Enhancing Deepfake Content Detection Through Blockchain Technology

Qurat-ul-Ain Mastoi[1], Muhammad Faisal Memon[2], Salman Jan[3], Atif Jamil[4],
Muhammad Faique[5], Zeeshan Ali[6], Abdullah Lakhan[7], Toqeer Ali Syed[8]
School of Computer Science and Creative Technologies, University of the West of England, Bristol, BS16 1QY, United Kingdom[1]
Department of Cybersecurity, Dawood University of Engineering and Technology, Karachi, Sindh, Pakistan, 74800[2, 7]
Faculty of Computer Studies, Arab Open University, Kingdom of Bahrain[3]
Dean of ICS, Dawood University of Engineering and Technology, Pakistan[4]
Department of Telecommunication, Dawood University of Engineering and Technology, Pakistan[6]
Department of Computer Science, Mehran University of Engineering and Technology, Pakistan[5]
Faculty of Computer and Information System, Islamic University of Madinah, Madinah, Saudi Arabia[8]

***Abstract*—Deepfake technology poses a growing threat to the authenticity and trustworthiness of digital media, necessitating the development of advanced detection mechanisms. While AI-based methods have shown promise, they generally face limitations in terms of generalization and scalability. We present a blockchain-enabled watermarking technique, characterized by its immutable, transparent, and decentralized nature, which offers a robust complementary approach for enhancing media authentication through methods such as cryptographic watermarking, decentralized identity, and content provenance tracking. To train and evaluate blockchain-based watermarking and deepfake detection systems, a variety of large-scale datasets are utilized. Video datasets include UADFV (49 real, 49 fake), Deepfake-TIMIT (320 real, 640 fake), DFFD (1000 real, 3000 fake), Celeb-DF v2 (590 real, 5639 fake), DFDC (23,564 real, 104,500 fake), DeeperForensics-1.0 (50,000 real, 10,000 fake), FaceForensics++ (1000 real, 5000 fake), and ForgeryNet (99,630 real, 121,617 fake). Image datasets include DFFD (58,703 real, 240,336 fake), FFHQ (70,000 GAN-generated), iFakeFaceDB (87,000 fake), 100k AI Faces, and over 2.8 million samples in ForgeryNet. Despite integration challenges such as scalability, computational cost, and standardization, blockchain-based solutions show promise in tracking content origin and enhancing verification. Simulation results demonstrate that the proposed blockchain-enabled watermarking achieves a higher accuracy in detecting fake content compared to existing machine learning methods.**

***Keywords*—Blockchain; deep fake; convolutional neural network (CNN); long short-term memory (LSTM); RNN (recurrent neural network); video and image**

I. Introduction

Deepfake technology represents a significant advancement in artificial intelligence, enabling the creation of highly realistic, manipulated, or entirely synthetic media, including videos, photos, and audio recordings [1]. These digital forgeries are crafted using sophisticated machine learning techniques, often employing algorithms such as Generative Adversarial Networks (GANs) [2]. A GAN typically involves two competing neural networks: a generator, which produces fake content, and a discriminator, which attempts to distinguish between real and synthetic media, resulting in increasingly convincing results [3]. Beyond GANs, other AI models, such as autoencoders and diffusion models, also contribute to the creation of deepfakes, highlighting the evolving landscape of this technology. This technology enables various manipulations, ranging from swapping faces and altering facial expressions to synthesizing entire identities and voices [4].

The proliferation of deepfakes is not solely due to technological advancements but also to the increasing accessibility of user-friendly tools and open-source platforms. Individuals with even basic computer skills can now generate convincing deepfakes, leading to a widespread presence of manipulated content online [5]. This ease of creation has amplified the urgency for robust and effective detection methods [6]. The potential for deepfakes to undermine the authenticity of information poses a considerable threat, capable of large-scale misinformation and manipulation [7].

The impact of deepfakes extends across various domains, eroding public trust in digital media and complicating the ability to discern genuine content from fabrications. Malicious applications range from spreading fake news and non-consensual pornography to identity theft, exploitation of public figures, financial scams, influencing elections, inciting social unrest, and even psychological warfare [8]. Notable examples illustrate potential harm, such as the circulation of a false White House tweet and a fake image of an explosion near the Pentagon, causing public anxiety and market fluctuations. Manipulated political speeches and videos have also surfaced, demonstrating the technology's ability to distort public opinion and the political landscape. Furthermore, deepfake fraud has resulted in significant financial losses for both individuals and organizations, as seen in cases of CEO impersonation leading to multi-million-dollar fraudulent transactions. Given the increasing sophistication of these forgeries, relying on human discernment alone is becoming increasingly ineffective, thereby highlighting the critical need for advanced technological solutions in deepfake detection [9].

The above studies have many limitations. Many machine learning methods, such as CNN and other techniques, have a high rate of false positives and consume significantly more time, requiring high scalability of resources during fake detection based on generated GAN methods-enabled data.

- To develop a robust deepfake detection framework that accurately identifies manipulated video and audio content using advanced AI models.
- To integrate blockchain technology for secure, tamper-proof verification and traceability of multimedia content origins and authenticity.
- To ensure decentralized trust and transparency in content-sharing platforms by utilizing smart contracts and immutable watermarking techniques.
- We proposed a novel hybrid system that combines deep learning-based deepfake detection with blockchain-enabled watermark verification.
- We developed a decentralized content validation mechanism using smart contracts to track and verify video authenticity across digital platforms.

The study is organized as follows: Section II is about related work. Section III is the proposed methodology. Section IV discusses the result analysis, and Section V presents the conclusion and future work.

II. Literature Review

In this direction, Kumar et al. [1] combined a blockchain-based AKA mechanism with explainable artificial intelligence (XAI) to secure smart city-based consumer applications. Initially, the involved entities securely communicate with one another to share data via a blockchain-based AKA mechanism. Conversely, they employed the SHapley Additive explanations (SHAP) mechanism to explain and interpret the leading features that significantly impact the decision.

In this context, Muneer et al. [2] emphasizes the prediction of prices of three top cryptocurrencies: Bitcoin, Ethereum, and Litecoin. They employed three machine learning (ML) algorithms: LSTM, SVM, and RF, and utilized the LSTM-RF ensemble to enhance the predictive accuracy of the predicted cryptocurrencies. In comparison to all the models studied, the hybrid LSTM-RF model was the most accurate in predicting the respective performance indices and outperformed the other conventional models and standalone machine learning techniques. Additionally, this work employs Explainable Artificial Intelligence (XAI) techniques to develop AI-based, human-friendly, and interpretable visualizations. This method makes insights accessible to basic users, as it enables investors to make informed decisions based on the outcomes provided by the proposed model.

In addition, Sachan et al. [3] strive to strike a balance between the responsible use and governance of human legal professionals in content generated by Generative AI through periodic audits. The research model conceptualizes the two algorithms. Strategic on-chain (on blockchain) and off-chain (off blockchain) storage of data in adherence to the data protection laws and key demands of stakeholders within a legal firm. Second, comparison auditing of the singular signature as Merkle roots of off-chain stored files to their immutable blockchain equivalent. The system preserves the integrity of data repositories that hold the choices made by an XAI model and a textual explanation of legal cases obtained through a prompt issued to the API of Generative AI, enabling automated auditing.

In addition, Kumar, P. et al. [4] introduced an explainable AI (XAI)-enabled blockchain to enhance the decision-making capabilities of cyber threat detection in the field of Smart Healthcare Systems. Initially, they employ blockchain to secure and store information across various cloud providers through the application of Clique Proof-of-Authority (C-PoA) consensus. Second, a new threat-hunting model based on deep learning is constructed by integrating Parallel Stacked Long Short-Term Memory (PSLSTM) networks and a multi-head attention mechanism to enhance attack detection. The large-scale experiment validates its potential to serve as an advanced decision support system for cybersecurity analysts.

Furthermore, Chen, H. Y et al. [5] give a comprehensive bibliometric description and visualization of the integration of two leading and promising technologies, explainable AI and Blockchain, into Smart Agriculture. Through the integration of artificial intelligence, blockchain technology, and "smart" agriculture, researchers can create a more efficient, more open, and more sustainable food system. They posited that the use of blockchain technology in smart agriculture can enhance transparency by providing an immutable record of a product's origin, production processes, and distribution. It can potentially create a food system that is more beneficial to farmers, consumers, and the environment by improving efficiency, transparency, and sustainability.

Moreover, Salama, R et al. [6] develop a new Blockchain with Explainable Artificial Intelligence-Driven Intrusion Detection for IoT-driven ubiquitous computing systems (BXAI-IDCUCS) model. The primary goal of the BXAI-IDCUCS model is to achieve energy efficiency and security in the IoT setting. The BXAI-IDCUCS model first groups the IoT nodes with an energy-aware duck swarm optimization (EADSO) algorithm to accomplish this. Additionally, a deep neural network (DNN) is employed to detect and classify intrusions within the IoT network. Finally, blockchain technology is used for safe inter-cluster data transmission procedures. To guarantee the productive operation of the BXAI-IDCUCS model, an extensive experimentation study is utilized, and results are measured under various parameters. The comparison study highlighted the superiority of the BXAI-IDCUCS model over existing state-of-the-art methods, achieving a packet delivery ratio of 99.29%, a packet loss rate of 0.71%, a throughput of 92.95 Mbps, an energy consumption of 0.0891 mJ, a lifetime of 3529 rounds, and an accuracy of 99.38%.

Additionally, El Houda, Z. et al. [7] introduce a new framework, referred to as FedIoT, which relies on Explainable Artificial Intelligence (XAI) methods and Blockchain to protect FL-based IDS from IoT networks. FedIoT employs sophisticated XAI methods for detecting local model manipulations and preventing FL-based attacks. Further, we

suggest a blockchain method that employs a lightweight reputation system that guarantees the reliability and credibility of the FL training process. We perform experiments to verify FedIoT, an FL-based intrusion detection system for IoT networks. We use the UNSW-NB15 dataset to verify that FedIoT can successfully identify malicious behavior and enable efficient FL collaboration among multiple users.

Sharma, N et al. [8] proposes a hybrid security system that integrates blockchain for decentralized data storage and an intrusion detection system (IDS) based on a multi-attention deep convolutional recurrent neural network (MA-DeepCRNN) model. The model enhances block creation time and throughput, with minimal transaction reversal probability. The IDS delivers high classification performance, enhancing accuracy and F1-score in IoMT systems, boosting security, scalability, and real-time threat detection in healthcare settings.

Further, Hasan, M et al. [9] attempts to bridge this gap by combining explainable artificial intelligence (XAI) methods and anomaly rules in tree-based ensemble classifiers for the identification of anomalous Bitcoin transactions. In addition, they introduce rules for explaining whether a Bitcoin transaction is anomalous or not. Furthermore, we propose an under-sampling algorithm called XGBCLUS, with the aim of balancing anomalous and non-anomalous transaction data. This algorithm contrasts with other widely employed under-sampling and over-sampling methods. Ultimately, the performances of several tree-based individual classifiers are contrasted with stacking and voting ensemble classifiers. Their experimental results show that: i) XGBCLUS improves true positive rate (TPR) and receiver operating characteristic-area under curve (ROC-AUC) scores over state-of-the-art under-sampling and over-sampling methods, and ii) their ensemble classifiers proposed in this study perform better than conventional single tree-based machine learning classifiers with respect to accuracy, TPR, and false positive rate (FPR) scores.

In addition, Dutta, J. et al. [10] utilizes SHapley Additive exPlanations (SHAP) to offer real-time, model-agnostic explanations of AI predictions to support personalized health monitoring and well-informed decision making in healthcare To enhance data security, a consortium blockchain is used, and AI is applied to recognize block data sensitivity at the edge in blockchain-integrated IoT architectures with Random Forest (RF) algorithm. This method achieves high precision in authenticating block sensitivity, enabling the effective selection of authentication methods in proof-of-authentication (PoAh) consensus within the consortium blockchain. This provides increased security for sensitive information and improves overall blockchain performance. The framework is tested in an edge computing system, exhibiting tremendous potential for promoting eHealth security and authentication, and marking a significant leap in medical technology.

We present Table I to define the findings and limitations of the studies.

TABLE I DEEPFAKE DETECTION BASED ON MACHINE LEARNINGS

Method	Mechanism	Key Strengths	Limitations
CNNs [11]	Analyze spatial hierarchies of features, detect inconsistencies in facial features, alignment, etc.	Effective at identifying spatial anomalies and inconsistencies within frames.	Can be fooled by high-quality deepfakes and may struggle with temporal inconsistencies.
RNNs [12]	Evaluate temporal consistency of movements, speech, and expressions.	Good at detecting unnatural movements and lip-sync issues.	May not be as effective for static images or subtle temporal anomalies.
Visual Artifact Analysis [13]	Detect inconsistencies in lighting, shadows, skin texture, blending boundaries, and other digital artifacts.	Can identify traces left by the deepfake generation process that may not be obvious to humans.	Effectiveness decreases as deepfake quality improves and fewer artifacts are present.
Biometric Analysis [14]	Analyzes unique physiological characteristics like blinking patterns and eye movements.	Exploits inherent human traits that are difficult for AI to replicate perfectly.	Requires high-quality video and may not be effective against deepfakes that accurately mimic these traits.
Multimodal Detection [15]	Combines analysis of audio, video, and potentially other data sources.	Provides a more comprehensive assessment of authenticity by leveraging information from multiple modalities.	More complex to implement and requires synchronized and high-quality data across all modalities.
Metadata Analysis [16]	Examines digital information embedded in media files (e.g., timestamps, editing history).	Can quickly identify inconsistencies that suggest tampering.	Metadata can be easily altered or removed. Relies on the assumption that original media has complete and accurate metadata.
Provenance-based Detection [17]	Examines metadata for signs of manipulation, often relying on digital watermarks.	It can provide strong evidence of authenticity if watermarks are reliably implemented and protected.	Suffers from lack of standardization, potential for removal or misuse of watermarks, and requires buy-in from content creators and platforms. Does not work for content without watermarks.
Inference-based Detection [18]	Analyzes the media content itself for statistical anomalies and patterns indicative of manipulation, without relying on external metadata.	Can detect deepfakes even without provenance information. Adapts to new deepfake techniques such as models are retrained.	Relies on the quality and diversity of training data. Can produce false positives or negatives. The "ground truth" is often assumed, not definitively known.

The above works and studies have limitations, as discussed in Table I. Therefore, we provide new solutions based on blockchain and watermarking techniques to detect deepfakes within data. These studies [19-22] suggest a fake news detection system based on blockchain and cybersecurity approaches, which has been deployed in practice. These studies [23-27] suggested an edge cloud-enabled distributed healthcare system to promote secure healthcare services through telemedicine applications for users.

III. Proposed Blockchain Architecture

Blockchain technology offers a novel and promising foundation for enhancing trust in digital information. Its fundamental principles and key features align well with the challenges posed by deepfakes. At its core, blockchain is a Distributed Ledger Technology (DLT) where data is not stored in a central repository but is instead distributed across numerous computers within a network. This decentralized nature enhances security and resilience by eliminating a single point of failure. A defining characteristic of blockchain is its immutability. Once a record, or block, is added to the blockchain, it becomes challenging to alter or delete. This is achieved through cryptographic hashing and the sequential linking of blocks. Any attempt to modify a block would necessitate recalculating the cryptographic hash for that block and all subsequent blocks. This computationally intensive task is practically infeasible for a malicious actor to accomplish without the consensus of the network. In the event of an error, a new transaction is added to the blockchain to rectify it, ensuring that both the original and the correction remain visible, thus maintaining an auditable history. This inherent immutability provides a robust guarantee of data integrity, making blockchain a strong candidate for verifying the authenticity of digital content. Transparency is another key principle of blockchain. While the level of transparency can vary depending on the type of blockchain (public versus permissioned), transaction records are typically visible to all network participants or authorized parties, fostering trust and accountability. All participants in the network have access to the distributed ledger, ensuring a shared and verifiable record of transactions. This transparency enables the open verification of content provenance, potentially increasing user trust in the authenticity of the media. The decentralized nature of blockchain means that no single entity has control over the network. This lack of central authority reduces the risk of censorship and eliminates single points of vulnerability. The network's distributed structure makes it more resilient to failures and cyberattacks. This decentralization enhances the security and robustness of any deepfake detection system built upon blockchain technology, as it avoids reliance on potentially compromised central authorities. Several key features of blockchain are particularly relevant to deepfake detection. Cryptographic hashing is a fundamental element where data within each block is transformed into a unique, fixed-length string of characters known as a hash, using cryptographic hash functions such as SHA-256 or Keccak-256. Even the slightest alteration to the original data will result in a completely different hash. Each block in the blockchain contains the hash of the preceding block, creating a secure and tamper-evident chain. These cryptographic hashes are designed to be irreversible and collision-resistant, further enhancing data integrity. This provides a mechanism for creating a unique and tamper-evident digital fingerprint for media content, which can be used to verify its integrity. Consensus mechanisms are the protocols by which network participants, or nodes, agree on the validity of new transactions and blocks before they are added to the blockchain. These mechanisms, such as Proof of Work or Proof of Stake, ensure the integrity of the ledger by requiring a majority of the network to validate any new addition. This ensures that any attempt to tamper with the blockchain record of media provenance would be rejected by the network, further strengthening trust in the system. Smart contracts are self-executing contracts where the terms of the agreement are directly written into code. These contracts can automate verification processes based on predefined rules, enabling efficient and trustless verification of content authenticity against established criteria or on-chain records.

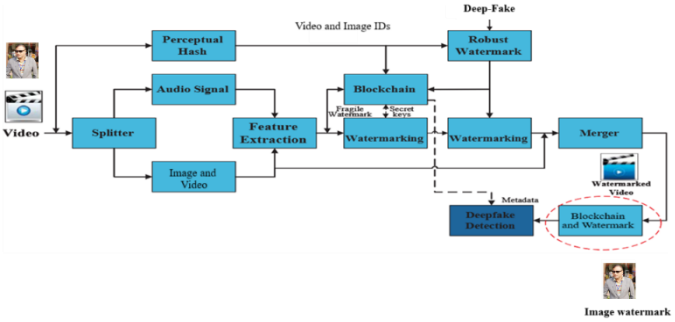

Fig. 1. Proposed blockchain-enabled deep fake detection for different data modes.

Fig. 1 presents a comprehensive framework for deep-fake detection and secure video authentication using blockchain and watermarking techniques. The process begins by splitting the input video into audio, image, and video components, followed by perceptual hashing and feature extraction to generate unique IDs. These features are then used to create fragile watermarks and secret keys, which are managed via blockchain for secure access and traceability. Robust and fragile watermarks are embedded into the video content to ensure both durability and sensitivity to tampering. The watermarked components are merged to produce a final authenticated video. Metadata from this process is passed to a deepfake detection module, which validates the authenticity of the content. The system ensures end-to-end integrity by combining blockchain verification with advanced watermarking, offering a reliable solution against manipulated media. Blockchain technology offers several specific applications that can significantly enhance the detection of deepfake content. One key application is the use of cryptographic hashing to create unique digital fingerprints of original media content, including images, videos, and audio. These hashes can then be securely stored on the blockchain. Subsequently, anyone can compare the hash of a potentially manipulated piece of content with the original hash recorded on the blockchain to detect any alterations. By storing these cryptographic hashes of original content on the immutable blockchain, a permanent and verifiable record of the content's integrity is established. When an original piece of media is

created, its unique hash is calculated and stored on the blockchain. If a deepfake is subsequently generated by manipulating this original content, the hash of the deepfake will invariably be different. Comparing this new hash to the original one stored on the blockchain provides irrefutable evidence of tampering. Blockchain also facilitates decentralized verification mechanisms where multiple participants within a network can independently verify the authenticity of digital content. This process can involve a consensus mechanism where network nodes vote on the authenticity of a given piece of media, potentially incentivized through rewards for accurate verification. Decentralized verification through blockchain can mitigate the risks of bias and single points of failure that are inherent in centralized verification systems. Instead of relying on a single organization or algorithm to determine if content is a deepfake, a blockchain-based system can harness the collective intelligence of a distributed network. This collaborative approach can lead to more accurate and unbiased verification outcomes, as multiple independent parties contribute to the process. Furthermore, blockchain can be used to secure blockchain-based digital watermarking. Digital watermarks containing crucial information about the content's origin and history can be embedded within the media itself, and their integrity can be protected by anchoring them to the blockchain [15]. This provides a tamper-proof method for tracking the provenance of the content [21]. Linking digital watermarks to the blockchain significantly enhances their security and makes them far more resistant to tampering or removal. While traditional digital watermarks may be susceptible to alteration or removal, anchoring the information within the watermark, such as the creator's ID, a timestamp, and the original hash of the content, to the blockchain safeguards its authenticity and integrity through the blockchain's immutable nature.

Ultimately, blockchain-based decentralized identity (DID) solutions can provide a secure and verifiable method for identifying content creators. Creators can cryptographically sign their content using their private key, and this signature can be verified by anyone using their corresponding public key, which is securely stored on the blockchain. This process effectively establishes ownership and attribution of the content. Utilizing blockchain-based decentralized identities can establish clear ownership and provenance for digital content, making it considerably easier to identify manipulated versions and to hold malicious actors accountable for their actions. When a content creator registers their identity on the blockchain and signs their work with their unique private key, it creates an irrefutable link between the creator and the content. This enables the verification of the media's origin and the tracking of its distribution across the internet, significantly aiding in the detection of any unauthorized alterations.

A. Blockchain-Based Approach

Employing blockchain technology for deep-fake detection offers several compelling benefits that address the limitations of current methods, as shown in Algorithm I. Blockchain provides an immutable and transparent record of the origin and history of digital content, significantly improving content provenance. Every modification or instance of sharing the content can be recorded on the blockchain, creating a comprehensive and traceable audit trail. This detailed history makes it considerably easier to identify the authentic version of the content and to detect any subsequent manipulations. The ability to trace the entire lifecycle of digital media on the blockchain significantly enhances content provenance and facilitates the identification of deepfakes. By recording every stage of a digital asset's existence, from its initial creation to any subsequent modifications and its distribution across the internet, a comprehensive historical record is established. This detailed provenance information allows for straightforward verification of whether a piece of content has been altered from its original, authentic state.

The inherent transparency and immutability of blockchain technology can substantially increase user trust in the authenticity of digital media. Users are empowered to independently verify the provenance and integrity of the content they encounter. The inherent trust mechanisms of blockchain can help combat the growing erosion of trust in digital media, caused by the increasing prevalence of deepfakes. In an online environment where distinguishing between real and fake content is becoming increasingly challenging, blockchain offers a technological solution that provides verifiable proof of authenticity. This transparency and the assurance of immutability can help rebuild trust in digital media by empowering users to verify information for themselves, rather than relying on potentially fallible centralized authorities.

B. Algorithm I: Blockchain and Watermark-Based Deepfake Detection Framework

1. Input a video file into the system.
2. Split the video into three components: audio, image frames, and video stream.
3. Generate perceptual hashes for the audio and image/video components to create unique content IDs.
4. Extract visual and audio features from the image and video components.
5. Use the extracted features to create fragile watermarks and secret keys.
6. Store the generated watermarks and keys on a secure blockchain ledger.
7. Embed the fragile and robust watermarks into the video components using watermarking techniques.
8. Recombine (merge) the watermarked audio, image, and video into a final watermarked video.
9. Send the metadata and watermarked video to the deepfake detection module.
10. Analyze the content for authenticity and detect any deep-fake manipulation using the embedded watermarks and blockchain metadata.
11. Verify the integrity and originality of the video using blockchain and watermark data.
12. Output the verified and watermarked video along with detection results.

Blockchain enables the creation of decentralized networks for content verification, allowing multiple independent parties to contribute to and benefit from shared knowledge and

verification efforts. This decentralized approach can lead to more robust and accurate detection outcomes through community-driven models of verification. Decentralized verification on the blockchain can leverage collective intelligence and reduce reliance on potentially biased or flawed centralized systems. By distributing the task of verifying content across a network of participants, a blockchain-based system can tap into a broader range of expertise and perspectives. This collaborative approach has the potential to enhance the accuracy and reliability of deepfake detection, as it mitigates the risk of errors or biases that might be inherent in a single centralized authority or algorithm.

The decentralized and cryptographically secured nature of blockchain enhances the security of provenance data and offers improved privacy. Furthermore, decentralized identity solutions built on blockchain can provide users with greater control over their data used in verification processes. Blockchain can provide a more secure and privacy-preserving infrastructure for verifying media authenticity compared to traditional centralized databases, which are often more vulnerable to breaches and unauthorized access. The cryptographic security inherent in blockchain, coupled with its decentralized structure, makes it a highly secure platform for storing sensitive information related to content provenance. Additionally, decentralized identity solutions built on blockchain can empower individuals to manage their identity data, enhancing privacy throughout the verification process.

Blockchain technology enables the accurate timestamping of data, thereby establishing a precise chronological order of content creation or publication. This timestamping can be particularly crucial for identifying manipulated content, as it provides a verifiable timeline. The timestamping feature of blockchain provides valuable temporal context that can help distinguish original content from later manipulations. Knowing the exact time when a piece of digital media was created and immutably recorded on the blockchain can be a critical factor in identifying deepfakes. For example, suppose a video purports to depict an event that allegedly occurred before the recorded creation time of the video on the blockchain. In that case, it will strongly suggest that the video has been manipulated or is entirely fabricated.

IV. Performance Evaluation

In the simulation of multimodal data for deepfake detection based on blockchain watermarking, a comprehensive set of parameters is configured to ensure robust performance and security. The simulation involves a dataset comprising both authentic and manipulated video and image content, sourced from publicly available benchmarks such as FaceForensics++, Celeb-DF, and Deepfake Detection Challenge datasets. Each video is processed at a resolution of 720p with a frame rate of 30 frames per second (FPS). At the same time, images are standardized to 224x224 pixels for consistency during training and inference using a MobileNetV2-based deep learning model. For watermarking, a unique, tamper-proof digital watermark is embedded into each frame and image using a frequency domain technique, such as the DWT-DCT hybrid or singular value decomposition (SVD), ensuring both imperceptibility and robustness. The embedded watermark is linked to a blockchain network using smart contracts deployed on an Ethereum testnet via Web3.py. Each transaction logs metadata, including the hash of the media content, timestamp, owner ID, and verification key. During detection, real-time feature extraction and watermark verification are performed concurrently, using OpenCV and TensorFlow to compare the embedded watermark against the blockchain-registered hash. The simulation also accounts for multiple sharing scenarios (e.g., social media platforms such as Facebook, Twitter, YouTube, and LinkedIn) and evaluates watermark resilience against compression, resizing, cropping, and format conversion. Performance metrics, including accuracy, precision, recall, F1-score, PSNR, SSIM, and blockchain transaction latency, are monitored to assess the system's capability to detect tampered content. Additionally, a GUI interface (built with Streamlit or Tkinter) visualizes results and logs verified or fake media instances into a local SQLite database or cloud-based storage, ensuring traceability and trust in media authenticity. All the details are shown in Table II, III, and IV.

TABLE II Simulation Parameters

Simulation	Parameters
Language	Ethereum and WEB3
Blockchain	Public Blockchain
Workload	Images, Videos
Training Dataset	Based on Watermarking and Blockchain
Nodes	Client and Social Media
Metrics	Accuracy, Precision, Recall, and F1-Score
Simulation Time	24 Hours

TABLE III Dataset Descriptions

Dataset	Real Videos	Fake Videos
UADFV	49	49
DFFD	1000	3000
Celeb-DF	590	5639
DFDC	23,564	104,500
DeeperForensics-1.0	50,000	10,000
FaceForensic++	1000	5000

TABLE IV Image Datasets

Dataset	Real Images	Fake Images
DFFD	58,703	240,336
FFHQ	60,000	70,000 GAN
iFakeFaceDB	80,000	87,000 GAN
100k Faces	10,0000	100,000 (GAN)
ForgeryNet	1,438,201	1,457,861

A. Result Analysis

Fig. 2 shows the blockchain-enabled watermarking training and validation for "Deepfake Dataset — Real versus Fake Image Counts", offers a comprehensive comparison of real and fake image samples across several widely used deepfake datasets. This visualization provides critical insights into the scale and distribution of data used in deepfake detection research.

The blockchain-enabled weather-marketing trade training based on datasets varies significantly in size and structure:

- Smaller datasets like UADFV maintain a balanced ratio of real and fake images (49 each), useful for initial model testing and evaluation.
- Moderate-sized datasets such as Deepfake-TIMIT, DFFD, Celeb-DF (v2), and FaceForensic++ show a noticeable skew, with fake samples outnumbering real ones by up to 5x in some cases. For example, DFFD contains 3000 fake vs. 1000 real images, and Celeb-DF (v2) includes 5639 fake vs. 590 real images.
- Larger datasets like DFDC and ForgeryNet provide an extensive volume of training material, critical for deep learning. DFDC hosts over 104,000 fake images compared to ~23,500 real ones, while ForgeryNet features nearly 100,000 real and over 120,000 fake samples.
- DeeperForensics-1.0 stands out for its inverse distribution, containing more real (50,000) than fake (10,000) samples — offering an alternative dynamic for model training.

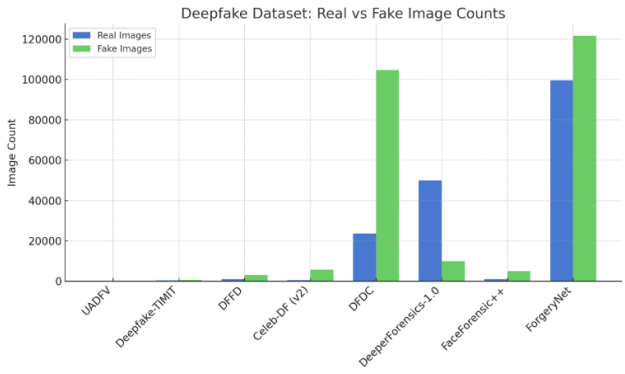

Fig. 2. Result analysis of trained real and fake images based on blockchain watermarking.

This graph highlights the prevalent imbalance between real and fake images in most datasets, underlining the need for careful model training techniques, such as data augmentation, sampling strategies, or synthetic data balancing. Overall, it provides a foundational reference for researchers developing and benchmarking deep-fake detection models.

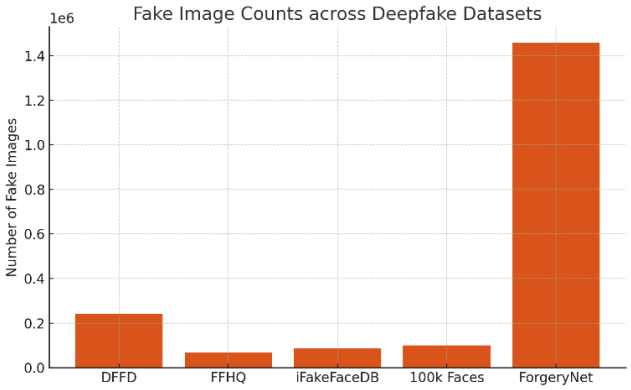

Fig. 3. Image video detection analysis based on blockchain watermarking.

Fig. 3 shows the video dataset training on the blockchain, and watermarking techniques based on the given data.

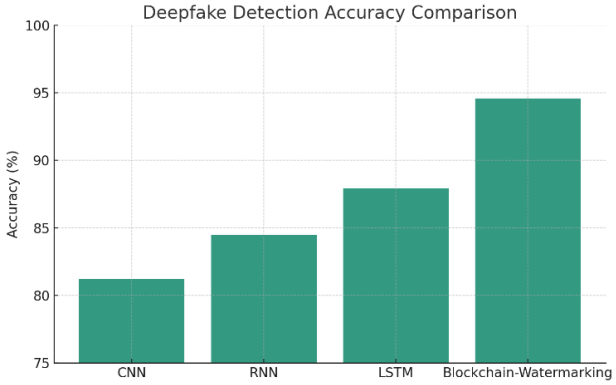

Fig. 4. Blockchain watermarking comparison with existing CNN, RNN and LSTM.

Fig. 4 shows the "Deepfake Detection Accuracy Comparison", which visually presents the accuracy performance of four different models or techniques used for detecting deepfake content. The models compared are CNN, RNN, LSTM, and a Blockchain-Watermarking-based approach. According to the chart, the CNN model achieves an accuracy of approximately 81%, while the RNN model performs slightly better at around 84.5%. The LSTM model shows further improvement with an accuracy close to 88%. However, the blockchain-watermarking approach significantly outperforms the others, reaching an accuracy of nearly 95%. This indicates that the integration of blockchain and watermarking methods greatly enhances the accuracy of deepfake detection compared to traditional machine learning models, suggesting the potential of hybrid systems in improving the reliability and security of deepfake detection frameworks.

V. Conclusion and Future Directions

In conclusion, deepfake technology presents a significant and evolving threat to the integrity of digital information, necessitating the development and implementation of robust detection methods. While current AI-based techniques have made considerable progress, they are not without limitations, including challenges in generalization, reliance on extensive datasets, and a continuous arms race with deepfake generation advancements. This landscape creates a compelling opportunity for complementary solutions, and blockchain technology, with its inherent characteristics of immutability, transparency, and decentralization, offers a promising avenue for enhancing deepfake detection capabilities.

The application of blockchain through cryptographic hashing, decentralized verification mechanisms, blockchain-based digital watermarking, and decentralized identity for content creators can significantly improve content provenance, increase user trust in digital media, and enable more collaborative and resilient verification processes. The benefits of this approach include a tamper-proof record of content origin and history, the empowerment of users to independently verify authenticity, the leveraging of collective intelligence for detection, enhanced data security and privacy, and the provision of crucial temporal context through timestamping.

However, the implementation of blockchain for deepfake detection is not without its challenges. Concerns regarding scalability, computational costs, the complexity of integration with existing systems, the critical need for widespread adoption and standardization, the constantly evolving nature of both deepfake and blockchain technologies, the potential for misuse if cryptographic keys are compromised, and the fundamental issue of ensuring the authenticity of data at the point of initial registration all need to be carefully addressed.

Despite these challenges, the emergence of several promising blockchain-based solutions and initiatives demonstrates the growing recognition of this technology's potential in combating deepfakes. These efforts often focus on leveraging blockchain's immutability for provenance tracking, its decentralized nature for verification, and cryptographic signatures for authentication.

Looking towards the future, several directions warrant further exploration. Research into hybrid approaches that seamlessly integrate the strengths of AI-based detection with the verifiable provenance and security offered by blockchain is crucial. The development of more scalable and cost-effective blockchain solutions tailored for media authentication will be essential for widespread adoption. Establishing industry-wide standards for blockchain-based content provenance and verification will foster interoperability and facilitate broader implementation. The integration of decentralized identity solutions for both creators and consumers of digital media can further enhance trust and accountability. Moreover, future efforts could explore blockchain's role not only in detecting but also in potentially preventing the creation and spread of malicious deepfakes through mechanisms like content registration at the point of creation. Finally, addressing the "garbage in, garbage out" problem by developing robust methods to ensure the authenticity of content before it is recorded on the blockchain will be critical for the overall integrity of these systems.

In conclusion, blockchain technology holds significant potential to play a vital role in safeguarding the integrity of digital media in an era increasingly challenged by sophisticated AI-generated forgeries. Continued innovation, research, and collaboration across various sectors will be essential to fully realize this potential and build a more trustworthy digital landscape where the authenticity of information can be reliably established.

References

[1] Kumar, R., Javeed, D., Aljuhani, A., Jolfaei, A., Kumar, P., & Islam, A. N. (2023). Blockchain-based authentication and explainable AI for securing consumer IoT applications. IEEE Transactions on Consumer Electronics, 70(1), 1145-1154.

[2] Muneer, Kinza, and Ubaida Fatima. "Cryptocurrencies Analytics with Machine Learning and Human-centered Explainable AI: Enhancing Decision-Making in Dynamic Market." International Journal of Computer Applications 975: 8887.

[3] Sachan, S., & Liu, X. (2024). Blockchain-based auditing of legal decisions supported by explainable AI and generative AI tools. Engineering Applications of Artificial Intelligence, 129, 107666.

[4] Kumar, P., Javeed, D., Kumar, R., & Islam, A. N. (2024). Blockchain and explainable AI for enhanced decision making in cyber threat detection. Software: Practice and Experience, 54(8), 1337-1360.

[5] Chen, H. Y., Sharma, K., Sharma, C., & Sharma, S. (2023). Integrating explainable artificial intelligence and blockchain to smart agriculture: Research prospects for decision making and improved security. Smart agricultural technology, 6, 100350.

[6] Salama, R., & Ragab, M. (2023). Blockchain with Explainable Artificial Intelligence Driven Intrusion Detection for Clustered IoT Driven Ubiquitous Computing System.

[7] Abou El Houda, Z., Moudoud, H., Brik, B., & Khoukhi, L. (2023, May). Securing federated learning through blockchain and explainable AI for robust intrusion detection in IoT networks. In IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 1-6). IEEE.

[8] Sharma, N., & Shambharkar, P. G. (2025). Multi-attention DeepCRNN: an efficient and explainable intrusion detection framework for Internet of Medical Things environments. Knowledge and Information Systems, 1-67.

[9] Hasan, M., Rahman, M. S., Janicke, H., & Sarker, I. H. (2024). Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis. Blockchain: Research and Applications, 5(3), 100207.

[10] Dutta, J., Eldeeb, H. B., & Ho, T. D. (2024, September). Advanced eHealth with Explainable AI: Secured by Blockchain with AI-Empowered Block Sensitivity for Adaptive Authentication. In 2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) (pp. 1-7). IEEE.

[11] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An improved dense CNN architecture for deepfake image detection. IEEE Access, 11, 22081-22095.

[12] Sastrawan, I. K., Bayupati, I. P. A., & Arsa, D. M. S. (2022). Detection of fake news using deep learning CNN–RNN based methods. ICT express, 8(3), 396-408.

[13] Gao, J., Micheletto, M., Orrù, G., Concas, S., Feng, X., Marcialis, G. L., & Roli, F. (2024). Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Engineering Applications of Artificial Intelligence, 133, 108450.

[14] Dudykevych, V., Yevseiev, S., Mykytyn, H., Ruda, K., & Hulak, H. (2024). Detecting Deepfake Modifications of Biometric Images using Neural Networks. Cybersecurity Providing in Information and Telecommunication Systems 2024, 3654, 391-397.

[15] Liu, X., Yu, Y., Li, X., & Zhao, Y. (2023). Mcl: multimodal contrastive learning for deepfake detection. IEEE Transactions on Circuits and Systems for Video Technology, 34(4), 2803-2813.

[16] Tran, V. N., Kwon, S. G., Lee, S. H., Le, H. S., & Kwon, K. R. (2022). Generalization of forgery detection with meta deepfake detection model. IEEE Access, 11, 535-546.

[17] Ai, J., Wang, Z., Huang, B., Han, Z., & Zou, Q. (2023, October). Deepfake Face Provenance for Proactive Forensics. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 2025-2029). IEEE.

[18] Hu, J., Liao, X., Liang, J., Zhou, W., & Qin, Z. (2022, June). Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 1, pp. 951-959).

[19] Qaisar, A., Abdullah, S., Ul Rehman, S., Ahmad, S., Syed, T. A., Ali, G., & Jan, S. (2025). Unmasking Fake News: COVAX Reality Dataset Enrichment and BERT Mastery. In Sustainable Data Management: Navigating Big Data, Communication Technology, and Business Digital Leadership. Volume 2 (pp. 623-632). Cham: Springer Nature Switzerland.

[20] Brohi, S., & Mastoi, Q. U. A. (2025). From Accuracy to Vulnerability: Quantifying the Impact of Adversarial Perturbations on Healthcare AI Models. Big Data and Cognitive Computing, 9(5), 114.

[21] Memon, M. F., Matlo, M. A., Siddiqui, A. A., Siddiqui, S. A., Mastoi, Q. U. A., & Lakhan, A. (2025). A Novel Blockchain Proof of Validation Scheme Based on Explainable AI for Healthcare Workload. VAWKUM Transactions on Computer Sciences, 13(1), 54-67.

[22] Brohi, S., & Mastoi, Q. U. A. (2025). AI under attack: Metric-driven analysis of cybersecurity threats in deep learning models for healthcare applications. Algorithms, 18(3), 157.

[23] Zhang, Y., Wang, X. A., Jiang, W., Zhou, M., Xu, X., & Liu, H. (2025). An Efficient and Secure Data Audit Scheme for Cloud-Based EHRs with Recoverable and Batch Auditing. Computers, Materials & Continua, 83(1).

[24] Lakhan, A., Mastoi, Q. U. A., Dootio, M. A., Alqahtani, F., Alzahrani, I. R., Baothman, F., ... & Khokhar, M. S. (2021). Hybrid workload enabled and secure healthcare monitoring sensing framework in distributed fog-cloud network. Electronics, 10(16), 1974.

[25] Lakhan, A., Mastoi, Q. U. A., Elhoseny, M., Memon, M. S., & Mohammed, M. A. (2022). Deep neural network-based application partitioning and scheduling for hospitals and medical enterprises using IoT assisted mobile fog cloud. Enterprise Information Systems, 16(7), 1883122.

[26] Jan, S., Musa, S., Syed, T. A., Nauman, M., Anwar, S., Tanveer, T. A., & Shah, B. (2020). Integrity verification and behavioral classification of a large dataset applications pertaining smart OS via blockchain and generative models. Expert Systems, 38(4), e12611. https://doi.org/10.1111/exsy.12611

[27] Shaikh, J., Syed, T. A., Shah, S. A., Jan, S., Ul Ain, Q., & Singh, P. K. (2024). Advancing DDoS attack detection with hybrid deep learning: Integrating convolutional neural networks, PCA, and vision transformers. International Journal on Smart Sensing and Intelligent Systems, 17(1), 0040. https://doi.org/10.2478/ijssis-2024-0040.