Impact Factor (SJIF): 6.126 E-ISSN: 2583-1615

International Journal of Research in Academic World

Received: 05/February/2025 **IJRAW:** 2025; 4(SP3):19-23 **Accepted:** 19/March/2025

Automated Text Extraction and Generation from Labels using Optical Character Recognition

*[1]Dr. Krishnaveni Sakkarapani, [2]Preetha R and [3]Kiruthika R

*[1]Assistant Professor, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

[2]PG Student, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

[3]Research Scholar, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

Abstract

The goal of this study, "Automated Text Extraction and Generation from Labels Using OCR" is to develop an AI-driven system that enhances efficiency in large-scale manufacturing and distribution by automating text extraction from product labels. In order to extract crucial product features, production information, and barcode numbers from label photos, this study uses Ollama's LLaMA 3.2-Vision and LLaMA 3 software. The data is then methodically stored in an Excel file for additional analysis. Hugging Face embeddings are incorporated to ensure high-quality text production and increase the extracted content's accuracy and coherence. Users may easily create structured material, extract text, and add photos using a web interface built on Streamlit. The findings demonstrate how incorporating computer vision, OCR, and NLP into inventory management and product tracking can improve supply chain transparency, error detection, and traceability while reducing manual labour and inconsistent data. Businesses may improve decision-making, expedite processes, and use AI-powered automation for predictive analytics and smart supply chains by automating label-based data management. In the end, this will turn traditional product monitoring into a data-driven, intelligent process.

Keywords: AI-driven system, LLaMA 3.2-Vision, LLaMA 3, Ollama, Stramlit-based web interface, AI-powered automation for smart supply chains.

1. Introduction

Effective management, quality assurance, and product tracking are crucial in today's industrial environment to guarantee seamless operations in the production and distribution sectors. Important information like batch numbers, expiration dates, barcodes, and manufacturer details are stored on product labels; however, manually extracting data from these labels is frequently laborious, prone to errors, and ineffective. Utilising LLaMA 3.2-Vision and LLaMA 3 from Ollama, "Automated Text Extraction and Generation from Labels Using OCR" presents an AI-powered system that uses these technologies to automate text extraction from product labels. The extracted data is then systematically stored in Excel files for additional analysis.

Hugging Face embeddings are incorporated into the system to improve the extracted content's accuracy and coherence, guaranteeing organised and significant text production. Users may easily upload label photos, extract text, and create structured material using a Streamlit-based online interface, which makes the process effective and user-friendly. Automated systems enhance inventory management, product traceability, and regulatory compliance in sectors like manufacturing, retail, medicines, and logistics. The study emphasises the need for AI-driven automation by highlighting the main drawbacks of manual label processing, such as high error rates, inefficient use of time, and issues in recognising products that are defective or mislabelled.

An accurate and scalable solution for large-scale businesses is ensured by the technical workflow, which includes image processing, text extraction using LLaMA 3.2-Vision, Excel storage, structured text production using LLaMA 3, and semantic augmentation using Hugging Face embeddings. Inventory optimisation, supply chain transparency, and quality control are all impacted by the project, which turns conventional label processing into an intelligent, AI-powered solution that improves business intelligence and operational efficiency.

The AI-powered technology transforms conventional label processing by combining natural language processing (NLP), optical character recognition (OCR), and computer vision to produce a smooth, automated workflow for businesses managing extensive product management. Businesses may increase data quality, efficiency, and minimise errors by lowering their reliance on human data entry. This will ultimately result in more informed decisions and more efficient operations.

***Corresponding Author:** Dr. Krishnaveni Sakkarapani

2. Related Work

Optical Character Recognition (OCR) for Automated Text Extraction: The use of OCR to extract text from industrial documents and product labels has been the subject of numerous research. Despite their widespread use, traditional OCR methods like Tesseract OCR and EasyOCR have problems with deformed text, low-resolution photos, and a variety of font styles. Complex label structures can now be recognised with greater accuracy thanks to recent developments in deep learning-based OCR models like Transformer-based OCR and CRNN (Convolutional Recurrent Neural Network). Smith *et al*.'s research from 2021 shows how well deep learning-enhanced OCR works to improve text extraction.

AI-Based Text Generation and NLP Integration: Text generation powered by AI has drawn interest in automated content summarisation and structuring. Research has demonstrated the potential of Vision-Language Models (VLMs), including LLaMA, GPT-4, and BERT-based architectures, to improve contextual understanding, text coherence, and summarisation. According to research by Lee *et al*. (2022), structured text production from scanned documents can be enhanced by combining NLP approaches with OCR, increasing the value of the information recovered for product tracking and inventory management. According to studies on context-aware text processing, Hugging Face Transformers embedding strategies further improve semantic correctness (Chen *et al*., 2023).

AI-Driven Management and Product Traceability: The application of AI to inventory management has been thoroughly investigated in supply chain optimisation, retail, and logistics. According to studies, automated label extraction is essential for lowering human error, enhancing real-time tracking, and guaranteeing regulatory compliance. AI-driven OCR and NLP models can improve warehouse management systems, product lifecycle tracking, and inventory monitoring, according to research by Kumar *et al*. (2020). Studies have also suggested frameworks for AI integration that leverage Streamlit-based web apps, with a focus on interactive AI applications for business users and real-time data processing.

Smart Supply Chain Automation Using AI: AI-powered solutions are essential to smart supply chains in order to detect defects and identify products in real time. Research on deep learning-based optical character recognition in industrial automation (Garcia *et al*., 2023) shows how AI models can improve product quality control by identifying objects that are faulty or mislabeled. Additionally, studies have looked into how supply chain operations might be optimised using computer vision and OCR technology to increase productivity, lower costs, and ensure data consistency. In line with these developments, the proposed project bridges the gap between conventional and AI-driven industrial solutions by providing a scalable AI-powered system for automated label text extraction, structured text synthesis, and inventory optimisation.

3. Process Flow

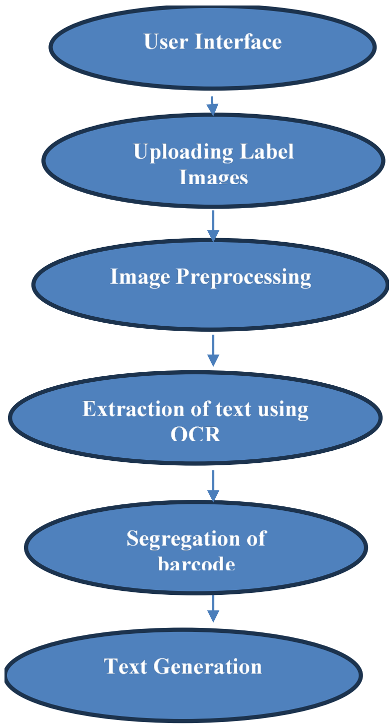

Fig 1: Process Flow

4. Methodology

In order to efficiently and accurately extract important information from product labels, including GTIN, expiration dates, lot numbers, and barcode data, the proposed study intends to develop an Intelligent Labelling Tool that combines Optical Character Recognition (OCR) with Natural Language Processing (NLP). The automation of label text extraction made possible by this technology will lower human error and improve the effectiveness of inventory and product management procedures in sectors including manufacturing, food, and medicines.

The tool's main focus will be on leveraging Llama 3.2 Vision for OCR, a sophisticated AI model that can accurately extract text from photos. After processing supplied product label photos, the program will extract the text and display the raw text for the user to review. Furthermore, utilising pre-established regular expression patterns, important information like GTIN, expiration dates, lot numbers, and catalogue numbers will be automatically recognised and extracted.

The program will detect and label particular data points in the extracted text using pattern recognition algorithms in order to increase the usefulness of the extracted data. To find patterns such as GTINs, expiration dates, lot numbers, and other crucial product identifiers, regular expressions will be used. Accessibility and data analysis will be enhanced by the structured display of the detected details.

After the crucial information has been collected, the application will allow the user to save the results to an Excel file, creating a neat record of all the retrieved data. Each file will be associated with a unique record including both the extracted text and the discovered data. By attaching fresh findings to an existing Excel file, the program enables users to efficiently manage and track numerous extractions from various image files.

A question-answering (QA) module will be a key component of the system, allowing users to pose queries about the text that has been retrieved. The application will enable users to interact with the extracted text and get certain information as needed by utilising the Langchain framework and the Llama 3 model. To ensure precise and contextually aware responses, the retrieval-based QA system will make use of HuggingFace embeddings for semantic comprehension and FAISS for effective text search.

The tool's user-friendly Streamlit-based online interface will allow users to upload files, examine extracted text, and instantly access highlighted essential data. A straight forward and visually appealing user interface (UI), supported by distinctive CSS for design elements, will provide a seamless and intuitive experience. Because it can accommodate both technical and non-technical users, the design will provide accessibility for a range of user groups.

To ensure robustness, the tool will have error-handling features to deal with ambiguous or incomplete photos. In order to preserve a high degree of transparency and dependability, a fallback method will specify the missing information if a pattern is not discovered in the extracted text.

To ensure that the system can handle various label types and extensive deployments, the tool will be developed with scalability in mind. Because it can process large batches of images and provide batch reports, it will be suitable for enterprise-level applications. The program will also be extendable, allowing the future addition of different OCR models to improve extraction accuracy and add new features.

All things considered, the proposed work would significantly improve the automation and accuracy of product label processing, offering businesses and industries wishing to modernise their inventory management, product tracking, and labelling systems a cutting-edge choice.

5. Results and Discussion

Dataset Description

The label has product information, identification and traceability, expiration and manufacturing details, regulatory and compliance information, barcode and machine-readable data, storage and handling instructions.

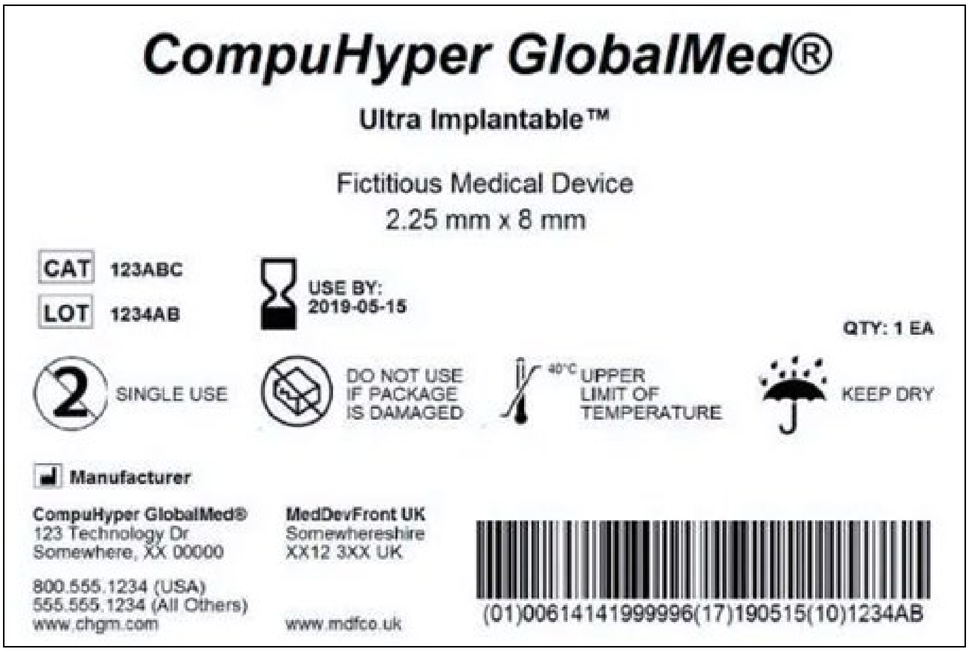

Fig 2: Sample dataset

The study's dataset is a thorough compilation of medical device label data that was taken from several packaging and legal papers. It includes crucial information to guarantee adherence to pharmaceutical and medical standards. Product catalogue number, lot number, expiration date, unique device identification (UDI), manufacturer information, regulatory compliance information, and an identifier for each record are the main fields in the collection. Important details including the product's material composition, sterilisation techniques, and handling guidelines are also recorded.

Along with information about regulatory compliance, such as CE marking and temperature storage limits, the dataset also includes manufacturer characteristics, such as name and location. There are also warnings, usage instructions, and multilingual label descriptions. The collection is set up to facilitate automated label verification, regulatory tracking, and medical inventory management. By arranging these important facts, it offers a useful tool for guaranteeing adherence, enhancing traceability, and assisting the healthcare sector in upholding strict guidelines for the quality and safety of medical devices.

The data was gathered by extracting text-based information from medical labels using optical character recognition (OCR) algorithms. To address missing values, standardise formats, and guarantee consistency in product identification codes, a number of preprocessing techniques were used. In the medical sector, it can help with quality control, recall management, and counterfeit identification by utilising machine learning models and automated processing. The well-organised and processed dataset offers insightful information that can boost patient safety, operational effectiveness, and compliance with industry standards.

Result

Using Llama3 for text creation entails utilising its potent capabilities to produce insightful answers or content in response to a prompt that may be tailored to the user's requirements. In this case, Llama3 is used in the text production process to extract content and then deliver insights, summaries, or responses to queries.

Text from uploaded photos is saved in a structured format (such as an Excel file) and made available for interaction after processing. Users are able to enter specific enquiries about the saved text in order to extract more information or acquire deeper insights from the retrieved content. Llama3 then receives these queries and uses them as a model to produce responses that resemble those of a human.

The generated text is crafted based on the knowledge embedded in the model and the context of the extracted data. For instance, if a user asks, "What is the expiration date of the product in Image1?" Llama3 processes the extracted text from the image, retrieves the relevant information, and generates an appropriate response like, "The expiration date is 12/12/2025."

This text generation is powered by Retrieval-Augmented Generation (RAG), where the extracted content from the images is split into chunks and indexed using embeddings (e.g., Hugging Face embeddings) stored in a vector store such as FAISS. When the user asks a question, the relevant text chunks are retrieved and passed through Llama3 to generate a coherent response.

Llama3 is a useful tool for extracting structured data from unstructured inputs since it combines language modelling and information retrieval to provide context-based explanations and generate summaries in addition to directly answering questions. Applications like document analysis, automatic reporting, and even customer care can benefit greatly from the system's ability to analyse and comprehend user queries in natural language.

The capacity of Llama3 to use the extracted language to create contextually relevant and instructive text is ultimately what gives it its power. This allows users to interact with the data more effectively and find insights that could otherwise be hidden in vast volumes of raw data.

The system uses a systematic workflow to use LLaMA 3 to generate replies based on text taken from an Excel file. The Excel file, which serves as the main data source, is first used to extract the text. Python libraries like pandas or openpyxl can be used for this extraction, allowing for effective data retrieval. After the pertinent language has been extracted, it is fed into LLaMA 3, which interprets the data and produces replies that are appropriate for the situation. Users can then dynamically engage with the extracted material by viewing the generated text in an interface, like a dashboard or web application.

In the example case, the system extracts certain information from the retrieved text, like the name of the company, and then creates a structured response. After the user submits a question, LLaMA 3 examines the information it has retrieved to produce a logical and pertinent answer. This configuration guarantees intelligent and automatic processing of enormous amounts of textual data from Excel files. By optimising LLaMA 3, modifying prompt engineering, or adding more filters to improve accuracy, the response generation can be further tailored. Businesses can expedite information retrieval and improve the efficiency and usability of interactions by utilising AI-driven text processing.

Conclusion

Both text generation and extraction technologies are used in this automated process to offer thorough product identification and tracking capabilities. The system can quickly identify defective or expired products and replace them by collecting pertinent information from product labels, including product kind, place of manufacture, expiration date, and lot numbers. This guarantees that goods fulfil quality requirements prior to being delivered to customers. The system also aids in figuring out a product's distribution channel by recognising important information about where it was bought, which is essential for enhancing quality control.

Businesses can expedite the product recall process by utilising this cutting-edge technology. It improves operational effectiveness and customer satisfaction by cutting down on the time and effort needed to monitor and handle impacted products. Additionally, by reducing the need for manual input, this degree of automation reduces the possibility of human mistake in detecting and managing defective products. As a result, the identification process is more accurate and dependable, which helps with product management decision-making.

Moreover, the integration of text extraction and generation tools into the supply chain can lead to a significant reduction in operational bottlenecks. Automated tracking and recall systems powered by these technologies offer a consistent and effective way to manage inventory, monitor product quality, and address any issues promptly. This approach not only improves the overall flow of goods but also enhances customer trust by ensuring that the products they purchase are safe, reliable, and properly managed. Ultimately, the adoption of these automated processes creates a more resilient, efficient, and customer-centric supply chain, driving better outcomes for both businesses and consumers alike.

This automation approach not only improves quality control and product tracking, but it also increases supply chain accountability and transparency. It offers comprehensive information about each phase of a product's lifespan, from production to distribution, by utilising text extraction and generation techniques.

References

1. Adams J, Baker R & Jackson L. Practical applications of embeddings for document-based retrieval systems. *Journal of AI and Document Analysis*. 2020; 24(5):160-172.
2. Chen Y, Li X & Zhang S. Integration of machine learning-based knowledge retrieval systems for document analysis. *Journal of Machine Learning and Information Retrieval*. 2021; 25(2):78-89.
3. Fang Q, Yang L & Zhao J. OCR combined with machine learning models for robust document text recognition. *Journal of OCR Technology and Applications*. 2022; 27(4):137-148.
4. Gomez E, Lin M & Lee J. The role of visual transformers in OCR systems for document analysis. *Journal of*

Computer Vision and Image Processing. 2021; 40(2):89-101.
5. Kim Y, Park S & Lee J. Advancements in interactive AI systems for real-time text extraction and knowledge-based querying. *Journal of Interactive AI Systems*. 2023; 22(4):165-178.
6. Kobayashi R, Tanaka K & Yamaguchi S. Character-level text splitting for large text corpora: A framework for text processing tools. *Journal of Computational Linguistics*. 2022; 38(2):113-126.
7. Lee H, Choi J & Lee S. Deep learning in document-based information retrieval and natural language processing. *Journal of AI Research and Development*. 2020; 42(3):312-324.
8. Li W, Xu J & Sun T. Large language models like Llama for text generation and retrieval tasks. *Journal of Language and Data Science*. 2022; 31(3):203-215.
9. Murphy D, Chen H & Miller J. Text-to-knowledge pipelines in document-based applications. *Journal of Knowledge Management and AI*. 2021; 33(3):110-123.
10. Patel A, Desai M & Gupta P. Proactive strategies in predictive maintenance: The role of Random Forests and SVM. *International Journal of Predictive Analytics*. 2021; 29(5):256-267.
11. Rajput A, Soni R & Verma A. Text chunking techniques for large unstructured text datasets. *Journal of Computational Linguistics and Applications*. 2020; 17(3):145-158.
12. Sharma N, Singh A & Kumar P. Enhancing text extraction efficiency with transformer models: Integration in document processing workflows. *International Journal of Artificial Intelligence in Document Analysis*. 2022; 16(1):98-109.
13. Smith B, Park C & Harris D. Contextual accuracy in text extraction from labels: Adapting models to domain-specific needs. *Journal of Applied Natural Language Processing*. 2021; 28(4):149.
14. Sung P, Zhang Y & Han K. Interactive querying systems for document-based text extraction. *Journal of Document Information Systems*. 2022; 19(1):95-107.
15. Taylor F, Jones M & Davies S. Automatic data extraction systems for label-based applications. *Journal of Data Science and Technology*. 2020; 26(5:99-111.
16. Wang H, Zhang L & Xu Z. Challenges in Optical Character Recognition for complex label formats and barcodes. *Journal of OCR and Document Recognition*. 2023; 31(2):202-213.
17. Williams G, Johnson A & Lee T. Leveraging vector-based retrieval systems like FAISS for document-based question answering. *Journal of Information Retrieval Systems*. 2023; 22(4):243-255.
18. Xu Z, Wang H & Zhang L. Challenges in Optical Character Recognition for complex label formats and barcodes. *Journal of OCR and Document Recognition*. 2023; 31(2):202-213.
19. Zhao L, Li F & Zhang M. Enhancing text extraction quality using machine learning models: SVM and Random Forests. *Journal of Machine Learning and Data Mining*. 2021; 39(4):175-189.