ESP-IJACT
ESP International Journal of Advancements in Computational Technology
ISSN: 2583-8628 / Volume 3 Issue 2 April 2025 / Page No: 1-9
Paper Id: IJACT-V3I2P101 / Doi: 10.56472/25838628/IJACT-V3I2P101

Original Article

Automated Document Pipelines: Deploying ML- Powered Workflows for End-to-End Archival and Retrieval

Tejas Dhanorkar[1], Shemeer Sulaiman Kunju[2], Swaminathan Sethuraman[3]
[1]Discover Financial Services, USA,
[2]HCL America Inc, USA,
[3]Visa Inc, USA.

Received Date: 14 February 2025 Revised Date: 25 March 2025 Accepted Date: 08 April 2025

Abstract *— Modern enterprises struggle with inefficient manual document processing, leading to productivity losses and compliance risks. In this research , we introduce the creation of an archival and retrieval of the document to an end endpoint by means of a ML driven pipeline. Specifically, we are merging these two components (ingest, pre_join, class) and (metadata generate and search) from NLP and Computer Vision. Federated learning for privacy, vector embedding for retrieval, semantic retrieval, adaptive learning (domain adaption). With the policy of the security of our sources our immediate action was to store and adhere to the policy of the security of our sources of information: it can read many kind of sources other than from other automated emails, APIs, scanners etc. Certainly a lot of these gain a great deal in terms of processing speed and accuracy. When generative AI gets involved, like, for instance, generation AI summarization or real time stream process, the generative AI algorithms would have been included in future calculations. Also, this work will be very helpful for a business that wishes to gather a scalable and clever solution to an automation of document workflow.*

Keywords *— machine learner, information retrieved (NLP based), metadata indexing, Document automation, compliance, semantic search (etc).*

I. INTRODUCTION

The high levels of documentation in the contemporary digital first enterprise world spans of its organizations[1]. These documents being readily available in practice are of huge volume and also are unstructured as well as semi structured data of all these documents such as contracts and invoices etc which he scanned these forms with or absent handwriting[2]. However, It is the necessary condition to achieve the very successful enterprise scaling with compliance to strict regulatory framework (this percentage value is often not known by startup leaders, and is subjective)[3]. Today, many accept that one of their obligations is to manage documents overnight in a routine and effective way in order to be able to make sure good regulatory compliance, ensure data integrity and finally to allow for timely decisions that can be as data driven as possible and as timely as possible.

Despite an expansion of the volume of the digital information, most [businesses] are still running on old, manual [4]. For example data entry, order files, marking of data's metadata, search and so on are such repetitive tasks. Today it is the case that 48 percent of employees find it difficult to locate documents, 47 percent fail to use company's recently made annoying and obstrusive digital file system. [5] Based on real numbers of cost, Foxit Software has calculated that bad document management costs the organization 21.3 percent less productivity [6]. In a 2022 article published in [7] it is mentioned that most of the business that bypasses the paper based workflow turns out to be 95% of the document management systems available on a digital front. For something to be adopted this fast, you know documents have been taken as a strategic enabler for the efficiency of business and business resilience in the digital document pipe. Furthermore, this market is expected to be worth USD 8,32 Billion in 2025 and USD 24,34 Billion in 2032 with CAGR of 16.6 percent from 2025 to 2032 [8].

While there's no ideal or perfect machine in its system of manual document processing, the organizations have now been made conscious of the automation and ML present in the current streams of documents processing. For example, in particular there are certainly not a doubt for the methods of data retrieval, data entry, and data retrieval as it is on manual classification etc [9]. In the opposite way, ML systems to which they suffer from are automation of document classification, key entity extraction from an article, and understanding the content of the article. Such capabilities also offer increased accuracy along with increased speed of processing in different formats and languages on the implementation of these capabilities [10]. Now the mathematical reason of why flexibility is not the way to be efficient. Not the ML, nor scale by adding ML, but just more ML (to correct more of what has happened). But being flexible scales better.

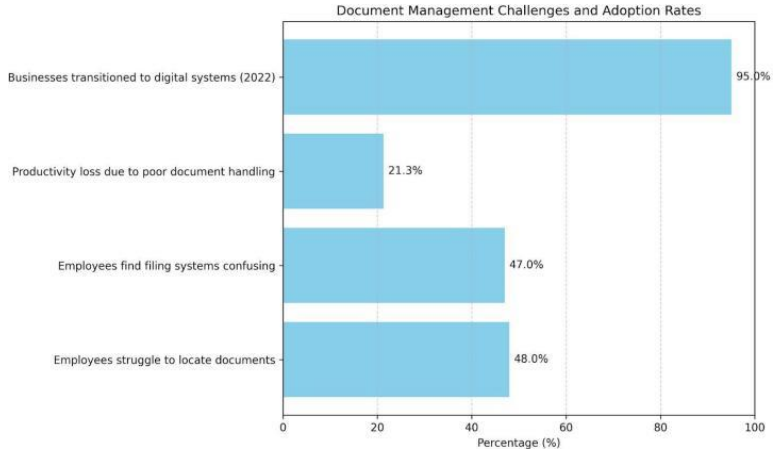

Figure 1: Document Management Challenges and Adoption rate

The rest of the paper is concerned with research on the feasibility of employing ML within the automated document document document pipeline for document archival and retrieval. It is an end to end process from ingestion level to one or more preprocessing, storage and search. Then it details the architecture of these automated systems, details in writing, of the materials that form a part of this system in relation with the initiation or starting of this system. In particular, the paper also describes requirements to designing such systems required to handle a sensitive, restricted data set in the industry.

A. Key Objective:

This research is guided by the following key objectives:

- To design and present an end-to-end architecture for automated document pipelines that integrates machine learning for efficient archival and retrieval processes.
- To investigate the role of ML techniques such as document classification, entity extraction and semantic search in enhancing the performance, accuracy and scalability of document management workflows.
- To analyze current challenges and future opportunities in deploying intelligent document systems, with emphasis on compliance, data governance and enterprise-wide integration.

II. RELATED WORK

It analyzes processing system documentation for manual and rule based schemes and extends it to more sophisticated machine learning (ML) based methods[2]. Secondly, it is directly connected with the fact that business documents of the modern businesses have become much more complex and voluminous

A. Overview of Existing Document Processing Pipelines

On the other hand, the traditional document processing segment can be split by the most elementary way that is by handcrafted rules and predefine heuristics, but in a portable way using the ML approach. Here, if the motor structure and the unstructured or semi structure data is taken as an input, system fails to be flexible[11]. In this area, the algorithm that predicts, understand, recognise, evaluate a thing with a data, instead of looking in similarity of the things and start predicting about the things with a data which is given to them. They are also interfaced to NLP as well as Computer Vision, where the different documents are further processed and understood [12].

B. Traditional Rule-Based vs. ML-Based Systems

Such systems are determined and transparent, and used when the data consistency and transparency requirements are needed. A base can be used for an API to grow and grow; document growing structures are really unable to remain appropriate and usable. To generate generalization from the examples for scalability and adapt capability of ML based systems. But they are 'black boxes' that demand a lot of training data; whereas human decision processes are much less transparent as explained in Table 1.

C. Gaps in current systems (e.g., siloed tools, lack of adaptability)

Nevertheless, the speedup in the processing of the document has not resolved all the problems. Due to the fact that such organizations are using siloed tools,.integration issues and data inconsistencies are usual [15]. The traditional rule based systems are not adaptable and a complete reprogramming is needed in case a new set of document types or formats are required to be handled. On the other hand, ML systems are more flexible, but they are more worried about data privacy owing to the basis on massive data sets. However, building and maintaining ML models is quite resource intensive having both computational power

Table 1: Summary of Traditional Rule-Based vs. ML-

Model	Methodology	Objective	Outcome
NLP4PBM (2024) [13]	Systematic review of NLP-based process extraction methods	Evaluate rule- based, ML, and deep learning techniques for process extraction	ML/DL methods often outperform rule-based approaches; lack of annotated datasets remains a challenge
MLRB (2024) [14]	Comparative analysis of ML algorithms and rule-based embeddings	Assess performance of ML classifiers vs. rule-based methods in text classification	ML models like SVM and CNN generally outperform rule- based embeddings
Scaling systematic review with ML Pipeline. (2020) [9]	Application of ML in systematic literature reviews	Automate data extraction from scientific documents	ML enhances efficiency in systematic reviews, reducing manual effort
Open Review (2024) [2]	Review of AI techniques in unstructured document analysis	Explore AI methods for information extraction from diverse documents	AI-based techniques show promise; challenges persist with varied document layouts

Based Systems and a lot of expertise. Such automated solution for document processing depends upon the solutions addressing these challenges.

D. How your approach differs or extends existing literature

And finally presented as part of this thesis work is a machine learning (ML) powered document processing pipeline to fix limitations in current systems existing. It is a unified process that captures in one many stages in document processing and decreases the number of fragmented tools and increases the workflow performance. Moreover, the pipeline supports adaptability to new types of documents with little annotated data through adaptive learning, in particular: semi supervised and transfer learning. It is an attempt to prevent the Loss of privacy of Sensitive information by incorporating federated learning into it such that the model training could also take place over from decentralized sources of data without compromising privacy. Furthermore, models used are lightweight and can help embed the system in cloud solutions to utilize computational resources efficiently to make it easily available and scalable for the different organizational needs.

III. ARCHITECTURE OF AN AUTOMATED DOCUMENT PIPELINE

Table 2: Comparison of Proposed Approach with existing Methodologies

Component	Proposed Methodology	Alignment with Existing Literature	Outcome
Unified Document Processing Framework	Consolidates various stages of document processing into a single pipeline, reducing tool fragmentation and enhancing workflow efficiency [16].	Reflects the trend towards integrated Intelligent Document Processing (IDP) systems that combine OCR, NLP, and ML for streamlined workflows.	Identified state-of-the-art architectures and technologies in document automation, highlighting the impact of generative AI and LLMs on the field.
Adaptive Learning Techniques	Employs semi-supervised and transfer learning to improve adaptability to new document types with minimal labeled data [17].	Aligns with approaches that utilize semi-supervised and transfer learning to enhance model performance when labeled data is scarce.	Highlighted the use of NLP and text mining in automating study selection, quality assessment, and data extraction; emphasized the need for further research in data synthesis automation.
Privacy-Preserving Federated Learning	Incorporates federated learning, allowing model training across decentralized data sources without compromising privacy [19].	Mirrors the adoption of federated learning in scenarios where data privacy is paramount, enabling collaborative model training without data sharing.	Found that while various tools exist, challenges remain in standardizing methodologies and integrating automation across different stages of SLRs.
Resource Optimization	Utilizes lightweight models and cloud-based solutions to optimize computational resources, making the system more accessible and scalable[18].	Consistent with the deployment of ML models in cloud environments to leverage scalability and resource efficiency.	The GPT-3.5-turbo model achieved the highest ROUGE-1 score of 0.364, outperforming other methods; a GUI was developed for the best-performing system.

Being a structured system, an automated document processing pipeline is auto ingests, auto analyzes and caters the doc using a combination of machine learning, natural language processing, computer vision etc...[20,19]. As a result, it will enable transforming unstructured information into structured one, which is more correct and scaleable for many document types.

A. Document Ingestion

The first step in the world of automated document processing is called document ingestion and is related to obtaining and first preparing the documents received from various sources for further processing (in downstream). *Emails, APIs, scanners, and also cloud drives are data sources:* Now the days have advanced quite a bit and the modern document processing systems were able to solve the document processing problem for several sources including emails, APIs, scanners, and cloud storage platforms on a smooth, unified workflow. If, however, your key customers do not communicate other than through email, emails are the main means of communication of key documents, like mail or contracts, which can be pulled automatically from some specific inboxes. It enables us to directly integrate with many of the applications and to have a real time exchange with these platforms (CRM, ERP...) for soliciting the document [22]. Physical Documents like books, film, documents and even maps or other physical entities can be scanned so that they can be machine readable and are then made machine readable using OCR technologies which converts such physical documents into machine readable formats on users' demand.

The most important benefit is that these platforms have cloud storage (Google Drive, Dropbox, One Drive) and the file is evenly stored in one place, it can be recovered automatically and synchronize, which means that all the files are up to date before processing begins. With technology, whereby the information is being consumed, information through different sources will actually be consumed and a good document intake solution is made in a way that it will receive all the information, process it, and classify it, and analyze it in a way that the operational process will flow quickly with the data which is available and it's approved quickly into its business processes. *File type normalization:*After documents are acquired, they are file type normalized (i.e. TXT to PDF) [22]. For the same reason mentioned above, there should be this standardization and it's important in ensuring that consistent and compatible processing stage can be reach at least to some extent, which means the same of achievement to maximum level more or less.

B. Preprocessing Layer

Preprocessing Layer [23] changes the input of raw raw document that is taken in automated document processing pipeline to structured and readable formats. The second layer is to be sure that documents are fully optimised to then be further accurately analysed in the next layer. *OCR and image cleanup:*At this stage, such documents and images are scanned in Optical Character Recognition to carry out conversion of visual text into machine readable. The purpose of cleaning up the current display text finds the removal of the de- skew matrix, removal of noise and increase contrast for the purpose of text recognition accuracy [24]. They used OCR engine and thus need to be preprocessed and have textual contents from different types of document.

*Text extraction:T*he structure of the sentence gets changed first and then we follow OCR, extract the text and implement them to get the appropriate information to extract. What we mean is that we will be able to parse the text without the uselesss like headers, footers, page numbers and end up with clean and well structured textual data [50]. Also work well on top of effective text extraction for the next stages of pipeline e.g. information retrieval, analysis. *Format unification (PDF, Word, images):* They differ in their nature and include PDFs, Word documents, image etc. These documents have a purpose of being transformed into a format that is then made consistent to be procided compatible. Once it's standardized like this, then the data extraction process, the prediction process, the storage process becomes much easier and much more streamlined and much more clear in a very simplified, standardised, digital document process.

C. ML - Powered Processing Layer

Without the ML Powered Processing Layer, it is impossible to have an automated document processing pipeline as it is based on machine learning to analyze and extract the structural data from the unstructured documents. It provides a layer that can handle the data and support making decisions on the above tasks such as document classification, entity recognition and information extraction conveniently and efficiently.

Document classification (e.g., SVM, CNNs, transformers): If we are talking about document classification, then, in such a case, machine learning is an automatic process of document classification and putting documents into a predefined class that is determined by some predefined algorithms. With support vector machines (SVM), CNNs and models based on Transformer, we analyze textual and visual features and using them we perform accurate classification of the document towards different document types [26]. *Entity Recognition (NER):* Named entity recognition (NER) is one of the information extraction sub tasks, used for identification and labelling of these named entities that are already present in unstructured text

in predefined categories such as person names, companies, locations, medical codes, time expressions, quantity, monetary value, percentages, etc. [27]. *Information Extraction (Key-Value Pairs, Tables):* Information extraction is one of the problems of extracting structured information from unstructured or semistructured document. It includes having this tabular data, one the one hand, extract a key value pair i.e. 'Invoice Number: 12345' and on the other hand use this for data analysis or decision making processes. Even how that information is rendered and structured in complex document layouts, it uses deep learning models that bring advanced techniques to do that precisely.

D. Metadata Generation and Indexing

In the context of medical document processing pipeline auto path, the ability by which the Metadata Generation and Indexing stage takes processed content and creates searchable content is of a critical nature. Metadata creation and organizing this well provides for better retrieval classification and semantic understanding to this layer. *Creating searchable metadata:* Indexing and searching the processed content of the original text through metadata will process the metadata into structured, searchable data, and is a vital step in creating such metadata. The metadata is the names, authors, dates and the keywords etc of the document, and allows the representation of content of the document in a concise form for fast indexing and retrieval here. Additional metadata enrichment techniques, such as applying LLMs, can be more advanced to carry out this process. LLM can help solve problems with filtering search results better, closer to relevance by reducing that to numbers which hold a fitting criteria such as location, rating, category etc.

Ontology/taxonomy mapping: The semantic consistency is also maintained and information retrieval made possible by mapping the extracted metadata and content to ontologies or taxonomies. It maps out the information in terms of the structured hierarchies and relationship of the document repository, thus making it more intuitive to navigate or find what is inside the document repository. *Embedding-Based Vector Storage (e.g., FAISS, Pinecone):* Thus, documents and their metadata are taken as vector embeddings of the semantic meaning for the advanced search capability. Vector databases can store the embeddings in FAISS or Pinecone, which performs efficient similarity search and handle use cases like semantic search or recommendation system.

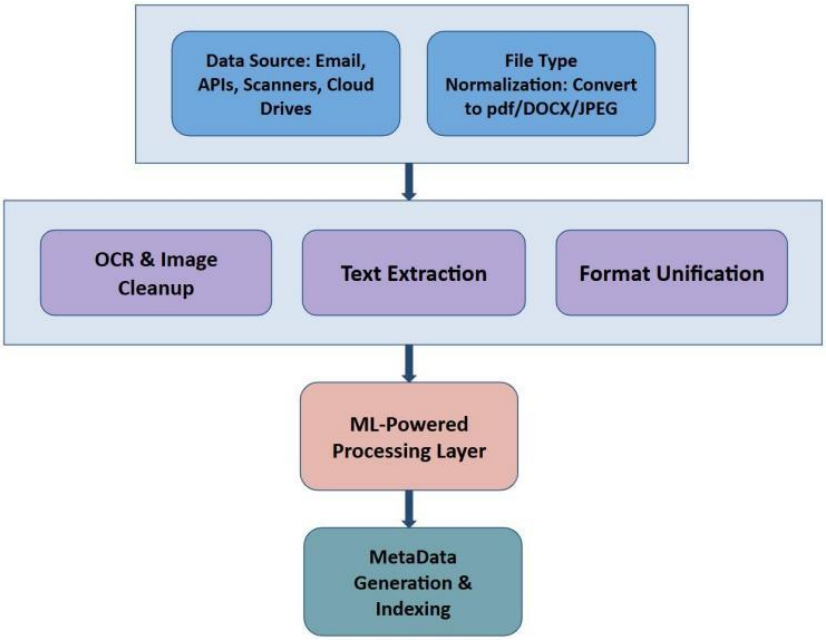

Figure 2: End-To-End Architecture of An ML-Powered Document Processing Pipeline, From Ingestion To Structured Retrieval

IV. ARCHIVAL STRATEGY

The Archival Strategy is a crucial component of an automated document processing pipeline, focusing on the secure, compliant, and efficient long-term storage of processed documents. This strategy ensures data integrity, facilitates regulatory compliance, and optimizes storage resources over time.

A. Secure, compliant long-term storage (e.g., WORM, blockchain audit trails)

To maintain data integrity and meet regulatory requirements, archival systems employ Write Once, Read Many (WORM) storage solutions. WORM storage ensures that once data is written, it cannot be altered or deleted, providing a tamper- proof environment essential for preserving sensitive information such as financial records and legal documents [28]. Additionally, integrating blockchain-basThen, they can attach themselves to a strong and brilliant audit trail with

another record of all the access and modification to which it has been subjected by them to its ledgers. This can be further enhanced security.

B. Compression and deduplication

The storage is made as efficient as possible by using data compression and deduplication techniques. There, each time the data file size gets reduced due to the data being compressed and deduped and there are the data entries of the data. Cumulatively these methods make the cost of storage less but hinder from retrieving the original data and data integrity as well.

C. Tiered Storage Options (Hot/Cold Archive)

Implementing a tiered storage architecture allows for the categorization of data based on access frequency. Frequently accessed data is stored in 'hot' storage tiers, offering rapid retrieval times, whereas infrequently accessed data is moved to 'cold' or 'archive' tiers, which are cost- effective but have longer access times. This stratification optimizes storage costs and ensures that data retrieval aligns with organizational needs. Governance and Retention Policies There are organizations that need to adopt well defined, comprehensive governance and retention policies to take control of their data, comply with legal and regulatory requirements and strive to minimise risks. They are how the data should be stored efficiently, kept compliant, store instead of risk, minimize an easy way for mismanagement [28]. The idea of a data retention policy comes into action when a company describes how long data of different categories should be kept and in what way they should be safely disposed when the retention period expires. More importantly, this approach ensures the data integrity and minimizes organizations from the surplus and obsolete data which negates the occurrence of data breaches and cyber attack. the data access so that the non tech users can use the information nicely.

D. Context-aware document ranking

Instead of prioritizing the results based on a search, ranking systems are itself context awareness (user's search history and behavior) based. The systems can learn from past user interactions, which lets them make more accurate user intent inference, and therefore will be able to provide documents that are more likely to fulfill the user's needs. Dynamic ranking increases the information retrieval relevance and efficiency.

E. Use of vector search engines

The basic idea of Vector search engines is to represent document and query representations as high dimensional vectors, where high dimensional means the number of dimensions is large, each dimension represents some semantic meaning and contextual nuances. Next, the system finds similarity of these vectors and is able to identify and return the documents associated with the user's query, though some of the keywords in the query are different. But this method will give you better accuracy and exactness of search results.

V. RETRIEVAL MECHANISMS

Retrieval Mechanisms component, which is very powerful, is vital for an efficient and accurate retrieval of stored information in the automated document processing pipeline. This layer also provides an advanced search layer that does the search on the basis of user query and returns the documents for the overall utility of this system.

A. Keyword vs semantic search

In the case of a traditional keyword search, terms used for a keyword search need to match words exactly from document content in order to get relevant using results if the user's intention is not cleared correctly. However, semantic search interprets the meaning of the queries that customers are typing and why, and using that information to make more relevant results for the end customer. It however helps in retrieving relevant documents for those cases where exactly the same keywords are not pulled.

B. Natural language queries using LLMs

With the integration of LLMs, the system can be used by natural language query; the user can interact with the system. For instance, such LLMs can be able to translate such queries into structured search commands that can access to such complex databases without the particular speaker knowing the query language. But this democatizes.

VI. CHALLENGES AND LIMITATIONS

The obvious evidences of these automated document processing pipelines is that these pipelines yield an immense amount of efficiency but at same time an extreme amount of challenges and limitation which is kept by the organization that is running them to achieve an effective and a secured running of the process itself.

A. Data Privacy and Security

The risks it must have without derogating is mandatory data privacy and security in the handling of sensitive such as financial documents, personal identifiers, legal document and so, and if not stopped, these include unauthorized

access, data breach, EU General Data Protection Regulation (EU GDPR) noncompliance, California's Privacy Rights Act (CCPA) and so on. In order to protect the data value in action during the data processing period within the processing pipeline, the encryption, access control and the compliance audit must be implemented.

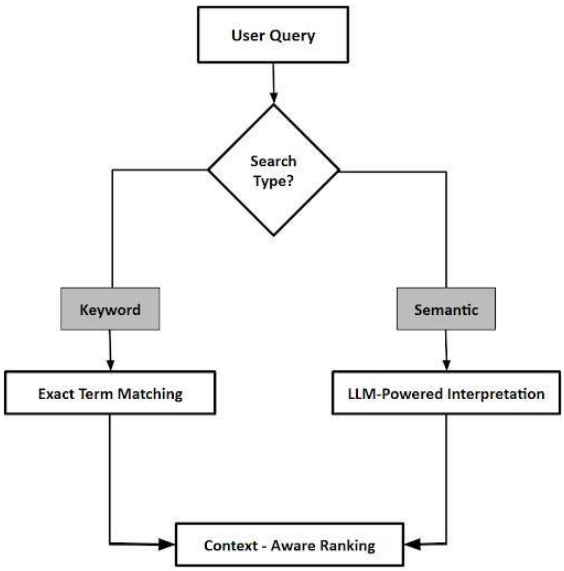

Figure 3: Retrieval Workflow in ML-Powered Document Pipelines

B. Training Data Quality and Labeling

In general, machine learning models for processing the documents are heavily tied to the quality of the training data. Nevertheless the data is not diverse enough, lacks of standardization on labeling, and containing human errors during the annotation work. In other words, such bias or wrong model outputs may exist. Overall, you need to outline very strict data labeling and Quality Assurance protocol rules to increase the model performance.

C. Domain-Specific Adaptation

The reason for that is that as a matter of fact, document processing systems in general are not generally suitable to be flexibly adapted to specific domains such as in the legal, medical or financial field. With the general purpose models, this term and structure of document is not understood, hence you get lower accuracy. More specifically, frequent training on a given domain specific way with 'knowledge' from 'an expert' can be of help only in a limited set of context.

D. Scalability and Cost

This is especially useful for making these systems scale to process such high volumes of data generated from the inside of a high throughput, low volume domain or more likely, a high throughput per document, low document volume space, with a low throughput, high volume document oriented processing. This is the face of challenges that are revolving around keeping costs of infrastructure to minimum, while not losing speed and quality of work. We need to implement something such as scalable architectures like cloud based or even to appeal to the reasoning of more efficient resource allocation to solve these issues in this way.

VII. FUTURE WORK

Finally, Future Work provides some ideas for using clever technologies and techniques in automated document processing pipelines in order to compensate shortcomings and help achieve functionality.

***A.* Integration with generative AI for summarization or Q&A**

Therefore, document processing systems are adopting more and more generative AI, and with this functionality of sumarization and question answering, the document processing systems should be more and more automated. The document learned on can furnish informative contents and enables choosing based on the documents learned on, which then facilitates compact summarization of documents as well as accurate responses to user queries for large amounts of documents. While generative AI is being used by platforms like amazon textex to aid in the data cleaning process and clean up data before being processed, generative AI is creating normalized key fields as well as generation of summary of the input data; reducing errors. Real-time document stream processing

These depend on that real time data stream processing since they can ingest and analyze real time generated documents. This has to be taken into consideration when the capabilities of an application that needs to process data fast comes into play. The group, however, faces data volume, data velocity and system scalability challenges in reality when the

time processing is actually implemented, and then Apache Kafka and Flink are applied to store and process continuous data stream.

B. Multilingual and cross-format support

In this, the file format and language required for it to become global, thus document processing systems should also be developed so that documents can be processed in different languages and file formats. Moreover, these could carry the ability to process the documentation in any language or for that matter in any format. To solve these needs we have built solutions for solving the needs of comprehensive multilingual content processing, most notably Amazon Bedrock in conjunction with Human in the loop mechanism.

C. Human-in-the-loop for continuous learning

Conversely, HITL provides continuous learning and model improvement and they could be integrated in the pipelines of document processing. Human reviewers can afford to provide feedback on model outputs, which can go some way to improving your algorithms where there is a degree of uncertainty or error. This collaborative framework is particularly important because of the documents processing systems that it makes available to the users of its users in complex or sensitive domains, whatever their value for their greatly enhanced accuracy and reliability.

VIII. CONCLUSION

However, automated document processing systems are now revolutionized information management tools, despite that they provide very important improvements in terms of accuracy, efficiency, and scalability. This has been a review of the architectures, user feedback, challenges and trends of the future of these systems. Real world implementation of Studies have shown that you can revolutionize the workflow, but the above implementations are neither that productive nor that accurate nor that scalable as of integration technologies like Optical Character Recognition (OCR), Natural Language Processing (NLP) and Machine Learning (ML). These technologies both let you perform an automatic data extraction, classification, processing from the different document types without much manual intervention and with the minimum errors. One of the main findings is that automated document processing emerges as the way to do so as a result of the ability to significantly reduce operations time due to greater speed in matching and processing data. However, there still exist problems in data privacy and security, quality of training data to maintain, adaptation to the domain specific needs and scaling and cost. However, in order to deploy and operate such systems so successfully such challenges have to be addressed.

This is very crucial for document management efficiency. These systems help the organizations to streamline the work flow, reduce processing time and also enhance the accuracy of data handling. It helps to reduce costs, ensure better compliance with the regulatory standards and makes better decisions. Additionally, it also allows thinking about how a company would shift from manual to automated procedures as well as how to allocate better the resources and their effort on more strategic aspects of the company. Organizations that may want to implement automated document processing systems are a focus on scalability. Since bigger data is not only available, but also coming in bigger, we need to design such systems that do not inconvenience the performance. Nevertheless, it also depends on user training, change management and integration with existing infrastructure. A phased implementation approach, with continuous evaluation and optimization, will have an implemented approach and will result in successful adoption and scalability for the new systems to satisfy the changing organizational needs.

REFERENCES

[1] A. Almeman, "The digital transformation in pharmacy: embracing online platforms and the cosmeceutical paradigm shift," J. Health Popul. Nutr., vol. 43, no. 1, p. 60, 2024.

[2] S. V. Mahadevkar et al., "Exploring AI-driven approaches for unstructured document analysis and future horizons," J. Big Data, vol. 11, no. 1, p. 92, 2024.

[3] S. B. Moore and S. L. Manring, "Strategy development in small and medium sized enterprises for sustainability and increased value creation," J. Cleaner Prod., vol. 17, no. 2, pp. 276–282, 2009.

[4] S. Jordan, S. S. Zabukovšek, and I. Š. Klančnik, "Document Management system–a way to digital transformation," Naše Gospod./Our Econ., vol. 68, no. 2, pp. 43–54, 2022.

[5] Business.com, "7 Statistics That Will Make You Rethink Your Document Management Strategy," 2023. [Online]. Available: https://www.business.com/articles/7-statistics-that-will-make-you- rethink-your-document-management-strategy

[6] Foxit Software, "10 Document Management Stats You Need to Know," 2023. [Online]. Available: https://www.foxit.com/blog/just-the-numbers-10-document-management-stats-you-need-to-kno

[7] PDF Reader Pro, "25 Document Management Statistics You Should Know," 2023. [Online]. Available: https://www.pdfreaderpro.com/blog/document-management-statistics

[8] Fortune Business Insights, "Document Management System Market Size, Share & COVID-19 Impact Analysis, By Component and Regional Forecast, 2025–2032," 2024. [Online]. Available: https://www.fortunebusinessinsights.com/document-management- system-market-106615

[9] D. Baviskar et al., "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic

literature review and future directions," IEEE Access, vol. 9, pp. 72894–72936, 2021.
[10] E. e Oliveira et al., "Unlabeled learning algorithms and operations: overview and future trends in defense sector," Artif. Intell. Rev., vol. 57, no. 3, p. 66, 2024.
[11] G. Chen, B. An, and S. Zeng, "A rule-based information extraction system for human-readable semi-structured scientific documents," in Proc. 4th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT), 2015, vol. 1.
[12] X. Chen, H. Xie, and X. Tao, "Vision, status, and research topics of Natural Language Processing," Nat. Lang. Process. J., vol. 1, p. 100001, 2022.
[13] W. Van Woensel and S. Motie, "NLP4PBM: a systematic review on process extraction using natural language processing with rule-based, machine and deep learning methods," Enterp. Inf. Syst., vol. 18, no. 11, p. 2417404, 2024.
[14] A. M. Aubaid, A. Mishra, and A. Mishra, "Machine learning and rule-based embedding techniques for classifying text documents," Int.
[15] J. Syst. Assur. Eng. Manag., 2024J. Patel, "Bridging data silos using big data integration," Int. J. Database Manag. Syst., vol. 11, no. 3, pp. 1–6, 2019.
[16] M. A. Achachlouei et al., "Document Automation Architectures: Updated Survey in Light of Large Language Models," arXiv e-prints, arXiv:2308, 2023.
[17] G. Sundaram and D. Berleant, "Automating systematic literature reviews with natural language processing and text mining: A systematic literature review," in Int. Congr. Inf. Commun. Technol., Springer, Singapore, 2023.
[18] N. F. Ali et al., "Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation," arXiv e-prints, arXiv:2411, 2024.
[19] F. Saeed et al., "Employing Federated Learning for the Implication of Digital Twin," in Digital Twins for Wireless Networks: Overview, Architecture, and Challenges, Cham: Springer Nature Switzerland, 2024, pp. 93–122.
[20] V. Bellandi et al., "Streamlining Legal Document Management: A Knowledge-Driven Service Platform," SN Comput. Sci., vol. 6, no. 2,
[21] pp. 1–17, 2025.
[22] H. A. R. I. P. Mandava, "Streamlining enterprise resource planning through digital technologies," J. Adv. Eng. Technol., ResearchGate, 2024.
[23] U. Kampffmeyer, Ed., Conversion & Document Formats: Backfile Conversion and Format Issues for Information Stored in Digital Archives, vol. 2, PROJECT CONSULT GmbH, 2002.
[24] J. S. Chu, Automated pipelines for information extraction from semi- structured documents in structured format, Ph.D. dissertation, Massachusetts Institute of Technology, 2023.
[25] K. M. O. Nahar et al., "Recognition of Arabic air-written letters: machine learning, convolutional neural networks, and optical character recognition (OCR) techniques," Sensors, vol. 23, no. 23, p. 9475, 2023.
[26] Q. Zhang et al., "Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction," arXiv preprint, arXiv:2410.21169, 2024.
[27] P.-Y. Hao, J.-H. Chiang, and Y.-K. Tu, "Hierarchically SVM classification based on support vector clustering method and its application to document categorization," Expert Syst. Appl., vol. 33, no. 3, pp. 627–635, 2007.
[28] S. R. Kundeti et al., "Clinical named entity recognition: Challenges and opportunities," in Proc. IEEE Int. Conf. Big Data, 2016.
[29] P. L. Bradshaw et al., "Archive storage system design for long-term storage of massive amounts of data," IBM J. Res. Dev., vol. 52, no. 4.5, pp. 379–388, 2008.