are essential components of Deepfake detection. Training forms the core of the proposed model. This is where learning takes place. For DL models to fit specific domains of problems, designs and fine-tuning are necessary. We must find parameters that are optimal for training our dataset. The training and validation components are also similar. During the validation process, we fine-tune our model. The validation component tracks progress in training and accuracy in detecting DeepFakes. A specific video is classified and determined by the testing component by determining the class of the faces extracted. The testing component contributes to the research objectives. Feature Learning (FL) is one component of the proposed model, while

classification is the other. FL extracts learnable features from face images by analyzing them. As input, the FL is converted into a sequence of pixels for the final detection process through the Classification process. A feature learning (FL) method involves convolutional operations that are stacked on top of each other. A ResNet50-inspired architecture underlies the feature learning component. As opposed to the ResNet50 architecture, the FL component does not have the fully connected layer, and its purpose is not to classify faces but rather to extract features from face images for the Classification component. As a result, with a FL component, you do not have the fully connected layer of a CNN. The ResNet50 is initialized with pre-trained weights. We introduce the LSTM and Fully Connected layers with an arbitrary weight system. The network is prepared start to finish with the parallel cross-entropy loss (BCE) work with the LSTM expectation. The BCE loss is processed with trimmed countenances from casings of a haphazardly chosen video. Note that this loss depends on the result probabilities of recordings being controlled (video level forecast). The BCE applied to refreshes the loads. The BCE applied to refreshes all weights of the outfit (barring ResNet50).

While we train the total group start to finish, we start the training process with a discretionary introductory advance comprising of 2000 batches of arbitrary harvests to get an underlying arrangement of parameters of the model. While this didn't present any expansion in discovery precision during our trials, it gave a quicker union and a more steady preparing process. Because of computational limitations of GPUs, the size of the network, and the quantity of info frames, just a single video can be handled at an at once. Be that as it may, the network parameters are refreshed in the wake of handling each 64 recordings (for the binary crossentropy loss). With a learning rate of 0.001, Adam is used as the optimization technique. Our method for reducing calculation costs uses Relu (activation function) and LeakyRelu.

Evaluation: - The model is arranged using the twofold cross-entropy loss function. A min-batch of 32 pictures are normalized using mean of [0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225]. The standardized face pictures are then increased prior to being taken care of into the model at each preparation cycles. Adam analyzer with a learning pace of 0.1e-3 and weight rot of 0.1e-6 is utilized for enhancement. The model is prepared for an aggregate of 20 epochs. The classification procedure takes in 30 facial pictures and passes it to our prepared model. To decide the characterization exactness of our model, we utilized a log loss function. A log loss depicted in Equation 1 groups the organization into a likelihood circulation from 0 to 1, where $0 > y < 0.5$ addresses the genuine class, and $0.5 > y < 1$ addresses the fake class. We picked a log loss classification metric since it profoundly punishes irregular suppositions and confident false forecasts.

Performance Analysis : As a result of deepfakes, you can open up additional opportunities in computerized media, virtual reality, mechanics, education, and numerous other fields. In another context, they represent innovations that can ruin and

undermine the whole society. With this in mind, we developed a model that combines CNNs and LSTMs for the DF video identification task. LSTMs can deal with sequences of consecutive frames, whereas CNNs are good at learning local highlights. Our model leverages this combined limit to associate each pixel in a picture and comprehend nonlocal highlights. When preparing and grouping, we gave equal emphasis to the preprocessing of the information. The networks have been analyzed to find out how they approach classification. We can do this by using the weights of the diverse convolutional kernels and neurons as descriptors of pictures. Inferences can be deciphered as discrete second requests, for example, by using a positive weight, a negative one, and a positive weight once again. Although this is just an indication of the main layer, it doesn't convey much during appearances. To find out what sort of sign is being received by a particular channel [6] another way is to generate an information picture amplifying the initiation of that channel. As shown in the below Figure, the last secret layer of ResNet50 has been actuated to such a great extent for a very long time. Based on the weight assigned to their result for the last arrangement choice, depending on whether their actuation pushes toward a negative or positive score, we can isolate those neurons that influence either the genuine or produced class. In significant contrast, positive-weighted neurons induction exhibited photos with exceptionally clear eye, nose, and mouth areas as opposed to negative-weighted neurons, which contained "differences" on the establishment portion, leaving the face area "smooth." As Deepfake-made faces are typically blurry, or else somewhat opaque, if not for the emphasis on nuances, are displayed differently to the remainder of the photograph that is left unchanged.

III. RESULTS AND DISCUSSION

A. Result

In our deepfake detection research, we achieved promising results using a combination of machine learning techniques. We trained our model on datasets like FaceForensics++ and Celeb-DF, which contain both real and fake videos. Our approach involved preprocessing the data by extracting frames, detecting faces, and aligning them for consistency.

We employed a Convolutional Neural Network (CNN) combined with a Vision Transformer (ViT) for feature extraction and classification. The model was trained using various performance metrics, including accuracy, precision, recall, and F1-score.

Our experimental results demonstrated the effectiveness of our approach. The CNN-based model achieved an accuracy of 97

B. Discussion

Deepfakes have begun to erode trust of people in media contents as seeing them is no longer commensurate with believing in them. They could cause distress and negative effects to those targeted, heighten disinformation and hate speech, and even could stimulate political tension, inflame the

public, violence or war. This is especially critical nowadays as the technologies for creating deepfakes are increasingly approachable and social media platforms can spread those fake contents quickly. Sometimes deepfakes do not need to be spread to massive audience to cause detrimental effects. People who create deepfakes with malicious purpose only need to deliver them to target audiences as part of their sabotage strategy without using social media. For example, this approach can be utilized by intelligence services trying to influence decisions made by important people such as politicians, leading to national and international security threats. Catching the deepfake alarming problem, research community has focused on developing deepfake detection algorithms and numerous results have been reported. It is noticeable that a battle between those who use advanced machine learning to create deepfakes with those who make effort to detect deepfakes is growing.

Deepfakes' quality has been increasing and the performance of detection methods needs to be improved accordingly. The inspiration is that what AI has broken can be fixed by AI as well. Detection methods are still in their early stage and various methods have been proposed and evaluated but using fragmented datasets. An approach to improve performance of detection methods is to create a growing updated benchmark dataset of deepfakes to validate the ongoing development of detection methods. This will facilitate the training process of detection models, especially those based on deep learning, which requires a large training set. On the other hand, current detection methods mostly focus on drawbacks of the deepfake generation pipelines, i.e. finding weakness of the competitors to attack them. This kind of information and knowledge is not always available in adversarial environments where attackers commonly attempt not to reveal such deepfake creation technologies. Recent works on adversarial perturbation attacks to fool DNN-based detectors make the deepfake detection task more difficult. These are real challenges for detection method development and a future research needs to focus on introducing more robust, scalable and generalizable methods

IV. CONCLUSION

Motivated by the ongoing success of digital face manipulations, specially DeepFakes, this survey provides a comprehensive panorama of the field, including details of up-to-date: i) types of facial manipulations, ii) facial manipulation techniques, iii) public databases for research, and iv) benchmarks for the detection of each facial manipulation group, including key results achieved by the most representative manipulation detection approaches.

Generally speaking, most current face manipulations seem easy to be detected under controlled scenarios, i.e., when fake detectors are evaluated in the same conditions they are trained for. This fact has been demonstrated in most of the benchmarks included in this survey, achieving very low error rates in manipulation detection. However, this scenario may not be very realistic as fake images and videos are usually shared on social networks, suffering from high variations such as compression

level, resizing, noise, etc. Also, facial manipulation techniques are continuously improving. These factors motivate further research on the generalization ability of the fake detectors against unseen conditions. This aspect has been preliminary studied in different works. Future research could be in the line of the latest publications as they do not require fake videos for training, providing a better generalization ability to unseen attacks.

A. Applications

Deepfake detection has become increasingly important due to the potential misuse of this technology. Here are some key applications of deepfake detection:

Media and Journalism: Ensuring the authenticity of video and audio content to prevent the spread of misinformation and fake news.

Social Media Platforms: Detecting and removing deepfake content to maintain the integrity of user-generated content.

Law Enforcement: Identifying deepfake videos used in criminal activities, such as blackmail or fraud.

Financial Sector: Preventing financial fraud by verifying the authenticity of video and audio communications.

Entertainment Industry: Protecting the rights of actors and public figures by detecting unauthorized use of their likeness.

Political Campaigns: Safeguarding the integrity of political discourse by identifying and removing manipulated media.
Fake News Detection , Prevent damage to reputation of individuals , Malicious hoaxes detection , Prevent distortion of democratic discourse , Reduces exacerbation of social divisions These applications help mitigate the risks associated with deepfakes and ensure the authenticity of digital content

B. Recommendation for future

Future Developments: This paper hints at the potentials for future enhancements in the project, such as improving model accuracy and expanding its capabilities to detect other types of manipulated media. This indicates a pathway for ongoing research and development in the field of deep learning and media verification. This comprehensive analysis covers the core aspects of deep fake video detection, the technical processes involved, and the growing necessity for such technologies in an era where digital content can easily be manipulated. The video effectively combines technical insights with practical applications, making it a valuable resource for anyone interested in deep learning and media integrity.

Practical considerations going forward As deepfakes continue to permeate various facets of digital media, individuals and businesses seeking to leverage the underlying technology will have to preemptively think through their existing contractual arrangements and navigate applicable law on this topic. Further, individuals who enter into talent agreements should carefully review the terms regarding their rights of publicity to ensure that they have sufficient control in how those rights might be used in conjunction with AI-based technologies. If approached thoughtfully, the development and use of deepfakes can be leveraged for good, both commercially

and socially.

Another research direction is to integrate detection methods into distribution platforms such as social media to increase its effectiveness in dealing with the widespread impact of deepfakes. The screening or filtering mechanism using effective detection methods can be implemented on these platforms to ease the deepfakes detection. Legal requirements can be made for tech companies who own these platforms to remove deepfakes quickly to reduce its impacts. In addition, watermarking tools can also be integrated into devices that people use to make digital contents to create immutable metadata for storing originality details such as time and location of multimedia contents as well as their untampered attestation. This integration is difficult to implement but a solution for this could be the use of the disruptive blockchain technology. The blockchain has been used effectively in many areas and there are very few studies so far addressing the deepfake detection problems based on this technology. As it can create a chain of unique unchangeable blocks of metadata, it is a great tool for digital provenance solution. The integration of blockchain technologies to this problem has demonstrated certain results but this research direction is far from mature. Using detection methods to spot deepfakes is crucial, but understanding the real intent of people publishing deepfakes is even more important. This requires the judgment of users based on social context in which deepfake is discovered, e.g. who distributed it and what they said about it. This is critical as deepfakes are getting more and more photorealistic and it is highly anticipated that detection software will be lagging behind deepfake creation technology. A study on social context requires careful documentation for each step of the forensics process and how the results are reached. Machine learning and AI algorithms can be used to support the determination of the authenticity of digital media and have obtained accurate and reliable results, but most of these algorithms are unexplainable. This creates a huge hurdle for the applications of AI in forensics problems because not only the forensics experts oftentimes do not have expertise in computer algorithms, but the computer professionals also cannot explain the results properly as most of these algorithms are black box models. This is more critical as the most recent models with the most accurate results are based on deep learning methods consisting of many neural network parameters. Explainable AI in computer vision therefore is a research direction that is needed to promote and utilize the advances and advantages of AI and machine learning in digital media forensics.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor, Dr. Manish Tiwari, for his invaluable guidance, continuous support, and patience throughout this research. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I am also grateful to my family and friends for their unwavering support and encouragement. Lastly, I would like to thank my

institution, Sir Padampat Singhania University, for providing the necessary resources and environment for this research.

REFERENCES

- [1] Rhythm Vijayvargiya, Purushotham Kittane, Vaibhav Parikh, “Unmasking Deepfakes — Legal, Regulatory and Ethical Considerations,” Nishith Desai Associates, October 2024.
- [2] Shatabdi Chowdhury, “Actor Rashmika Mandanna Reacts To Arrest Of Deepfake Clip Creator”, NDTV article, 2024.
- [3] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, Yogesh K. Dwivedi, “Deepfakes: Deceptions, mitigations, and opportunities,” ScienceDirect, Journal of Business Research, January 2023.
- [4] Vejay Lalla, Adine Mitrani and Zach Harned, “Artificial intelligence: deepfakes in the entertainment industry”, USA, WIPO Magazine, June 2022.
- [5] M. M. El-Gayar, Mohamed Abouhawwash, S. S. Askar & Sara Sweidan , “A novel approach for detecting deep fake videos using graph neural network”, Journal of Big Data, SpringerOpen
- [6] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales and Javier Ortega-Garcia, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection”.
- [7] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.
- [8] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “Using capsule networks to detect forged images and videos ”.
- [9] TackHyun Jung, SangWon Kim, and KeeCheon Kim.“Deep-Vision: Deepfakes Detection Using Human Eye Blinking Pattern”. IEEE Access, 8:83144–83154, 2020.
- [10] Park, T., Liu, M. Y., Wang, T. C., and Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2337-2346).