"CRYPTOGUARD: FORTIFYING SOCIAL MEDIA DEFENCES AGAINST DEEPFAKE ONSLAUGHTS THROUGH CUTTING-EDGE CRYPTOGRAPHIC MODALITIES"

AFREEN SIDDIQUI[1], SAAD AHMAD[2], SHWETA SINHA[3]

[1,2] STUDENT SCHOLAR, [3]ASSISTANT PROFESSOR.

[1,2,3] DEPARTMENT OF COMPUTER SCIENCE, NATIONAL POST GRADUATE COLLEGE, LUCKNOW, INDIA

[1]aafreens373@gmail.com

[2]muhammadshad2002@gmail.com

[3]sinha.shweta020776@gmail.com

KEYWORD

CRYPTOGRAPHY, DEEPFAKE, GENERATIVE ADVERSARIAL NETWORK, NEURAL NETWORKS, NON-REPUDIATION.

ABSTRACT

Deepfake is a term used to describe synthetic media or content generated using sophisticated artificial intelligence techniques, specifically employing deep learning algorithms. Deepfakes are a rapidly emerging subject at the intersection of multimedia and artificial intelligence that have experienced a significant surge in interest in recent years. Deep learning algorithms enable the manipulation and production of digital information that is incredibly realistic and difficult to distinguish from actual content, leading to the creation of fake media. An overview of the most recent detection methods, curated datasets for deepfake research, related obstacles, and potential directions for future study are also included in this work, which offers a general understanding of deepfakes and their development. This paper provides a quick overview of the different kinds of deepfake challenges, their development and operation. Moreover, this document presents a resolution to the challenges posed by deepfakes, achieved

through the application of cryptographic methods whose primary goals are confidentiality, integrity, authentication, and non-repudiation, specifically involving encryption and decryption to prevent screen-shots, screen-recording, or any malicious or third-party applications to access the images without the consent of the owner.

1. INTRODUCTION

An artificial intelligence technique called "deepfake AI" is used to produce realistic-looking photo, audio, and video hoaxes. The word, which is a combination of fake and deep learning, refers to both the technology and the phony information that results from it.

Although the practice of fabricating information is not new, deepfakes use sophisticated methods from artificial intelligence and machine learning to edit or produce visual and auditory elements that are easier to trick. There are many types of deepfakes like face-swapping, voice synthesis, text-based deepfakes etc.

Since technology is always changing, finding a long-term solution to the problems caused by deepfake technology is quite difficult. On the other hand, there are a number of steps that may be taken to lessen the dangers posed by deepfakes and to hinder their production and dissemination.

2. DEEPFAKE

The technology in which a picture or video of a person when their body or face has been digitally manipulated to make them look like someone else, usually done for malicious purposes or to disseminate misleading information is termed as DEEPFAKE.

Basically, Deepfake technology can seamlessly stitch anyone in the world into a video or photo they never actually participated in [1]. Deepfakes are created by editing already existing photos and movies to create information that looks real but is completely fake.

The development of sophisticated AI based tools and software that don’t require technical knowledge, has simplified the process of creating deepfakes. Neural networks and generative adversarial networks (GANs) are the two methodologies that are used into the creation process.

In order to stay ahead of this crucial arms race, the research community must stay up to date on the latest advancements in deepfake generation and detection

technologies, given the world's current unprecedented exponential advancement in generative artificial intelligence. [2]

FIGURE 1– THE ABOVE DEEPFAKE IMAGES ARE CREATED FROM THE 'MORPHME APP' – A DEEPFAKE APPLICATION. HERE THE FACES OF A- ARIANA GRANDE AND B- SELENA GOMEZ HAS BEEN SWAPED RESPECTIVELY FROM THE FACE OF THE ORIGINAL IMAGE.

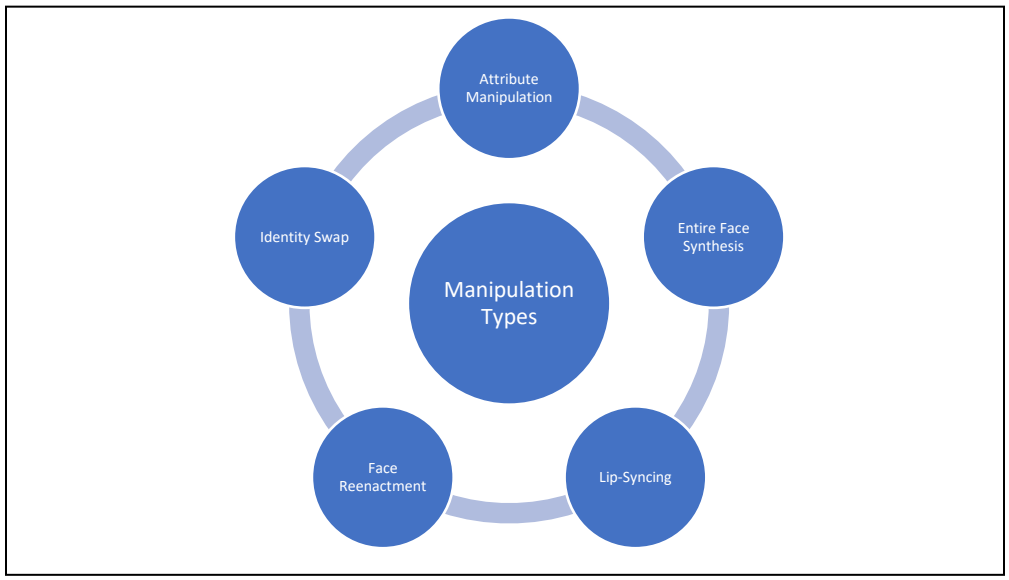


FIGURE 2- THE FIVE PRINCIPAL CATEGORY OF DEEPFAKE MANIPULATION

A deep neural network is fed endless hours of real video footage in order to create a deepfake movie. The network is then "trained" to identify minute rhythms and characteristics of a human. This is done in order to give the algorithm an accurate picture of that person's appearance from different angles.

The next step is to overlay real-time footage of a person with AI-generated facial and voice patterns derived from neural network input by combining the trained learning algorithm with computer graphics technologies.

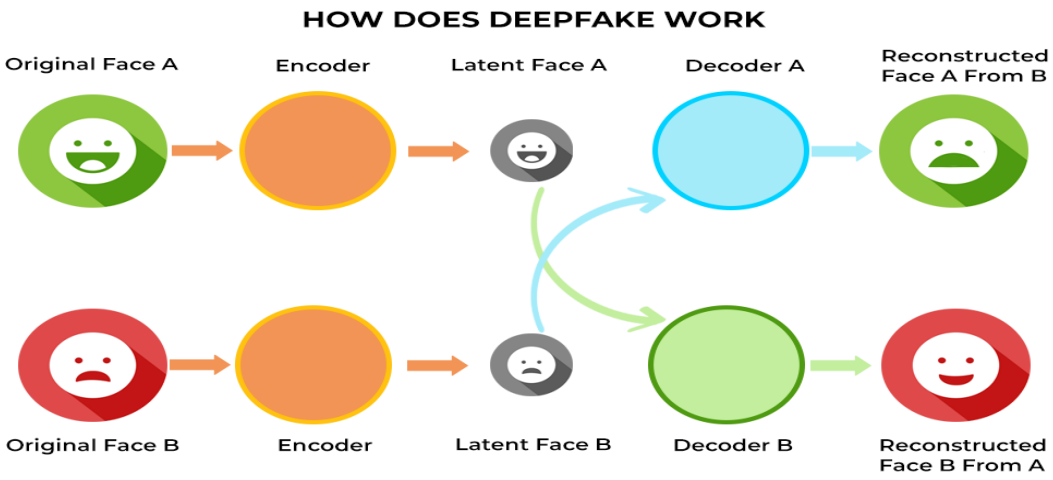
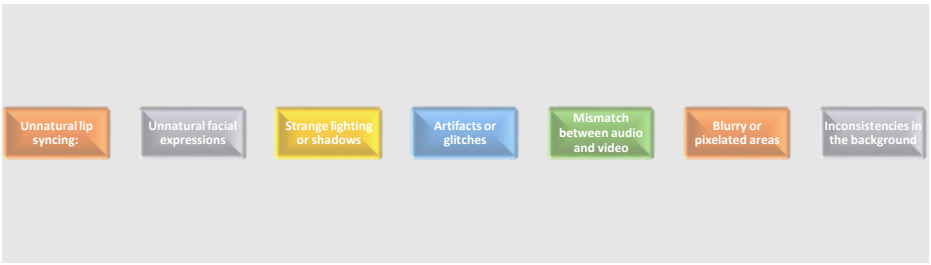


FIGURE 3- WORKING OF DEEPFAKE

FIGURE 4- DEEPFAKE CONTENT C HARACTERISTICS



Hyper realistic deepfakes involve complex techniques, but all the hackers need are access to someone's audio or video content.

This is especially possible given the wealth of original content that is now available, which will feed the algorithm and help it create a convincing deepfake.[3]

TABLE 1- DEEPFAKE FRAUD PREVENTION PRACTICES

TECHNIQUE	DESCRIPTION	SHORTCOMINGS
Use Anti-Fake Technology	Businesses can protect themselves from deepfake assaults by developing a trustworthy method of identifying them, particularly through automated technology. Software for detection fuelled by AI presents such a possibility. The same deep learning methods can also be used to create deepfakes, which are used to show signs that a photo or video has been altered. [3]	False negatives—which fail to recognize real deepfakes—or false positives—which mark authentic information as fraudulent—can result from anti-fake technologies. It can be difficult to balance sensitivity and specificity, and doing so may affect how reliable these systems are. As anti-fake technologies advance, those who create deepfakes might come up with increasingly complex ways to get around these safeguards.
Proper Training and Awareness	An additional line of Défense can be built by raising awareness and providing adequate training. The focus of training should be on how technology is utilized in hostile attempts and how to identify them, for example, by teaching staff members how to spot social engineering attempts that use deepfakes.[6][4]	People might not always be able to tell the difference between actual and deepfake content, even with training, especially as deepfake technology keeps getting more realistic. Certain deepfakes are so realistic that they are hard to spot without certain training or equipment. Awareness efforts may not stay up to current with the most recent developments, which could result in faulty identification of deepfakes and outdated material. Despite the fact that

		awareness and training might encourage users to exercise greater caution, bad actors may still take advantage of weaknesses in people and systems. It's possible that training by itself won't be enough to stop deliberate or focused deepfake attacks.
Enforce Robust Security Protocols	Talk to family members and coworkers about the problems that might arise from deep faking and how it works. Learn how to spot a deepfake for both yourself and other people. Be sure to use reliable news sources and to be aware of the media. Establish strong foundational protocols — "trust but verify." Although being wary of voicemails and videos won't keep you from falling for scams, it will help you stay clear of numerous dangers.[3]	More sophisticated deepfakes can imitate real-life actions and traits, making it difficult for traditional security procedures to identify them. Insider threats, in which people with access to systems mistakenly or purposely aid in the spread of modified content, can also be a component of deepfake attacks. Insider threat risk may not be totally eliminated by security mechanisms. It is possible for new vulnerabilities or exploits (zero-day exploits) to surface, and security procedures may not be ready to address them right away. Such flaws might be exploited by deepfake attacks before security updates are released.
Explore the Use of Blockchain	People can utilize blockchain technology, for example, to validate and confirm the authenticity of a private audio or video file. The likelihood that a film will be accepted as an authentic document increases with the number of persons who electronically sign it. This isn't	Once data is captured, blockchain effectively ensures its integrity and immutability. It does not, however, stop the production of deepfake content. Blockchain does not automatically prevent the creation of false content, but it can assist in confirming the legitimacy of content once it has been

	the ideal choice. Further steps will need to be taken in order to evaluate and account for the expertise of the individuals' casting votes on a file.[3]	recorded. When integrating blockchain-based solutions into current workflows and platforms, users who are not familiar with or opposed to blockchain technology may object. The effectiveness of blockchain-based deepfake prevention techniques may be impacted by adoption issues.
Employ reverse Image Search Tools	Reverse image searches on suspected deepfake photos can be carried out using web resources such as Google photos or TinEye. These resources can assist in locating the image's original source and detecting any manipulation. [7] [5]	Databases containing indexed images are necessary for reverse image search. There may be a false sense of security if the deepfake material is created using unique or never-before-seen data and is not found in the databases. Synthetic faces produced by AI algorithms are used in some deepfakes. Because these artificial faces don't resemble any previously published real-world photos, reverse image search engines might have trouble identifying them. While the primary goal of reverse image search tools is to locate visually comparable photos, they could not reveal the source or author of the content. To address the underlying cause of deepfake fraud and ensure proper attribution, it is imperative to comprehend the source.

Watermarking	It is suggested to use a hybrid watermarking technique that combines strong and weak speech watermarking approaches, offering tamper-proofing and copyright protection. Speech and video features are cross-referenced to offer Défense against potential copy assaults. For tamper-proof recording, the metadata for the content features and embedded watermarks is kept on the blockchain. The purpose of the simulations is to assess how resilient the embedded watermark is to typical signal processing and video integrity threats. [8]	Expert producers of deepfakes may devise methods to modify or eliminate watermarks from manipulated media, particularly if they possess the original watermark or are familiar with the watermarking algorithm. Watermarking is mostly useful for visual media. Other media, like text or audio, which can likewise be manipulated, might not respond well to it. Adversaries may use copy-paste attacks, in which they take a piece of an image that isn't watermarked and paste it over the watermarked section, where the watermarked portion of the image is obvious and recognized. In the event that the watermarking methodology is discovered or made available to the public, attackers may create countermeasures to more successfully erase or alter watermarks.

3. **OUR SOLUTION: THE INTRODUCTION OF THE NEW FEATURE "ALLOW ACCESS"**

Prevention is always better than cure, but these prevention techniques have one or more shortcomings which indeed make them insufficient in the efficient solution of deepfake problems. It is obvious that people can do harm through deepfake only if they have access to one's photo or audio sample. We all know the biggest source of the same is the social media – The photo that we post, the audio that we upload, somehow serve the attackers as the sufficient resource to harm anyone's privacy. Unknowingly we ourselves help the misuser by over sharing our photos, video. They

can save our photos or take screenshots or screen recordings and even can use third party apps in order to access our photos without our permission.

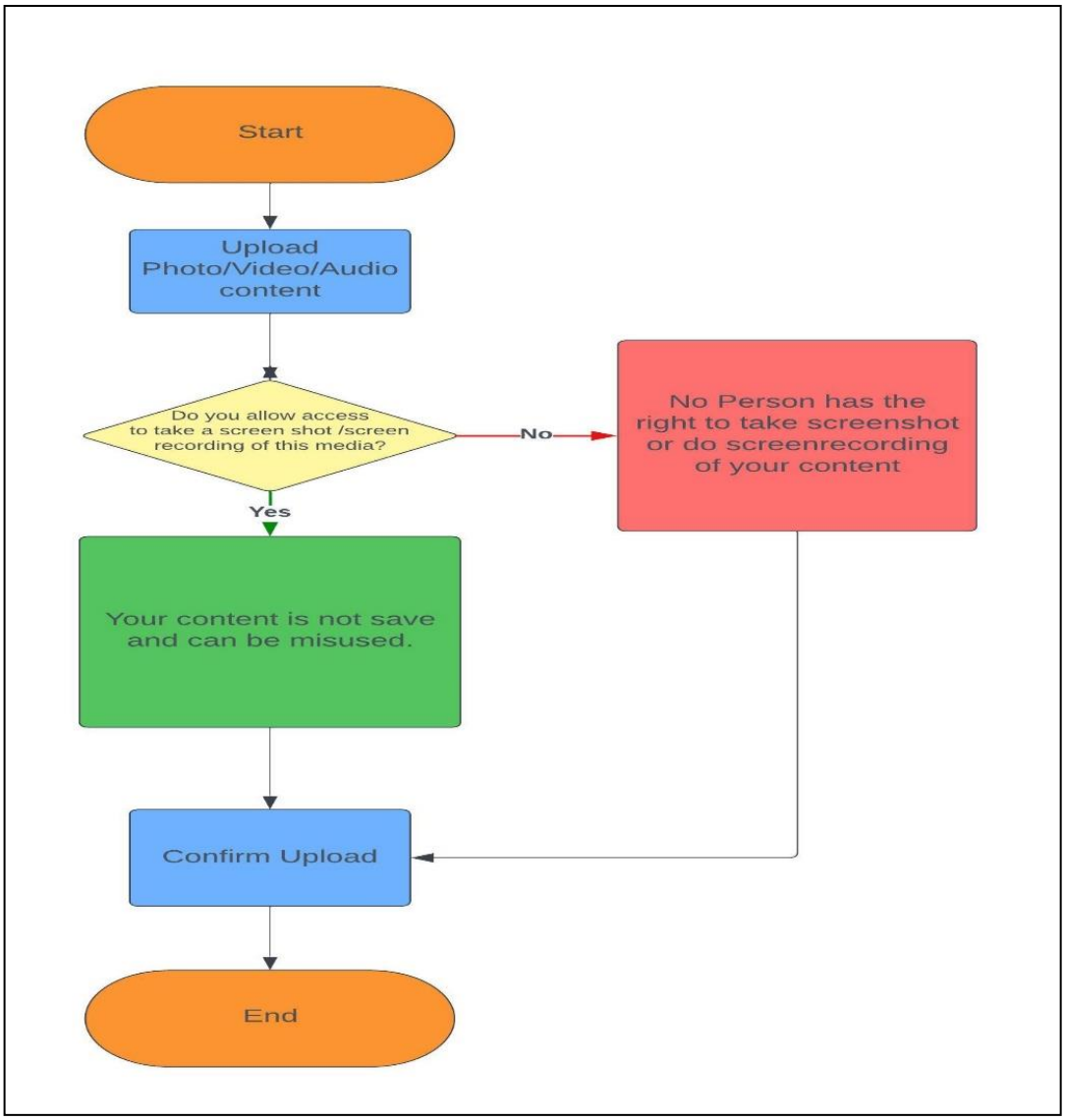


FIGURE 5-FLOWCHART OF THE NEW FEATURE "ALLOW ACCESS"

The Introduction of this new Option would definitely control the malicious activity as attackers will have the shortage of the resources through which they make the fake content. This step will serve as one of the ways to prevent the stealing of anyone's photo without their consent thus, providing the protection.

This option will protect the uploaded content with the help of the guard and protects the content you share, that does not allow the viewers to take screenshot or do screen recording or any third-party malicious application to access it. Therefore, this Option

"Allow Access?" before uploading the photos on any social media applications and if users deny this permission, then that particular content will be guarded and protected.

This mechanism involves the use of Cryptography techniques of encryption and decryption-

3.1 FOR SCREEN SHOT-

- Public Key Encryption: Let I be the original digital image, and(Kpublic,Kprivate) be a pair of public and private cryptographic keys. The encryption process using public key cryptography can be represented as: E=Encrypt_Public(I,Kpublic) This process involves encrypting the image with the recipient's public key, ensuring that only the corresponding private key can decrypt it.
- Private Key Decryption: To display the content, the decryption process using the private key is applied: I=Decrypt Private(E,Kprivate) Here, I is the original image, and E is the encrypted version obtained using the recipient's public key.
- Screenshot Capture: If a user attempts to take a screenshot, they capture the encrypted image E. Without the corresponding private key Kprivate, the captured image is meaningless: Captured Image=E The security of public key cryptography relies on the difficulty of deriving the private key from the public key, making it infeasible for an attacker to decrypt the content without the private key.

The encrypted image alone doesn't reveal the original content without the corresponding key for decryption.

3.2 FOR SCREEN RECORDING-

- Public Key Encryption for Screen Recording: Let *It* be the digital image at time t, and (public,private)(Kpublic,Kprivate) be a pair of public and private cryptographic keys. The encryption process using public-key cryptography for screen recording can be represented as: Et=Encrypt_Public(*It*,Kpublic) Here, *Et* is the encrypted frame obtained by applying the recipient's public key to It.
- Private Key Decryption for Displaying Content: To display the content, the decryption process using the private key is applied to each frame: *It* =Decrypt_Private(*Et*
  ,Kprivate) The original image *It* is obtained by decrypting the encrypted frame *Et* with the private key.
- Public Key Screen Recording Capture: During screen recording, an attacker captures the encrypted frames continuously. The captured screen recording,

denoted as SR, is a sequence of encrypted frames: SR=[E1,E2,…,Et,…] Without the corresponding private key Kprivate, this screen recording captures only encrypted content, making it challenging to understand or reconstruct the original information.

3.2 FOR THIRD-PARTY APPLICATIONS

- Public Key Encryption for Display in Original Application: Let *It* be the digital image at time t, and (Kpublic,Kprivate) be a pair of public and private cryptographic keys. The encryption process using public-key cryptography within the original application can be expressed as: *Et* =Encrypt_Public(*It*,Kpublicapp) Here, *Et* is the encrypted frame obtained by applying the application-specific recipient's public key to *It*.
- Private Key Decryption for Display in Original Application: The decryption process within the original application using the private key: It=Decrypt_Private(*Et*,Kprivateapp) The original image *It* is obtained by decrypting the encrypted frame *Et* with the application-specific private key.
- Public Key Communication with Third-Party Application: For secure communication with a third-party application, the original application can use public-key cryptography to securely share a communication key Kcomm. Let Kpubliccomm represent the public key associated with Ecomm=Encrypt_Public(Kcomm,Kpubliccomm)
- Private Key Decryption by Third-Party Application: The third-party application receives Ecomm and can decrypt it using its private key Kprivatecomm: Kcomm=Decrypt_Private(Ecomm,Kprivatecomm)
- Access Control Mechanisms with Public Key: Introduce access control mechanisms through a function AccessControl using public key cryptography to verify the authenticity of the third-party application: Authorized=AccessControl(AppID,Kpublicaccess)
- Secure Key Management: Ensure secure key management practices, including protecting public and private keys associated with application-specific keys (Kpublicapp, Kprivateapp, Kpubliccomm, Kprivatecomm, Kpublicaccess) from unauthorized access or tampering.
- Terms of Service and Policies: Mathematically express the enforcement of terms of service and policies through a function EnforcePolicy: PolicyEnforced=EnforcePolicy(Policy,AppID,UserAgreed) where Policy represents the terms of service or policies, AppID is the third-party application identifier, and UserAgreed indicates whether the user has agreed to the terms. These mathematical representations aim to capture the cryptographic processes,

access control decisions, and policy enforcement relevant to protecting content in the context of both the original application and potential third-party applications.

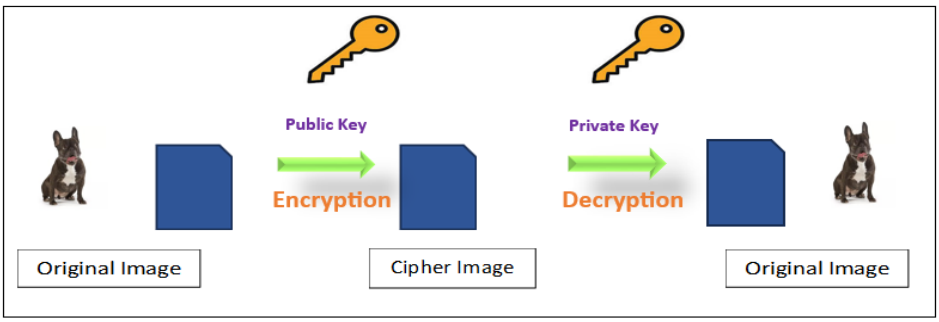


FIGURE 6- THE ACTUAL TECHNIQUE OF CRYPTOGRAPHY OF IMAGE

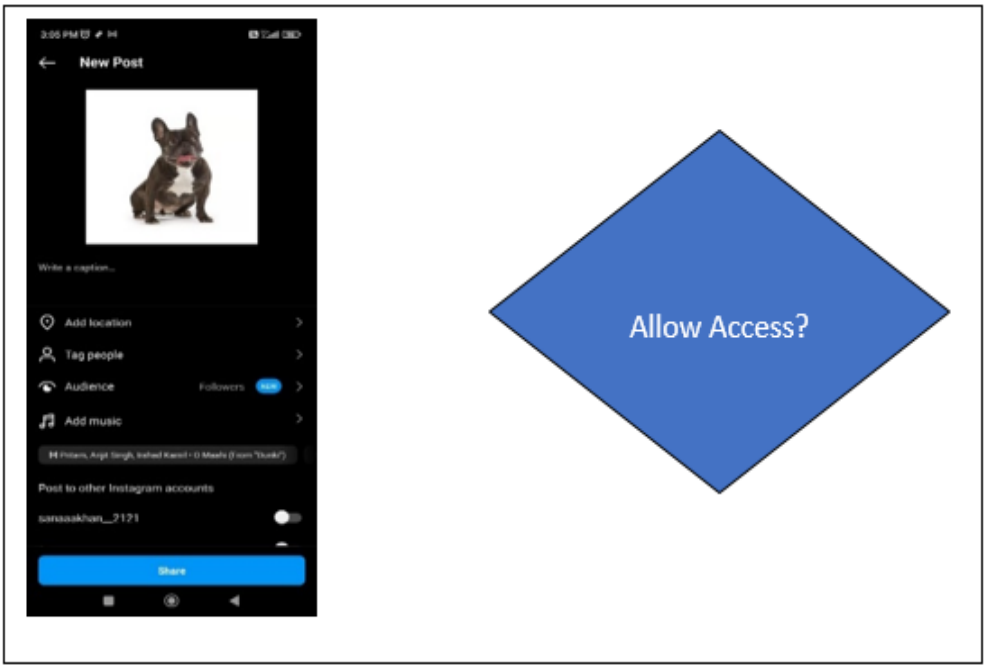


FIGURE 7- INTRODUCE THIS OPTION WHILE UPLOADING THE PHOTO

4. CONCLUSION

Deepfake technology poses significant challenges and concerns to various aspects of society, including privacy, security, and the integrity of information. The potential for abuse and injury is a serious concern, despite the fact that technology presents novel opportunities for creative expression and amusement.

While a permanent solution to completely eradicate deepfakes may be elusive due to the rapid evolution of technology, but an effort can be made to completely eradicate the deepfake problems using Cryptographic techniques discussed in this paper.

While these measures can significantly reduce the impact of deepfakes, it is essential to recognize that technological advancements will likely continue to present new challenges. The public, lawmakers, educators, and technology specialists must work together to develop a proactive, multidisciplinary strategy that addresses the growing threat of deepfake content and promotes safety in the digital realm.

However, only content submitted to social media platforms is protected by this encryption method. This method would not function if the attacker had access to the victim's photo from another source.

The modality will not aid in prevention, for instance, if the attacker attempts to take a picture with a different phone or record a video with a different device, but it will still be highly beneficial and efficient to a considerable degree.

10. References

- IEEE Spectrum - https://spectrum.ieee.org/what-is-deepfake
- Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions
- https://www.spiceworks.com/it-security/cyber-risk-management/articles/what-is-deepfake/
- (https://www.loginradius.com/blog/identity/how-to-identity-mitigate-deepfake-attacks/)
- https://www.linkedin.com/pulse/prevention-deepfakes-deebase-c-1fn7c/
- Vishal Sharma: Blog- Emerging Threads of Deepfakes
- Deebase C : Article on Prevention of Deepfakes
- Amna Qureshi, David Megías and Minoru Kuribayashi: Detecting Deepfake Videos using Digital Watermarking