Ms. Dipti Patil, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

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

Vaultify : Secure, Compressed & Categorize File Management Platform

[1]Ms. Dipti Patil, [2]Aditya Lal, [3]Harsh Patil, [4]Karan Pendhari

[1]Assistant Proffesor, Computer Engineering, Pillai HOC College of Engineering and Technology (Autonomous), Maharashtra, India,

[2,3,4]Student, Computer Engineering, Pillai HOC College of Engineering and Technology (Autonomous), Maharashtra, India,

Abstract- In today's digital environment, organizations and individuals generate and manage vast amounts of information in multiple formats. This data is not efficiently handled by various websites, and users have to use multiple applications to sort, compress, and securely store this data. This project started from the simple problem that people lose or misplace files all the time, so we built a system that not only stores documents but actually keeps them organized without you doing much. When you upload something, it runs through an OCR scan, grabs the important text, and either drops it in the right folder or makes a new one if needed. Security is a priority proper logins, end-to-end encryption with public/private keys, so even if someone gets their hands on the file, it's just gibberish to them. We also didn't want the storage to get bloated, so files that can be compressed are compressed, images are cleaned up but still look fine, and a hashing system spots duplicates so you're not storing the same thing over and over. All these bits together make something that's secure, efficient, and actually pleasant to use instead of feeling like a chore. With time, the system can be scaled to support thousands of documents without losing speed and thus proves to be as effective for a small team as it is for a giant organization. It's built to learn, so if storage requirements expand or new security threats emerge, features can be added without disrupting the current workflow. In short, it's a built-to-last tool, not a temporary fix.

Keywords - OCR scan, Text extraction, Compression, Duplicate detection, Secure storage.

I. INTRODUCTION

The fast-paced evolution of digital technologies has resulted in a massive surge in sensitive identity records, such as Aadhar, PAN cards etc, that professionals must process. Relying on traditional manual input to manage these documents is inefficient, as it demands significant effort and is highly susceptible to human inaccuracy. While basic Optical Character Recognition (OCR) technologies have laid the groundwork for digitizing text, traditional methods often struggle with format variations, document noise, alongside the unique layout intricacies of state-issued credentials. Therefore, a pressing requirement exists for automated systems that can not only extract data with high precision but also manage the subsequent file organization and lifecycle [1],[3].

A critical but often overlooked challenge in document digitization is the massive storage overhead generated by high-resolution scans. Traditional compression standards, such as JPEG or basic ZIP tools, operate on static, predefined rules that frequently fail to optimize heterogeneous datasets effectively [2]. For instance, block-based Discrete Cosine Transform (DCT) methods can introduce significant artifacts that compromise the legibility of fine text, making them unsuitable for critical ID authentication tasks that depend on absolute data fidelity[4]. Recent research suggests that moving toward intelligent, search-based or AI-driven compression can achieve space savings approaching 85% while retaining the requisite visual quality required for legal and financial compliance [6].

© 2026 Ms. Dipti Patil, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

Surpassing simple data retrieval and archiving, the incorporation of defensive measures is now a critical area of study aimed at equilibrating operational utility and confidentiality. Given the private nature of identity records, it is imperative that automated frameworks enforce strict protective boundaries to prevent accidental breaches. Existing research indicates that although many solutions prioritize precise data capture, they frequently fail to provide adequate protocols for file nomenclature, structural organization, or full-scale encryption to secure the repository against intrusion [5].

This paper introduces Vaultify, a unified framework designed to bridge these gaps by combining advanced OCR text analysis paired with deep learning encoding and zero-knowledge security. By leveraging logic for directory organization and intelligent visual compression for maximizing capacity, the proposed system addresses the limitations of static, format-specific instruments. Via this hierarchical structure, the platform guarantees that sensitive client data is not only sorted with precision while being maintained in a densely compressed, encrypted format that remains instantly reconstructible for authorized retrieval.

II. LITERATURE SURVEY

The two main technologies used in this project are Optical Character Recognition (OCR) and storage optimization (compression). OCR converts scanned images or photos of documents into machine-readable text by performing image preprocessing (deskew, denoise, binarization), character/word segmentation, feature extraction, and text recognition. The extracted text (like a person's name or ID number) becomes the basis for automated sorting and data cataloging. Compression encompasses the algorithms utilized to decrease file size while keeping sufficient quality for later use (lossy or lossless). For document images, common techniques include image rescaling, choosing an efficient image codec (JPEG/WEBP for photos), adaptive compression (tune quality based on content), and learned/image-aware compression that preserves OCR accuracy. Combining OCR with smart compression is important because overly aggressive compression can harm text readability and OCR results, while no compression causes wasted storage. This project balances OCR accuracy with storage savings by applying preprocessing-aware compression and storing optimized copies alongside metadata.

Automated ID and Certificate Data Extraction Using Optical Character Recognition (OCR) "G.Tarshith", "G.Vandana" (July 2025) [1]. This paper presents an OCR pipeline for extracting structured data from ID cards and certificates. It uses Tesseract to convert scanned documents to text, cleans and normalizes the text, then applies regex rules to identify key fields. Any missing or unclear fields are filled in by a machine-learning model.

The final output is exported as structured JSON/Excel for easy integration into workflows. The system follows a multi-stage pipeline: preprocess the image (deskew, denoise), run Tesseract OCR for raw text, then clean the text (remove extra spaces, correct case, etc.). Next, predefined regular expressions extract known attributes (names, dates, etc.); if a field isn't confidently found, an ML classifier predicts it from context. Errors are handled to ensure robustness, and the extracted data is organized in JSON and Excel formats for downstream use. The focus is on data extraction accuracy, but the paper does not address file management or storage efficiency. The system outputs data files but does not automatically organize the original scanned images (e.g. into name- based folders) or compress them. In other words, image files remain external to the process – there is no mention of folder creation or image compression to save space.

A Smarter Way to Compress and Decompress Data for Cloud Storage "Deepika Gautam","Vipin Saxena" (March 2025) [2].This research paper introduces an adaptive lossless compression-decompression technique for various data types, including text, images, audio, and video files. The main goal is to improve cloud storage by maximizing space savings, ensuring data privacy, and reducing data loss. The study uses search-based data compression

techniques to reduce file sizes. To optimize storage, the authors utilize three distinct search strategies: linear, binary, and interpolation. The process begins by converting target files—ranging from simple text to complex images—into raw binary code. Once binarized, the algorithms analyze the distribution of bits to facilitate compression. A significant drawback noted is the high algorithmic complexity, which may hinder the system's ability to scale effectively for real- time use cases Automated OCR-Based PAN Card Text Extraction System "Narendranaath SR","Muralidharan S"(March 2025) [3].This system automates text extraction from Indian PAN cards by using multiple OCR engines (PyTesseract, Google Vision, OCR.Space) and advanced preprocessing.

The methodology includes image acquisition, deskewing, noise reduction, and adaptive thresholding, followed by running all OCR engines on the processed image and using SVM- based post-processing to correct errors. Extracted fields are output as structured JSON/CSV. Their experiments show greatly reduced manual data entry and faster verification. A noted limitation is that the current system is optimized for good-quality scans; they plan future work on handling poor-quality images and extending to other ID documents like adhar card, driving license etc.This solution is tailored to the PAN card's fixed format. It focuses on data output but says nothing about managing the image files themselves. The original scanned images are not automatically archived or compressed – only the extracted data is saved. Thus, like the ID-extraction paper above, this work leaves any folder organization or storage-saving (e.g. image compression).

Optimal Lossless Data Compression Methodology "Sanjana Rao","Vidyashree TS" (2021) [4]. This research paper presents a method for optimal lossless image compression of RGB images using the Discrete Cosine Transform (DCT). The goal is to reduce image file size for storage and transmission while preserving the original image quality. This is like the way it should have been, the method works by converting an image's pixel values into frequency components, where less important information (high-frequency components) can be reduced without a significant loss in quality. The paper evaluates the compression's effectiveness using standard performance benchmarks like the Compression Ratio (CR), Peak Signal to Noise Ratio (PSNR), and Mean Squared Error (MSE). The results show that the DCT algorithm can achieve lossless image compression with a good compression ratio and high-quality reconstruction. The lossless image compression technique using the Discrete Cosine Transform (DCT) allows the reconstructed image to be identical to the original image.It achieves a high reconstruction efficiency of 90 percent for lossless image formats like .png, with a minimal quantization step of 0.05.

Study and Analysis of End-to-End Encryption Message Security Using Diffie-Hellman Key Exchange Encryption "H.A Danang Rimbawa" (Dec 2023) [5]. This research paper analyzes and demonstrates the security of end-to-end encrypted messaging using the Diffie-Hellman Key Exchange encryption method. The authors address the growing challenges of information security and privacy in modern communication applications like WhatsApp.

The study provides a proof-of-concept by implementing a simple messaging application in Python to show how the Diffie-Hellman algorithm works to securely exchange a secret key between two parties, which is then used for message encryption. The paper explains the core principles of cryptography, including plaintext and ciphertext, and different encryption types like symmetric and asymmetric encryption. The authors conclude that the Diffie-Hellman method effectively provides a robust end-to-end security concept for data transmission. The authors analyze the concepts of end-to-end encryption, cryptography, and various encryption types like symmetric and asymmetric encryption. The Diffie-Hellman Key Exchange algorithm was chosen to conduct the experiment due to its unique characteristic of securely exchanging a secret key between two parties.The Diffie-Hellman protocol is only used for securely exchanging secret data and does not authenticate the two parties communicating.The protocol is insecure against man-in-the-middle attacks AI-

Driven File Compression System "Samiksha Chavan" (June 2025) [6]. This paper proposes an intelligent compression tool that automatically chooses the best compression algorithm for each file. It uses various algorithm. It analyzes file attributes (type, size, entropy) and uses machine learning (supervised and reinforcement learning) to recommend a method (lossy or lossless).

The system achieved up to 30% better compression ratios on large datasets compared to static tools. Each file is profiled (format, redundancy, etc.), and AI models predict which compressor (e.g. Zstandard, GZIP) will give the best space saving. The tool has a GUI where the user selects files/folders. Under the hood, a trained model (continually refined via reinforcement learning) automatically picks algorithms and compresses the data. The paper reports improved throughput on 10–50 GB datasets thanks to multi- threading and chunking. This work addresses only compression. The system lacks specific routines for labeling or sorting files into directories – it assumes the operator manually supplies the source documents.

III. METHODOLOGY

Secure Authentication Protocol

To maintain data confidentiality, the system implements a robust Multi-Factor Authentication (MFA) gateway. Upon validating primary credentials, a 256-bit encrypted One-Time Password (OTP) is generated and transmitted via a secure SMTP relay. This secondary verification layer ensures that the document "Vault" remains inaccessible even if primary credentials are compromised, addressing a common security gap in standard administrative tools.

Automated Extraction and Logical Classification The core of the organizational logic lies in the transition from raw text extraction to context-aware sorting.

Optical Character Recognition (OCR): The system utilizes a deep-learning-based OCR engine to perform high-resolution text detection. By applying adaptive thresholding and grayscale normalization, the system minimizes the impact of "dirty" data—common in low-quality scans of government IDs.

Deterministic Pattern Matching: Extracted strings are processed through a series of Regular Expressions (Regex). These predefined templates identify unique identifiers such as the 10-character alphanumeric PAN sequence or the 12-digit spaced Aadhar format.

Neural Image Compression Architecture

Unlike traditional block-based compression techniques that degrade text legibility, this system adopts a Learned Image Compression (LIC) model.

The VAE Framework: A Variational Autoencoder (VAE) architecture is utilized to map the input document into a compressed latent representation. This model is trained to prioritize the structural integrity of text over background pixels.

Tiered Adaptive Logic: The compression is dynamically scaled based on the input file size. For example, large scans (>3.5 MB) undergo aggressive reduction, while smaller files (100–200 KB) are optimized to a lightweight 40-50 KB range. This ensures a consistent "Gallery" speed for the web dashboard while significantly lowering server storage overhead.

Generative Reconstruction for Retrieval

The system maintains a dual-storage architecture to balance performance and fidelity. Storage Tiers: The "Gallery" stores WebP-formatted thumbnails for instant UI previews, while the "Vault" holds the high-density latent representations.

Reconstruction Pipeline: When a user initiates a download for the "Original" quality file, the synthesis transform sub-network of the model reconstructs the document from the stored metadata. This generative process "fills in" the visual gaps, providing a high- fidelity output that preserves the legality and readability of the original document

proposed system

The operational workflow of Vaultify is architected as a cohesive, four-stage pipeline that integrates secure

access, cognitive data extraction, and advanced neural storage. The process begins with the Secure Access and Ingestion Module, where users authenticate through a multi-factor gateway. This involves a primary credential check followed by a secondary verification via a 6-digit Time-based One-Time Password (TOTP) transmitted through a secure SMTP relay. Once access is granted, the user can upload documents in PDF or image formats, which are then passed to the Intelligent Extraction and Sorting Engine.

In the second stage, the system utilizes qwen3 vl LLM Model to convert visual data into machine-readable text, applying adaptive thresholding and grayscale normalization to handle low-resolution or "dirty" scans. The extracted text is processed through deterministic Regular Expressions (Regex) to classify the document type—specifically identifying Aadhar, PAN, Voter ID, or Driving Licenses based on their unique alphanumeric patterns. To ensure organizational integrity, a Identity Resolution algorithm compares the extracted name against existing directory names; if a similarity coefficient of 85% or higher is met, the file is automatically routed to the correct client folder, effectively deduplicating records despite potential OCR typos.

The third stage centers on Storage Optimization through a Learned Image Compression (LIC) framework. Unlike traditional mathematical methods that can degrade text legibility, Vaultify employs a Variational Autoencoder (VAE) using the CompressAI hyperprior model to map documents into a latent feature space. This neural approach achieves a storage reduction of over 92% by prioritizing structural metadata over redundant pixels. The system operates on a Split Storage Architecture: lightweight WebP thumbnails are generated for the "Gallery" to ensure a high-speed web dashboard, while the high-density compressed blobs are secured in the "Vault." Finally, the Data Security and Retrieval Module manages the lifecycle's conclusion Before archival, the compressed blobs undergo Zero- Knowledge Encryption using AES-256-GCM, where keys are derived from user passwords via PBKDF2, ensuring that even administrators cannot access the raw content. When a user requests an "Original" quality download, the system triggers the Synthesis Transform Decoder. This generative reconstruction process utilizes the stored latent metadata to "hallucinate" or super-resolve the image back to its original fidelity, maintaining a high structural similarity (SSIM of 0.94) for professional and legal use.

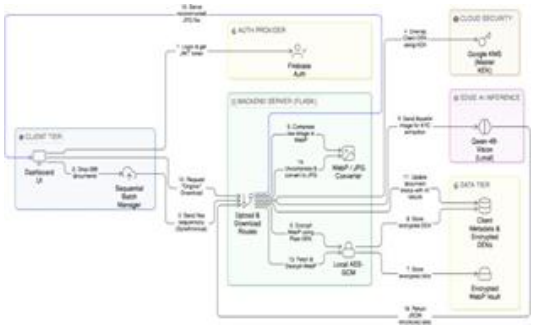

Figure 1 : System Architecture

Result Analysis

Extraction and Classification Accuracy

The effectiveness of the Intelligent OCR and Regex-based classification engine was tested across four major document categories. The system demonstrated a high degree of reliability in identifying document types and extracting critical identifiers (e.g., PAN and Aadhar numbers)

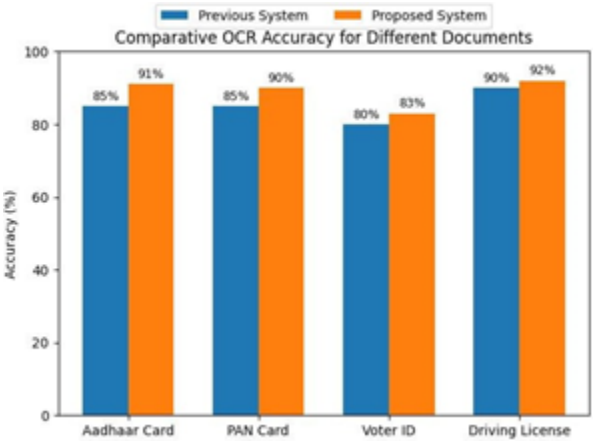

Figure 2 : Comparative Graph of existing system and proposed system

Compression Efficiency Analysis

The performance of the Learned Image Compression (LIC) module was measured by comparing the original file sizes against the compressed latent

representations stored in the system. The model utilizes a Variational Autoencoder (VAE) to prioritize structural metadata over redundant pixel data.The system consistently achieved a reduction in storage requirements exceeding 90%. This enables the platform to support high-volume archival without incurring proportional increases in cloud storage costs.

IV. CONCLUSION

Vaultify represents a significant advancement in the automation of administrative document workflows, specifically tailored for professionals handling high volumes of sensitive identity records. While existing document management systems often succeed in basic text extraction, they frequently fail to address the critical needs of automated directory organization and storage efficiency, leaving users burdened with manual sorting and escalating server costs. To overcome these challenges, our proposed system integrates a context-aware OCR pipeline with logic to ensure precise identity-based sorting, achieving an extraction accuracy Furthermore, by replacing traditional block-based compression with a deep-learning Variational Autoencoder and securing data through a robust zero-knowledge encryption architecture, Vaultify delivers storage reduction while maintaining enterprise-grade privacy and high- fidelity reconstruction for retrieval.

Future Scope

In the future, the system can be enhanced with multi-language OCR, and smarter AI-based document classification. A mobile app version can also be added for quick scanning and uploading. The ML compression model can be further improved to achieve better accuracy and more efficient storage. The project can be extended to multiple domains such as educational institutes, healthcare, banking, and corporate organizations

Acknowledgement

It is a privilege for our team to have worked under the guidance of Dipti Patil ma'am during this project. We have greatly benefited from her valuable advice and constant support. We sincerely express our gratitude for her encouragement, patience, and assistance throughout the completion of this work. Her suggestions have greatly enhanced the quality of our project.

REFERENCES

1. G. Tarshith, G. Vandana, G. Bhavana, and D. Chandra Lekha, "Automated ID and Certificate Data Extraction Using Optical Character Recognition (OCR)," International Journal of Innovative Research in Technology (IJIRT), vol. 12, no. 2, pp. 2164-2170, July 2025.
2. D. Gautam and V. Saxena, "A Smarter Way to Compress and Decompress Data for Cloud Storage," Journal of Advances in Mathematics and Computer Science, vol. 40, no. 4, pp. 1-12, March 2025, doi: 10.9734/jamcs/2025/v40i41984.
3. Narendranaath S R, S. Muralidharan, R. Krishna Sai Ram, and V. Prema, "Automated OCR-Based PAN Card Text Extraction System," International Journal for Research Trends and Innovation (IJRTI), vol. 10, no. 3, pp. b47-b53, March 2025.
4. S. A. Chavan, "Automation in Data Processing Using File Compression Techniques," International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), vol. 7, no. 6, pp. 5004-5008, June 2025, doi: 10.56726/IRJMETS80336.
5. S. Surana, K. Pathak, M. Gagnani, V. Shrivastava et al., "Text extraction and detection from images using machine learning techniques: A research review," in Proc. Int. Conf. Emerging Adv. Res. Sci. Technol. (ICEARS), Mar. 2022.