**Invoice Information Automation**

**Literature Review**:

**Introduction:**  
Invoice management serves as the core financial element which guides many organizations particularly those working in finance logistics sectors and e-commerce. Organizations handle many invoices received from various formats which results in manual processing taking longer time with higher error rates. The extraction of raw text from invoices has been done using Traditional Optical Character Recognition (OCR) systems since many years ago. The inability of these systems to achieve precise and organized data extraction becomes notable because invoice formats along with layouts and necessary contextual interpretation for accurate information retrieval vary widely.

Artificial Intelligence (AI) through its Small Language Models (SLM) demonstrates effective performance improvement for OCR systems when applied to functions like text structuring and field identification and contextual understanding. The designed capability of Small Language Models (SLM) to process information efficiently matches that of GPT-3 while remaining smaller yet retaining low Computational complexity. The combination of OCR with Small Language Models can boost invoice processing efficiency because it enables more precise accurate results which in turn eliminates manual work while maintaining data reliability.

The present literature review examines scientific research about combining OCR tools and Small Language Models, transformer-based techniques for invoice data extraction. The research will review different machine learning strategies which boost OCR performance by applying Small Language Models to improve text organization and information retrieval. This review will focus on existing business challenges with misclassification and data error problems alongside invoice layout diversity and present solutions from innovative AI systems. Research from both OCR and NLP technology allows the foundation to be built for developing a new hybrid model which uses both systems to enhance invoice processing efficiency.

**Scope:**

Challenges in Traditional OCR Systems:

The text extraction from invoices along with structured documents uses two primary OCR methods known as Tesseract and ABBYY FineReader. Traditional OCR systems encounter multiple obstacles while extracting information from invoices with detailed designs or handwritten entries or documents containing background noise. Numerous studies indicate that distorted text remains an obstacle for OCR models which results in subpar results when extracting critical financial information fields (Martínek et al., 2020). The semantic understanding capabilities of OCR systems remain limited because they struggle to recognize invoice-relevant elements such as invoice numbers dates and amounts according to Xue et al. (2020).

Small Language Models (SLM) for Enhanced Data Structuring:

Text extraction along with data structuring benefits from integration with Small Language Models through improved accuracy in OCR workflows. SLMs represent a lightweight solution for powerful text understanding that offers effective contextual awareness at a minimal computation cost. The study demonstrates that SLMs successfully process extracted text content to find patterns that lead to accurate invoice detail classification (Fukuda et al., 2025). SLMs excel specifically when the system calls for minimal training data combined with high performance levels (Kuzman & Ljubešić, 2025).

NER technology functions as a key information extraction system:

NER models alongside the integrated system now help invoice processing staff identify essential key points that include invoice numbers and dates and payment information. Digitized information becomes more precise when NER joins forces with OCR thus achieving elevated performance across finance and healthcare domains according to Xue et al. (2020) and Martínek et al. (2020). The NER model uses contextual information to pinpoint structured data precisely which leads to enhancement of system performance.

Hybrid Models: Combining OCR and SLM for Improved Efficiency:

Modern advancements have generated new models which unify OCR programming with SLM technology capabilities. Text extraction occurs through OCR while SLMs conduct automated data organization and remediate errors in these systems. The joint utilization of zero-shot learning and fine-tuning techniques makes hybrid models ready to process multiple invoice formats without significant retraining according to Lowe et al. (2024). Orthogonally to basic OCR outputs the OCR2SEQ model implements sequence-to-sequence learning to enhance text accuracy together with specialized domain data extraction in finance and healthcare applications (Lowe et al., 2024).

Industry Applications and Future Directions:

The use of OCR alongside SLMs brings extensive advantages to multiple industries in their operations. Automated invoice processing in finance operates through these tools which healthcare institutions similarly use to digitize medical information. SLM systems will gain more capability through transformer-based architecture advancements to better understand complex invoice structures according to Ponnurua et al. (2024) and Rehm et al. (2024).

Agentic Model in OCR and Invoice Processing:

When OCR and invoice data extraction systems work together using the Agentic Model their efficiency and accuracy improve because they leverage both OCR and NLP while performing contextual reasoning. The system demonstrates superior performance when processing documents with different layouts and multiple invoice styles and complicated patterns of information.

**DATA SOURCE IDENTIFICATION**:

1). From Hugging Face:

<https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1>

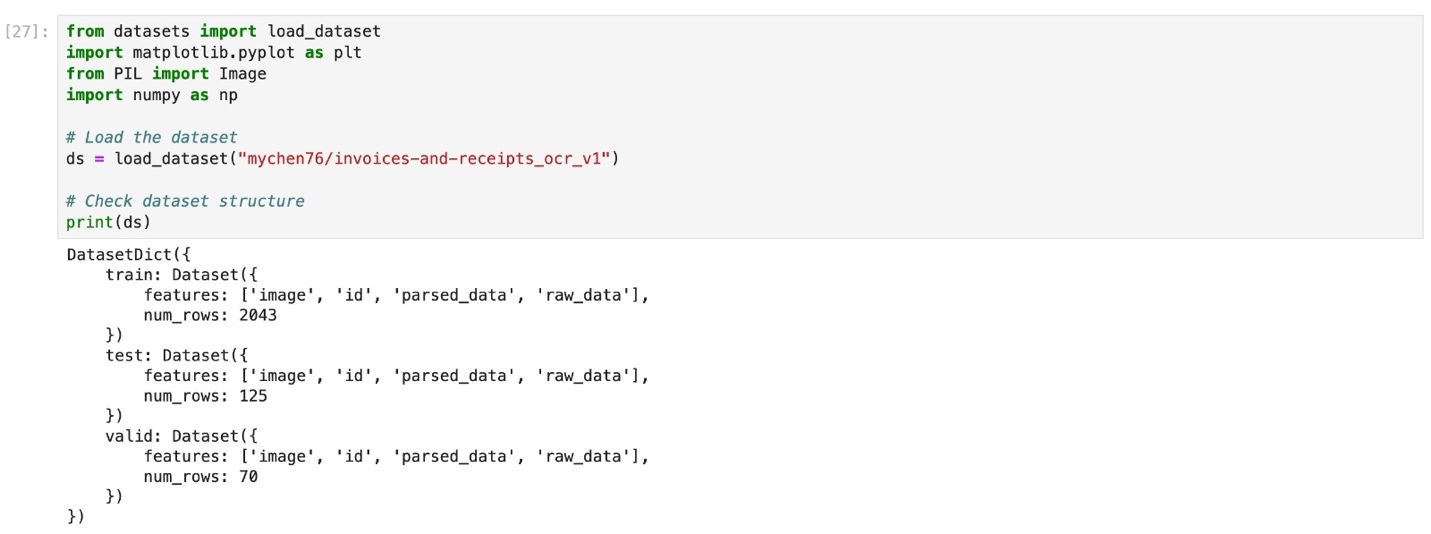
2). From GIT HUB:

<https://github.com/femstac/Sample-Pdf-invoices/blob/main/1000%2B%20PDF_Invoice_Folder.zip>

**Phase 1 Coding**:

Data Preprocessing:

Loading necessary libraries and dataset:



Function to display the sample image in the data set:

A screenshot of a computer program

Description automatically generated

Output:

A screenshot of a invoice

Description automatically generated

OCR using pytesseract:

A screenshot of a computer

Description automatically generated

**References:**

* Fukuda, N., Nozue, H., & Oishi, H. (2025). *Small Language Model Agent for the Operations of Continuously Updating ICT Systems*. *IEEE Access*.
* Kuzman, T., & Ljubešić, N. (2025). *LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification*. *IEEE Access*.
* Lowe, M., Prusa, J. D., Leevy, J. L., & Khoshgoftaar, T. M. (2024). *Advancing Machine Learning with OCR2SEQ: An Innovative Approach to Multi-Modal Data Augmentation*. *Journal of Big Data*.
* Martínek, J., Lenc, L., & Král, P. (2020). *Building an Efficient OCR System for Historical Documents with Little Training Data*. *Neural Computing and Applications*.
* Ponnurua, M., Ponmalar, S., Likhitha, A., Sree, T., & Chaitanya, G. (2024). *Image-Based Extraction of Prescription Information Using OCR-Tesseract*. *Procedia Computer Science*.
* Rehm, G., Dietze, S., Schimmler, S., & Krüger, F. (2024). *Natural Scientific Language Processing and Research Knowledge Graphs*. *Lecture Notes in Artificial Intelligence*.
* Xue, W., Li, Q., & Xue, Q. (2020). *Text Detection and Recognition for Images of Medical Laboratory Reports With a Deep Learning Approach*. *IEEE Access*.