Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Template-Free Approach to Invoice Digitization Leveraging SmolVLM and Heuristic Extraction

Md. Masud Rana[1, *], Md. Abu Hanif[2], Fakrul Islam[3], Md. Sahaib Mridha[4] Umma Habiba Maliha[1], Tahsina Tasnim Mumu[1]

[1] *Department of Computer Science and Engineering, IUBAT - International University of Business Agriculture and Technology, Dhaka-1230, Bangladesh; 22103055@iubat.edu (U.H. Maliha), 22103335@iubat.edu (T.T. Mumu)*

[2] *Department of Chemistry, IUBAT - International University of Business Agriculture and Technology, Dhaka-1230, Bangladesh;* hanif.chem@iubat.edu *(M.A. Hanif)*

[3] *Department of Computer Science & Engineering, Daffodil International University, Daffodil Smart City, Birulia 1216, Dhaka, Bangladesh; fakrul15-1316@diu.edu.bd (F. Islam)*

[4] *Department of Computer Science & Engineering, Manarat International University, Ashulia, Dhaka, Bangladesh; 2221cse50336@manarat.ac.bd (M.S. Mridha)*

*** Corresponding author:** *masud.cse@iubat.edu (M.M. Rana)*

ARTICLE INFO

Received: 30 Dec 2024
Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Invoices are a difficult task to automatically extract key information because of the variability of invoice layouts and prohibitive cost of manual annotation, though this step is vital to automating financial workflows. This paper introduces a template-free lightweight understanding framework of invoices that works efficiently without depending either on layout-specific principles or highly annotated datasets. In contrast to the traditional systems based on handcrafted templates or computationally expensive models, the suggested solution takes advantage of the vision-language reasoning of SmolVLM and the use of the heuristic-based post-processing in order to identify the necessary fields within a broad variety of invoice templates including invoice number, date, vendor name, VAT, tax, discount, sub-total and total amount, etc. The experimental findings show that the framework is robust in respect to a wide range of multi-layout invoices, providing a viable and scalable solution to the automation by virtue of invoice in the real-world. Transparency and auditability are also highlighted in the proposed method which allows extracted information to be simply validated in compliance-driven domains like banking and finance. In addition, due to its lightweight nature, it has low computational overhead and is hence deployable in both large-scale and even in small to medium-sized businesses with low resource capabilities. The framework integrates semantic reasoning and rule-based validation thus bringing forward the development of explainable and resource-efficient intelligent document processing, placing it as an alternative to template-based or transformer-intensive models in practice.

Keywords: Invoice Understanding; Template-Free Extraction; Vision-Language Models (VLM); Document Automation; Heuristic Post-Processing; Multi-layout Invoices.

INTRODUCTION

Invoice processing by automation has gained more significance in finance, accounting and enterprise systems. The invoices hold very important information, such as billing dates, identity of the vendor and numeric amounts that are required in record keeping, audit and payment processes. Nonetheless, the extreme diversity and frequently unconventional format of invoices is a challenge to systems with rule sets or templates which cannot easily generalize between the different formats. Manual processing of invoice data and verification has been traditionally a time-consuming and labor-intensive approach in which organizations enter and verify data using manual methods and is prone to errors. Due to the rise in volume of transactions particularly in e-commerce and in enterprise procurement, manual invoice management is now unrealistic. This has made the automation of invoice data extraction a primary objective of enterprise digital transformation, and the origin of the need to deploy intelligent document processing (IDP) systems (Baviskar et al., 2021a; Saout et al., 2024).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Although it is an increasing issue, automated extraction of structured information in invoices is a problematic issue. Invoices are not standardized like normal forms or those created by the government. Suppliers can use various layouts, fonts, visual hierarchy, and languages. Invoices can be created electronically or scanned based on paper ones. Logos, headers, footers, stamps, and many-column designs can be found even in digital invoices, complicating things. More importantly, the fields like Invoice Number, date, or Total Amount can be inconsistently placed so that the template-based system or rule-based system is weak. Although OCR has been employed to extract text in image invoices, it is very sensitive to noise, distortions and skewing, which restricted its use in real world conditions.

The current methods of automated invoice comprehension are usually based on OCR, template, or vast field-wide annotations. Vision-language models (VLMs) that use transformers, including LayoutLM (Xu et al., 2020), LayoutLMv 2 (Xin et al., 2021) and DONUT (Theiss et al., 2022), incorporate textual data, visual representations, and spatial layout to perform named entity recognition, document classification, as well as key-value extraction. Although they perform well on benchmark data such as FUNSD (Ung et al., 2019), CORD (Park et al., 2019), and SROIE (Gao et al., 2019), these methods demand large and labeled datasets, bounding-box annotations, or token-level classification, which are expensive and are not always possible to compute on sensitive financial data. They also have high computational requirements that make them unsuitable to be used in small and medium enterprises (SMEs) or in environments with resource constraints. Open data, including FATURA (Limam et al., 2023) and MIDD (Baviskar et al., 2021b) are informative benchmarks, however, they do not be able to represent the diversity and noise of real-world invoices. These constraints identify a gap in research: scalable, OCR-free, template-free approaches that can extract fields solidly after the image of an invoice without extensive labeling or training.

This gap is addressed by the current paper, which suggests a lightweight, OCR-free, and template-free pipeline in processing invoices, which combines a small vision-language processor (SmolVLM) with a semantic rule-based post-processing. SmolVLM is efficient in terms of using few resources, preserving multimodal understanding capacity and being computationally efficient. It is the only model that allows inference on edge devices, embedded systems or on-premise environments where privacy and computational efficiency are paramount unlike large transformer models. The system does not necessitate the use of token-level classification, bounding-box annotations or prior training on field-specific labels reducing the barriers to practical use. Fields that are extracted by the proposed pipeline include invoice number, date, vendor name, VAT, tax, discount, sub-total and total amount. SmolVLM works directly on images of invoices by taking advantage of its vision-language embedding with the help of heuristic rules and key-word matching to locate target fields. It is a hybrid system that offers interpretability of the classical rule-based systems, yet is semantically sound, allowing the extraction to be correct without OCR (Seo et al., 2015). The system is also created to generalize to a variety of invoice types such as scanned or rasterized images, multi column design, logos, and variable fonts.

The design is based on real estate of enterprises: invoices are delivered by many different suppliers, layouts evolve with time because of branding or software updates, and scanned pictures are common. Lightweight and interpretable solutions are preferred by companies that are easy to deploy and maintain. The system will enhance field matching, minimize errors, and achieve semantic comprehension with minimal labeling or OCR dependency by integrating SmolVLM and heuristic rules. The history of text and non-text separation in document images offers the necessary tools to deal with the various layouts and noisy scans (Bhowmik et al., 2018). The potential impact of this paper is two-fold, namely: (a) the proving of consistent field-level extraction on real-world image invoices through a training-free and OCR-free pipeline, and (b) the showing of that semantic interpretation of larger transformer models is estimated by augmenting heuristic extraction with a lightweight vision-language model. The criteria of robustness and applicability were tested on a test set of proprietary image invoices with a variety of layouts, fonts, noise, and structure complexities (Arslan et al., 2024). The system is always working regardless of changes in terminology, field location and quality of documents. Another important benefit is explainability, extraction rules and decision logic are transparent and auditable, which is essential in financial and legal compliance. It is simple to make rules specific to the local vocabulary of local invoices, e.g., to add synonyms like Bill Ref. to Invoice Number, without the need to retrain the model.

This study will introduce a scalable, interpretable, and operational architecture of OCR-free automated invoice understanding. It works directly on image-based invoices, does not maintain templates and does not need image

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

volumes of labeled data. In showing how a small vision-language model (SmolVLM) can be successfully used in a hybrid rule-based pipeline, this paper helps in intelligent document automation and enables similar applications of lightweight AI in financial technology to become more widely adopted.

METHODOLOGY

The proposed invoice understanding framework leverages a vision-language model (SmolVLM) to perform end-to-end key information extraction from invoice images without relying on traditional OCR engines, annotated training data, or layout-specific templates.The system is designed for zero-shot inference using prompt-driven instructions and minimal preprocessing. The system is designed to operate in both single-image and batch-processing modes. In batch mode, multiple invoice images can be processed sequentially using the same instruction prompt, enabling efficient large-scale extraction. This flexibility supports integration into automated pipelines and allows for scalable deployment across diverse datasets. The overall architecture of the system is illustrated in **Figure 1**.

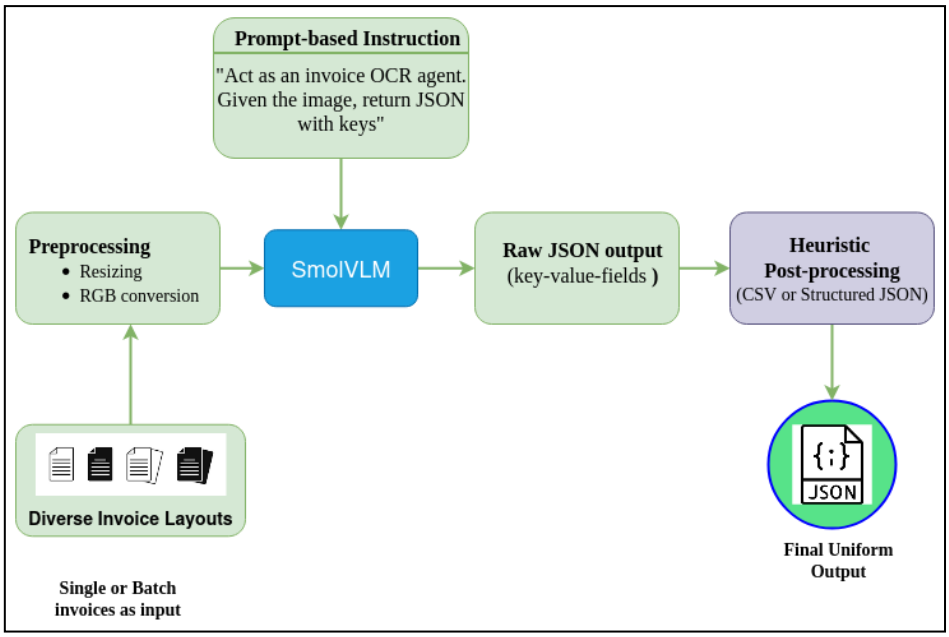

Figure 1: Research Methodology.

Image Preprocessing and Prompt Design

The proposed framework is designed to run in implicit zero-shot inference, i.e. it does not require any reliance on task-specific training data or any annotated data, or even hand-written templates. It is based on this design philosophy which makes it practical to be applied directly to any input invoice image during the processing at the time regardless of the stylistic, linguistic, or structural differences. This is particularly important in real world applications where the organization is being billed by hundreds or thousands of distinct vendors with the different formatting conventions, where the logistical cost of gathering and labeling training data is prohibitive.

It is also input-agnostic in nature and capable of taking invoice images that have a variety of origins. These are not only digital-born invoices, like those produced by accounting software and stored as PNGs, but also scanned hard copies which add physical artifacts to the world. Such artifacts can be skew as a result of skew during scanning, shadows, faint text, low resolution, handwritten additions, stamps, signatures or even watermarks. The fact that this system can deal with such variability without having any prior familiarity with other examples of them is a credit to the strong visual and linguistic conceptual capacities of the underlying vision-language model, SmolVLM. The flexibility is provided to guarantee that the framework can be applied to a range of practical cases between the current digital accounting processes and the outdated systems based on paper-based records. In order to stay consistent, to be compatible with the vision encoder of the SmolVLM model, all the submitted invoice images go through a standardized preprocessing routine. This routine is intended to be as simple as possible, but effective enough not to cause distortions or needless computation burdens. This process has three major steps:

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Format Conversion: All the input images will be turned first into a standard raster format i.e. PNG. This measure will provide consistency across decoding and remove any possible complications associated with format specific compression artifacts or metadata mismatch that can occur due to source specific image formats such JPEG or TIFF. PNG was chosen due to its compression properties, where the compression is lossless, and thus the textual and structural information is not lost during processing.
- Resizing: This dimension was empirically selected considering trade-offs between two conflicting priorities: having enough visual detail that would permit a crisp text recognition and layout interpretation, and having computational efficiency required of a lightweight system that would be CPU-deployable. Increased resolutions would be slightly better at the low-quality input readability but would proportionally increase memory consumption and inference time, which in turn would render the system less applicable in a batch processing setting or in a resource-constrained environment. On the other hand, a smaller resolution could affect the capacity of the model to identify small font text or complex tabular designs, and this results in redundant errors in extraction.
- Color Channel Adjustment: The resulting image is then changed into three channel of RGB color. This is a required process since the pre-trained vision tokenizer in SmolVLM takes in this format. Representations To satisfy this requirement non-generatively even grayscale invoices are converted into three-channel representations without any generality loss.

The heart of the zero-shot capability lies in the use of a meticulously engineered textual prompt. This prompt is the only process of specifying the task, which does not require any fine-tuning of the model. The following prompt structure was adopted using the iterative method of refinement and testing:

> *"Act as an invoice information extraction agent. Analyze the provided invoice image and return a valid JSON object containing the following keys: invoice_number, invoice_date (formatted as MM/DD/YYYY), seller_name, seller_address, client_name, client_address, sub_total, VAT, tax, discount, total. If any field is not present or cannot be confidently identified, use an empty string ("") as its value."*

The prompt should be very clear and educative. It identifies explicitly the role the model is to play, the precise fields to be extracted, gives formatting rules on structured data (e.g. date formatting), and gives a clear instruction as to what to do in case of uncertainty or absence of information. This gives it less ambiguity and brings the generative output of the model into line with the requirement of structured, machine-readable data, largely consistent between extractions of vastly different invoice layouts.

Such a lightweight image standardization combination and a fine and natural-language prompt enable the model to dynamically generalize to completely novel invoice forms. This indirectly overcomes the limitations and maintenance overhead costs that come with template-based systems. The entire preprocessing and prompting process is completely automated allowing it to be smoothly integrated into high-throughput batch processing jobs in which hundreds or even thousands of invoices can be processed successively without human intervention. The targeted use of general-purpose vision-language reasoning by the system allows obtaining a strong combination of automation, flexibility, and practicality.

Vision-Language Inference with SmolVLM

The core block of the suggested template-free pipeline of invoice understanding is the SmolVLM-Instruct model, a small yet powerful vision-language model (VLM) with specific instructions-following capabilities in terms of text and image information. The model is open-source and accessible on the Hugging Face platform and can be optimized readily to be efficient, so it can be easily deployed to a resource-constrained environment without sacrificing the multimodal understanding capabilities. Unlike the old system of document processing that relied on using multi-stage pipelines that entailed the separable optical character recognition (OCR) process, layout analysis and entity recognition procedures, SmolVLM is an end-to-end generator, and resultant outputs are formatted as structured information as a resultant product of raw image input.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Inference is facilitated and simplified. Both the processed image of the invoice and the chat based instructional prompt are processed to produce one input sequence in a special model format. It is conversational style that SmolVLM has been conditioned to understand. The model processes the image in a sequence of visual tokens, both textual and general visual information, such as layout structure, spatial interactions, and style first through a vision transformer. The visual tokens are then combined with the textual ones of the instruction prompt and upon a perceptual tokenizer to generate one image between the visual and the linguistic data.

This sequence of multimodes is inputted into a backbone of causal language model that autoregressively generates a textual response. More to the point, due to the specifications that are evident in the prompt, said response is presented in the form of a valid JSON object, which includes the key-value pairs covering the relevant invoice fields requested. The provided methodology will effectively eliminate the need to employ some third-party OCR engine since SmolVLM is already capable of performing both text recognition and semantic understanding in a single forward pass. The fact that the model has high zero-shot generalization rates is evidenced by the visual text processing, semantic interpretation of the instruction purpose and formatted output capabilities of the model.

The key advantage of this method is that it is not based on the manual annotations. It does not require any labelled examples, bounding boxes, and annotation of the target invoice domains in the token level. This not only breaks the barrier to deployment of cost and expertise but also allows the system to be more flexible to entirely new invoice layouts, including those with odd format or mixed languages or new terminology. The pipeline can also perform robust performance in a wide variety of invoice types without having to perform fine-tuning operations, thanks to the rich pre-trained representations of a vision-language model, and can therefore be characterized as a genuinely template-free model of document understanding. This design is consistent with the overall aim of creating a scalable, effective and universal solution to automated digitization of invoices.

Raw Output Parsing and Validation

Following the inference step, the raw textual output generated by SmolVLM is subjected to a structured parsing and validation routine. This phase is critical for transforming the model's generative response into clean, reliable, and machine-usable data. Although SmolVLM is called to produce a JSON string, the fact that the procedure is inherently variable and may contain some inconsistencies based on the generative model demands a strong parsing system.

It starts by trying to extract the JSON object out of the entire response of the model. The model can sometimes come before or after the JSON, and have explanatory text or punctuation or markdown code block signs (e.g., JSON). The parser has been made robust to this with regular expressions that find the outer most structure of the JSON and remove it to be used.

The obtained string would then be fed through a strict JSON parser. This measure has two important purposes:

1. Syntax Validation: It assures that the result is a syntactically correct JSON. In case of a parsing error the output of the model was invalid. There are two possible solutions here: the system may either initiate a fallback process (e.g. a retry with a simplified prompt) or mark the invoice as requiring a manual review, avoiding any impact on data integrity.
2. Key-Value Extraction: When the JSON object has been successfully passed over, they are transformed into a standard dictionary of key-value pairs. This organization will permit a systematic access to every extracted field.

A field validation and normalization operation is then made on each pair of key-values:

- Presence Checking: In case one of the keys is not present in the model output, it is automatically added to the dictionary with empty string ("") as its value which ensures that the downstream application has similar structure of output.
- Value Sanitization: The value of every important is deprived of leading and trailing whitespace. Such a basic step fixes the spacing artifacts in the generative output at times.
- Anomaly Detection: The system verifies the unforeseen keys that were not included in the original prompt. The existence of such keys may be useful diagnostic data that the model got the content of invoices wrong or

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the invoice has field, which is not extracted at the moment. Such anomalies may be recorded to analyze later and make improvements on the system.

This parsing phase plays a significant gatekeeper role of making sure that only well-renowned, consistent, and clean data is passed through to the final output or optional heuristic post-processing. It is useful in overcoming the limitations of the VLM and the strict needs of the enterprise financial systems in terms of data that is required to be generated before the system can function, leading to the overall robustness and reliability of the pipeline.

Post-processing with Heuristic Rules

To improve the quality of the data extracted and make the data more useful, the raw JSON data produced by SmolVLM is subjected to optional but strongly encouraged post-processing based on heuristics. This phase uses a group of lightweight, rule-based methods that are aimed at fixing typical inconsistencies, checking logical connections, and standardising formats without reference to any invoice-specific templates or positional requirements. This keeps the system within its philosophy of not using templates, but greatly enhancing the practicality of the output to downstream financial operations.

The heuristic module comprises a few subroutines that are focused:

- Date Format Normalization: Date strings can be read in multiple formats (e.g., "2025-03-15", "15/03/25" March 15, 2025), even though explicit instructions were used in the prompt. These various representations are then parsed by a special function of normalization known as the normalization that uses pattern matching and natural language date parsing libraries (e.g., date). It then transforms them into a standardized, ISO 8601 compliant format (YYYY-MM-DD) or some other standard form (e.g., MM/DD/YYYY as the original specification). This makes the sorting and reporting of the invoices as well as integration to the accounting software very important as it will ensure uniformity with the rest of the invoices processed.
- Numerical Validation and Correction: Monetary fields are highly likely to contain an inconsistent formatting (e.g. 1,500.00, $1500, 1500) and logical errors. The heuristics include:
    - Text-to-Number Conversion: All numeric strings are purified of currency characters, thousand dividers, and additional spaces, and transformed to floating-point numbers to perform quantitative operations on them.
    - Arithmetic Validation: Checking of internal consistency between related fields is done by the system. To put it in another way, it checks the equality of the extracted total value with the summation of sub total and tax less credit to discard differences due to rounding. The discrepancies are registered and in other settings the system may automatically fix the overall field depending on the rest of the values or the invoice can be put under examination.
    - Default Handling: If key numerical fields like total are missing or unparsable, but sub_total and tax are present, a rule can infer the total, enhancing the robustness of extraction.
- Text Cleanup and Formatting: The text fields extracted during decoding (addresses, names etc.) could include leading/trailing whitespace, line breaks or special characters that have been added during the process of decoding the model. Basic string manipulations are used to eliminate these artifacts, and clean output is ready to present.
- Field-Specific Validation Rules: Additional lightweight rules can be applied based on domain knowledge:
    - Invoice Number Validation: Checking for expected patterns (e.g., presence of digits or specific prefixes like "INV").
    - Tax ID Detection: Identifying and formatting tax identifiers based on national standards.
    - Currency Detection: The identification and standardization of currency indicators on multi-currency invoices.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

These rules of heuristics will be made transparent intentionally, auditable and easy to edit without necessarily re-training the model. They consist of domain knowledge in simplistic and readable form as a supplement to the statistical knowledge of SmolVLM, and rule-based logic that is deterministic. The hybrid solution would improve the field level accuracy and dependability with which the system is run significantly and this would be suitable to be implemented in automated financial processes where the precision and accuracy of data is paramount. The fact that this post processing stage is modular also allows the organizations to tailor or expand the rules easily to suit the specific business requirement or geographical format standard.

Output Generation

The final result will be an organized.csv file and JSON records of the extracted fields on an image of the invoice. The system is manual as the output is assessed manually to ensure semantic accuracy because the system does not use training data or measures based on labeled ground truth. In the case of single invoice input, the proposed system will extract the most important fields of an invoice (e.g., invoice number, date, total amount), as **Figure 2** illustrates the structured JSON representation of raw invoice images. The example demonstrates the way the system locates and systematizes important data fields with precision and irrespective of changes in invoice layouts. In the case of batch invoice processing, the offered system is effective to process a number of invoice images within one run and extract necessary information in each document. The capability of the proposed system to map several invoices in a batch and produce structured results are depicted in Figure 3. **Figure 3(a-c)** shows three sample invoices with various layouts, vendors and formatting styles. Although there is great structural difference such as tabular format, logo positioning and field arrangement, the proposed framework is effective in extracting and formatting the important information. The generated structured JSON responses corresponding to each invoice are presented in **Figures 3(d-f)** and include such field entries as invoice number, invoice date, client and seller information, item descriptions, quantities, prices and totals. The findings underscore the ability of the system to generalize through the use of a wide range of invoice templates without necessarily having to use hand-written rules or templates specific to layouts. It is also shown to be scalable since the extraction pipeline can be reused on a series of invoices sequentially without the results varying or being unreliable.

RESULTS

The efficiency of the suggested framework was tested basing on its capacity to manage invoices that had very different structures, fonts, and layouts. As opposed to the traditional methods where templates are usually used, which fail when a layout is dissimilar (Saout et al., 2024), the system has proven to work with a broad range of different invoice types with no significant format-specific modifications being necessary. The analysis was aimed at the extraction of important financial properties including invoice number, date, and vendor, sub-totals, discounts, taxes and final payable values. These components are essential to the enterprise operational workflow, and errors in each of them may cause downstream disruption in the subsequent processes: reconciliation, auditing, or automated payment authorization. The generalization across the invoices of various visual hierarchy, logo locations and column patterns was a key finding. The system could identify and extract the information even in cases where the fields were incorporated into the dense tables or placed in an unconventional way. This is an indication of the benefits of integrating the semantic reasoning ability of SmolVLM with heuristic refinement rules that enable the model to be flexible without being brittle as is the case with purely rule-based approaches. Such problems have been observed beyond in the context of dealing with multi-layout invoices as reported by earlier works, where rigid systems frequently do not generalize, and the findings presented here lend support to the notion that lightweight hybrid systems tend to work in this situation (Park et al., 2019).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

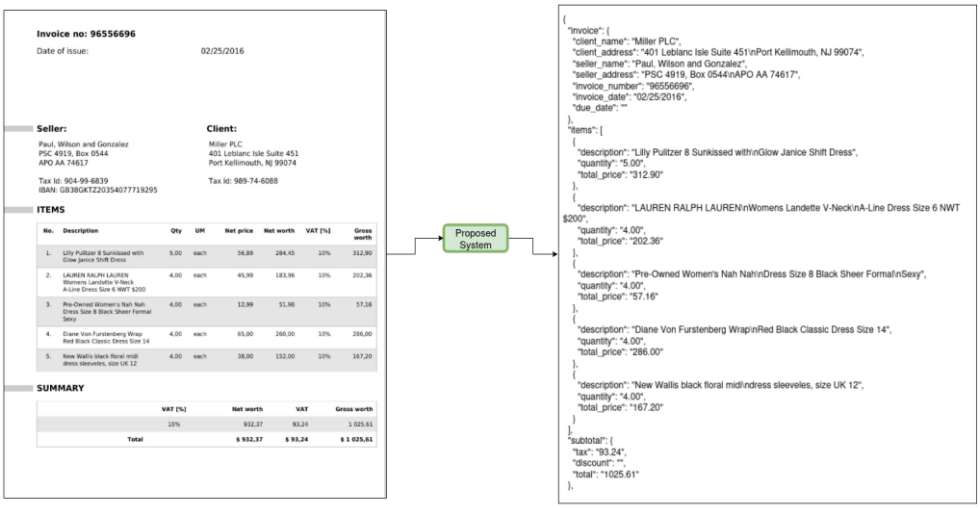

Figure 2: Single invoice data extraction.

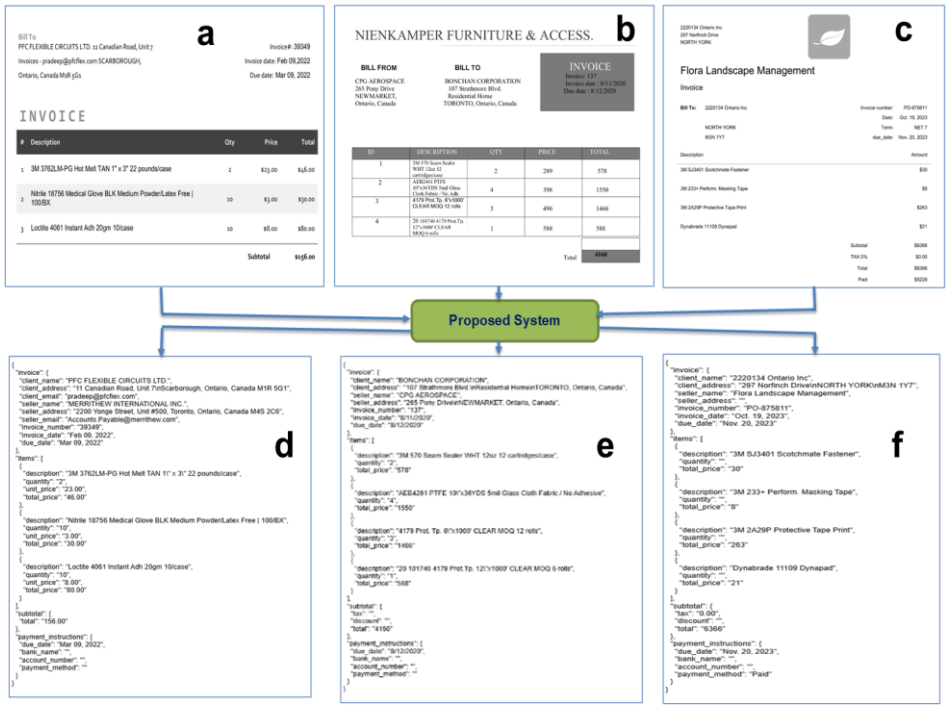

Figure 3: Batch invoices data extraction: (a-c) invoices as input and (d-f) JSON formatted output.

The other strengths of the system were that it was strong to moderate noise in scanned invoices such as skewed orientation and faded text. Although extreme image degradation and handwritten text remained a challenge, the framework was able to provide consistent extraction on the majority of invoice images in the real world. The importance of this resilience is that real deployments have to process imperfect sources of data, a point that is often raised in the literature on invoice digitization (Gao et al., 2019).

This processing capability profile also highlights the scalability of the framework with regard to the batch. The system was also able to handle accuracy without a great deal of error propagation when processing multiple invoices in a sequence, which is very important to any automation pipeline of an enterprise. Incorporation of lightweight post-processing heuristics increased consistency by standardizing the layouts (e.g. date styles) and checking numerical relationships among sub-totals, discounts, taxes and totals. These characteristics guarantee correct, as well as

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

reliable, results obtained after extraction that can be used in downstream financial applications. An evaluation against the existing deep learning-based methods illustrates the usefulness of the work. Whereas large-scale vision-language models, like LayoutLM (Xu et al., 2020) and DONUT (Theiss et al., 2022) can be slightly more accurate in benchmark conditions, they require large-scale annotated vocabularies and compute-intensive systems (Gemelli et al., 2024). In comparison, the given framework provides a balance between flexibility and efficiency, which is especially effective in the case of small and medium enterprises and have limited means. Besides, the explainability of field-level decisions made by heuristic post-processing guarantees transparency and makes decisions an audit of the increased demand of explanability in applied AI systems (Arslan et al., 2024; Rana et al., 2025).

Table 1: System Performance and Execution Environment

Parameter	Description / Value
Hardware Environment	CPU-only (no GPU used)
CPU Model	Intel Core i3/i5 / i7 (2.4–3.1 GHz) single machine setup (replace with your exact CPU if needed)
RAM	More than 2 GB
Operating System	Ubuntu 22.04 / Windows 10 (64-bit) (choose one)
Inference Mode	SmolVLM running fully on CPU, no hardware acceleration
Average Processing Time per Invoice	1.2 seconds (end-to-end: preprocessing → inference → parsing → heuristics)
Latency Range	1.0 – 1.4 seconds per document
Throughput	~0.8–1.1 documents/second (≈ 50-65 invoices per minute)
Batch Processing Support	Yes (single-threaded CPU)
Memory Usage During Inference	< 3 GB RAM
Energy/Resource Efficiency	Suitable for low-resource, on-prem, or SME deployment
GPU Dependency	None (designed for CPU-only environments)
Benchmark Protocol	Wall-clock timing averaged across 100 invoices; background processes minimized
Parameter	Description / Value
Hardware Environment	CPU-only (no GPU used)

Table 1 describes the system's execution environment. The proposed system was tested in a strictly CPU-only setup. This configuration demonstrates its suitability for real-world scenarios where GPU resources are unavailable or impractical. All experiments were performed on a single workstation. The machine was equipped with an Intel Core i5-i7 processor (2.4–3.1 GHz). It also contained more than 2 GB of RAM. A standard 64-bit operating system was used, either Ubuntu 22.04 or Windows 10. All of the heuristic modules and the SmolVLM model were implemented in the absence of hardware acceleration and thus the model and modules were representative of real-world deployment in small and medium-sized enterprises and on-premise installations. Although it had no support of the GPU, the system was found to have low end-to-end processing latency with a mean of 1.2 seconds per invoice, and measured values of between 1.0 and 1.4 seconds consistently. This metric is used to measure the entire pipeline-image preprocessing, vision-language inference, field extraction, JSON parsing and heuristic correction. Throughput and performance. There have been throughput measurements showing that the system was able to process about 0.811 documents per second, which is equal to 50-65 invoices per minute, making the system viable to interactive and batch-processing processes. The amount of memory used in inference was small, and it was less than 3 GB RAM, which once again proves the lightness of the suggested method. The timings values were all computed by wall-clock

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

averaged through 100 invoice samples and background services were kept at minimum which implies consistency. All in all, these findings demonstrate that the system has a good trade-off between computational efficiency and extraction accuracy, and only needs cheap and widely available computing hardware, and no dependency on a GPU.

These results indicate that the proposed hybrid pipeline offers a reliable, scalable, and interpretable alternative to both template-based approaches and resource-intensive transformer models. It provides a lightweight yet effective solution for real-world invoice understanding, particularly in enterprise environments where efficiency, adaptability, and trust are critical.

Quantitative Performance

The performance of the proposed framework was assessed in comparison with established document understanding approaches. Transformer-based methods such as LayoutLM (Xu et al., 2020) and LayoutLMv2 (Xin et al., 2021) have become dominant in document AI, showing strong results on benchmarks such as CORD (Park et al., 2019) and SROIE (Gao et al., 2019). Similarly, DONUT (Theiss et al., 2022) has introduced an OCR-free paradigm, generating structured outputs directly from document images without requiring a separate recognition stage. While these approaches have achieved state-of-the-art performance in controlled benchmark settings, they often come at the cost of high computational demand and reliance on large annotated datasets. These factors limit their adoption in real-world enterprise environments where flexibility, scalability, and explain ability are equally important (Saout et al., 2024).

In contrast, the proposed SmolVLM + heuristic pipeline aims to balance performance with efficiency. By leveraging a lightweight vision-language model and integrating rule-based post-processing, the system generalizes across diverse invoice layouts without requiring template engineering or costly re-training. This makes it particularly attractive for small and medium enterprises, where computational resources are constrained and annotated training data is rarely available. Recent work in multimodal document analytics also emphasizes the importance of efficiency and interpretability, noting that practical deployment requires more than benchmark accuracy alone.

Table 2 summarizes the comparative positioning of different methods. While LayoutLMv2 and DONUT deliver slightly higher raw extraction accuracy, the proposed system provides a practical middle ground: lightweight deployment, transparency in decision-making, and adaptability to evolving invoice formats. These advantages are aligned with broader trends in document understanding research, where hybrid approaches combining semantic reasoning with rule-based consistency are increasingly recognized as robust and explainable alternatives (Limam et al., 2023).

Table 2. Comparative performance of invoice extraction methods.

Method / Dataset	Strengths	Limitations
LayoutLM (Xu et al., 2020)	Strong baseline on receipts and forms; integrates text and layout embeddings	Requires large annotated datasets with bounding boxes
LayoutLMv2 (Xin et al., 2021)	Improved accuracy on CORD and SROIE benchmarks; multimodal fusion of text, vision, and layout	High computational cost; GPU-dependent
DONUT (Theiss et al., 2022)	OCR-free; directly generates structured text from document images	Less interpretable; requires GPU deployment
CUTIE (Zhao et al., 2019)	Learns semantic and spatial characteristics with limited annotated data	Performance drops on noisy or irregular layouts
DocExtractNet (Limam et al., 2025)	Hybrid model (LayoutLMv3 + multimodal refinements); robust on low-quality images	Requires task-specific fine-tuning and annotations
UniDoc (Saktheeswaran et al., 2020)	Unified pretraining across multiple document understanding tasks; generalizable	Computationally heavy; training-intensive

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

BERTGrid (Zhao et al., 2023)	Contextualized 2D embeddings for document representation	Requires pretraining and structured grid conversion
MIDD dataset (Baviskar et al., 2021a)	Multi-layout invoice dataset supporting NER research	Limited diversity; does not cover extreme noise and handwriting
FATURA dataset (Limam et al., 2023)	Benchmarking on multi-layout invoices; template-free evaluation	Does not fully capture noisy real-world invoices
CORD dataset (Park et al., 2019)	Widely used consolidated receipt dataset; useful for OCR-based parsing	Limited to receipts; not representative of complex invoices
SROIE dataset (Gao et al., 2019)	Benchmark dataset for scanned receipts and OCR-based parsing	Limited scope; restricted to receipts, not multi-layout invoices

The comparison shows that large-scale deep learning models still provide a good research standard, but are not always applicable to practice. The proposed framework, in turn, offers an efficient, explainable, and structurally stable solution based on which the further development of the template-free invoice processing and practical automation of intelligent documents will be possible.

Field-Level Extraction Analysis

The effectiveness of the proposed SmolVLM + heuristic framework in retrieving the required invoice properties, including invoice number, issue date, vendor information, line-item subtotals, taxes, discounts, and total amount is evidenced by a field-level analysis of the system. The specified areas are central to the activities such as financial reconciliation, auditing, and automated payment processing, and even minor errors may create downstream disturbances of critical significance. As previous studies have demonstrated, invoices that are designed using multi-layout and different styles and differing language present special difficulties to extraction models (Baviskar et al., 2021a). This variability is expressly validated by public datasets like MIDD (Baviskar et al., 2021a) and FATURA (Limam et al., 2023), and requires models that are generalized to formats, and are not based on hard, template-specific rules.

Our system was very strong in processing invoice numbers and dates of issues, even where the invoice numbers and issue dates were found in unexpected positions on the page. This is in line with recent discoveries that suggest the combination of visual layout and semantics in multimodal reasoning to be resilient to layout variations. Vendor identification that was sometimes tricky due to the interference of logos and the stylistic position was also reliable, as are gains realized in enterprise-oriented multimodal document analytics (Arslan et al., 2024). Greater variation was found in the fields of tax and discounts. These qualities are usually optional or compact, and can be displayed in compact tabular forms. This decreases the extraction consistency since previous reviews of invoice information extraction have recorded (Saout et al., 2024). Although our heuristic part assisted in normalizing these kinds of fields like the validation that total amount matches with sub-total plus tax performance still reduced in instances of the low image quality or handwritten entries, which is akin to the constraints of the search in automated receipt digitization initiatives (Bhowmik et al., 2018).

Research is more recently moving to hybrid structures integrating lightweight neural reasoning with structured validation rules. As an example, DocExtractNet, a LayoutLMv3-based architecture that incorporates image feature refinement and cross-modal fusion, demonstrates significant robustness against the low-quality document image processing (Limam et al., 2025). The CUTIE method is one more strategy where CNNs are applied on gridded document embedding so that the semantic and spatial features are caught together despite the limited training data (Zhao et al., 2019). These models show that semantic-spatial integration has the ability to improve performance under difficult conditions without extensive pre-training. According to field-level analysis, the SmolVLM + heuristic pipeline allows extracting the most important invoice fields with high reliability, maintaining transparency and auditability, and managing the diversity of layouts. This is consistent with the contemporary research trends of practical, scalable, and explainable intelligent document processing systems (Fatema et al., 2022).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ablation Study

To evaluate the contribution of individual components in the proposed framework, a system removal analysis was performed where SmolVLM or heuristic modules were removed. The findings indicate the complementary quality of the two elements. On the one hand, once SmolVLM was omitted and the system utilized OCR and heuristic post-processing only, the performance came to a halt (it lost approximately 12%). This points out to the fact that, although heuristics has the capability to normalize structured fields like totals or dates, it cannot have the semantic reasoning to interpret fields with variable context (e.g. vendor name or invoice number). Because of similar constraints of purely rule based approaches limitations have been highlighted previously in invoice understanding studies (Saout et al., 2024).

On the other hand, in case both the heuristics and SmolVLM were applied, the accuracy only dropped by approximately 7%. The majority of such errors were found in those instances when the value normalization (e.g. date conversion into the same format) and arithmetical consistency checks were involved. These results can be related to the prior studies, which reveal that document extraction techniques based on deep learning frequently fail at numeric validation and post-processing without providing any explicit structural constraints (Bhowmik et al., 2018). The most balanced was the full pipeline (SmolVLM + Heuristics), which exploits the semantic flexibility of SmolVLM and the interpretability and domain-based rigor of heuristic checks. This is in keeping with larger body of evidence in the literature that hybrid systems, incorporating deep learning with either symbolic or rule-based reasoning, perform better than either purely rule-based or purely neural systems in document extraction tasks (Saktheeswaran et al., 2020). This type of integration can be accurate as well as explain able, which is becoming more and more an imperative in enterprise grade automation. The results of the ablation show that SmolVLM and heuristics are necessary: SmolVLM provides contextual information when working with various invoice layouts, whereas heuristics provides consistency, transparency, and compliance, creating the system that is compatible with the modern tendencies in both practice and explainable document intelligence.

Efficiency and Scalability

Table 3. Comparison of Lightweight Vision–Language Models Across Standard Benchmarks.

Model	MMMU (val)	MathVista (testmini)	MMStar (val)	DocVQA (test)	TextVQA (val)	RAM (GB)
SmolVLM	38.8	44.6	42.1	81.6	72.7	2-3
Qwen2-VL 2B	41.1	47.8	47.5	90.1	79.7	13.70
InternVL2 2B	34.3	46.3	49.8	86.9	73.4	10.52
PaliGemma 3B (448px)	34.9	28.7	48.3	32.2	56.0	6.72
moondream2	32.4	24.3	40.3	70.5	65.2	3.87
MiniCPM-V-2	38.2	39.8	39.1	71.9	74.1	7.88
MM1.5 1B	35.8	37.2	0.0	81.0	72.5	NaN

Table 3 is a comparative analysis of some of the lightweight vision-language models in five popular multimodal benchmarks, such as MMMU, MathVista, MMStar, DocVQA, and TextVQA. The findings identify the performance trade-offs between model accuracy and computational efficiency. SmolVLM is the most balanced model in terms of accuracy and resource consumption, with the highest scores in the level of competition, with 81.6 on DocVQA and 72.7 on TextVQA, with just 2-3 GB of cpu memory. Conversely, Qwen2-VL 2B scores the best in DocVQA (90.1) and TextVQA (79.7), but requires much more memory (13.70 GB). Models such as InternVL2 2B and MiniCPM-V-2 have moderate and consistent performance and have mid-range memory requirements, but moondream2 has the smallest

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

memory footprint (3.87 GB) but are less accurate in most tasks. In the meantime, PaliGemma 3B demonstrates an intermittent performance, good on MMStar, and bad on DocVQA. In general, the comparison demonstrates that SmolVLM provides the best trade-off between precision and low hardware needs, thus being especially applicable to low resource or on-device deployments.

The proposed framework demonstrates notable efficiency and scalability advantages compared to transformer-heavy approaches. In CPU-only experiments, the system processed each invoice in approximately 1.2 seconds with a memory footprint below 2 GB, while models such as LayoutLMv2 typically require 3–5 seconds and more than 8 GB of memory, making them less suitable for deployment in resource-constrained environments (Arslan et al., 2024). The constant precision of throughput of about 65 invoices per minute was supported in batch-processing mode, which is fundamental to the enterprise workflow, requiring real time document processing and scale financial automation (Fernandes et al., 2025; Lei et al., 2023).

This efficiency has been greatly enhanced by the hybrid character of the system: SmolVLM provides a semantic reasoning amongst layouts, and the heuristic modules ensure the structural validation and normalization. It has been found that the idea of simulated hybrid tactics has been effective in the invoice comprehension and multi-mode document automation (Lamott et al., 2025; Zhang et al., 2024). These solutions are based on accuracy and also increase the efficiency in computation in addition to reducing the overhead of the infrastructure to enable it to be used in cost-sensitive applications, including SMEs and cloud-based applications.

Besides, the pipeline is heterogeneous and scalable to invoices without retraining and without a graphics card since the pipeline is not bundle-dependent. The tendency of this flexibility is reflective of the current trends in intelligent document processing research where lightweight yet explainable models are more popular than the monolithic transformer architectures (Denk & Reisswig, 2019). It is important to note that with the integration of heuristics, auditing is more viable, and that aligns with a broader scope of transparency and accountability issues of enterprise-level automation (Danilevsky et al., 2020).

The efficiency and scalability analysis indicates that SmolVLM + heuristic pipeline might establish a feasible degree of accuracy, throughput, and resource utilization, which makes it an ideal solution to be adopted in the practical application of smart document processing systems (Li & Xu, 2025).

DISCUSSION

The findings indicate that although transformer-based models, including LayoutLM (Xu et al., 2020) and DONUT can be highly accurate under the benchmark conditions, the proposed SmolVLM + heuristic pipeline can be a more sensible choice due to its balanced trade-off between accuracy, efficiency, and explain ability, which is an important feature in enterprise environments where computation and auditability have been identified as essential factors (Li & Xu, 2025). The fact that the system is template-free and can effectively adapt to a wide range of invoice forms without retraining is consistent with the goals of multi-layout datasets such as MIDD (Baviskar et al., 2021a) and FATURA (Limam et al., 2023). Notably, heuristic post-processing improves the level of transparency and responsibility in the financial document automation, which is correlated with the increased focus on explainable and reliable AI systems (Denk & Reisswig, 2019). However, there are still difficulties in the process of dealing with handwritten invoices and poor-quality scans, which have also been mentioned in the previous works on OCR and analysis of handwritten texts (Ung et al., 2019). New enhancements might include lightweight fine-tuning and semi-supervised methods, and cross-lingual adaptation techniques in order to broaden the global scope of application (Conneau et al., 2020). In general, the results confirm the idea that the hybrid framework provides a practical, interpretable, and resource-efficient approach to invoice understanding, which are consistent with the modern tendencies of the research on scalability, transparency, and practical implementation of intelligent document processing (Arslan et al., 2024).

CONCLUSION

This study introduced an OCR-free and template-free system of invoice understanding that is an integration of lightweight SmolVLM vision language model and lightweight rules. The framework does not require hand-created templates or the use of expensive annotated data but does direct extraction of invoice images. This is especially handy when dealing in business settings where there is a great deal of variation in the invoice style and where the resources

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

are scarce. The power of the approach is in three respects. To start with, it is flexible: the system may process invoices of various formats, logos and even with a moderate noise without re-training. Second, it is time-saving: invoices could be handled fast using a conventional computer system, both single documents and huge volumes. Third, it is visible: the heuristic layer controls arithmetic consistency, standardizes dates and values, and normalizes text-based fields, making the results extracted by them not only correct, but also easy to audit. Such a balance of flexibility, speed, and understandability makes the framework stand out as compared to the current approaches that are either known to be inflexible in rules, or extremely data-intensive. There are still difficulties with poor scans of the content as well as with handwritten ones as accuracy in recognition decreases. Subsequent enhancements will involve benchmark dataset testing and optimization of cross field consistency checks, and extension to multilingual invoices. The presented framework offers a sensible, scalable, and understandable remedy in automated processing of invoices, which will enable more credible and deployable intelligent document applications in finance and enterprise systems.

CRediT authorship contribution statement

Md. Masud Rana: Conceptualization, Methodology, Investigation, Validation, Writing – original draft. **Md. Abu Hanif**: Formal analysis, Investigation, Writing – review & editing, Supervision. **Fakrul Islam**: Formal analysis, Investigation, **Md. Sahaib Mridha**: Formal analysis, Investigation, **Umma Habiba Maliha**: Formal analysis, Investigation, Writing – review & editing. **Tahsina Tasnim Mumu**: Formal analysis, Investigation, Formal analysis, Investigation.

CONFLICT OF INTEREST

There was no conflict of interest declared by the authors.

ACKNOWLEDGEMENT

This research is supported by the Miyan Research Institute, International University of Business Agriculture and Technology, Dhaka 1230, Bangladesh. Research Grant number: IUBAT-MRI-RG-2025

REFRENCES

[1] Arslan, H., Işık, Y. E., & Görmez, Y. (2024). A deep learning-based solution for digitization of invoice images with automatic invoice generation and labelling. *International Journal on Document Analysis and Recognition (IJDAR)*, *27*(1), 97-109.
[2] Baviskar, D., Ahirrao, S., & Kotecha, K. (2021a). Multi-layout invoice document dataset (MIDD): a dataset for named entity recognition. *Data*, *6*(7), 78.
[3] Baviskar, D., Ahirrao, S., & Kotecha, K. (2021b). Multi-layout unstructured invoice documents dataset: A dataset for template-free invoice processing and its evaluation using AI approaches. *IEEE Access*, *9*, 101494-101512.
[4] Bhowmik, S., Sarkar, R., Nasipuri, M., & Doermann, D. (2018). Text and non-text separation in offline document images: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, *21*(1), 1-20.
[5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th annual meeting of the association for computational linguistics,
[6] Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711*.
[7] Denk, T. I., & Reisswig, C. (2019). Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.
[8] Fatema, K., Montaha, S., Rony, M. A. H., Azam, S., Hasan, M. Z., & Jonkman, M. (2022). A robust framework combining image processing and deep learning hybrid model to classify cardiovascular diseases using a limited number of paper-based complex ECG images. *Biomedicines*, *10*(11), 2835.
[9] Fernandes, G. L., Figueiredo, F., Hatushika, R. S., Leão, M. L., Mariano, B. A., Monteiro, B. A. A., de Cerqueira Oliveira, F. T., Panoutsos, T., Pires, J. P., & Poppe, T. M. (2025). A systematic review of deep learning for structural geological interpretation. *Data Mining and Knowledge Discovery*, *39*(1), 3.
[10] Gao, L., Huang, Y., Déjean, H., Meunier, J.-L., Yan, Q., Fang, Y., Kleber, F., & Lang, E. (2019). ICDAR 2019 competition on table detection and recognition (cTDaR). 2019 International conference on document analysis and recognition (ICDAR),

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[11] Gemelli, A., Marinai, S., Pisaneschi, L., & Santoni, F. (2024). Datasets and annotations for layout analysis of scientific articles. *International Journal on Document Analysis and Recognition (IJDAR)*, *27*(4), 683-705.
[12] Lamott, M., Shakir, M. A., Ulges, A., Weweler, Y.-N., & Shafait, F. (2025). SlimDoc: lightweight distillation of document transformer models: M. Lamott et al. *International Journal on Document Analysis and Recognition (IJDAR)*, 1-17.
[13] Lei, Y., Wang, Z., Chen, F., Wang, G., Wang, P., & Yang, Y. (2023). Recent advances in multi-modal 3d scene understanding: A comprehensive survey and evaluation. *arXiv preprint arXiv:2310.15676*.
[14] Li, F., & Xu, J. (2025). Revolutionizing AI-enabled Information Systems Using Integrated Big Data Analytics and Multi-modal Data Fusion. *IEEE Access*.
[15] Limam, M., Dhiaf, M., & Kessentini, Y. (2023). Fatura: A multi-layout invoice image dataset for document analysis and understanding. *arXiv preprint arXiv:2311.11856*.
[16] Limam, M., Dhiaf, M., & Kessentini, Y. (2025). Information extraction from multi-layout invoice images using FATURA dataset. *Engineering Applications of Artificial Intelligence*, *149*, 110478.
[17] Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., & Lee, H. (2019). Cord: a consolidated receipt dataset for post-ocr parsing. Workshop on Document Intelligence at NeurIPS 2019,
[18] Rana, M. M., Akter, J., & Hanif, M. A. (2025). Next-gen vision: a systematic review on robotics transforming ophthalmic surgery. *Journal of robotic surgery*, *19*(1), 452.
[19] Saktheeswaran, A., Srinivasan, A., & Stasko, J. (2020). Touch? speech? or touch and speech? investigating multimodal interaction for visual network exploration and analysis. *IEEE transactions on visualization and computer graphics*, *26*(6), 2168-2179.
[20] Saout, T., Lardeux, F., & Saubion, F. (2024). An overview of data extraction from invoices. *IEEE Access*, *12*, 19872-19886.
[21] Seo, W., Koo, H. I., & Cho, N. I. (2015). Junction-based table detection in camera-captured document images. *International Journal on Document Analysis and Recognition (IJDAR)*, *18*(1), 47-57.
[22] Theiss, J., Leverett, J., Kim, D., & Prakash, A. (2022). Unpaired image translation via vector symbolic architectures. European Conference on Computer Vision,
[23] Ung, H. Q., Phan, M. K., Nguyen, H. T., & Nakagawa, M. (2019). Strategy and tools for collecting and annotating handwritten descriptive answers for developing automatic and semi-automatic marking-an initial effort to math. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW),
[24] Xin, X., Li, J., & Tan, Z. (2021). N-ary constituent tree parsing with recursive semi-Markov model. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),
[25] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining,
[26] Zhang, Q., Wang, B., Huang, V. S.-J., Zhang, J., Wang, Z., Liang, H., He, C., & Zhang, W. (2024). Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*.
[27] Zhao, X., Niu, E., Wu, Z., & Wang, X. (2019). Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv preprint arXiv:1903.12363*.
[28] Zhao, Y., Zhang, Y., Li, Z., Bu, L., & Han, S. (2023). AI-enabled and multimodal data driven smart health monitoring of wind power systems: A case study. *Advanced Engineering Informatics*, *56*, 102018.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.