BHODM: A Blockchain-Based Heuristic Deduplication Model for Secure, Transparent, and Redundancy-Aware Cloud Storage

Goutham Kunamalla[1], Buddineni Vinitha[2], Bairi Sai Vardhan[2], Gootla Ganesh[2], Chaparthi Sravani[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

To Cite this Article
Goutham Kunamalla, Buddineni Vinitha, Bairi Sai Vardhan, Gootla Ganesh, Chaparthi Sravani, "BHODM: A Blockchain-Based Heuristic Deduplication Model for Secure, Transparent, and Redundancy-Aware Cloud Storage", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 461-469, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp461-469
Submitted: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

ABSTRACT

Cloud storage systems have become an integral part of modern data management by enabling users to store and access data remotely. However, traditional cloud storage architectures rely on centralized servers, which introduce several critical challenges such as single-point failure, redundant data storage, increased storage costs, and security vulnerabilities. In earlier systems, data was stored in centralized data centers where duplicate files were repeatedly maintained, leading to inefficient utilization of storage resources. Although basic deduplication methods were implemented, they often compromised data confidentiality and lacked transparency in metadata management. Furthermore, failure of the central server could result in permanent data loss. To overcome these issues, this research proposes the Blockchain-enabled Heuristic Optimized Deduplication Model (BHODM), which integrates blockchain technology, InterPlanetary File System (IPFS), Convergent Encryption (CE), and heuristic-based chunking techniques. In this model, files are divided into optimized chunks based on file size using a heuristic approach. Each chunk is encrypted using CE, where the encryption key is derived from the hash of the data itself, allowing secure deduplication without exposing plaintext data. Duplicate chunks are identified through hash comparison, ensuring that only unique data is stored. The encrypted chunks are stored in IPFS, a decentralized peer-to-peer storage network, while metadata such as file names, block numbers, and hash values are securely maintained in an Ethereum blockchain smart contract, ensuring immutability and transparency. The system is implemented using Django, Web3, IPFS API, and AES-CTR encryption. Experimental results based on storage utilization and computation time demonstrate improved efficiency over traditional approaches.

Key words: Blockchain, InterPlanetary File System (IPFS), Deduplication, Convergent Encryption, Heuristic Chunking, Ethereum Smart Contract, Decentralized Storage, Data Security, Storage Optimization.

1.INTRODUCTION

The rapid increase in digital data produced by individuals and organizations has significantly strained traditional cloud storage systems as shown in Figure 1. Although centralized cloud architectures are widely used, they are associated with major limitations such as single points of failure, lack of transparency, and susceptibility to data breaches [1]. These challenges have encouraged the exploration of decentralized cloud storage models, where data is distributed across multiple nodes to enhance fault tolerance, privacy, and system reliability [2].

Blockchain technology has gained considerable attention as a strong foundation for decentralized storage systems due to its key features, including immutability, transparency, and decentralized consensus mechanisms. By removing the dependency on a central authority, blockchain-based storage solutions improve data security and establish greater trust among users [3]. However, despite these advantages, decentralized systems introduce additional challenges, such as increased data duplication, inefficient utilization of storage resources, and uneven distribution of workloads across network nodes. One significant challenge in decentralized storage systems is handling duplicate data efficiently. Data deduplication, a technique that eliminates redundant copies of data, is essential for optimizing storage capacity and reducing network overhead [4]. Ensuring reliable and privacy-preserving deduplication in a blockchain environment requires sophisticated cryptographic methods that safeguard data confidentiality while enabling effective redundancy detection.

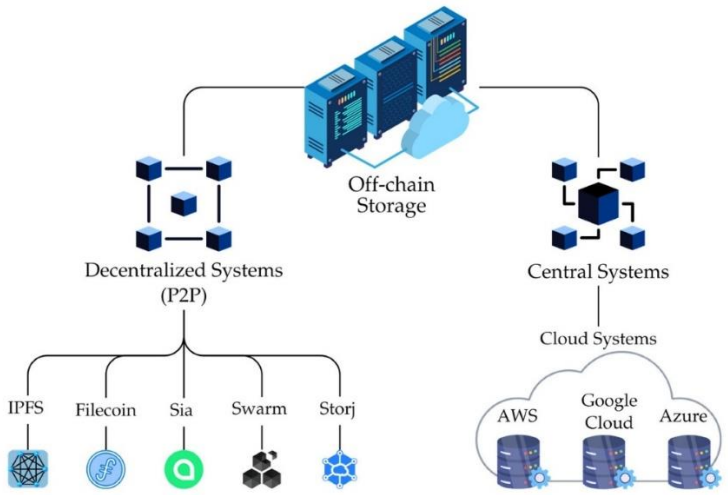


Figure 1: Blockchain tamper-proof, deduplicated, decentralized cloud storage.

Furthermore, to maintain system performance and reliability, it is crucial to balance storage loads dynamically across distributed nodes. Without proper storage balancing, some nodes may become overloaded, causing bottlenecks and reduced availability, whereas others remain underutilized. Effective load distribution algorithms must consider node capacity, network conditions, and data access patterns. This research aims to address these challenges by designing and implementing a BHODM framework that integrates reliable deduplication and intelligent storage balancing [5]. The proposed system not only enhances storage efficiency and security but also ensures equitable resource utilization, contributing to the scalability and robustness of proposed BHODM system solutions.

2. LITERATURE SURVEY

Merlec, et al. [6] proposed a systematic comparative analysis of decentralized storage systems, emphasizing their potential to enhance sustainable DSS. By highlighting the integral role of blockchain technology, this study critically examines various decentralized storage platforms, including Arweave, BitTorrent, Dat Protocol, Filecoin, Hypercore Protocol, IPFS, MaidSafe, Sia, Storj, and Swarm. The analysis covers the key architectural features of these systems, their performance metrics, and their contribution to user data sovereignty. The study aims to comprehensively explain how these decentralized storage solutions allow users to maintain complete control over their data, thus offering a viable alternative to traditional centralized storage methods. Therefore, this paper contributes to ongoing data sovereignty research and guides future developments in decentralized storage technologies. Meng, et al. [7] proposed a decentralized storage system combining Hyperledger Fabric and Inter Planetary File System (IPFS). In addition, from the perspective of security and availability of the decentralized storage system, they study the partitioning and the k-r allocation scheme of the stored

data, propose the allocation function about the stored files, derive the mathematical formula of file security and availability based on the allocation function, and discuss the optimal parameter setting of the allocation function based on the formula to guarantee the high security and availability of the stored files. The experimental results show that the performance of the k-r allocation policy based on the minimum number nodes (MNN) is better than that of the k-r allocation policy based on the minimum slices number (MSN); however, with the same security and availability guarantees, the MNN policy will have more copies relative to the MSN policy, which is relatively wasteful of space.

Li, et al. [8] proposed a blockchain-based decentralized storage system with reliable deduplication and storage balance strategy to provide reliability for deduplicated outsourced data. Encrypted data is split into chunks by a ramp secret sharing scheme, and it is distributed to multiple independent cloud servers. States of the chunks are recorded on the tamper-proofed blockchain, and they can be used to recover the raw data or support the verification of user identity. To balance the distribution of data among storage servers, a heuristic matching algorithm is designed to efficiently allocate the available storage space. The allocation services are published by autonomous smart contracts and other participants gain rewards by giving the best matching results of data chunks and storage servers. Formulation analysis demonstrates the correctness of the proposed scheme in terms of data consistency, integrity, and reliability. Experimental results show that the proposed scheme preserves the confidentiality of outsourced data with acceptable computational consumption. Wang, et al. [9] explored blockchain storage optimization from the perspective of data management, analyzed current techniques such as pruning technique, IPFS optimization, sharding, erasure coding, deduplication, and data compression. It also discusses the challenges in blockchain scalability and provides directions and prospects for future research.

Belfqih, et al. [10] proposed a decentralized authentication protocol leveraging blockchain technology and the IPFS data management framework to provide secure and real-time communication between IoT devices. Using the Ethereum blockchain, smart contracts, elliptic curve cryptography, and ASCON encryption, the proposed protocol ensures the confidentiality, integrity, and availability of sensitive IoT data. The mutual authentication process involves the use of asymmetric key pairs, public key registration on the blockchain, and the Diffie–Hellman key exchange algorithm to establish a shared secret that, combined with a unique identifier, enables secure device verification. Additionally, IPFS is used for secure data storage, with the content identifier (CID) encrypted using ASCON and integrated into the blockchain for traceability and authentication. Feng, et al. [11] proposed an Information-Centric Networking-based blockchain storage architecture. The architecture uses the enhanced resolution system for community division to build blockchain node partitions and store blockchain ledgers in the underlying network. It introduces virtual chain for rapid blockchain indexing and adopts a collaborative block replica deletion algorithm across neighboring partitions, including replica number decision based on blockchain access decay characteristics and replica deletion based on resource relationship. Finally, they compare and analyze the proposed blockchain storage architecture with BC-store and KASARASA, and the results demonstrate that this architecture has significantly lower average access time than others. The replica data volume of this method is reduced by 57.2% compared to the full replica policy, but the access time is only 5.2% slower when compared to the full replica policy, which substantially increases the replica storage utilization.

Kandpal, et al. [12] Analysed academic literature on blockchain technology, emphasized three key aspects: blockchain storage, scalability, and availability. These are critical areas within the broader field of blockchain technology. The study employs CiteSpace and VOSviewer to understand the current state of research in these areas comprehensively. These are bibliometric analysis tools commonly used in academic research to examine patterns and relationships within scientific literature. Thus, to visualize

a way to store data with scalability and availability while keeping the security of the blockchain in sync, the required research has been performed on the storage, scalability, and availability of data in the blockchain environment. The goal is to contribute to developing secure and efficient data storage solutions within blockchain technology. Musa, et al. [13] proposed a thorough review of prior research and identified critical research gaps in the field. Android's dominant position in the mobile market justifies their focus on this platform. Additionally, they delve into the historical evolution of blockchain and its relevance to modern mobile app security in a dedicated section. Their examination of encryption techniques and the effectiveness of blockchain in securing mobile app data storage yields important insights. They discussed the advantages of blockchain over traditional encryption methods and their practical implications. The central contribution of this paper is the Blockchain-based Secure Android Data Storage (BSADS) framework, now consisting of six comprehensive layers. They address challenges related to data storage costs, scalability, performance, and mobile-specific constraints, proposing technical optimization strategies to overcome these obstacles effectively. To maintain transparency and provide a holistic perspective, they acknowledge the limitations of their study. Furthermore, they outline future directions, stressing the importance of leveraging lightweight nodes, tackling scalability issues, integrating emerging technologies, and enhancing user experiences while adhering to regulatory requirements.

Gnana, et al. [14] aimed a flexible direct decentralized symmetry deduplication architecture in a cloud scenario. It first distributes application logic to the contents of the directory through implementation-oriented steering to maintain a deployment location and also attributes the same kind of information to the cloud backup node with the storage node specificity by means of a hand printing-based network model to attain adequate global deduplication performance. They build up a new ownership mechanism during file deduplication to ensure continuity of tagging and symmetrical modeling and verify shared ownership. In addition, they plan an effective ownership policy maintenance plan. In order to introduce a probabilistic key process and reduce key storage capacity, a user-helped key is used for in-user block deduplication. Habib, et al. [15] developed a thorough analysis of blockchain technology in the paper, paying particular attention to its evolution, applications and benefits, the specifics of cryptography in terms of public key cryptography, and the challenges of blockchain in distributed transaction ledgers, as well as the extensive list of blockchain applications in the financial transaction system. This paper presents a detailed review of blockchain technology, the critical challenges faced, and its applications in different fields. Blockchain in the transaction system is explained in detail with a summary of different cryptocurrencies.

3. PROPOSED SYSTEM

The proposed system presents a BHOD model aimed at delivering a secure, decentralized, and efficient cloud storage environment. It addresses the limitations of traditional centralized cloud storage systems by integrating blockchain technology, IPFS-based decentralized storage, CE, and heuristic-based file chunking as demonstrate in Figure 2. The overall operation of the proposed model is carried out in a systematic and sequential manner, as described in the following steps.

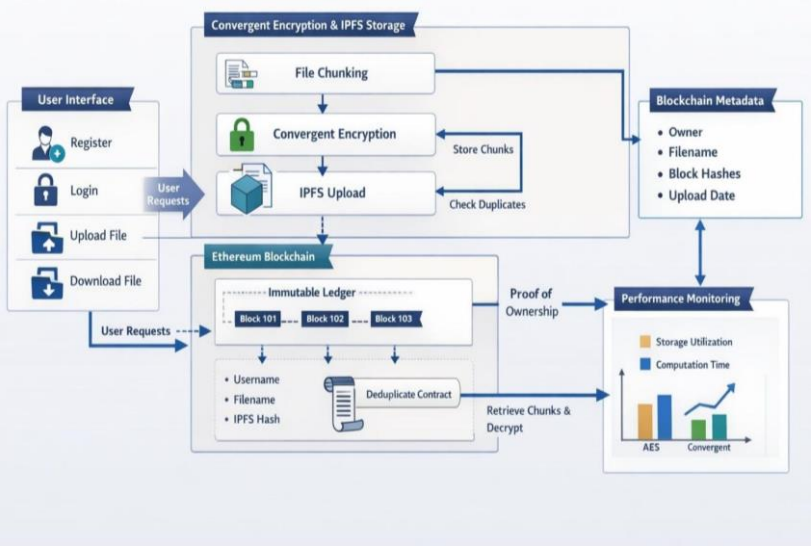


Figure 2: Proposed system architecture of BHODM.

User Registration and Authentication: The system allows users to register and log in through a web-based interface developed using the Django framework. User credentials and profile information are securely stored on the blockchain to ensure integrity and prevent unauthorized modifications.

File Upload Initiation: After successful authentication, the user uploads a file through the application interface. The system reads the file content and determines its size to decide the optimal chunking strategy.

Heuristic-Based File Chunking: The uploaded file is divided into multiple chunks using a heuristic-based approach. The number and size of chunks depend on the total file size, ensuring optimal processing efficiency and reduced computational overhead.

CE of Chunks: Each file chunk undergoes CE, where the encryption key is derived from the hash of the chunk itself. This ensures that identical chunks generate identical ciphertexts, enabling secure deduplication while maintaining data confidentiality.

Duplicate Detection: The system compares the hash of each encrypted chunk with existing hashes stored in the blockchain metadata. If a matching hash is found, the chunk is identified as a duplicate and is not stored again.

Decentralized Storage Using IPFS: Unique encrypted chunks are stored in the IPFS. IPFS provides distributed, content-addressed storage, eliminating dependency on a centralized storage server and improving fault tolerance.

Blockchain-Based Metadata Storage: Metadata related to the uploaded file including filename, chunk identifiers, hash values, and upload timestamp is stored in a smart contract deployed on the Ethereum blockchain. This ensures immutability, transparency, and traceability of data records.

Performance Monitoring and Analysis: The system records computation time and storage utilization during the upload process. These metrics are later used to analyze and compare system efficiency against traditional storage approaches.

Secure File Retrieval: During download, metadata is fetched from the blockchain, encrypted chunks are retrieved from IPFS, and convergent decryption is applied to reconstruct the original file accurately.

Fault Tolerance and Reliability: By distributing storage and metadata across decentralized platforms, the system effectively mitigates single-point failure, ensuring continuous data availability and reliability.

4. RESULTS ANALYSIS

This section presents the results obtained after implementing and testing the proposed BHODM system. The results focus on evaluating the effectiveness of secure deduplication, encryption performance, decentralized storage reliability, and blockchain-based metadata management. Key performance indicators such as storage utilization, computation time, data security, and retrieval accuracy are analyzed to validate the efficiency of the proposed system.

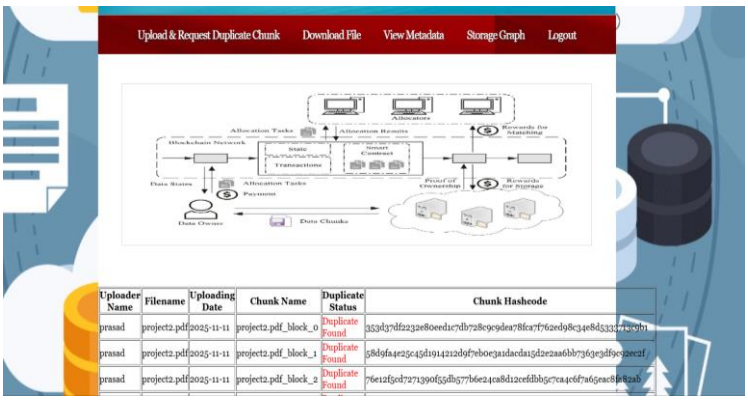


Figure 3: Upload & request duplicate chunks for the proposed BHODM system.

Figure 3 demonstrates the file upload interface along with the duplicate detection mechanism. When a file is uploaded, the system generates a SHA-256 hash for CE and checks for duplicate content. If duplicate chunks are detected, the system avoids re-uploading data to IPFS and instead references the existing CID. This confirms the successful implementation of secure deduplication.

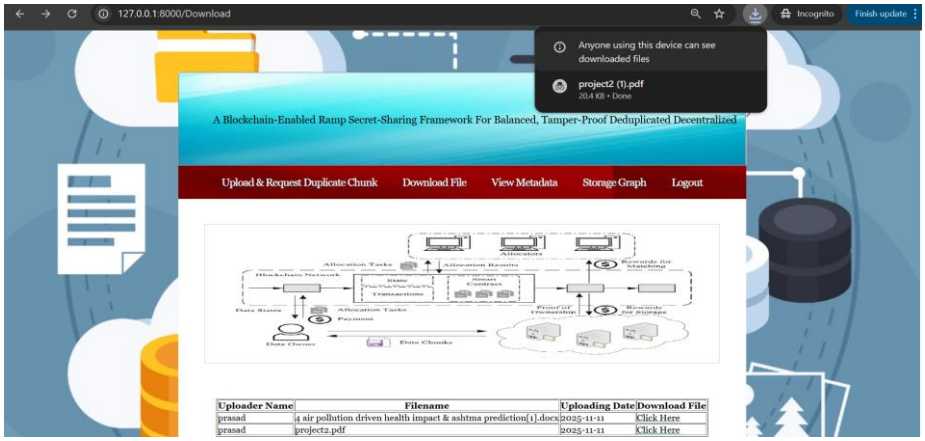
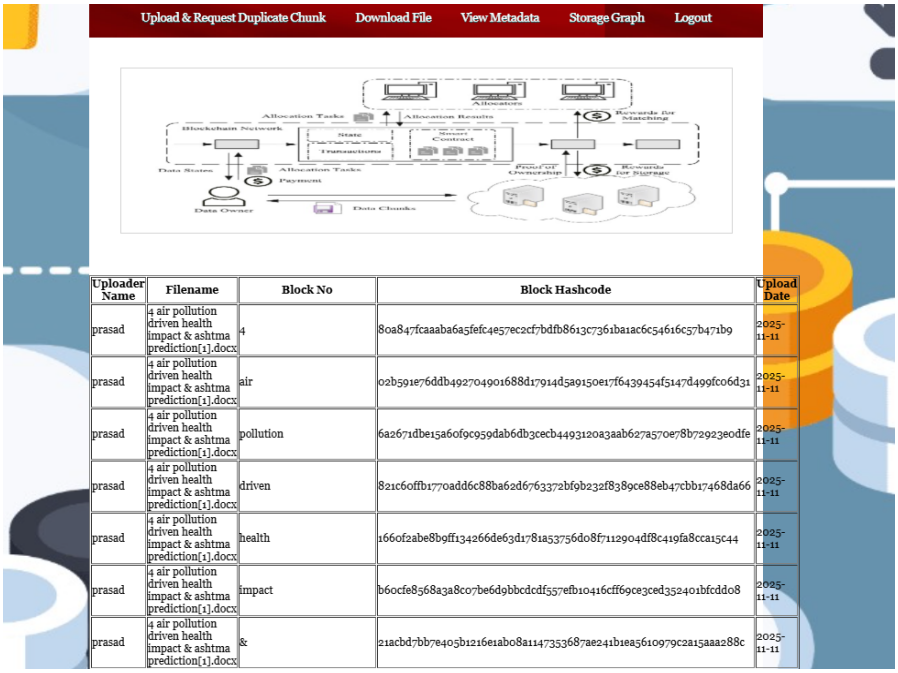


Figure 4: Download file from proposed BHODM system.

Figure 4 shows the file download process from IPFS. Upon request, the system retrieves metadata from the Ethereum blockchain, fetches encrypted file chunks using the CID, and performs AES-CTR decryption. The original file is reconstructed and delivered to the user securely. This verifies data integrity and successful decentralized retrieval.



Uploader Name	Filename	Block No	Block Hashcode	Upload Date
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	4	80a847fcaaaba6a5fefc4e57ec2cf7bdfb8613c7361ba1ac6c54616c57b471b9	2025-11-11
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	air	02b591e76ddb4927049016881d7914d5a9150e17f6439454f5147d499fc06d31	2025-11-11
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	pollution	6a2671dbe15a60f9c959dab6db3cecb4493120a3aab627a570e78b72923e0dfe	2025-11-11
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	driven	821c60ffb1770add6c88ba62d6763372bf9b232f8389ce88eb47cbb17468da66	2025-11-11
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	health	1660f2abe8b9ff134266de63d1781a537756d08f7112904df8c419fa8cca15c44	2025-11-11
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	impact	b60cfe8568a3a8c07be6d9bbcdcdf557efb10416cff69ce3ced352401bfcdd08	2025-11-11
prasad	4 air pollution driven health impact & ashtma prediction[1].docx	&	21acbd7bb7e405b1216e1ab08a1147353687ae241b1ea5610979c2a15aaa288c	2025-11-11

Figure 5: View metadata for the proposed BHODM system.

Figure 5 illustrates the metadata viewing section of the system. Users can view file-related information such as file hash, IPFS Content Identifier (CID), owner address, and transaction details stored on the blockchain. Since metadata is stored via smart contracts, it remains immutable and tamper-proof. This feature ensures transparency and auditability.

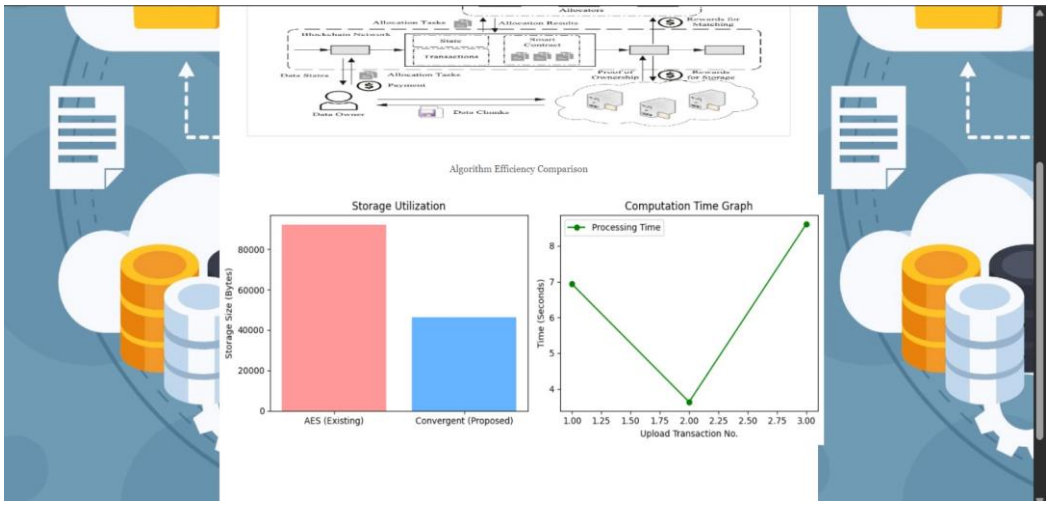


Figure 6: Storage graphs for the proposed BHODM system.

Figure 6 presents a graphical analysis of storage utilization and deduplication performance. The graph compares storage usage before and after deduplication. Results show reduced storage consumption due to CE and chunk-level deduplication. This confirms that the proposed system optimizes storage space while maintaining strong encryption and decentralized architecture.

5. CONCLUSION

The proposed BHODM system successfully integrates CE, IPFS-based decentralized storage, heuristic-based chunking, and Ethereum blockchain technology to provide a secure and efficient cloud storage solution. The system effectively minimizes data redundancy through content-based deduplication, where identical data chunks are identified using hash-based comparison and stored only once. CE

ensures data confidentiality by deriving encryption keys directly from the data, enabling secure deduplication without exposing plaintext information. By storing encrypted chunks in IPFS and maintaining metadata such as file details, block information, and hash values in the blockchain, the system eliminates dependency on centralized cloud servers and removes the risk of single-point failures. The use of smart contracts ensures secure, transparent, and immutable management of metadata and user interactions. The heuristic-based chunking mechanism further improves storage efficiency by optimally dividing files based on size, enhancing deduplication performance. The implementation results demonstrate reduced storage utilization and efficient computation time during file upload and retrieval operations. The system also ensures reliable file reconstruction through secure chunk retrieval and decryption processes. The decentralized architecture improves data integrity, availability, and security compared to traditional cloud storage systems. The research successfully achieves its objective of developing a secure, scalable, and storage-efficient BHODM-based decentralized storage platform.

REFERENCES

[1] Wilkinson, S., et al. (2022). Storj: A Secure Decentralized Cloud Storage Network. Proceedings of the IEEE International Conference on Cloud Computing Technology and Science (CloudCom), 2022.

[2] Xu, X., & Chen, J. (2019). Blockchain-based Secure Data Deduplication Scheme for Cloud Storage. Journal of Network and Computer Applications, 128, 25-34.

[3] Liu, L., & Pu, C. (2021). Decentralized Storage Systems: A Survey. ACM Computing Surveys, 50(3), Article 41.

[4] Wang, F., & Liu, Y. (2020). Load Balancing in Peer-to-Peer Storage Networks. IEEE Transactions on Parallel and Distributed Systems, 28(4), 1234-1247

[5] Faris, H., & Moustafa, N. (2020). A Blockchain-Based Storage Model with Efficient Deduplication and Access Control. IEEE Access, 8, 140733- 140745.

[6] Merlec, M.M.; In, H.P. Blockchain-Based Decentralized Storage Systems for Sustainable Data Self-Sovereignty: A Comparative Study. Sustainability 2024, 16, 7671. https://doi.org/10.3390/su16177671

[7] Meng, L.; Sun, B. Research on Decentralized Storage Based on a Blockchain. Sustainability 2022, 14, 13060. https://doi.org/10.3390/su142013060

[8] Li, J.; Li, Y.; Wu, J.; Zhang, Z.; Jin, Y. Blockchain-Based Decentralized Cloud Storage with Reliable Deduplication and Storage Balancing. IEEE Trans. Netw. Sci. Eng. 2024, 11, 3289–3304, https://doi.org/10.1109/tnse.2024.3369630.

[9] Wang, Y.; Wang, H.; Cao, Y. Comprehensive Review of Storage Optimization Techniques in Blockchain Systems. Appl. Sci. 2025, 15, 243. https://doi.org/10.3390/app15010243

[10] Belfqih, H.; Abdellaoui, A. Decentralized Blockchain-Based Authentication and Interplanetary File System-Based Data Management Protocol for Internet of Things Using Ascon. J. Cybersecur. Priv. 2025, 5, 16. https://doi.org/10.3390/jcp5020016

[11] Feng, H.; Wang, J.; Li, Y. A Blockchain Storage Architecture Based on Information-Centric Networking. Electronics 2022, 11, 2661. https://doi.org/10.3390/electronics11172661

[12] Kandpal, M.; Goswami, V.; Priyadarshini, R.; Barik, R.K. Towards Data Storage, Scalability, and Availability in Blockchain Systems: A Bibliometric Analysis. Data 2023, 8, 148. https://doi.org/10.3390/data8100148

[13] Musa, H.S.; Krichen, M.; Altun, A.A.; Ammi, M. Survey on Blockchain-Based Data Storage Security for Android Mobile Applications. Sensors 2023, 23, 8749. https://doi.org/10.3390/s23218749

[14] Gnana Jeslin, J.; Mohan Kumar, P. Decentralized and Privacy Sensitive Data De-Duplication Framework for Convenient Big Data Management in Cloud Backup Systems. Symmetry 2022, 14, 1392. https://doi.org/10.3390/sym14071392

[15] Habib, G.; Sharma, S.; Ibrahim, S.; Ahmad, I.; Qureshi, S.; Ishfaq, M. Blockchain Technology: Benefits, Challenges, Applications, and Integration of Blockchain Technology with Cloud Computing. Future Internet 2022, 14, 341. https://doi.org/10.3390/fi14110341