BlockShield: A TPM-Integrated Blockchain-based Framework for Shielding Against Deepfakes

Venkata K. V. V. Bathalapalli
Dept. of Computer Science and Engineering
University of North Texas.
Email: vb0194@unt.edu

Aakarshan Kumar
Texas Academy of Mathematics and Science
University of North Texas.
Email: aakarshankumar@my.unt.edu

Saraju P. Mohanty
Dept. of Computer Science and Engineering
University of North Texas.
Email: saraju.mohanty@unt.edu

Elias Kougianos
Dept. of Electrical Engineering
University of North Texas
Email: elias.kougianos@unt.edu

Venkata P. Yanambaka
School of the Sciences
Texas Woman's University
vyanambaka@twu.edu

***Abstract*—The increasing threat to individual privacy and personalized digital content on social media posed by Deepfakes has highlighted the importance for a secure and reliable multimedia content integrity mechanism. In this paper a novel Blockchain driven hardware secure video attestation scheme, BlockShield is proposed for Deepfake mitigation. The proposed system includes a novel approach that ensures digital content traceability and privacy using Blockchain smart contracts and TPM's digital signature mechanism. The proposed work explores the scope of hardware-assisted security for Deepfake mitigation through a hardware TPM working together with Blockchain for enhanced digital media protection and sharing. The proposed work is experimentally validated and presented which validates the scope of hardware-assisted and blockchain integrated Deepfake mitigation framework.**

***Index Terms*—Deepfake, Blockchain, Trusted-Platform-Module (TPM)**

I. Introduction

The growing era of Artificial Intelligence (AI) enhances the capability for fake content creation and manipulation. Deepfake, an emerging AI application threat uses neural network architectures to perform audiovisual content modification. Initially these techniques were deployed in entertainment industry and referred to as Synthetic media. As the usage and influence of social media and its content is increasing along with readily available applications that use deep learning techniques for content modification, the AI based digital content modification creation and sharing on social media is increasing at an alarming rate. To counter this, various approaches were considered as potential solutions ranging from regulatory frameworks to digital signature based content attestion. Blockchain emerges as a feasible solution for digital content attestation and verification due to its immutability, decentralized working flow and potential in emerging global technological advancements in AI, Machine learning, Internet-of-Things (IoT) [1], [2].

Any person uploading digital content on social media cannot be sure of its authenticity as an adversary can access it and use various techniques to perform gesture swapping, facial attribute manipulation, or face swapping, and lip syncing. Facial attribute manipulation could be performed by manipulating the original face in the video/image with the target's facial attributes. Deepfake detection and mitigation emerged as alluring research areas for countering the impact of fake content creation and sharing. Deepfake detection and mitigation have become more challenging with advancing Deep learning enhanced techniques and mobile applications enabling a user to perform modification of digital media more realistically. This research work focuses on tracebility of digital content and a presents a sustainable approach for digital media sharing and copyright protection using Blockchain and TPM hardware security module. The conceptual overview of proposed research work is presented in Fig. 1

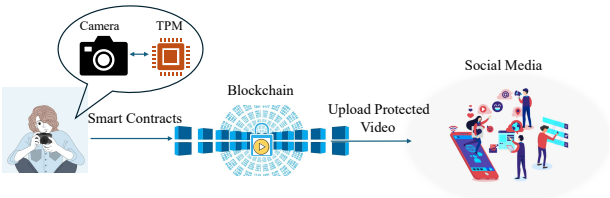


Fig. 1: Overview of BlockShield

Deepfake detection involves using AI-assisted techniques to identify the authenticity of uploaded digital content on social media. Various techniques include monitoring breathing pattern of source in video, analyzing eye blinking rate, and audio inconsistencies. These techniques help in verifying the legitimacy of digital content. However, mitigation can also be a sustainable solution to counter Deepfakes for ensuring trustability. Due to the emergence of various applications and easy access to AI, super realistic synthesis of digital content becomes easy which make mitigation more suitable option than detection to address the digital content integrity issues.

Video-based modification techniques at the frame level, pixel, and block levels could really question the integrity of videos. Video broadcasting can also impact the authenticity due to the unreliable transmission channel. Furthermore, the video tampering also could be done within a specified region of video frame called a block through various techniques like resizing and pixel intensity variation [3]. Due to this, the temporal consistency in a video and connection between consecutive frames is disrupted, which can lead to inconsistency.

Deepfake mitigation involves combating unauthorized access and modification to personalized digital content. The approaches include regulation, digital rights protection, security through watermarking and forensic techniques, and trusted ownership attestation of multimedia content. Digital rights management (DRM) and protection techniques have been the most sought research topics in Deepfake mitigation. Trusted attestation techniques include watermarking, encryption of multimedia, and authenticated communication. However, the computational and performance overhead make it challenging to sustain these resource-intensive cryptographic and watermarking techniques. Furthermore, compression of digital media make it challenging to decode watermarks and could pose a question on watermarking solutions. To address these issues, trusted attestation of video/image through a secure and reliable hardware-integrated distributed ledger framework can guarantee data protection and sharing ensuring authenticity.

This research work presents a broad vision of ensuring a hardware-assisted secure and decentralized framework to attest the digital content with a Blockchain-enhanced trusted content-sharing framework. The TPM hardware security module can be used to perform various tasks, ranging from remote attestation, cryptographic key generation, random number generation, and secure booting [4] in computing applications. It is a secure microprocessor which offers default security and privacy to emerging computing applications and can be an effective approach for ensuring root-of-trust in digital media applications [5]. The proposed research idea to the best of authors knowledge is the first work on Deepfake mitigation that proposes a TPM and Blockchain-integrated framework. The proposed TPM-integrated Blockchain framework securely attests the digital content source by attesting the device capturing video/image using TPM and then securely storing the captured video metadata and the device's TPM-generated digital signature for video inside the blockchain. The proposed work can ensure the integrity of digital content at the hardware source capturing the video/image. The device can be a smartphone or camera with an embedded TPM chip for security.

The rest of this paper is organized as follows. Section II discusses related works on Blockchain based Deepfake mitigation. Novel contributions of the proposed work are presented in Section III. A detialed explanation of the proposed Deepfake mitigation approach is discussed in section IV. Experimental validation results are presented in section V and finally, the conclusion and future research directions are discussed in Section VI.

II. Related Research

This section discussed state-of-art research on blockchain-based Deepfake mitigation approaches. Table. I presents various research works for Deepfake detection and mitigation.

A blockchain-based framework as a proof-of-concept for Deepfake mitigation is presented in [2] where a blockchain-based approach is used for securely uploading videos on social media and to verify video integrity using hash value and video metadata. A Deepfake mitigation approach through Blockchain smart contract is proposed in [6], where a video is uploaded onto a decentralized file storage system and can be edited by secondary artists using smart contract which would be approved by the primary artist. The above work claims to counter Deepfake with a robust smart contract-based video editing process.

Video source tracking and integrity verification using blockchain which allows video source tracking using timestamp is presented in [7] which works based on the Inter-Planetary file system-based approach for video source verification. Machine learning and Blockchain integrated framework to perform Deepfake detection, similarity, and fake news using IPFS storage is proposed in [8]. A smart contract-based video sharing and content copyright protection framework is presented in [9] which works on storing video in a decentralized file storage system that has video metadata and can be used to trace back the original artist. A secondary artist can connect to the smart contract and request video editing permission. This work claims to counter unauthorized access and copyright protection using Blockchain and Hyper ledger fabric. The smart contract based approach in this work is inspired from the above cited works and proposes a TPM-based hardware attestation of video for sustainable content sharing and traceability.

In [10], a Deepfake news mitigation framework based on watermarking and blockchain is proposed which presents a watermarking framework to attest the video. Blockchain supports the forensic analysis of the videos to perform verification of video's legitimacy by retrieving stored video data from decentralized IPFS file storage system. For authenticity of digital media, a Hyperledger Fabric 2.0 and CNN LSTM model as a Proof of Concept is presented in [11] which works on obtaining a hash of digital media captioning which is encoded with CNN LSTM model and securely stores data inside Blockchain.

Blockchain and Non-Fungible Token-based digital content traceability system in [12] works on smart contract-based digital content creation and minted NFT tokens are created for media in Non-Fungible Token Marketplace (NFTM), thereby guaranteeing authenticity to digital media. Furthermore, movie streaming can also be facilitated, video metadata and artifacts can also be stored on Blockchain through smart contracts as discussed in [13].

III. Novel Contributions

The objective of the paper aims to address the video and its metadata integrity ensuring Deepfake resistant personalized

TABLE I: Related Research On Deepfake Detection and Mitigation

Work	Technique	Methodology	Tools
[8]	Deepfake Detection	ML and Blockchain integrated Fake news detection	Efficient Net, Smart Contracts
[14]	Mitigation(Image)	PUF and ML framework for facial feature attestation	Dlib 68 (Facial detection and keypoint prediction), PUF
[10]	Fake news mitigation	Watermarking and Blockchain for Deepfake Video protection	IPFS, MTCNN algorithm, and Face Alignment Network (FAN) algorithm
[15]	Audio Deepfake Mitigation	Fragile speech watermarking with Blockchain	MTCNN, Wav2Lip
BlockShield (Current Work)	Deepfake Video Mitigation	Blockchain and TPM-based video attestation	Hardware TPM, Smart Contracts

digital content sharing through hardware-based multimedia attestation.

A. Research Problems

- Authenticity of digital content
- Reliable digital content traceability mechanism.
- Lack of trusted multimedia content sharing mechanism
- Conventional video protection mechanism using Blockchain incurring high storage and transaction fees.
- Lack of trusted camera authentication and hardware-assisted content attestation mechanism

B. Contributions of the Paper

- A sustainable Deepfake mitigation approach using state-of-art TPM and Blockchain technologies.
- A secure visual Deepfake mitigation approach for individual content privacy and security on social media.
- An energy efficient solution that integrates TPM and Blockchain using smart contracts.
- A secure digital content sharing framework using Blockchain to provide integrity and authenticity.
- An approach based on TPMs digital signature mechanism facilitating hardware root-of-trust for the video/image.

IV. Methodology

This section includes details of the proposed Blockchain and TPM based video attestation technique.

Peer-to-Peer decentralized Inter-Planetary File storage system (IPFS) can be used to the store video and its metadata which is more efficient than Blockchain-based storage which is expensive and computationally infeasible. IPFS generates a unique identifier or hash for a bundle of files and the unique identifier can be used to access and retrieve files from IPFS. Furthermore, Trusted Platform Module (TPM) is a hardware security primitive that performs video attestation by digitally signing the video file before uploading to IPFS. It generates primary root keys like the endorsement key (EK) and attestation identity keys (AIK) which can further generate RSA key pairs for digitally fingerprinting TPM. Furthermore, the primary RSA key pair is generated and used to perform TPM attestation operations using public portion of AIK. The RSA key pair is generated for digitally attesting the video file and stored inside TPMs NVRAM where the sig.rssa contains the generated digital signature of video. The video and its corresponding digital signature file from TPM are stored in IPFS and are included in smart contract as attributes. Proposed attestation framework is illustrated in Fig. 2.

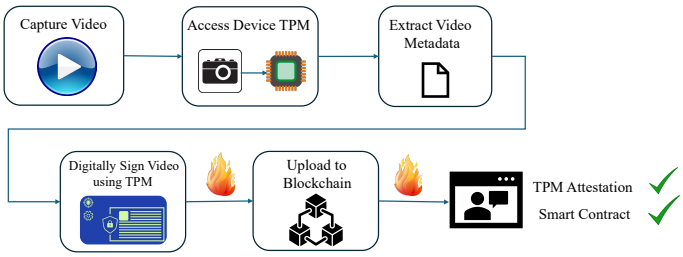


Fig. 2: Secure Video Attestation in BlockShield

Additionally, the video metadata of IPFS is also digitally signed and stored in IPFS for verification. The uploaded video on social media if prone to Deepfake and modified can be easily traced back and verified using TPM since TPM guarantees tamper-proof fingerprinting of digital content. The working flow of TPM-Digital content attestation is illustrated in Algorithm 1, and 2.

Algorithm 1 Performing TPM attestation of digital content

Input: Video File Fi
Output: Digital Signature of Video D_{Fi}
1: Access TPM hardware security module at the camera
Video file Fi *.mp4* $\rightarrow$ *TPM 2.0* $\rightarrow$ *Response* D_{Fi}
2: *tpm2_createprimary -C e -c primary.ctx*
Create Primary Key
3: tpm2_evictcontrol -C o -c primary.ctx 0x81010001
Assign a unique identifier in TPM NV RAM to make it persistent
4: tpm2_create -G rsa -u rsa.pub -r rsa.priv -C 0x81010001
5: tpm2_load -C 0x81010001-u rsa.pub -r rsa.priv -c rsa.ctx.
6: tpm2_evictcontrol -C o -c rsa.ctx 0x81010002
Create RSA keys using primary key and make it persistent
7: Load the video file and Hash it $Fi\rightarrow$ SHA256(Fi)$\rightarrow$ $Fi.hash$
Hash the video file
8: tpm2_sign -c 0x81010002 -g sha256 -o sig.rssa $Fi.hash$
Digitally sign the video hash file using TPM
9: tpm2_hash -C e -g sha 256 -o sig.rssa.hash -t ticket.sig.rssa sig.rssa
Genrate SHA 256 hash of Digital signature for video file

The video, its extracted metadata, and the TPM signature file are then uploaded to IPFS for secure and public access. This action also triggers the smart contract to store the video

and generate the IPFS hash along with the video editing permissions file. From here, an editor can obtain the video from IPFS and request editing permissions to the video by accessing the contract address of original source, triggering a '*requestPermission'* event from the smart contract. When permission is requested, the source user is notified and can either accept or deny permissions to edit. If permission is given, the '*grantPermission*' event of the smart contract is triggered which adds the editor address to the list of editors for the video. Then, the editor can send the edited video to the original artist for authorization. The original source will generate a digital signature of the edited video uploaded to IPFS.

Finally, the uploaded IPFS unique identifier is shared with the editor or secondary user. The edited video's metadata will contain a hash that points back to the original artist's address. This is a crucial step in verifying the chain of edited videos and their integrity. Videos that are not approved by the owner will not contain this point-back hash and are not considered as trusted videos. The working flow of BlockShield is presented in Fig. 3.

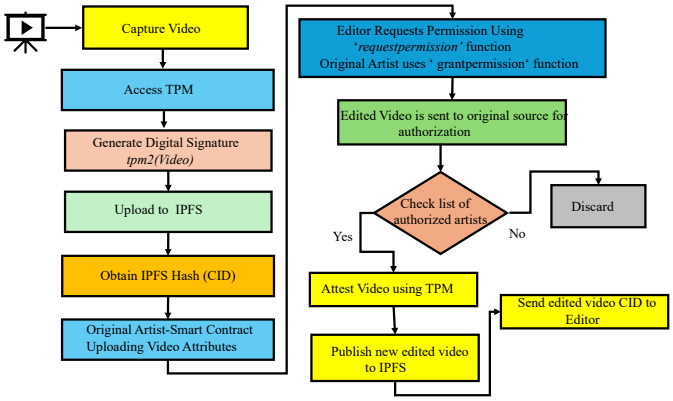


Fig. 3: Working flow of BlockShield

V. Experimental Evaluation

For hardware evaluation, GeeekPi TPM 2.0 module based on Infineon Optiga™ SLB 9670 and compatible with Raspberry pi through is considered. Once a primary key is created, a unique handle in TPM non-volatile memory storage is allocated to the primary key to make it persistent. The primary key in primary.ctx file is assigned 0x81010001 as identifier for accessing in future. Once the primary key is created and persisted, a signing key pair is generated from the primary key and persisted by assigning 0x81010002 as identifier at NVRAM. Once the TPM signing keys are made persistent, they can be used to verify signature since the signing key and primary key have unique identifiers in TPM non-volatile memory. The TPM attestation and digital signature for video are shown in Fig. 4

The tools required for the implementation include metamask, Ganache, Remix IDE, and the Volta Test Network. The implementation used Solidity 0.8.0 inside Remix IDE

Algorithm 2 Blockchain video access control framework

Input: Digital Signature of Video file D_{Fi} and Video File Fi
Output: Digital content is securely stored in Blockchain and secure accessed using smart contract

for Each Primary artist **do**
  Individual contract is called by Primary artist to manage access
  Primary Artist attested Video file information on to IPFS system
  for Each Video File **do**
    Upload video file Fi and Digital signature D_{Fi} on IPFS.
    IPFSfile $Ii \leftarrow$ IPFS.upload(Fi, D_{Fi})
  end for
  ArtistContract.addIPFSHash(Ii)
  Return hash is added as an attribute in the newly created Primary artist contract
end for
Share the contract address and IPFS hash of the video provide access to video file
Editor initiates a 'requestpermission' function to primary artist contract address which is accessible online.
Update list of artists at the primary artist side using 'grantpermission' function
Editor submits a new edited version of video file Fi to primary artist
Edited video file VFi is attested using TPM
$VFi \rightarrow$TPM$\rightarrow F = D_{FVi}$
Upload Edited and attested video file VFi and D_{FVi} to IPFS and share it with secondary artist or editor.

to write the smart contract, the metamask wallet was used to import the two test accounts. These test accounts were imported from Ganache, a local environment that allows for simulated execution of the blockchain environment locally and smart contract testing. and transactions were made through the Volta Test Network. The two addresses used were '0x250805540e2978852a011 237a2bb77e33a0729a4' for the editor and '0x6c27c94191c630438ace1 2e123164a1b628882a6' for the original artist or source.

The front-end deployment of proposed work which involves artists connecting to meta mask wallet uploading video attributes and deploying smart contract is shown in Fig. 5. The performance analysis of BlockShield is evaluated for 5 open-source videos which initially have TPM attestation and then the digital signature file along with video uploaded to IPFS. As soon as it is uploaded, a CID is generated and finally, a unique transaction hash is obtained for each video. Table II presents the obtained evaluation results for BlockShield.

The source user initiates the smart contract with 4 functions: '*uploadMedia', 'requestPermission','grantPermission', and 'editMedia*'. The uploadMedia function takes in the parameters video title, video IPFS hash, and video metadata IPFS hash and stores this data in a hash map with the key as the video IPFS hash. This triggers the MediaUploaded event. Fig. 6 shows smart contract function calls and the transaction validation outputs from viola test network. The 'requestPermission' function takes in the IPFS hash of the video and triggers the 'PermissionRequested' Event The function notifies the source user of the permission request. The 'grantPermission' function takes in the IPFS hash of the video

TABLE II: BlockShield Performance Analysis

Video	Duration (s)	Frame Rate	Bitrate	TPM-Signature	TX Hash	IPFS CID
48255-453831896.mp4	16	24	1633.922	e206fc334fa 48ae5917cac93dff 260d0fc0f0535f4e2 25c932466c 2291833df9	0x4c99d2c8f26 5498b09c53b372 e94f3cefc89d17be 12b401de25bf2b db892609b	QmQjxbmrjFbS7Xz4WXSigtkP msAKRi5FEXdDDpPK1Zn5g1
61299-498228517.mp4	26.559867	29.97	3526.986	fa60e6faf0f64 c50846ac74ca185ffc d83d89fbd68fb 9d2985a6bb5a454eab1a	0x631719a0744dd 4d880924ac7ad 57b98d5d 385a73af7e8e5 4e55039d 4612e723b	QmS5PjDzYqGtfTCYYCm9QrW Fam6ZHZVKU2uw5vJgN6abqf
61706-500316063.mp4	15.65	29.97	937.89	e206fc334fa48a e5917cac93dff260d0fc0f 0535f4e225c93 2466c2291833df9	0x539dfe6fad 4015a6b4ed84 85717614f b9091f7ec0ff1 aa5fc7c93 3a76821b0fe	QmTtS4J4GHG4n2tMEPsfNbVT MuTEmw3uKni5x4djUPEqSN
73711-549547411.mp4	25.2	29.27	1098.274	9c1f7e38f1528cc1876 5b79e28fa76f3fab662d01 63cd309260939 b208d7dcee	0xbc61d12a 52c0815799da10 e0ea8806 28c287a538eb aae65fe857e8 aa0b1435cc	QmNw2PUDAtZ Mt8twnVQ4Kjw7Py 1P4NsZ7dWnak71eVHAK
44645-439940290.mp4	10.88	25	5596.436	9bee3bef81c8ac 3f596a 6f4c44b b218cb713d2 ba2541e89c487a 59362729f60f	0x067ee279182e24 372e49ecd7e5 22551169d23 1a30da152f7 651c1c4ac2 945a6d	QmWVAzf4EJJS 7NHtVweTDyk7gq724y RyqGFCQqTkrVeABY

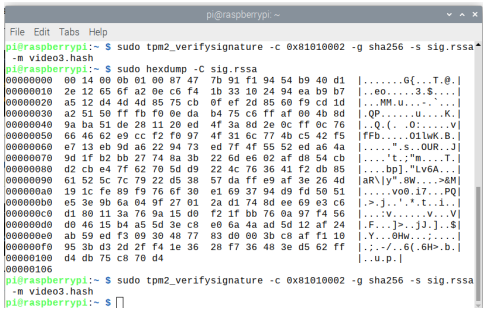


Fig. 4: TPM-Attestation

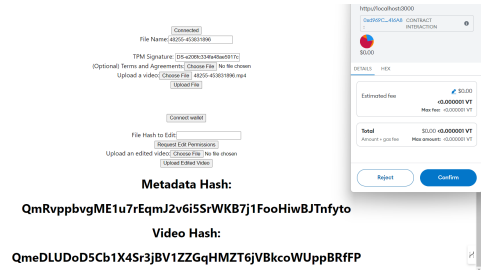


Fig. 5: Frontend Deployment of BlockShield

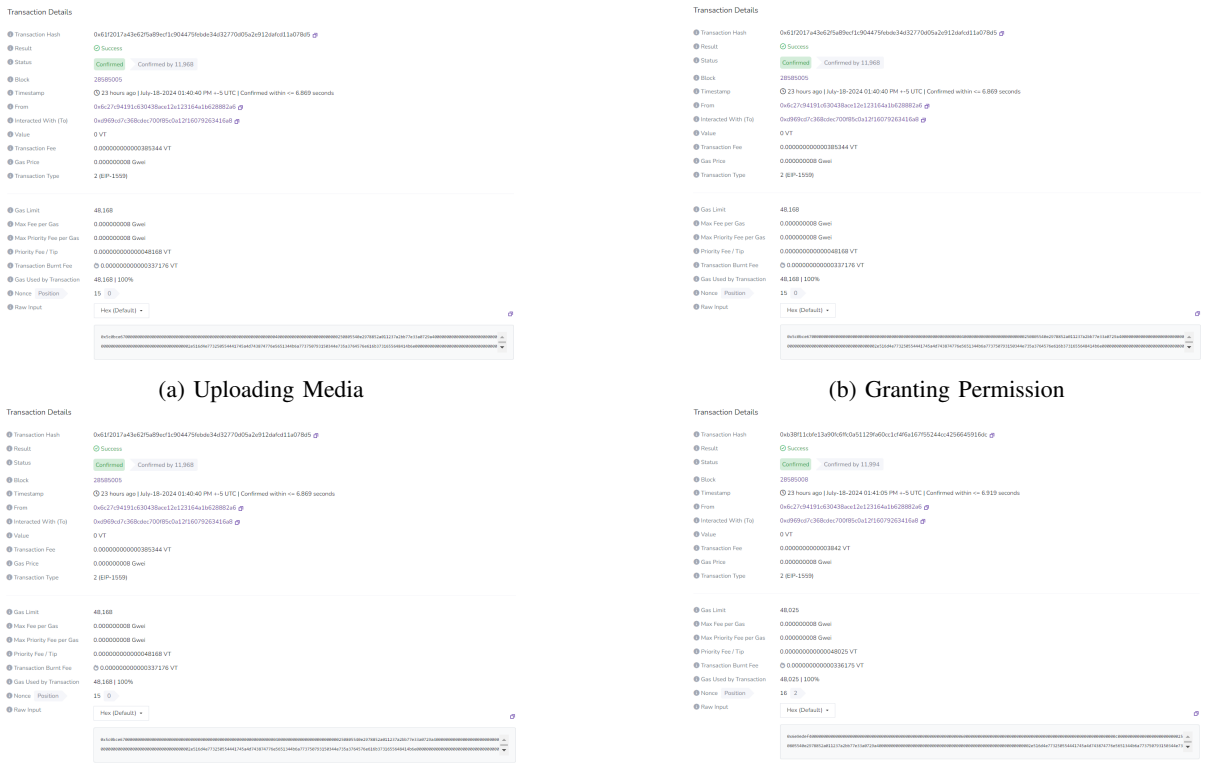
and the editor address, sets the editor as a valid editor for the video, and triggers the 'PermissionGranted' event. The editMedia function takes in the IPFS hash of the video, the IPFS hash of the edited video, and the editor's deployed blockchain contract address and appends the edited video's hash to the chain of edits for the original video. It also triggers the MediaEdited event. If the permissions are given, then the editor can upload a video to IPFS, and once again the video metadata will automatically be extracted from the video and uploaded to IPFS, triggering the grantPermission event of the smart contract and adding the editor address to the list of editors for the video. The metadata will include a hash that points back to the original owner of the video. This is a crucial step in verifying the chain of edited videos and their integrity.

Threat analysis of proposed BlockShield shows that the editing and modifying the video contents can be done only after the authorization of the original source. The editor or secondary user can only get the modified video attested at the original user or source through TPM attestation. The original source generates the digital signature and securely uploads video contents to IPFS, and video metadata is extracted and updated with the secondary user smart contract deployment address. This shows the robustness of proposed approach in enhancing the trustworthiness of digital content source. Furthermore, when uploaded on social media, the media links can also be included in IPFS and accessed using smart contract to ensure source integrity.

VI. Conclusion and Future Research

This research work presented and experimentally validated a Blockchain and TPM integrated approach for Deepfake mitigation through TPM-based hardware digital content attestation. The proposed work with state-of-art TPM-digital signature approach ensures hardware based digital content source attestation facilitated through Blockchain smart contract-based access control approach ensuring digital content authenticity. This is a novel work with TPM attestation and blockchain smart contract for access control and digital content sharing with the substantial performance indicators showcasing the robustness of the proposed Deepfake mitigation approach.

Furthermore, proposed research work could be further applied to Deepfake mitigation for images with effective mechanism for facial feature and biometric-based user authentication. Additionally, this work could be extended to smart cities surveillance applications which work in untrusted environ-



(a) Uploading Media

(b) Granting Permission

(c) Request Permission

(d) Editing Media

Fig. 6: Smart Contract Deployment

ments to guarantee privacy, security and traceability to digital content.

REFERENCES

[1] D. Garg and R. Gill, "Deepfake Generation and Detection - An Exploratory Study," in *Proc. 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, vol. 10, 2023, pp. 888–893.

[2] U. Patil, P. Chouragade, and P. Ambhore, "An Effective Blockchain Technique to Resist Against Deepfake Videos," in *Proc. Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 1646–1652.

[3] U. Patil and P. Chouragade, "Deepfake Video Authentication Based on Blockchain," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2021.

[4] V. K. V. V. Bathalapalli, S. P. Mohanty, E. Kougianos, V. Iyer, and B. Rout, "iTPM: Exploring PUF-based Keyless TPM for Security-by-Design of Smart Electronics," in *Proc.IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2023, pp. 1–6.

[5] S. P. Mohanty, V. P. Yanambaka, E. Kougianos, and D. Puthal, "PUFchain: Hardware-Assisted Blockchain for Sustainable Simultaneous Device and Data Security in the Internet of Everything (IoE)," 2019.

[6] H. R. Hasan and K. Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts," *IEEE Access*, vol. 7, pp. 41 596–41 606, 2019.

[7] J. A. Costales, S. Shiromani, and M. Devaraj, "The Impact of Blockchain Technology to Protect Image and Video Integrity from Identity Theft using Deepfake Analyzer," in *Proc. International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 2023, pp. 730–733.

[8] M. Taeb, H. Chi, and S. Bernadin, "Targeted Data Extraction and Deepfake Detection with Blockchain Technology," in *Proc. 6th International Conference on Universal Village (UV)*, 2022, pp. 1–7.

[9] M. M. Rashid, S.-H. Lee, and K.-R. Kwon, "Blockchain Technology for Combating Deepfake and Protect Video/Image Integrity," *Journal of Korea Multimedia Society*, vol. 24, pp. 1044–1058, 08 2021.

[10] A. Alattar, R. Sharma, and J. Scriven, "A System for Mitigating the Problem of Deepfake News Videos Using Watermarking," *Electronic Imaging*, vol. 32, no. 4, pp. 117–1–117–10, 2020.

[11] C. C. Ki Chan, V. Kumar, S. Delaney, and M. Gochoo, "Combating deepfakes: Multi-lstm and blockchain as proof of authenticity for digital media," in *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*. IEEE, Sep. 2020.

[12] H. R. Hasan, K. Salah, R. Jayaraman, I. Yaqoob, and M. Omar, "NFTs for combating deepfakes and fake metaverse digital contents," *Internet of Things*, vol. 25, p. 101133, 2024.

[13] A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, "Making Sense of Blockchain for AI Deepfakes Technology," in *2020 IEEE Globecom Workshops (GC Wkshps*. IEEE, 2020.

[14] V. K. V. V. Bathalapalli, V. P. Yanambaka, S. Mohanty, and E. Kougianos, "PUFshield: A Hardware-Assisted Approach for Deepfake Mitigation Through PUF-Based Facial Feature Attestation," in *Proceedings of the Great Lakes Symposium on VLSI 2024*, ser. GLSVLSI '24. ACM, 2024.

[15] A. Qureshi, D. Megías, and M. Kuribayashi, "Detecting Deepfake Videos using Digital Watermarking," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1786–1793.