International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Cloud Storage With AI-based Intelligent File Management

Achyuth Totteti
Computer Science and Engineering
Geethanjali College of Engineering and Technology
Hyderabad, Telangana

Mokshagna Ponnada
Computer Science and Engineering
Geethanjali college of Engineering and Technology
Hyderabad,Telangana

Manirathnam Dornala
Computer Science and Engineering
Geethanjali College of Engineering and Technology
Hyderabad, Telangana

D. Swaroopa
Sr. Assistant Professor,Computer Science and Engineering Geethanjali college of Engineering and Technology Hyderabad,Telangana

***Abstract*—The rapid growth of digital data and cloud-based applications has increased the need for intelligent file management and efficient storage optimization. Traditional cloud storage systems rely on manual organization and keyword-based search, which often results in inefficient file retrieval and redundant storage usage. This paper presents an AI-driven personal cloud storage framework for intelligent file management and semantic search. The proposed system incorporates file preprocessing, content extraction, and embedding-based feature generation to capture contextual relationships between stored files. A semantic similarity engine is used to retrieve contextually related files, while a duplicate detection module identifies redundant data using hash comparison and cosine similarity. Additionally, the framework includes a storage optimization and analytics module to monitor usage patterns and improve storage efficiency. The system also provides visual insights and recommendations to enhance user understanding and data organization. Experimental evaluation demonstrates improved retrieval efficiency and reduced redundant storage compared to conventional cloud storage methods. The proposed approach offers a scalable and reliable solution for intelligent personal cloud storage, semantic search, and data management applications.**

***Index Terms*— AI-driven Cloud Storage, Semantic Search, Duplicate Detection, File Optimization, Embedding Techniques, Storage Analytics, Artificial Intelligence, Data Management.**

I. INTRODUCTION

Artificial intelligence is transforming data management and cloud storage technologies at a rapid pace. Intelligent file organization, semantic search, and automated storage optimization are some of the key developments driving this transformation. Advances in embedding models, vector similarity techniques, and deep learning–based retrieval systems are enabling computers to understand file content more effectively and manage data with greater accuracy. These technologies allow systems to capture contextual relationships between documents, images, and multimedia files while improving search efficiency and reducing manual effort. As a result, such innovations have been applied in various beneficial areas including personal cloud storage, enterprise document management, collaborative platforms, and intelligent digital archives. At the same time, the rapid growth of data and reliance on cloud storage also introduce challenges related to efficient organization, duplicate data management, and optimal utilization of storage resources.

The rapid growth of cloud storage has resulted in the accumulation of large volumes of digital files that often contain redundant or poorly organized data. Users frequently upload similar or duplicate documents, images, and media files, either unintentionally or due to lack of effective management tools. Such redundancy leads to inefficient storage utilization and makes it difficult to locate relevant information quickly. The widespread availability of cloud platforms and easy file-sharing mechanisms further increases the chances of duplicate data generation, thereby reducing system performance and increasing storage costs. In addition, traditional search methods that rely on filenames or basic metadata are often unable to identify semantically related files, making retrieval less effective. Consequently, intelligent storage management solutions are required to detect duplicate content, improve organization, and enhance overall data accessibility within cloud environments.

Conventional cloud storage and file retrieval techniques mainly depend on manual organization and keyword-based search mechanisms, such as folder hierarchies, filename matching, and metadata filtering. Although these methods are useful for basic file management, they often struggle to identify contextually related content, especially when filenames do not reflect the actual file information. As data volume increases, these traditional approaches become less effective, leading to inefficient retrieval and redundant storage usage. Therefore, there is a growing need for intelligent systems capable of learning contextual relationships automatically from file contents. AI-driven semantic search techniques address this issue by generating embedding representations and modeling similarities between files. These approaches enable the system to detect meaningful relationships, improve retrieval accuracy, and reduce redundancy within large cloud storage environments.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

To address these challenges, the present study introduces an AI-driven personal cloud storage framework that integrates semantic search, duplicate detection, and storage optimization within a unified architecture. The framework is designed to improve retrieval efficiency, storage utilization, and overall data organization. The semantic search component generates embedding representations to capture contextual relationships between file contents, enabling accurate similarity-based retrieval. At the same time, the duplicate detection module identifies redundant data using hash comparison and cosine similarity, which helps reduce unnecessary storage consumption. In addition, an analytics module monitors storage usage and provides insights to enhance system performance. By combining semantic understanding, intelligent duplicate detection, and analytics-driven optimization in a single architecture, the proposed system delivers improved efficiency and transparency, making it suitable for personal cloud storage, enterprise data management, and intelligent file organization applications.

II. LITERATURE REVIEW

Early cloud storage and retrieval systems relied on manually defined metadata features such as filenames, timestamps, file size, and directory hierarchy. These features were processed using traditional information retrieval techniques including rule-based filtering and keyword matching. Although these methods provided basic organization and search capabilities, they often failed when handling large-scale storage systems containing unstructured and heterogeneous data. Such approaches were unable to capture contextual relationships between files, leading to inefficient retrieval and difficulty in identifying semantically similar or redundant content.

With the advancement of artificial intelligence, embedding-based representations have become a widely adopted solution for semantic file retrieval. Embeddings provide vector representations of file content, enabling systems to understand contextual similarities beyond simple keyword matching. Deep learning–based embedding models have demonstrated strong capability in extracting semantic relationships and identifying meaningful patterns within documents. These models automatically learn hierarchical feature representations, reducing reliance on manually engineered rules and improving retrieval accuracy in cloud storage environments.

However, file content within cloud storage environments often exhibits contextual relationships that cannot be fully captured using simple keyword-based or metadata-driven approaches. To address this limitation, embedding-based similarity techniques are employed to represent contextual dependencies between files. These embedding models preserve semantic information by mapping file content into vector space, allowing related files to be grouped based on meaning rather than filenames. Additionally, similarity computation methods such as cosine similarity help maintain contextual relationships and identify relevant files more effectively. Therefore, combining embedding-based semantic search with similarity-based comparison techniques has become a powerful and widely adopted approach for intelligent file retrieval and organization.

Nevertheless, despite the progress made in intelligent cloud storage and semantic retrieval techniques, several challenges still remain in existing systems. Many traditional storage solutions do not scale efficiently when handling large volumes of heterogeneous data, which leads to reduced retrieval performance. Additionally, similarity-based search methods may struggle to generalize across different file types when content extraction is limited. Computational overhead is another concern, as embedding generation and similarity computation can be resource-intensive in large storage environments. Furthermore, many existing platforms lack interpretability and analytics support, providing only basic search outputs without meaningful insights or visualization. These limitations highlight the need for an integrated framework that combines semantic retrieval, duplicate detection, and analytics-driven optimization.

The proposed system addresses these challenges through a modular and interpretable intelligent storage framework. The feature processing mechanism combines content extraction and embedding-based representation, which together with semantic similarity computation enables the system to capture contextual relationships between files. The duplicate detection module identifies redundant data using hash comparison and similarity measures, while the optimization component reduces unnecessary storage consumption. In addition, the system provides explainable outputs such as storage analytics, usage statistics, and recommendations. These features enhance transparency, improve usability, and help users manage data efficiently in personal cloud storage, enterprise file management, and intelligent data organization scenarios.

III. EXISTING SYSTEM

Mostly, current cloud storage systems are based on manual file organization combined with traditional keyword-based retrieval methods such as filename matching, metadata filtering, and directory hierarchy navigation. These approaches rely on user-defined descriptors, such as file names, tags, timestamps, and file type information, to differentiate between stored files. Even though these techniques can perform reasonably well in controlled environments, their effectiveness heavily depends on consistent file naming and proper organization, making them less versatile for handling large volumes of unstructured and continuously growing data.

Recent advances in artificial intelligence have introduced embedding-based semantic search and similarity learning techniques for automated file retrieval and organization. Embedding-based methods are effective at capturing contextual relationships in file content, which helps in identifying semantically related documents within large storage environments. Similarly, similarity computation techniques analyze relationships between stored files to detect redundant or related content. Hybrid approaches that combine semantic embeddings with duplicate detection integrate contextual understanding and similarity analysis for more accurate file retrieval and efficient storage management.

Even with these improvements, intelligent cloud storage and semantic retrieval methods still face several practical

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

challenges. Many existing systems are able to generalize only to a limited extent when handling diverse file types or large heterogeneous datasets. Additionally, embedding generation and similarity computation require significant computational resources, which increases processing time and affects real-time performance. The lack of interpretability is another limitation, as many solutions provide only basic search outputs without offering meaningful analytics or visualization to help users understand storage usage and retrieval decisions.

Such limitations emphasize the need for a scalable, efficient, and interpretable intelligent cloud storage framework. Inspired by these challenges, the proposed system is designed to combine semantic feature extraction, similarity-based retrieval, and analytics-driven optimization in order to enhance retrieval efficiency, storage transparency, and practical usability for personal cloud storage and data management applications.

IV. Proposed System

The proposed AI-driven personal cloud storage system is a complete end-to-end intelligent file management platform that integrates user interaction, automated file processing, semantic search analysis, and analytics-based visualization. While traditional storage systems mainly focus on basic file upload and retrieval, the proposed framework provides a full workflow starting from secure user access to intelligent storage insights and recommendations. The system architecture operates through several interconnected modules, each corresponding to functional operations within the system interface. The entire workflow ensures usability, security, scalability, and efficient storage management performance.

A. User Authentication and System Access

The entry point of the proposed platform is secured through user authentication modules such as registration and login. Users accessing the home page can utilize system features including project overview, account creation, login, and file management operations. During registration, users provide credentials that are validated and securely stored using authentication protocols. Similarly, the login module verifies user identity and allows access only to authorized users, preventing unauthorized usage and ensuring secure handling of uploaded files. After successful authentication, users are redirected to their personalized dashboards.

B. Dashboard and Analysis Management

The dashboard acts as the main control panel for users to monitor the system in real time or review summary reports of storage activities performed by the platform. Users can view the total number of uploaded files, history of search operations, duplicate detection results, and storage usage statistics, along with visual summaries of analytics such as file distribution and storage consumption. These details can be browsed together with corresponding timestamps of operations. This centralized interface allows users to track past activities easily and compare different storage insights in a user-friendly manner, which helps in efficient data organization and management.

C. File Upload and Preprocessing Module

The file management module is designed to allow users to upload their files from supported formats such as documents, images, and text-based content. As soon as a user uploads a file, the system automatically initiates preprocessing operations to standardize the input data. The preprocessing pipeline consists of content extraction for relevant information, metadata generation for consistent indexing, normalization to remove unnecessary characters, and format validation to ensure compatibility. These preprocessing steps help reduce variability caused by different file formats and sources, thereby ensuring reliable feature representation and improving the efficiency of semantic search and duplicate detection.

D. Feature Extraction and Representation

After preprocessing, the system captures meaningful content features that are essential for semantic search and intelligent storage analysis. To further improve retrieval efficiency, complementary feature representations are generated from processed file content. Embedding vectors are created to capture contextual relationships and semantic meaning within documents. These embeddings encode structural information and thematic similarity between files. These embeddings encode content structure and thematic similarity between files. On the other hand, metadata-based features such as file type, keywords, and size distribution provide additional descriptive information, which can be used for identifying related content patterns. A combination of semantic embeddings and metadata representations allows comprehensive analysis of both contextual relationships and structural characteristics of stored files..

E. AI Semantic Search Engine

The central intelligence of the proposed system is achieved using an embedding-based semantic search mechanism that is capable of analyzing contextual relationships between file contents. The embedding generation component extracts semantic representations from processed files and maps them into vector space. It identifies contextual similarities and content relationships while reducing dimensional complexity through vector encoding. As a complementary component, similarity computation analyzes relationships between stored embeddings by comparing vectors using cosine similarity in both forward and backward contextual space. Thus, the system is capable of detecting semantically related files, content similarities, and contextual associations within storage. Such a combination of embedding representation and similarity-based comparison enables the system to learn contextual and structural relationships simultaneously, resulting in improved retrieval performance compared to traditional search methods.

F. Duplicate Detection Output

After performing similarity analysis, the system produces results that indicate whether the uploaded file is unique or duplicated within the storage. The output interface displays the duplicate label together with similarity scores, percentage match, and file details obtained during comparison. Duplicate files are typically characterized by highly similar content

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

representations and minimal variation in semantic structure. On the other hand, unique files exhibit distinct contextual embeddings and varied content patterns that differentiate them from existing stored data.

The duplicate detection results are further integrated with the analytics dashboard to provide detailed visualization of redundancy within the storage system. Users can view grouped duplicate files along with similarity percentages, file sizes, and timestamps, which helps in understanding how redundant data is distributed. The system also provides recommendations such as removing exact duplicates or reviewing semantically similar files before deletion.

G. Analytics Visualization Module

To evaluate storage behavior and system performance, the platform includes a module that visually represents different aspects of storage usage and retrieval efficiency. Through these visualizations, users can track changes in storage consumption over time, observe duplicate percentage variations, and analyze file distribution patterns. The displayed graphs confirm that storage utilization is optimized and redundancy is minimized, providing evidence of the system's efficiency and its ability to manage data effectively.

H. Analytics and Explainability

One of the major capabilities of the proposed system is analytics-based explainability, which makes storage management decisions transparent to users. The system analyzes file characteristics, identifies redundancy patterns, and highlights differences between unique and duplicate content through comparative visualizations such as similarity scores, storage distribution, duplicate percentages, and usage trends. These interpretability features provide users with insight into how retrieval and optimization decisions are made; in other words, storage recommendations become more credible and easier to understand for efficient data management.

I. System Workflow Summary

The entire process of the proposed system begins with user authentication and access to the main dashboard, followed by file upload and validation. The uploaded file is first preprocessed and standardized, after which semantic embeddings and metadata features are extracted for further analysis. The similarity-based semantic search engine performs contextual comparison to identify related files and detect duplicates with confidence scores. Finally, the system presents analytics visualizations along with detailed storage insights that assist users in understanding retrieval results. With this integrated workflow, the proposed system provides an efficient, transparent, and reliable solution for intelligent personal cloud storage and data management.

V. SYSTEM ARCHITECTURE

The overall structure of the proposed AI-driven personal cloud storage system is designed as a layered architecture to maintain modularity, improve scalability, and enable efficient handling of file data. Different layers of the architecture are responsible for executing specific operations that together provide intelligent storage management and semantic retrieval. The architecture consists of four major layers described as follows.

1) User Interaction Layer

The User Interaction Layer acts as the entry point of the proposed system and provides users with a way to upload and manage files. It enables smooth interaction between the end user and the storage platform through graphical or web-based interfaces.

Users can upload files in supported formats, initiate semantic search operations, and view duplicate detection and analytics results.
Input validation procedures play an important role in verifying uploaded files against system requirements such as supported formats, file size limits, and data integrity. This layer also performs functions such as request handling, session management, and secure data transmission to prevent unauthorized access. By abstracting technical complexity, the User Interaction Layer improves usability and allows users to efficiently manage storage operations without requiring advanced technical knowledge.

2) Processing Layer

The Processing Layer is responsible for preparing uploaded file data for semantic analysis and similarity computation. It performs multiple preprocessing tasks such as content extraction, text normalization, metadata generation, format validation, and indexing. These steps help reduce variability across different file formats and maintain consistent data quality for files originating from various sources.
Additionally, this layer transforms raw file content into structured representations suitable for embedding generation and similarity-based retrieval. Data refinement operations such as keyword extraction and content filtering may also be applied to improve retrieval performance. The Processing Layer acts as a bridge between user input and computational analysis, ensuring that subsequent modules receive standardized and optimized data for efficient storage management.

3) Analysis Layer

The Analysis Layer forms the core of the proposed system and is responsible for semantic embedding generation and similarity-based retrieval. It utilizes embedding techniques to analyze contextual relationships within file content. The embedding component extracts semantic patterns and converts processed data into vector representations suitable for similarity computation.
The generated vectors are then passed to similarity comparison modules. These modules capture relationships between files by evaluating contextual similarity in both direct and related content spaces. As a result, the system is capable of identifying related documents, redundant files, and contextual associations within storage.
This layer also includes indexing, similarity scoring, and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

ranking mechanisms along with optimization techniques to improve retrieval accuracy and reduce computational overhead. These components help enhance system efficiency and ensure consistent performance during large-scale storage analysis.

4) **Output Layer**

The Output Layer is responsible for generating the final storage analysis results and presenting them in a human-readable format

The similarity and analytics modules produce outputs that indicate whether files are unique, duplicated, or semantically related. Confidence scores and similarity percentages are calculated to quantify relationships between stored files.
This layer not only displays retrieval results but also generates visualization outputs such as storage usage graphs, duplicate percentage charts, and analytics summaries that improve transparency and usability. Moreover, the results may be stored as reports or integrated into external applications for monitoring and data management. The Output Layer ensures that system decisions are not only accurate but also understandable for practical deployment.

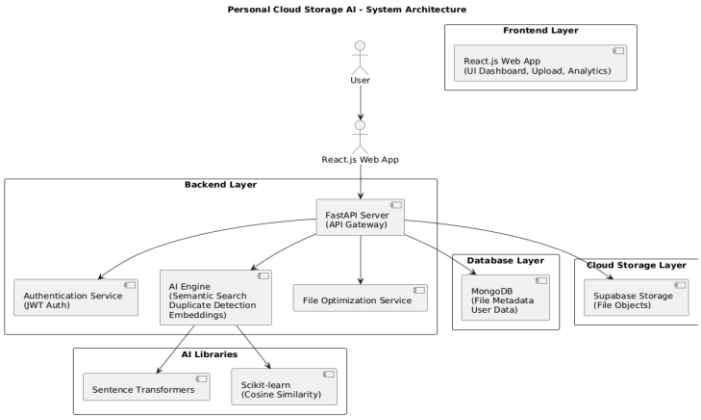

Fig. 1: system architecture

VI. IMPLEMENTATION

The proposed AI-driven personal cloud storage system was designed as a comprehensive intelligent file management platform that combines a user-friendly interface, secure authentication modules, and semantic search with storage analytics. The implementation involves user interaction, file preprocessing, similarity-based analysis, and generation of storage insights and optimization results.

A. Dataset Description

The proposed AI-driven personal cloud storage system is evaluated using a dataset consisting of various file types such as documents, PDFs, text files, and sample media files uploaded by users. To analyze system performance, the dataset is divided into testing and validation sets to evaluate retrieval accuracy and duplicate detection efficiency.

B. System Configuration Parameters

The proposed intelligent cloud storage system operates using standard configuration parameters for semantic search and similarity analysis. The configuration process includes selecting embedding methods, defining similarity thresholds, setting indexing parameters, and specifying storage optimization rules to achieve efficient performance.

C. System Interface and Workflow

The home page serves as the main gateway of the system and offers links to project information registration login, dashboard, and analysis modules. A user-friendly interface has been created to make it easy and secure for all users to access the system.

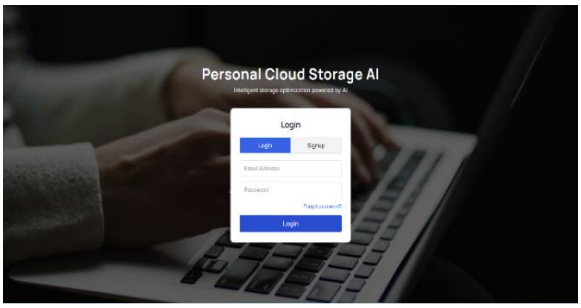

Fig. 2: AudioShield Home Page Interface

Users must first create an account through the registration module, which validates user inputs and securely stores credentials.

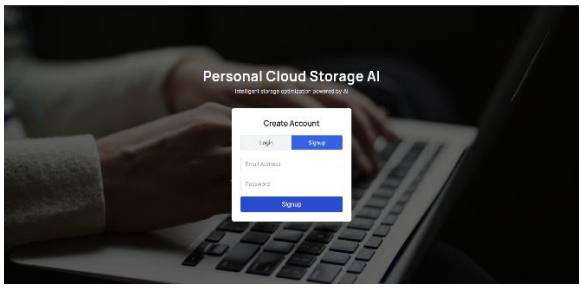

Fig. 3: User Registration Page

After registration, authenticated users access the system through the login interface, ensuring restricted and secure usage.

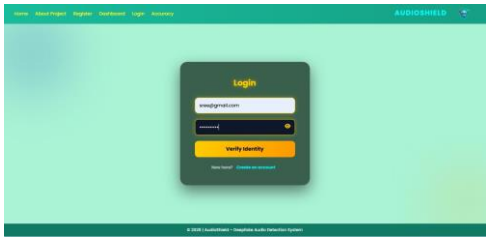

Fig. 4: User Login Page

D. Dashboard and Analysis History

After successful login, users are redirected to their personalized dashboard where they can view system statistics and file activity history. The dashboard displays the number of uploaded files, duplicate detection results, storage usage, similarity scores, and timestamps of operations. This

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

information allows users to track previous file uploads, monitor storage analytics, and review retrieval results efficiently.

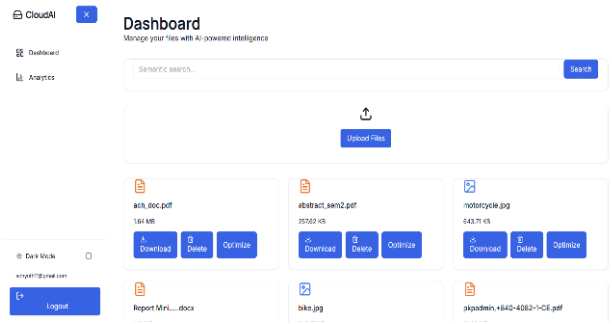

Fig. 5: User Dashboard with file uploading and semantic search

E. File Upload and Preprocessing

The file management module allows users to upload files in supported formats such as documents, text files, and media content. Uploaded files undergo preprocessing operations including format validation, content extraction, normalization, and metadata generation before semantic feature extraction.

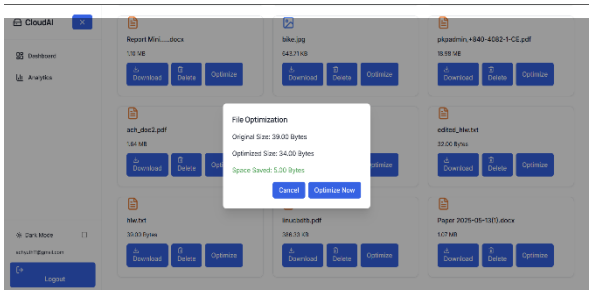

Fig. 6: Voice Analysis Module – Audio Upload Interface

F. Duplicate Detection Case Study

A sample file containing similar content was analyzed using the semantic similarity and duplicate detection module. The system identified the file as a duplicate and displayed similarity percentages along with contextual comparison results such as content overlap and embedding similarity.

Fig. 7: Deepfake Detection Result

G. Storage Analytics Case Study

A group of uploaded files was analyzed to evaluate storage usage and redundancy patterns. The system calculated storage consumption, identified duplicate percentages, and generated analytics visualizations. Based on these observations, the platform provided recommendations for removing redundant files and optimizing storage utilization.

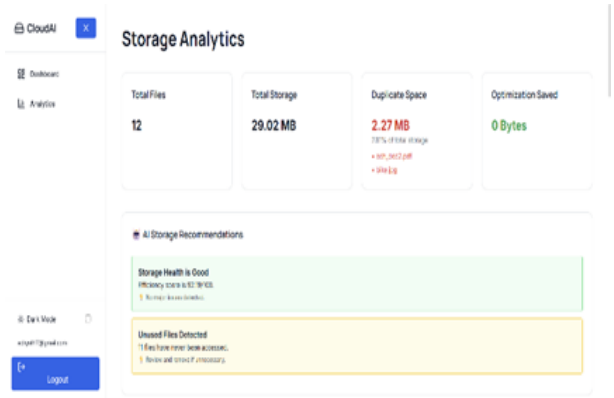

Fig. 8: Analytics page

H. Storage Efficiency Evaluation

The proposed system demonstrated consistent improvement in storage utilization during evaluation. Duplicate detection and optimization mechanisms effectively reduced redundant files, resulting in better storage efficiency. Analytics trends indicated balanced storage usage and improved retrieval performance over time. The system handled increasing file volumes without significant performance degradation. These results confirm the scalability and effectiveness of the intelligent cloud storage framework.

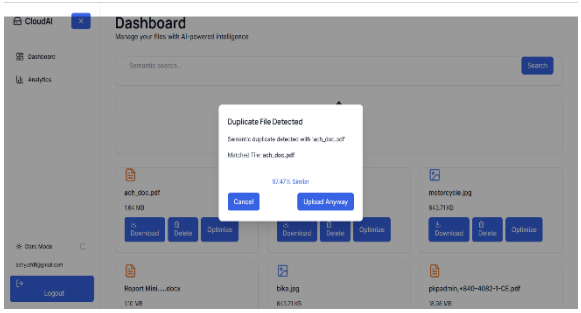

Fig. 9: Model Accuracy and Performance Analysis

I. Analytics and Feature Comparison

The proposed system provides interpretable analytics visualization that highlights differences between unique and duplicate file patterns. Typically, unique files show varied content structure and distinct contextual relationships. However, duplicate files display high similarity scores and nearly identical content representations.

VII. RESULTS AND DISCUSSION

The proposed intelligent cloud storage framework improves file management by integrating semantic search, duplicate detection, and analytics-based optimization. Traditional keyword-based search methods retrieve files based only on names or metadata, which often leads to inefficient results. The proposed system overcomes this limitation by using embedding-based semantic similarity, allowing users to retrieve contextually related files.

The duplicate detection module further enhances storage

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

efficiency by identifying redundant files using similarity comparison. This reduces unnecessary storage consumption and improves overall organization. Additionally, the analytics module provides visual insights such as storage usage, duplicate percentage, and file distribution. These analytics help users understand storage behavior and optimize file management effectively.

The experimental evaluation shows that semantic search improves file retrieval compared to traditional methods, while duplicate detection reduces redundant data. The system demonstrates stable performance when handling multiple file types and increasing file volumes. Analytics visualization confirms improved storage utilization and efficient data organization.

The system also demonstrates improved usability through its interactive dashboard and analytics visualization. Users can easily monitor storage usage, duplicate statistics, and file distribution in real time. This transparency helps users understand how their data is organized and identify redundant files quickly. The visualization components simplify complex storage information and improve decision-making for file management and cleanup operations.

Another important observation is the system's ability to handle multiple file formats effectively. The preprocessing and embedding generation modules standardize input data, allowing semantic search to work consistently across documents, text files, and other supported formats. This capability ensures that the system remains flexible and scalable when deployed in real-world environments where heterogeneous data is common.

The duplicate detection module plays a significant role in optimizing storage utilization. By identifying both exact and semantically similar files, the system reduces redundancy and prevents unnecessary storage consumption. This leads to better organization and faster retrieval performance. The combination of similarity comparison and analytics-driven insights further enhances the efficiency of storage management.

The system also maintains stable performance when the number of stored files increases. The semantic search engine efficiently compares embeddings and retrieves relevant files without significant delay. This indicates that the proposed framework can scale to handle larger datasets while maintaining retrieval efficiency. The indexing and similarity computation mechanisms contribute to faster search operations and improved responsiveness of the system

Overall, the results demonstrate that the integration of semantic search, duplicate detection, and analytics visualization provides a comprehensive solution for intelligent cloud storage. The system improves retrieval accuracy, reduces storage redundancy, and enhances user experience. These improvements highlight the effectiveness of the proposed framework and its suitability for scalable personal cloud storage applications.

Furthermore, the analytics module provides meaningful insights that assist users in optimizing storage usage. By highlighting duplicate percentages, storage consumption trends, and file distribution, the system enables informed decision-making. These insights help users remove unnecessary files, organize data effectively, and maintain optimal storage utilization over time.

VIII. CONCLUSION AND FUTURE WORK

The proposed intelligent cloud storage framework presents an efficient and scalable solution for semantic file management by integrating preprocessing, embedding-based feature representation, and similarity-driven retrieval mechanisms. The preprocessing stage standardizes uploaded files and extracts meaningful content, while embedding techniques capture contextual relationships between documents. The combined semantic search and duplicate detection modules improve file discovery and reduce redundant storage. Additionally, analytics visualization enhances transparency by providing insights into storage usage, duplicate statistics, and file distribution patterns.

Experimental observations indicate that the system effectively retrieves contextually related files and minimizes redundancy within storage. The inclusion of analytics dashboards and similarity-based recommendations improves usability and supports better storage organization. Compared to traditional keyword-based approaches, the proposed framework offers improved automation, enhanced retrieval efficiency, and better scalability. These capabilities make the system suitable for personal cloud storage, intelligent file management, and data organization applications.

Despite these advantages, there are several opportunities for future enhancement. Future work may focus on improving semantic understanding for multimedia content such as images, audio, and video files. Optimization techniques can also be explored to reduce computational overhead and enable faster similarity computation for large-scale storage systems. Additionally, implementing adaptive indexing and incremental learning methods can further improve retrieval performance.

Further research may include integration with distributed cloud platforms and real-time storage monitoring systems. Advanced analytics such as predictive storage recommendations and automated cleanup mechanisms can also be incorporated. These enhancements will improve scalability, efficiency, and intelligent data management capabilities, making the system more suitable for large-scale cloud storage environments.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology, for guidance and support.

REFERENCES

[1] Y. Zhang, X. Liu, and H. Wang, "Semantic search for intelligent document retrieval using deep learning embeddings," IEEE

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Access, 2024.
[2] A. Kumar, R. Sharma, and P. Singh, “AI-based personal cloud storage system with duplicate detection and optimization,” International Journal of Advanced Computer Science and Applications, 2025.
[3] M. Chen, L. Zhao, and K. Li, “Efficient duplicate file detection using content similarity and hashing techniques,” IEEE Transactions on Cloud Computing, 2023.
[4] S. Patel and N. Shah, “Semantic similarity-based document retrieval using transformer embeddings,” arXiv preprint, 2024.
[5] J. Brown, T. Green, and P. Adams, “Content-aware storage optimization in cloud environments,” Journal of Cloud Computing, 2023.
[6] T. Mikolov et al., “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems, 2013.
[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019.
[8] Y. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” EMNLP, 2019
[9] H. Wang, J. Li, and Y. Chen, “Cloud storage optimization using machine learning-based redundancy removal,” IEEE Access, 2025.
[10] L. Xu, Z. Wang, and Q. Sun, “Embedding-based semantic retrieval for large-scale document systems,” Journal of Information Science, 2024.
[11] K. Aggarwal and S. Gupta, “Intelligent file management using semantic similarity and clustering,” International Journal of Computer Applications, 2023.
[12] D. Lin, “An information-theoretic definition of similarity,” International Conference on Machine Learning, 1998.
[13] S. Singh and R. Verma, “AI-driven storage analytics and optimization for personal cloud platforms,” IEEE Cloud Computing, 2025.
[14] A. Radford et al., “Learning transferable visual models from natural language supervision,” ICML, 2021.