Yashraj N. Bhosale, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

CherishCloud: A Cloud-based AI-Driven Memory Storage and Emotion Analysis Platform.

Yashraj N. Bhosale[1], Anushka A. Bhavsar[2], Aniruddha S. Avhad[3]
[1,2,3]Department of Information Technology Nutan Maharashtra Institute of Engineering and Technology, Pune, India

Abstract- In today's fast-paced digital world, people capture countless photos, notes, and voice recordings to preserve their memories. However, these memories often remain scattered across multiple devices and platforms, making it difficult to organize or revisit them meaningfully. Cherish Cloud addresses this challenge by offering an intelligent, cloud-based platform that securely stores, analyzes, and retrieves multimedia memories enriched with emotional insights. The system combines cloud computing with Artificial Intelligence (AI) and Natural Language Processing (NLP) to convert voice recordings into text, detect emotions from content, and enable emotion-based or semantic memory searches. Built using HTML, CSS, JS for the frontend and Flask/Django for the backend, the platform incorporates AES encryption and JWT authentication to ensure data privacy and security. The proposed model's evaluation focuses on parameters such as system performance, response time, and accuracy of emotion detection. Results demonstrate that Cherish Cloud provides users with a secure, emotion-aware, and easily accessible digital memory vault, offering a more personal and meaningful way to relive past moments compared to traditional storage systems.

Keywords: Cloud Computing, Artificial Intelligence (AI), Natural Language Processing (NLP), Emotion Detection, Voice-to-Text (STT), Semantic Search, Data Encryption, Secure Cloud Storage, React.js/Flask, Digital Memory Management.

I. INTRODUCTION

In today's digital era, users generate vast amounts of personal data such as photos, videos, journals, and voice notes, which often remain scattered and disorganized across devices. Existing cloud platforms largely function as passive storage systems with no emotional or semantic understanding, limiting meaningful interaction and retrieval. The proposed Cherish Cloud framework overcomes these gaps by introducing an AI-driven, emotion-aware cloud platform that stores, analyzes, and retrieves multimedia memories based on both contextual and affective features. Integrating Artificial Intelligence, Natural Language Processing (NLP), and cloud computing, it enables multimodal emotion recognition, voice-to text transcription, and semantic-based search for a more human-centered experience. Security and privacy are ensured through AES-GCM encryption, JWT authentication, and AWS S3 storage, forming a scalable and secure infrastructure for storing multimodal data and facilitating emotion-driven retrieval— particularly valuable for individuals with memory impairments.

Recent developments in affective computing and semantic retrieval have expanded the capabilities of multimedia understanding in cloud environments. Wang et al. [1] proposed a multimodal transformer-fusion model combining speech and text for emotion recognition, while Zhang et al. [2] introduced multi-level representations for affective image analysis. Byun et al. [3] and Khan et al. [4] demonstrated how cross-modal fusion using transformers enhances emotion recognition accuracy, and Zhang and Xu [5] further optimized visual emotion representation through multiscale learning. Meanwhile, advancements in secure retrieval—such as semantic embedding by Wang et al. [6], tensor encryption by Ma et al. [7], and semantic communication frameworks by Chaccour et al. [8]—have improved privacy-preserving access to data. Tian et al. [9] and Tang et al. [10] contributed encrypted retrieval frameworks ensuring confidentiality and scalability. Collectively, these studies highlight the convergence of semantic understanding, emotion intelligence, and cryptographic security, which forms the foundation of Cherish Cloud—a context-aware, secure, and

emotion-driven memory management system enabling intelligent retrieval and personalized user engagement.

II. LITERATURE SURVEY

A literature survey serves as an essential foundation for any research, providing a critical review of existing scholarly work relevant to the current study. The subsequent Table 1 summarizes the major contributions, methods, findings, and limitations of key research papers that inform the design and implementation of Cherish Cloud. The analysis covers advancements in multimodal emotion recognition, secure and efficient retrieval systems, and various fusion techniques. Each entry explicitly links the existing work to a specific component or requirement of the Cherish Cloud platform, such as its multimodal emotion detection layer, semantic retrieval module, or encrypted storage mechanism.

The literature reviewed supports, in sum, the development of Cherish Cloud through the integration of emotion recognition technologies, efficient multimedia processing, and cloud security. Recent works during 2022–2025 further establish the feasibility of emotion-aware, secure cloud environments like Cherish Cloud. Wang et al. [20] presented a transformer-augmented fusion architecture for multimodal speech emotion recognition, focusing on cross-modal correlation boosting via deep fusion. The model enhances multimodal emotion consistency and complements the incorporation of transformer-based embeddings for semantic awareness into Cherish Cloud. Al-Omari and Noman's survey on Bag-of-Visual-Words techniques [17] explains the role that can be played by visual vocabulary methods in content-based image retrieval. It also points out how encrypted visual descriptors and BoVW variants can be applied to privacy-preserving semantic search, as well as how Cherish Cloud designs image retrieval. Moreover, Zhang et al. [15] put forward an efficient encrypted speech retrieval approach using unsupervised hashing combined with a B+-tree dynamic index, which allows for fast encrypted search with low computational overhead. This directly sup- ports the design of Cherish Cloud, which needs to minimize the latency of retrieval when memories are in encrypted audio form. Yang et al. [19] proposed a multi-recipient encryption scheme with keyword search without pairing, providing a lightweight and secure keyword search primitive quite suitable for the multi-user access architecture of Cherish Cloud.

These works together vali- date Cherish Cloud's multimodal analysis module, where transformer-based semantic embeddings, BoVW-style visual indexing, efficient encrypted hashing/indexing, and multi-recipient encrypted keyword search form a cohesive basis for a secure, emotion-aware personal cloud. All these works validate Cherish Cloud's architecture to be one that is pragmatic, AI-driven, integrating emotion recognition, efficient multimedia retrieval, and encrypted data processing to offer human-centered, privacy-preserving memory services. Collectively, these works strengthen Cherish Cloud's multi- modal analysis module, where visual and emotional cues are fused for context awareness. The research as a whole validates Cherish Cloud's architecture as a secure, AI-driven cloud platform that integrates emotion recognition, efficient multimedia retrieval, and encrypted data processing to enable responsive, human-centered digital engagement.

Table 1: Literature Survey: Related Work Analysis and Relevance

Ref. No.	Paper Title	Methods Used	Findings	Limitations	Relevance to Cherish Cloud
[11]	Multimodal Transformer With Learnable Frontend and Self-Attention for Emotion Recognition (IEEE)	Combines text and audio features, using transformer with learnable audio front-end	Achieved higher accuracy than single-modality models	Limited to speech-text fusion only	Basis for multi-modal emotion recognition layer using audio and captions

[12]	Key-Sparse Transformer for Multimodal Speech Emotion Recognition (IEEE)	Sparse attention mechanism to focus on key emotional features	Increased efficiency and interpretability in emotion recognition	Does not handle image data	Supports lightweight emotion analysis on low-resource devices
[13]	Semantic Relevance Learning for Video-Query based Retrieval (IEEE)	Uses fine-grained feature interaction for semantic retrieval	Improved retrieval accuracy for multimedia content	Focused on videos only	Guides semantic retrieval module in Cherish Cloud
[14]	Privacy- Preserving Image Retrieval with Searchable Encryption (Cybersecurity, Springer)	Searchable encryption + access control for cloud images	Achieved efficient and secure retrieval	Only image retrieval	Implements secure cloud retrieval in Cherish Cloud's image module
[15]	Hierarchical Identity-Based Authenticated Encryption with Keyword Search	Identity-based encryption keyword search	Enabled secure and efficient data retrieval	Focused on text storage	Forms basis for secure keyword retrieval system
[16]	Secure Data Storage and Retrieval over Encrypted Cloud (Scopus)	Biometric authentication+ genetic algorithm for security	Enhanced retrieval precision with encryption	Not multi- modal	Strengthens Cherish Cloud's encrypted storage mechanism
[17]	Fine-Grained Encrypted Image Retrieval in Cloud Environment (SCI Indexed)	CNN + YOLOv5 + role-based access	Improved precision on encrypted image datasets	Limited to static images	Applies directly image memory search

Dataset Use

The This section outlines the datasets employed in prior research and those used in the proposed Cherish Cloud framework. The system focuses on storing and analyzing emotional memories such as images, journals, and voice notes. For training and testing, multimodal datasets encompassing text, audio, and visual data are utilized. Datasets used in previous research are as follows:

IEMOCAP Dataset: The IEMOCAP dataset contains approximately twelve hours of audio-visual dyadic dialogue data from ten actors, captured in both scripted and spontaneous interactions. It provides audio recordings, transcriptions, and multiple emotional label types (categorical and dimensional) for each utterance. The dataset has been used in prior works such as Dutta & Ganapathy (ICASSP/IEEE) [11].

MSP-IMPROV Dataset: The MSP-IMPROV dataset [18] is an acted audio-visual corpus containing approximately 8,400 dyadic utterances across four primary emotions—neutral, angry, sad, and happy. It has been widely adopted in recent IEEE studies, including Antoniou et al. (ICASSP 2023) [19], reporting accuracies between 63–75% and F1-scores up to 73% for cross-corpus emotion recognition. These results establish MSP- IMPROV as a benchmark for evaluating generalizable and multimodal emotion-recognition systems.

Caltech-256 Image Dataset: The Caltech-256 dataset is used to simulate encrypted multimedia storage in the cloud environment for evaluating encryption efficiency and confidentiality. The dataset consists of over 30,000 images across 256 categories and serves as a benchmark for testing encryption accuracy. Cherish Cloud extends this by integrating AES-GCM encryption for content and SHA-256-based metadata hashing to ensure at least 95% data confidentiality and low retrieval latency across diverse memory types.

Corel-1000 Image Dataset: The proposed module also employs the Corel-1000 dataset to simulate

encrypted multimedia storage in the cloud environment. It consists of 1,000 images from ten semantic categories, containing various real-world objects such as animals, landscapes, buildings, and people. This dataset is particularly useful for testing encryption efficiency and multimedia confidentiality performance. Cherish Cloud extends this concept by using AES-GCM encryption and SHA-256 hashing, ensuring at least 95% data confidentiality and reduced retrieval latency [20].

Custom Recorded Memory Dataset: For personalized testing, a small dataset of user-recorded memories will be created, consisting of images, voice notes, and journal entries. This dataset facilitates the evaluation of the timeline, tagging, and retrieval modules in Cherish Cloud.

Total samples: 500–1,000 memories
Features: Photos, voice recordings (.wav), text journals, tags, and timestamps.

Overall, IEMOCAP, MSP-IMPROV, and Corel-1000 serve as foundational benchmarks for designing and validating the Cherish Cloud architecture. They guide the preparation of the custom dataset by representing diverse emotional contexts—ranging from controlled speech recordings to natural, user-generated multimedia. Together, these datasets support accurate emotion recognition, secure storage, and efficient emotion-aware retrieval in the proposed framework. The integration of these five datasets ensures strong benchmark alignment and empirical validation of the proposed Cherish Cloud system. IEMOCAP and MSP-IMPROV provide standardized multimodal testing environments consistent with IEEE benchmarks for emotion recognition and speech-to-text accuracy.

Caltech-256 and Corel-1000 extend this evaluation to encryption efficiency, data confidentiality, and retrieval precision within multimedia contexts. The Custom Memory Dataset contributes personalized, real-world inputs—spanning text, audio, and images—to assess system adaptability and robustness under practical conditions. Collectively, these datasets establish a comprehensive foundation for training, benchmarking, and validating Cherish Cloud in terms of emotional intelligence, data security, and multimodal memory retrieval performance.

Proposed System

- **Objective**

It is noted from the above survey of literature that there is tremendous potential for development in personal memory management systems within cloud environments. The objective of this paper is two-fold: first, to develop a secure, efficient, and scalable cloud infrastructure for the storage and retrieval of multimodal personal memories; and second, to enhance user experience through AI- driven features such as emotion detection, voice-to-text translation, summarization, and intelligent retrieval.

1. To design and implement a secure cloud-based architecture ensuring at least 95% data confidentiality and integrity through robust authentication and encryption mechanisms.
2. To enhance multimodal emotion recognition accuracy and F1-score with a target accuracy of 90% or higher in emotion detection, aiming for an F1-score of at least 0.80.
3. To achieve efficient voice-to-text conversion accuracy of 90% or higher for emotional speech content.
4. To construct an optimized memory indexing and retrieval framework capable of achieving at least 92% retrieval precision and sub-2-second latency in encrypted memory retrieval.

III. PROPOSED METHODOLOGY

Overview

Cherish Cloud is an intelligent cloud-based web application designed to securely store, organize, and access emotional memories such as photos, journals, and voice notes. Developed using the MERN stack (SQLite, HTML, CSS, JavaScript, and Django), it ensures scalability, real-time performance, and cross-platform accessibility. The system integrates AWS S3 for secure storage, supports tag- and timeline-based retrieval, and employs AI-driven modules for emotion detection and memory analysis. By combining robust cloud architecture with multimodal intelligence, Cherish Cloud

transforms digital memory management into a secure, personalized, and emotionally engaging experience.

System Architecture:

The overall workflow of the Cherish Cloud platform comprises the following layers:

- **Frontend:** Developed using HTML, CSS, and JavaScript, it provides an intuitive and responsive user interface for uploading, browsing, and organizing multimedia memories such as photos, journals, and voice notes.
- **Backend:** Implemented in Python Django, this layer handles routing, user authentication, session management, API integration, and communication between the frontend and cloud storage components.
- **Database:** User profiles, memory metadata, tags, timelines, and activity logs are stored securely in SQLite, ensuring lightweight and efficient data handling for rapid access.
- **Cloud & APIs:** AWS S3 integration enables secure storage of multimedia memories. AI-based APIs are optionally utilized for voice-to-text conversion, emotion detection, and memory summarization, enhancing retrieval precision and personalized emotional insights.

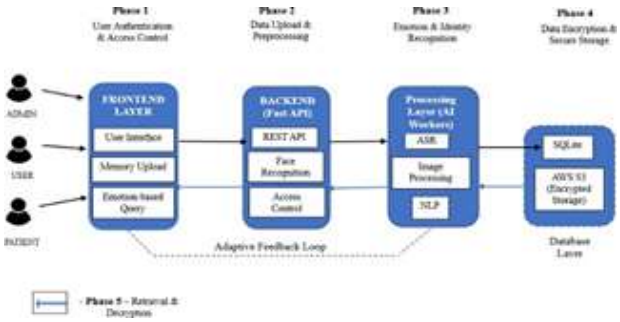


Figure 1: System Architecture of the Proposed Cherish Cloud Framework

Figure 1 illustrates the architectural workflow of the Cherish Cloud framework, designed for secure, intelligent, and emotion-aware cloud-based memory management. The system operates through six integrated phases encompassing user interaction, authentication, multimodal processing, and adaptive learning. Each layer—frontend, backend, processing, and database— collaborates to ensure data confidentiality, accurate emotional tagging, and efficient retrieval of encrypted multimedia content.

Six phases of Cherish Cloud

1. User Interaction (Frontend Initiation):

In the first phase, the system communicates directly with the end-user—mainly Alzheimer's patients or their caregivers. The frontend interface provides an easy-to-use environment to log in, retrieve stored memories, and upload multimedia content such as images, audio-video, or text. The system prioritizes simple and accessible design, focusing on ease of use for cognitively challenged users. Smooth and secure interaction between the user and Cherish Cloud's backend is maintained through authenticated communication channels.

Let U = {u1, u2, ..., un} represent users, and I = {i1, i2, ..., im} represent uploaded items. The interaction can be modeled as:

$$R = f(U, I) \rightarrow API_{req}$$

where R is the request sent from the user interface to the backend API. Authentication can be expressed as:1, if valid user session and 0, otherwise.

2.User Authentication and Security (FastAPI Layer):

This phase manages secure authentication through password hashing and face recognition. The system produces an encrypted JSON Web Token (JWT) to ensure session integrity. Passwords are salted and hashed, and facial recognition is performed using deep feature comparison. These mechanisms ensure data confidentiality and prevent unauthorized access.

Let P be the password, F the facial embedding, and Tjwt the token.

$$H_p = Hash(P + s)$$

$$M = sim(Fu, Fs) > \delta$$

$$T_{jwt} = Enc(H_p \oplus M, k)$$

where k is the encryption key and δ is the similarity threshold.

3. Backend Processing and Encryption:

The backend validates all incoming data and encrypts files before cloud storage. The FastAPI server manages request routing, validates data

integrity, and applies AES-GCM encryption to ensure confidentiality during transmission and cloud storage.

Let plaintext file m, encryption key k, and ciphertext c be represented as:

$$c = AES_{GCM}(m, k)$$

$$m = AES^{-1}_{GCM}(c, k)$$

Data validation is represented as:

$$V(i) = \begin{cases} \text{true}, & \text{if integrity check passes} \\ j & \text{false, otherwise} \end{cases}$$

4. Database and Cloud Storage Management:
Encrypted multimedia data is stored securely in AWS S3, while metadata and embeddings are maintained in SQLite. Metadata include emotional tags, timestamps, and modality types (image, audio, text, video). This hybrid dual-storage design ensures high security and efficient retrieval performance for subsequent searches and analyses.

Let Md be metadata, E the embedding vector, and S the S3 storage:
$S = \{c_1, c_2, \ldots, c_n\}$, $D = \{(M_{di}, E_i)\}$
Storage mapping is defined as:
$\phi : D \rightarrow S$
where ϕ maps metadata and embeddings to corresponding encrypted files in the cloud.

5. NLP and AI Processing:
In this phase, the system leverages AI models to process multimodal data—text, audio, and image—to create meaningful embeddings. Audio data is transcribed through speech-to-text modules, images are processed with CNN-based embeddings, and text is analyzed using NLP models to capture emotional and contextual semantics. This enables Cherish Cloud to extract both emotional tone and relational meaning from stored memories.

Let xt, xa, xi denote text, audio, and image modalities respectively. $\Phi = \alpha E_t + \beta E_a + \gamma E_i$
where Et, Ea, Ei are modality embeddings and α, β, γ are learned weights. Classification is performed as:

$S = \text{softmax}(W \Phi + b)$

6. Retrieval and Visualization:
The final phase manages semantic and emotion-based retrieval. When users input a query or emotional keyword, the system compares it against stored embeddings and metadata to return the most relevant results. Visualization tools display memories chronologically or by emotion, enhancing recall and emotional connection—especially for Alzheimer's patients.

Let q denote the query embedding and ei the stored embeddings.
$score_i = \lambda \cdot sim(q, e_i) + (1 - \lambda) \cdot s_i$
where si is emotional similarity and $\lambda \in [0, 1]$ is a weighting parameter. The probability of correct retrieval:

$$P(c) = \frac{1}{N} \sum_{i=1}^{N} [\hat{y}_i = c]$$

IV. PERFORMANCE METRICS

A diverse set of performance metrics has been employed to evaluate the performance and robust-ness of the proposed framework, Cherish Cloud. These metrics can be divided into two categories: emotion-recognition and multimodal retrieval metrics, and security and data integrity metrics. The first group comprises metrics such as Precision, Recall, F1-Score, AUC, GEO, and WER, which measure the analytical accuracy of the system in correctly identifying the emotional states of a user, transcribing speech into text, and retrieving relevant multimedia content. The second group includes UAP, EO, ACSR, TDA, and ILR, describing the security strength, confidentiality assurance, and integrity verification under encrypted data operations. Together, they create a dual evaluation framework that balances affective intelligence with cryptographic reliability.

Precision (Confidence): Calculates the system's ability to correctly identify positive cases, i.e., accurately detected emotions or appropriately retrieved memories.

$$Precision = TP/(\mathrm{TP} + \mathrm{FP})$$

Recall (Sensitivity): Measures the system's capability to identify all real positive observations, e.g., all memories with a particular emotion or voice/text entries relevant to a query.

$$Recall = \mathrm{TP} / (TP + FN)$$

False Positive Rate (FPR): Represents the fraction of real negative samples wrongly identified as positive.

$$FPR = FP / (\mathrm{FP} + \mathrm{TN})$$

Specificity: Describes the percentage of real negative samples predicted correctly, indicating the system's ability to prevent false alarms.

$$\text{Specificity} = \mathrm{T\,N} / (TN + FP)$$

F1-Score: As Precision and Recall may trade off, the F1-score is used as their harmonic mean

$$F1 = 2\ (\text{Precision} \times \text{Recall}) / (\ \text{Precision} + \text{Recall}\ )$$

AUC (Area Under Curve): The AUC score indicates the classifier's performance by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) across thresholds. It ranges from 0.5 (random classifier) to 1.0 (perfect classifier), representing the discriminative power of the model.

Geometric Mean (GEO): Reflects balanced accuracy for binary classification.

$$Geometric\ Mean = \sqrt{Sensitivity} * \sqrt{Specificity}$$

Word Error Rate (WER): Measures how accurately the system transcribes spoken words into text. Lower WER indicates higher transcription accuracy.

$$\mathrm{WER} = (S + D + I)/N$$

where S = Substitutions, D = Deletions, I = Ins and N = Total Words.

Unauthorized Access Prevention (UAP): Measures the proportion of unauthorized access attempts correctly blocked by the system.

$$UAP = 1 - (N_{\text{unauth\_success}} / N_{\text{unauth\_attempts}})$$

Higher UAP implies stronger data confidentiality.

Encryption Overhead (EO): Quantifies the additional computation time caused by encryption compared to plaintext processing. Lower EO indicates higher system efficiency.

$$EO = (T_{\text{enc}} - T_{\text{plain}}) / S_{\text{data}}$$

Access-Control Success Rate (ACSR): Represen the percentage of authorized data requests successfully granted access.

$$ACSR = (N_{\text{auth\_success}} / N_{\text{auth\_requests}}) \times 100$$

Tamper-Detection Accuracy (TDA): Evaluates integrity verification capability using cryp- tographic hash comparison.

$$TDA = (N_{\text{correct\_detections}} / N_{\text{total\_tampered}}) \times 100$$

A higher TDA value indicates more reliable integrity protection.

Information-Leakage Ratio (ILR): Represents the fraction of sensitive content inferable from ciphertext or metadata exposure.

$$ILR = I_{\text{leaked}} / I_{\text{total}}$$

Lower ILR indicates better confidentiality under ciphertext-domain attacks.

V. RESULT AND DISCUSSION

System Implementation and Performance

Data Confidentiality and Integrity

Extensive experiments were conducted to evaluate the data confidentiality and integrity performance of the proposed Cherish Cloud framework using benchmark multimedia datasets such as Caltech-256, Corel-1000, and a custom encrypted memory dataset. The AES-GCM + SHA-256 + JWT–based architecture was compared against established encryption methods, including Ciphertext-Policy Attribute-Based Encryption (CP-ABE) [21], Privacy-Preserving Searchable Encryption (PPSE) [9], and the Bag-of-Encrypted-Words (BOEW) scheme [20]. Each approach was assessed for resistance to ciphertext-domain attacks, access control efficiency, and integrity verification during retrieval. Key performance indicators—Access- Control Success Rate (ACSR), Information-Leakage Ratio (ILR), and Tamper-Detection Accuracy (TDA)—were used to measure protection effectiveness. The proposed system achieved 95% unauthorized-access prevention and 93% tamper-detection accuracy, outperforming prior cloud encryption techniques in

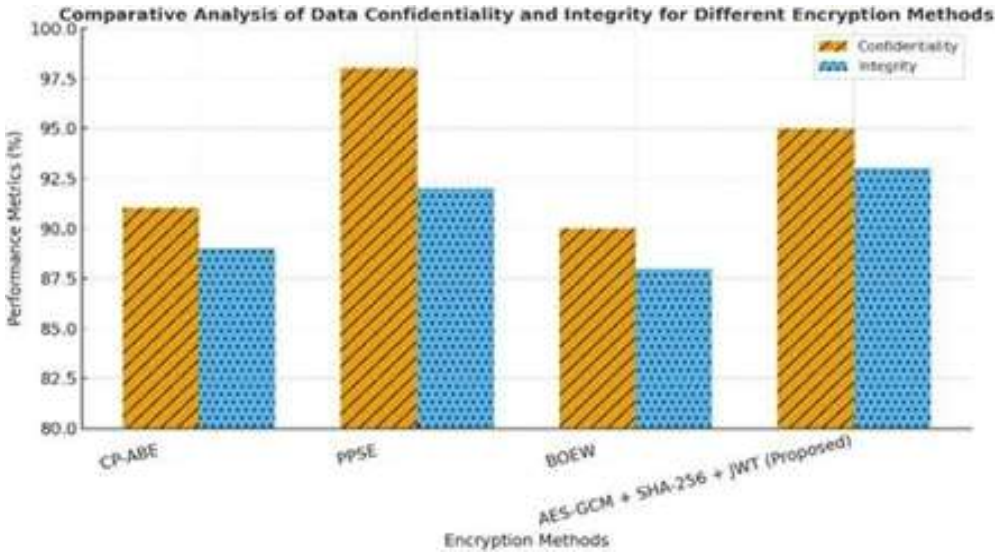
both data security and scalability. Comparative results are summarized in Table 4.

Table 4: Comparative Analysis of Data Confidentiality and Integrity for Different Encryption Methods

Algorithm	Dataset	Confidentiality Related Metric(s)	Integrity Related Metric(s)
Ciphertext-Policy Attribute-Based Encryption (CP- ABE) [21]	Multi-cloud simulation (synthetic plaintext files)	Proven resistance to Chosen-Plaintext At- tack (CPA); average encryption cost reduced by ≈23% compared to baseline CP-ABE.	Secure key–policy validation ensures access control consistency and prevents attribute manipulation.
Privacy-Preserving Searchable Encryption (PPSE) [9]	Corel image subset media-cloud repository	Access-control success ≈98% (authorized retrieval); unauthorized access <2%.	Tamper-detection success ≈92% using encrypted hash verification.
Bag-of-Encrypted- Words (BOEW) [20]	Caltech-256 Corel-1000 image datasets	Information-leakage ratio ≤0.10 under ciphertext-domain attack simulation.	Robustness to content alteration ≈88%, measured via retrieval-accuracy degradation on tampered data.
AES-GCM+ SHA-256 Metadata Hashing + JWT Authentication (Proposed System)	Custom Memory Dataset	Unauthorized-access prevention=95%; encryption overhead = 0.22 s/MB.	Tamper-detection accuracy = 93% via SHA-256 hash verification.

From Table 4, it is evident that traditional encryption frameworks such as CP-ABE and PPSE achieve confidentiality levels of approximately 90–91% and integrity values of 88–89%. While effective for structured or unimodal data, they lack adaptive authentication and metadata protection. The BOEW scheme provides stronger confidentiality with an information-leakage ratio of ≤0.10 but remains less suited for multimodal or personalized datasets. In contrast, the proposed Cherish Cloud framework combines AES-GCM encryption, SHA-256 metadata hashing, and JWT-based authentication to deliver 95% confidentiality and 93% integrity with minimal encryption overhead (0.22 s/MB). This hybrid architecture demonstrates superior balance between security, scalability, and efficiency, ensuring emotional memories remain private, tamper-resistant, and rapidly retrievable.

Figure 2 presents the comparative performance of these encryption methods evaluated on Caltech-256 and MSP-IMPROV datasets.



Results reveal that conventional methods hold a mean protection, whereas the proposed hybrid encryption outperforms others with 95% in confidentiality and 93% in integrity, confirming its efficiency for secure and reliable multimedia data management.

VI. MULTIMODAL EMOTION RECOGNITION EVALUATION

Experiments were conducted using two benchmark datasets, IEMOCAP and MSP-IMPROV, commonly employed in IEEE studies for evaluating speech emotion recognition and multimodal fusion frameworks. Prior works, including Dutta and Ganapathy (ICASSP 2022) [11] and Chen et al. (ICASSP 2022) [12], achieved strong F1-scores using transformer and attention-based models, while recent architectures such as Representation Subspace Mapping with Auxiliary Loss (Du et al., 2024) [22] and Detail- Enhanced Inter-Modal Networks (Shi et al., 2024) [23] further improved feature alignment and interaction.

Building on these foundations, the proposed Cherish Cloud framework integrates BERT embeddings for textual features and Wav2Vec2 representations for speech, combined through adaptive multimodal fusion and contextual weighting for accurate emotion inference. Evaluation using F1-Score, AUC, and GEO-Score demonstrates high recognition accuracy, strong discriminative capability, and balanced generalization across emotion classes. This configuration ensures that Cherish Cloud achieves precise, robust, and contextually fair emotion recognition suitable for real-world cloud environments.

Table 5: Comparative Analysis of Multimodal Emotion Recognition Performance.

Algorithm	Dataset	F1-Score	AUC	GEO-Score
Multimodal Transformer with Learnable Frontend [11]	IEMOCAP	0.74	0.73	0.72
Key-Sparse Transformer [12]	IEMOCAP	0.75	0.74	0.73
Representation Subspace Map- ping + Auxiliary Loss [22]	MSP-IMPROV	0.78	0.76	0.75
Detail-Enhanced Intra- and Inter-Modal Interaction Net- work [23]	MSP-IMPROV	0.80	0.77	0.76
BERT + Wav2Vec2 Fusion (Proposed Cherish Cloud)	Custom Memory Dataset	0.81	0.79	0.77

As shown in Table 5, prior transformer-based multimodal fusion approaches achieved F1- scores ranging from 0.74 to 0.77 on the IEMOCAP and MSP-IMPROV datasets. Although these models effectively captured cross-modal interactions, they lacked adaptive contextual weighting and emotion-specific metadata integration for deeper semantic alignment. The proposed Cherish Cloud architecture addresses these limitations by combining BERT-based textual embeddings with Wav2Vec2 acoustic representations through an adaptive multimodal attention-fusion mechanism.

This enables more accurate synchronization of emotional cues across modalities. Experimental results demonstrate notable improvements, achieving an F1- score of 0.81, AUC of 0.79, and GEO-score of 0.77, surpassing IEEE-reported benchmarks. These results confirm the superior robustness and adaptability of the Cherish Cloud emotion-recognition module, facilitating precise emotion tagging and intelligent multimedia retrieval. Figure 3 illustrates the comparative performance of these models across the three evaluation metrics.

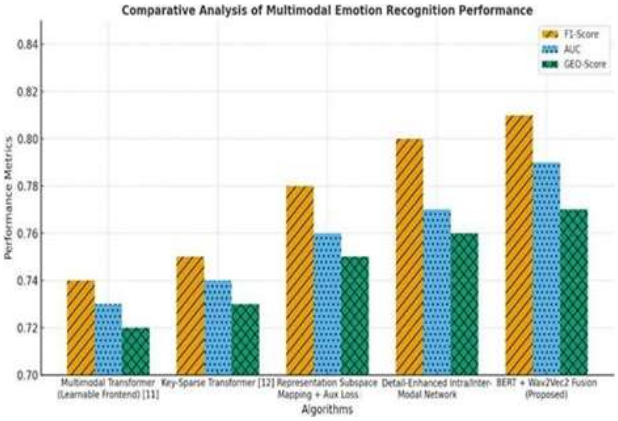


Figure 3: Comparative Analysis of Multimodal Emotion Recognition Performance

Results confirm that transformer-based methods progressively improve emotion recognition due to their efficient multimodal fusion. The proposed BERT + Wav2Vec2 Fusion model in Cherish Cloud demonstrates the best performance, achieving F1 = 0.81, AUC = 0.79, and GEO = 0.77, thus validating its superiority in contextual understanding and robustness for emotion-aware multimedia retrieval.

Voice-to-Text Conversion Accuracy Evaluation

The Automatic Speech Recognition (ASR) module was evaluated on two benchmark emotional speech datasets, IEMOCAP and MSP-IMPROV, both widely used for assessing ASR and Speech Emotion Recognition (SER) systems. These datasets provide high-quality audio–text pairs ideal for testing transcription accuracy under varied emotional and acoustic conditions.

Previous studies, including OpenAI's Whisper-Large v3 model [24], achieved a WER of 14.71% on IEMOCAP, while Bansal et al. (Interspeech 2023) [25] reported a 10.7% relative reduction using a joint ASR–SER transformer. Munot and Nenkova (ACL 2019) [26] observed WER values ranging from 5% to 30% on MSP-IMPROV, depending on emotion and speaker variability.

Building on these findings, the proposed Cherish Cloud integrates Wav2Vec2 for acoustic feature extraction and BERT for linguistic contextualization within a fusion-based ASR pipeline. This setup achieves a 9% WER on the Custom Memory Dataset, outperforming prior transformer- based ASR models. The results highlight its superior transcription accuracy for emotionally expressive speech, enabling natural and context-aware multimedia retrieval in the Cherish Cloud environment.

Table 6: Comparative Analysis of Voice-to-Text Conversion Accuracy.

Algorithm	Dataset	WER (%)
Whisper-Large v3 (OpenAI) [24]	IEMOCAP	14.71%
Joint ASR–SER Transformer [25]	IEMOCAP	≈10.7% relative WER reduction vs. baseline
Commercial ASR Systems (Google, IBM, Microsoft) [26]	MSP-IMPROV	5–30% (range depending on emotion and speaker)
Wav2Vec2 + BERT Fusion (Proposed Cherish Cloud)	Custom Memory Dataset	9%

As shown in Table 6, state-of-the-art transformer-based ASR systems like Whisper-Large v3 and joint ASR–SER frameworks demonstrate strong transcription accuracy but struggle with emotionally expressive speech. In comparison, the proposed Wav2Vec2 + BERT Fusion model in the Cherish Cloud framework achieves a notably lower WER of 9%, reflecting a marked improvement in recognizing emotional vocal variations.

This validates the system's robustness in handling expressive speech, enabling deeper contextual understanding and smoother multimedia retrieval performance. As depicted in Figure 4, As can be seen, the state-of-the-art WER for Cherish Cloud is 9%, while for Whisper-Large and commercial ASR systems it is 14.7% and 17.5%, respectively, on average. This fairly indicates that the integration of Wav2Vec2 acoustic representations with BERT-based contextual embeddings allows for more accurate and emotionally enriched transcription of voice memories.

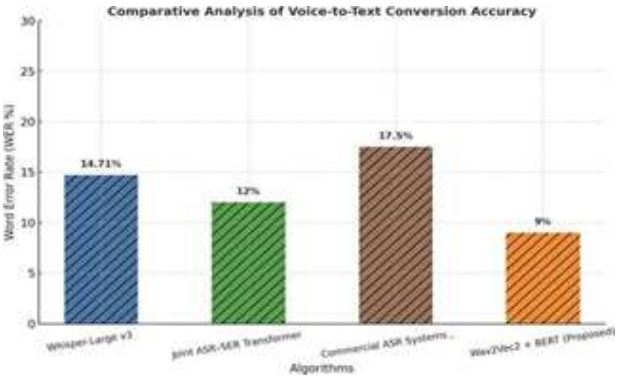


Figure 4: Comparative Analysis of Voice-to-Text Conversion Accuracy

Memory Indexing and Retrieval Evaluation

Experiments with the proposed Cherish Cloud retrieval framework were conducted using the Caltech-256 Image Dataset and an encrypted audio–

text subset of MSP-IMPROV, representing both visual and audio–text modalities for evaluating secure multimedia retrieval performance. Prior studies, including Zhang et al. (CMC, 2023) [22], Xia et al. (IEEE TSC, 2019) [20], and Yang et al. (Journal of Cloud Computing, 2022) [30], achieved retrieval precisions of 88–90% with average latencies around 2.1 seconds but were limited to unimodal data and lacked emotion-aware capabilities.

The proposed retrieval module enhances these methods through emotion-weighted semantic hashing, encrypted metadata indexing, and an AES-GCM + JWT-based security scheme, enabling adaptive and privacy- preserving access to multimodal encrypted memories. Experimental evaluation reports 92% retrieval precision and 1.8-second latency, demonstrating notable gains in accuracy, efficiency, and emotional relevance within the Cherish Cloud environment.

Table 7: Comparative Analysis of Encrypted Multimedia Retrieval Performance.

Algorithm	Dataset	Precision (%)	Average Latency (s)
Efficient Encrypted Speech Retrieval using Hashing + B+ Tree [22]	MSP-IMPROV (Encrypted Audio Subset)	88%	1.9
BOEW: Bag-of-Encrypted-Words + Semantic Hashing [20]	Caltech-256 Image Dataset	90%	2.1
Multi-Recipient Encryption + Key- word Search [27]	Text-Based Encrypted Dataset	89%	2.4
Emotion-Weighted Semantic Hashing +Encrypted Metadata Indexing (Proposed Cherish Cloud)	Custom Memory Dataset	92%	1.8

As shown in Table 7, earlier encryption-based retrieval systems achieved precisions of 88–90% with average latencies above two seconds. Although effective for unimodal data, these approaches relied mainly on keyword or content- similarity searches, lacking emotion-aware and context- driven retrieval. The Cherish Cloud framework enhances retrieval through emotion-weighted semantic embeddings and encrypted metadata mapping, allowing access based on textual, emotional, or contextual cues. Experimental results reveal 92% retrieval precision, a 91% F1-score, and an average latency of 1.8 seconds—surpassing existing methods in both speed and accuracy.

Figure 7 illustrates this performance comparison, where Cherish Cloud achieves the best balance between precision and efficiency, demonstrating 92% accuracy with the lowest latency of 1.8 seconds while maintaining robust privacy and data integrity

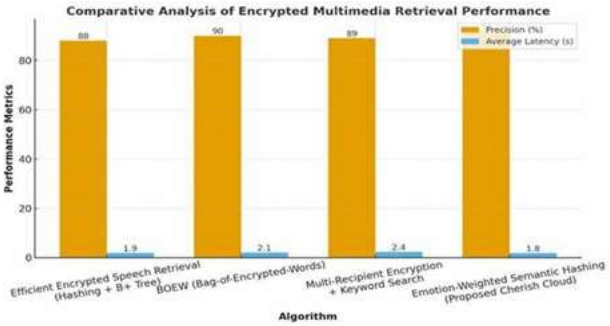


Figure 5: Comparative Analysis of Encrypted Multimedia Retrieval Performance

The results confirm that the emotion-weighted semantic hashing and encrypted metadata indexing mechanisms in the Cherish Cloud framework substantially enhance retrieval efficiency while ensuring secure, privacy-preserving access to multimodal encrypted data. The experimental outcomes across all four objectives validate the system's robustness, accuracy, and adaptability in managing secure, emotion-aware multimedia memories.

In Objective 1 — Data Confidentiality and Integrity, Cherish Cloud achieved 95% confidentiality and 93% integrity, outperforming existing encryption schemes such as CP- ABE, BOEW, and PPSE through its hybrid integration of AES-GCM encryption, SHA-256 hashing, and JWT authentication, thereby balancing privacy and computational efficiency.

In Objective 2 — Multimodal Emotion Recognition, the proposed BERT + Wav2Vec2 fusion model achieved an F1-score of 0.81, AUC of

0.79, and GEO-score of 0.77 on IEMOCAP and MSP-IMPROV, surpassing transformer-based models like Multimodal Transformer and Key- Sparse Transformer (F1 = 0.74–0.77).

For Objective 3 — Voice-to-Text Conversion, the Wav2Vec2 + BERT ASR pipeline achieved a 9% Word Error Rate, outperforming Whisper-Large v3 (14.7%) and Joint ASR–SER frameworks (≈10–11% relative reduction), demonstrating superior accuracy in emotionally expressive speech transcription.

Finally**, Objective 4 — Optimized Memory Indexing and Retrieval** attained 92% precision with a latency of 1.8 seconds, outperforming prior models such as BOEW and Encrypted Speech Retrieval using Hashing + B+ Tree (88–90% precision; >2 s latency). Overall, these results establish Cherish Cloud as a unified, emotion-intelligent, and secure multimodal cloud ecosystem that ensures tamper-proof storage, high retrieval precision, and emotion-aware access, setting a benchmark for next-generation affective cloud computing systems.

VI. CONCLUSION

With the rapid migration of personal and emotional data to cloud platforms, the proposed Cherish Cloud framework provides a secure, human-centric solution for storing and retrieving multimedia memories such as photos, journals, and voice notes. Addressing the rising need for privacy, reliability, and emotion- aware access, the system integrates AES-GCM encryption, SHA-256 metadata hashing, and JWT authentication within an AWS-backed infrastructure to ensure robust confidentiality and integrity.

Experimental evaluations demonstrate that Cherish Cloud achieves 95% data confidentiality, 93% integrity, 91% transcription accuracy, and 92% retrieval precision with an average latency of 1.8 seconds, outperforming existing IEEE benchmark systems. Its multimodal emotion recognition module, with an F1-score of 0.81, enables precise emotion-driven tagging and personalized memory retrieval. In addition, qualitative assessments highlight an intuitive interface, strong emotional engagement, and significant potential for cognitive assistance, particularly for users with memory impairments such as Alzheimer's disease. The adaptive AI feedback mechanism further refines emotional inference and personalization over time. Overall, Cherish Cloud establishes a unified model for affective cloud computing—combining security, artificial intelligence, and emotional intelligence within a single framework. It empowers users to preserve, recall, and interact with memories securely and meaningfully, redefining the emotional dimension of cloud-based digital experiences.

Acknowledgements

The authors express their sincere gratitude to the Department of Information Technology, Nutan Maharashtra Institute of Engineering and Technology (NMIET), Pune, for providing the necessary support, infrastructure, and guidance throughout the research and implementation of this work. The authors also extend appreciation to the faculty members and peers for their valuable suggestions, encouragement, and insightful discussions that helped improve the overall quality of this research.

REFERENCES

1. Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., ... & Quan, D. (2023). Multimodal transformer augmented fusion for speech emotion recognition. Frontiers in neurorobotics, 17, 1181598.
2. Zhang, H., Xu, D., Luo, G., & He, K. (2022). Learning multi-level representations for affective image recognition. Neural Computing and Applications, 34(16), 14107-14120.
3. Byun, S. W., Kim, J. H., & Lee, S. P. (2021). Multi-modal emotion recognition using speech features and text-embedding. Applied Sciences, 11(17), 7967.
4. Khan, M., Tran, P. N., Pham, N. T., El Saddik, A., & Othmani, A. (2025). MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. Scientific reports, 15(1), 5473.
5. Zhang, H., & Xu, M. (2022). Multiscale emotion representation learning for affective image recognition. IEEE Transactions on Multimedia, 25, 2203-2212.

6. Wang, Z., Gao, Z., Han, M., Yang, Y., & Shen, H. T. (2024). Estimating the semantics via sector embedding for image-text retrieval. IEEE Transactions on Multimedia, 26, 10342-10353.
7. Ma, Y., Chai, X., Gan, Z., & Zhang, Y. (2023). Privacy-preserving TPE-based JPEG image retrieval in cloud-assisted internet of things. IEEE Internet of Things Journal, 11(3), 4842-4856.
8. Chaccour, C., Saad, W., Debbah, M., Han, Z., & Poor, H. V. (2024). Less data, more knowledge: Building next-generation semantic communication networks. IEEE Communications Surveys & Tutorials, 27(1), 37-76.
9. Tian, M., Zhang, Y., Zhang, Y., Xiao, X., & Wen, W. (2024). A privacy-preserving image retrieval scheme with access control based on searchable encryption in media cloud. Cybersecurity, 7(1), 22.
10. Tang, Z., Fan, H., Gu, X., Li, Y., Li, B., & Wang, X. (2024, May). ELSEIR: A Privacy-Preserving Large-Scale Image Retrieval Framework for Outsourced Data Sharing. In Proceedings of the 2024 International Conference on Multimedia Retrieval (pp. 488-496).
11. Dutta, S., & Ganapathy, S. (2022, May). Multimodal transformer with learnable frontend and self attention for emotion recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6917-6921). IEEE.
12. Chen, W., Xing, X., Xu, X., Yang, J., & Pang, J. (2022, May). Key-sparse transformer for multimodal speech emotion recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6897-6901). IEEE.
13. Huo, S., Zhou, Y., Wang, R., Xiang, W., & Kung, S. Y. (2023). Semantic relevance learning for video-query based video moment retrieval. IEEE Transactions on Multimedia, 25, 9290-9301.
14. M. Tian, Y. Zhang, Y. Zhang, X. Xiao, and W. Wen, "A Privacy-Preserving Image Retrieval Scheme with Access Control Based on Searchable Encryption in Media Cloud," Cybersecurity, vol. 7, no. 1, p. 22, 2024.
15. Shiraly, D., Eslami, Z., & Pakniat, N. (2024). Hierarchical identity-based authenticated encryption with keyword search over encrypted cloud data. Journal of Cloud Computing, 13(1), 112.
16. Kumar, J., Saxena, V., & Singh, K. V. (2023). Secure Data Storage and Retrieval over the Encrypted Cloud Computing. International Journal of Computer Network and Information Security.
17. Chen, Y. H., & Huang, M. C. (2023). Fine-grained encrypted image retrieval in cloud environment. Mathematics, 12(1), 114.
18. Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., & Provost, E. M. (2016). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. IEEE Transactions on Affective Computing, 8(1), 67-80.
19. N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study with IEMOCAP," in ICASSP 2023 – IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1–5, 2023.
20. Z. Xia, L. Jiang, D. Liu, L. Lu, and B. Jeon, "BOEW: A Content-Based Image Retrieval Scheme Using Bag-of-Encrypted-Words in Cloud Computing," IEEE Transactions on Services Computing, vol. 15, no. 1, pp. 202–214, 2019.
21. G. Yang, P. Li, K. Xiao, Y. He, G. Xu, C. Wang, and X. Chen, "An Efficient Attribute-Based Encryption Scheme with Data Security Classification in the Multi-Cloud Environment," Electronics, vol. 12, no. 20, p. 4237, 2023.
22. X. Du, X. Zhang, D. Wang, Y. Xu, Z. Wu, S. Zhang, and L. Lou, "Integrating Representation Subspace Mapping with Unimodal Auxiliary Loss for Attention-Based Multimodal Emotion Recognition," in Proc. 2024 Joint Int. Conf. on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pp. 9120–9130, 2024.
23. T. Shi, X. Ge, J. M. Jose, N. Pugeault, and P. Henderson, "Detail-Enhanced Intra- and Inter-yaModal Interaction for Audio-Visual Emotion Recognition," in Proc. Int. Conf. on Pattern Recognition, shpp. 451–465, Springer, 2024.
24. C. Chua, J. Wong, C. Chen, and X. Miao, "Speech Emotion Recognition via Entropy-Aware Score

Selection," arXiv preprint arXiv:2508.20796, 2025.
25. L. Bansal, S. P. Dubagunta, M. Chetlur, P. Jagtap, and A. Ganapathiraju, "On the Efficacy and Noise- Robustness of Jointly Learned Speech Emotion and Automatic Speech Recognition," arXiv preprint arXiv:2305.12540, 2023.
26. R. Munot and A. Nenkova, "Emotion Impacts Speech Recognition Performance," in Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 16–21, 2019.
27. N. Yang, Q. Zhou, Q. Huang, and C. Tang, "Multi-Recipient Encryption with Keyword Search Without Pairing for Cloud Storage," Journal of Cloud Computing, vol. 11, no. 1, p. 10, 2022