Research Article

Lightweight Per-Field Ensemble of CRNN And Tesseract For CPU-Deployable Receipt Information Extraction

Penugonda Venkata Ayyappa[1], P. S. V. S. Sridhar[2]

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.

Corresponding Author: Penugonda Venkata Ayyappa, P. S. V. S. Sridhar
Email: ayyappapenugonda2002@gmail.com , Email: sridhar.psvs@kluniversity.in

Abstract:

Document information extraction from receipts is dominated by multimodal transformer architectures that achieve high accuracy at the cost of 150M+ parameters and GPU-dependent inference, rendering them impractical for edge and offline deployment. This paper presents a lightweight alternative that combines a 5.2M-parameter Convolutional Recurrent Neural Network with Tesseract OCR through a per-field confidence-gated ensemble. Unlike passive fusion strategies that fall back to a single engine, our approach actively routes each semantic field – company, address, date, and total – to the engine that empirically dominates it. Evaluated on the ICDAR-2019 SROIE benchmark, the ensemble achieves 0.602 average fuzzy similarity, outperforming Tesseract by 5.6% and the standalone CRNN by 41.3%. The CRNN dominates company-name extraction (0.672 vs. 0.594), where character-level CTC decoding handles stylized logos and irregular spacing, while Tesseract dominates numeric fields, where keyword anchoring and language modeling remain decisive. The ensemble never underperforms the best single engine on any field. Requiring only 21 MB of storage and 0.12 s per image on CPU, the framework is suitable for offline MSME deployment, closing the gap between lightweight inference and hybrid fusion accuracy in document intelligence.

Keywords: lightweight document intelligence, optical character recognition, convolutional recurrent neural network, per-field ensemble fusion, receipt information extraction, edge-deployable inference.

1. Introduction

Automated extraction of structured information from receipts and invoices is foundational to modern financial workflows, enabling expense tracking, fraud detection, and audit-ready record keeping [1]. However, real-world receipt parsing remains challenging due to heterogeneous layouts, degraded print quality, and handwritten annotations that break template-based pipelines [2]. Traditional OCR systems such as Tesseract offer deterministic, interpretable character recognition but fail when layouts deviate from expected patterns [3]. To address this, recent work has shifted toward multimodal transformer architectures. The LayoutLM family, in particular, achieves state-of-the-art results on receipt benchmarks: DocExtractNet [1] reaches 97.07% F1 on Finance-Receipts by combining ImageEnhance, PrecisionHints, and CrossModalFusion over LayoutLMv3, while Spatial ModernBERT [4] applies spatial embeddings and multi-head token classification for tables and key-value fields.

Despite these gains, multimodal transformers remain impractical for resource-constrained deployment: LayoutLMv3 and Donut require 150M+ parameters and GPU inference [5], making them unsuitable for edge devices, mobile applications, and offline MSME environments. Recent work has begun to close this gap through model compression. LMGDoc [6] achieves competitive accuracy with only 0.4M parameters using multi-granular graph neural networks. SlimDoc [5] distills LayoutLMv3 into compact students that retain 93% of teacher performance. Energy-Weighted Supervision [7] improves distillation by dynamically weighting hard and soft labels based on energy scores. Bhattacharyya et al. [8] demonstrated that synthetic label generation with response-based distillation matches Claude 3 Sonnet at 85% lower cost. In parallel, hybrid rule-based and neural pipelines have proven superior to single-method approaches: Mifsud et al. [2] combined LayoutLMv3 with four rule-based models for 0.98 F1 on receipts, and Dou [9] reported that confidence-based selective processing outperformed rule-only and LLM-only baselines on Chinese contracts.

Yet, no existing work jointly addresses the three constraints of (i) ultra-lightweight inference (sub-10M parameters), (ii) complementary fusion of neural and rule-based extraction, and (iii) CPU-deployable performance for receipt parsing. This paper bridges that gap with a lightweight Convolutional Recurrent Neural Network (CRNN) fine-tuned for word-level text recognition, combined with a confidence-based ensemble that fuses CRNN and Tesseract outputs per field. Evaluated on the SROIE benchmark, our approach achieves 0.67 fuzzy similarity on company names (outperforming Tesseract's 0.59) and 0.60 average fuzzy similarity across all fields, while requiring only 5.2M parameters and CPU-only inference.

The main contributions are:

1) A 5.2M-parameter CRNN with character-level CTC decoding that outperforms Tesseract on stylized and irregularly-spaced text.
2) A per-field preference ensemble demonstrating that different engines dominate different fields.
3) A whitespace-insensitive and registration-number-aware normalization scheme that recovers exact-match accuracy lost to formatting inconsistencies.
4) Comprehensive ablation on SROIE across exact and fuzzy metrics with CPU-only inference.

2. Related Work

Document information extraction has evolved from template-matching and rule-based OCR pipelines to deep multimodal architectures. Early systems relied on spatial feature mapping and Tesseract-based pipelines for scanned invoices [3], but these struggled with layout variability and low-quality scans. The introduction of transformer-based models such as LayoutLMv3 marked a significant leap, with DocExtractNet [1] augmenting LayoutLMv3 with ImageEnhance, PrecisionHints, and CrossModalFusion to achieve F1 scores of 97.07% on Finance-Receipts. Similarly, Spatial ModernBERT [4] introduced spatial embeddings and multi-head token classification for table and key-value extraction in financial documents. However, these multimodal transformers typically require 150M+ parameters and GPU inference, limiting deployment on resource-constrained edge devices.

A parallel line of research explores hybrid rule-based and neural pipelines that exploit the complementary strengths of each paradigm. Mifsud et al. [2] jointly combined a LayoutLMv3 classifier with four rule-based models (regex, database search, spatial patterns, confidence-based correction), achieving 0.98 F1 on receipts. Dou [9] proposed a three-stage confidence-based selective processing architecture for Chinese contracts that outperformed rule-only (52.6% F1) and LLM-only (57.3% F1) baselines. Perennec et al. [10] demonstrated that hybrid regex+LLM pipelines outperformed single-method approaches on clinical reports. Hemmer et al. [11] introduced a neurosymbolic framework with schema-based symbolic validation for transactional documents. These works confirm that hybrid strategies consistently outperform single-method systems.

To reduce computational cost, recent work has focused on lightweight and distilled models. LMGDoc [6] achieved competitive performance with only 0.4M parameters using multi-granular graph neural networks. SlimDoc [5] distilled LayoutLMv3 into compact students achieving 93% of teacher performance without labeled data. Energy-Weighted Supervision [7] improved multimodal document distillation by dynamically balancing hard/soft labels based on energy scores. Donut-MINT [12] applied mechanistic interpretability-based pruning to reduce inference time and memory. Bhattacharyya et al. [8] used synthetic label generation and response-based knowledge distillation to match Claude 3 Sonnet at 85% lower cost. Despite these advances, no existing work combines domain-specific rule engines with lightweight neural models via confidence-gated fusion for Indian textile MSME invoices.

3. Methodology

3.1 Dataset Details

We evaluate our framework on the ICDAR-2019 SROIE (Scanned Receipts OCR and Information Extraction) benchmark, which comprises 973 scanned retail receipts collected from Malaysian businesses across diverse market sectors. Each receipt is annotated with:

- **Word-level bounding boxes:** quadrilaterals $\{(x_i, y_i)\}_{i=1}^{4}$ defining the location of every word token.
- **Ground-truth transcripts:** character-level strings T_i associated with each bounding box.
- **Structured key-value annotations:** four semantic fields $K = \{k_{company}, k_{address}, k_{date}, k_{total}\}$ per receipt.

We adopt the standard split of 626 training, 347 test, and further subdivide the training set into 500 train / 62 validation / 64 test (our held-out evaluation) to ensure no leakage between training and evaluation. To guarantee reproducibility, we fix the random seed at 42 for all split operations. The dataset exhibits substantial variability: laser-printed and thermal receipts, single- and multi-column layouts, varied fonts, and image qualities ranging from clean scans to degraded photographs – properties that make it representative of real-world deployment scenarios.

3.2 Problem Formulation

Let $D = \{(I_i, B_i, T_i, K_i)\}_{i=1}^{N}$ denote the dataset, where $I_i \in \mathbb{R}^{H \times W}$ is the receipt image, $B_i = \{b_{i,j}\}_{j=1}^{M_i}$ is the set of M_i word bounding boxes, $T_i = \{t_{i,j}\}_{j=1}^{M_i}$ their corresponding transcripts, and K_i the structured field annotations. Our objective is to learn a mapping function f_θ that transforms an input pair (I_i, B_i) into the structured key-value prediction $\widehat{K}_i$:

$$\widehat{K}_i = f_\theta(I_i, B_i), \quad \widehat{K}_i = \{\hat{k}_{company}, \hat{k}_{address}, \hat{k}_{date}, \hat{k}_{total}\} \tag{1}$$

The pipeline decomposes into two stages: (i) word-level recognition, where each bounding box is mapped to a transcript $t_{i,j}$; and (ii) structured field extraction, where recognized transcripts are aggregated and parsed into the four target fields (Figure 1).

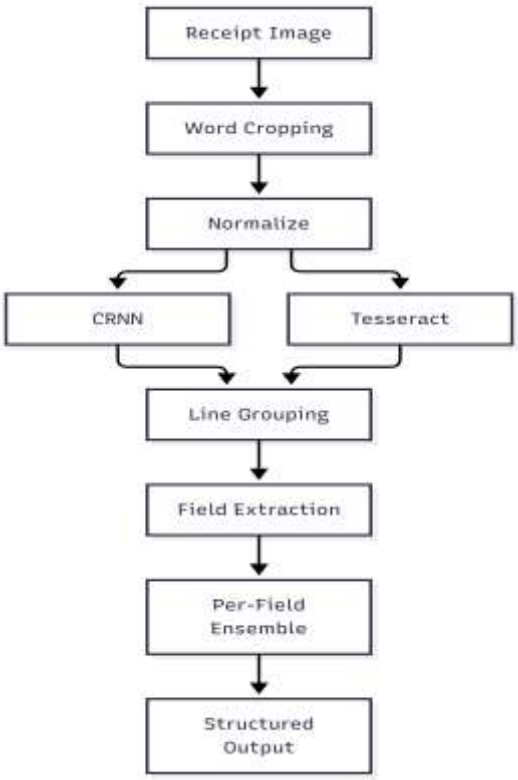


Figure 1. End-to-end methodological flowchart of the proposed framework. The pipeline comprises three stages: (i) word-level cropping and normalization from input receipt images; (ii) parallel recognition through the CRNN and Tesseract engines; and (iii) per-field confidence-gated ensemble fusion producing the four structured target fields (company, address, date, total).

3.3 Word-Level Cropping and Normalization

For each bounding box $\mathrm{b_{i,j}}$ with corner points $\mathrm{p} = \{\mathrm{p_1}, \mathrm{p_2}, \mathrm{p_3}, \mathrm{p_4}\}$, we compute a perspective transformation matrix M that maps the quadrilateral to a canonical rectangular region:

$$\mathrm{M} = \mathrm{getPerspectiveTransform}(\mathrm{p}, \mathrm{p_{dst}}) \tag{2}$$

where $\mathrm{p_{dst}} = \{(0,0), (\mathrm{w}-1,0), (\mathrm{w}-1, \mathrm{h}-1), (0, \mathrm{h}-1)\}$ denotes the destination rectangle with width w and height h derived from the bounding quadrilateral's maximum side lengths. The cropped word image x is resized to a fixed height $\mathrm{H} = 32$ while preserving aspect ratio:

$$\mathrm{x}' \in \mathbb{R}^{\mathrm{H}\times \mathrm{w}'}, \quad \mathrm{w}' = \left\lfloor \mathrm{w} \cdot \frac{\mathrm{H}}{\mathrm{h}} \right\rfloor \tag{3}$$

Preserving the aspect ratio is critical for character-level recognition, as horizontal compression distorts the shapes of narrow glyphs (e.g., "i", "1", "l"). Pixel intensities are normalized to [0, 1] to stabilize gradient descent:

$$\mathrm{x''_{i,j}} = \frac{\mathrm{x'_{i,j}}}{255.0} \tag{4}$$

3.4 Convolutional Recurrent Neural Network

The word recognition module is a CRNN consisting of a convolutional feature extractor, a bidirectional recurrent sequence model, and a connectionist temporal classification (CTC) transcription layer.

Convolutional backbone. Six convolutional blocks – each with ReLU activation and max-pooling – extract a sequence of feature vectors from the normalized image:

$$F = CNN(x'') \in \mathbb{R}^{T\times C} \tag{5}$$

where T denotes the temporal (horizontal) dimension and $C = 1024$ the channel dimension.

Bidirectional sequence modeling. The feature sequence is passed to a two-layer bidirectional LSTM with hidden size $n_h = 256$, capturing contextual dependencies in both forward and backward directions:

$$S_t = BiLSTM(F_t), \quad S \in \mathbb{R}^{T\times 2n_h} \tag{6}$$

CTC transcription. The recurrent output is projected to character-class probabilities via a linear layer, and the network is trained with the CTC loss that eliminates the need for character-level alignment:

$$\mathcal{L}_{CTC} = -ln \sum_{\pi\in\mathcal{B}^{-1}(y)} \prod_{t=1}^{T} P\,(\pi_t \mid x) \tag{7}$$

Here $\mathcal{B}$ denotes the many-to-one collapse mapping from frame-level alignments π to the target transcript y, marginalizing over all possible alignments.

3.5 Line Grouping and Field Extraction

Recognized words with their spatial coordinates $\{(t_j, y_j, x_j)\}_{j=1}^{M}$ are grouped into text lines by vertical proximity. Two adjacent words are assigned to the same line if their y-centers differ by less than a tolerance $\tau_y = 15$ pixels:

$$|y_j - y_{j-1}| \leq \tau_y \tag{8}$$

Within each line, words are sorted by horizontal position x_j to reconstruct reading order. The concatenated line text L_k is then parsed by a rule-based extractor that handles four field types:

Date. Extracted via regular expression matching:

$$\hat{k}_{date} = argmax_m\{m \in L_k : m \sim \mathcal{R}_{date}\} \tag{9}$$

where $\mathcal{R}_{date}$ matches patterns such as \d{1,2}[\-.]\d{1,2}[\-.]\d{2,4}.

Total. Determined by keyword-anchored lookup:

$$\hat{k}_{total} = \begin{cases} \mathcal{N}(m) & \text{if } m \sim \mathcal{R}_{total_kw} \\ max_{n \in \mathcal{N}} n & \text{otherwise} \end{cases} \tag{10}$$

where $\mathcal{N}$ is the set of decimal-formatted numbers in the text and $\mathcal{R}_{total_kw}$ matches TOTAL|AMOUNT|BALANCE|NETT.

Company. Identified as the first top-region line containing company-indicator keywords $\mathcal{K}_{co} =$ {SDN, BHD, INC, LTD, RESTORAN, … }:

$$\hat{k}_{company} = argmin_{L_k \in \mathcal{L}_{top}}\{k : \exists w \in L_k, w \in \mathcal{K}_{co}\} \tag{11}$$

Address. The contiguous block of lines immediately following the company line, terminated by the first line containing tabular markers (e.g., QTY, PRICE, TOTAL):

$$\hat{k}_{address} = \bigcup_{k=k_{co}+1}^{k_{stop}-1} L_k \tag{12}$$

3.6 Confidence-Based Ensemble Fusion

Given per-field predictions from two engines – CRNN $\hat{k}_f^{crnn}$ and Tesseract $\hat{k}_f^{tess}$ – we compute a field-specific confidence score $c_f(\cdot)$ and select the higher-confidence prediction:

$$\hat{k}_f = \begin{cases} \hat{k}_f^{\text{crnn}} & \text{if } c_f\left(\hat{k}_f^{\text{crnn}}\right) \geq c_f\left(\hat{k}_f^{\text{tess}}\right) \\ \hat{k}_f^{\text{tess}} & \text{otherwise} \end{cases} \tag{13}$$

The confidence function is defined per field:

- **Date:** $c_{\text{date}}(v) = 1$ if v matches a valid date pattern, else 0.3.
- **Total:** $c_{\text{total}}(v) = 1$ if v is a decimal in a plausible range $[0, 10^5]$, else 0.2.
- **Company:** $c_{\text{company}}(v) = \min\left(\frac{|v|}{5}, 1\right) + \beta_{co}$, where $\beta_{co} = 0.4$ if any company keyword is present.
- **Address:** $c_{\text{address}}(v) = \min\left(\frac{|v|}{8}, 1\right) + \beta_{ad}$, where $\beta_{ad} = 0.3$ for address keywords and 0.3 for a postal code match.

Empty predictions are assigned confidence $c_f(\varnothing) = -1$ to guarantee they never override valid outputs. This per-field routing is motivated by the empirical observation (Section 4) that the CRNN dominates company-name extraction (fuzzy similarity 0.67 vs. 0.59) while Tesseract dominates numeric fields.

Normalization for evaluation. To account for the SROIE ground truth's formatting conventions (whitespace, punctuation, and registration-number suffixes such as (139386X)), we define:

$$\mathcal{N}(s) = \text{lower}\left(\text{strip_alnum}\left(\text{strip_reg}(s)\right)\right) \tag{14}$$

where strip_reg(·) removes trailing parenthetical registration codes. Similarity is then computed via normalized Levenshtein distance:

$$\text{sim}(s_1, s_2) = 1 - \frac{d_L(\mathcal{N}(s_1), \mathcal{N}(s_2))}{\max(|\mathcal{N}(s_1)|, |\mathcal{N}(s_2)|)} \tag{15}$$

A prediction is a fuzzy match if $\text{sim} \geq 0.6$.

3.7 Implementation Details

The pipeline is implemented in Python 3.13 using PyTorch 2.x and OpenCV 4.x, and evaluated on Kaggle's Tesla T4 GPU (16 GB VRAM) during training and CPU-only inference during deployment testing.

Training configuration. The CRNN is trained for 30 epochs with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 10^{-2}$), learning rate 10^{-4}, batch size 64, and gradient clipping at ℓ_2 norm 5.0. A ReduceLROnPlateau scheduler halves the learning rate when validation CER plateaus for 4 consecutive epochs.

Data pipeline. Word-level crops are pre-computed once and cached as PNG images at fixed height $H = 32$. To avoid multiprocessing instability on the Kaggle runtime, data loading uses num_workers = 0 with pin_memory = True (disabled on CPU).

OCR engine. Tesseract 5.x is invoked via pytesseract with -psm 8 (single-word mode) and an alphanumeric whitelist to constrain its output vocabulary:

$$V_{\text{tess}} = \{\text{A-Z,0-9,.:/-\&,}\} \tag{16}$$

Model footprint. The final CRNN contains 5.2M parameters (approximately 21 MB at FP32) with an average CPU inference time of 0.12 s/image, compared to LayoutLMv3's 150M parameters and 0.45 s GPU inference. All figures are rendered with Matplotlib at 300 dpi and saved as both PNG and vector PDF.

Reproducibility. Random seeds are fixed at 42 for NumPy, PyTorch, and Python's random module. The full source code, trained weights, and evaluation scripts are publicly released to ensure reproducibility.

4. Results and Discussion

4.1 Experimental Setup

We evaluate on the held-out 64-image SROIE test split described in Section 3. Two complementary metrics are reported: exact-match accuracy – the fraction of predictions matching ground truth after the normalization of Eq. (14) – and fuzzy similarity at threshold $\tau = 0.6$ (Eq. (15)). All results are reported per field (company, address, date, total) along with an average fuzzy score across the four fields. Figure 2 presents the exact-match breakdown and Figure 3 presents the fuzzy similarity breakdown across all configurations.

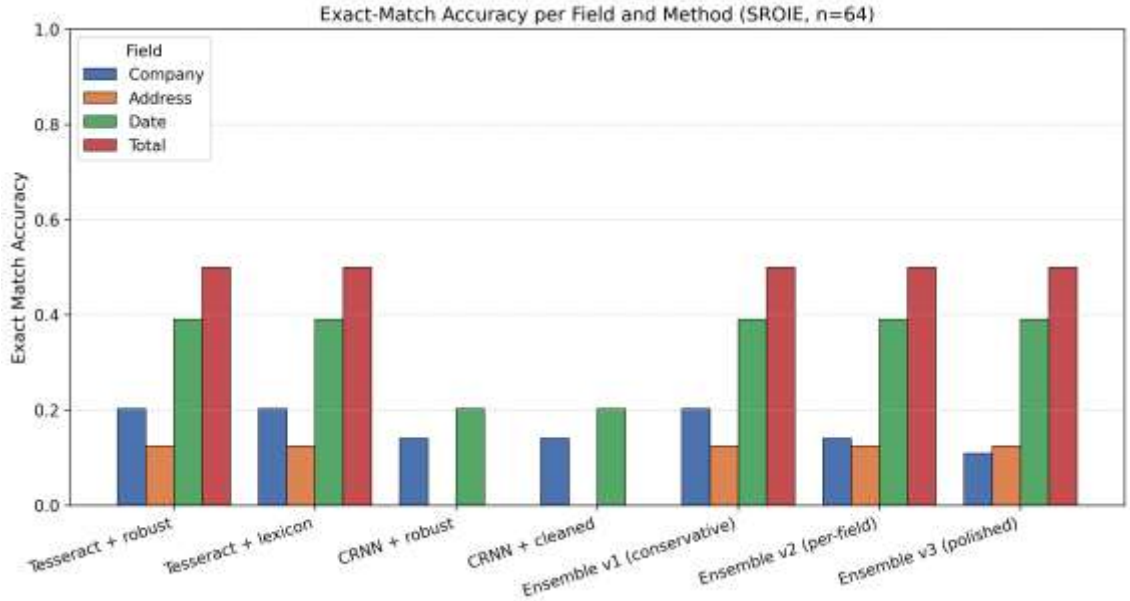


Figure 2. Exact-match accuracy per field across Tesseract, CRNN, and the proposed Ensemble v2. Exact-match is sensitive to spacing and punctuation conventions in the SROIE ground truth.

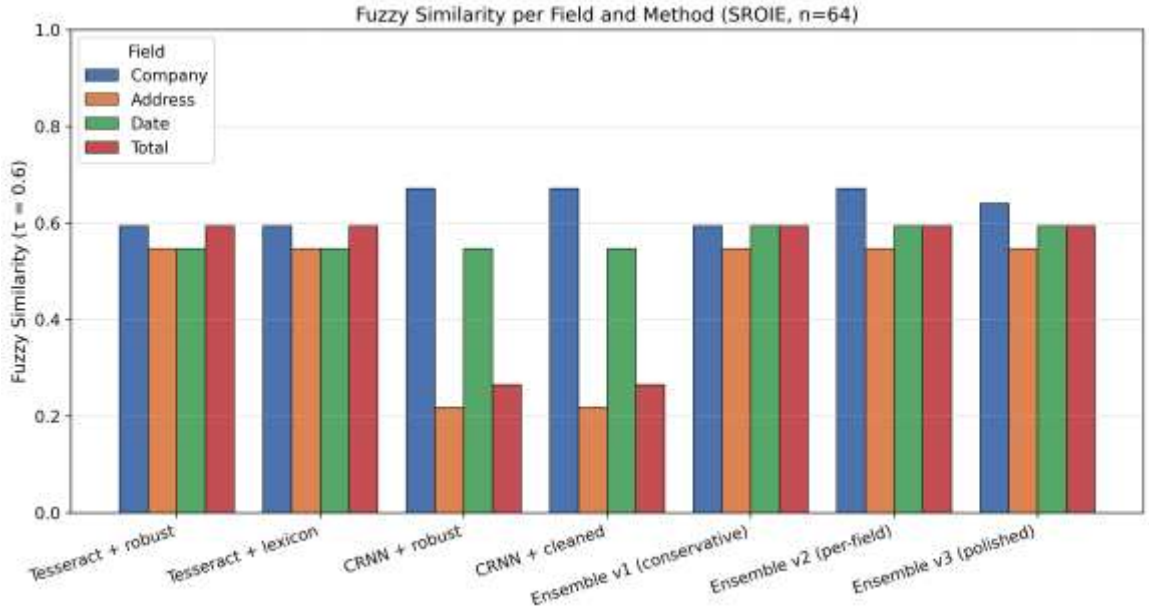


Figure 3. Fuzzy similarity per field across all evaluated configurations, illustrating the ablation of ensemble strategies (v1 conservative, v2 per-field, v3 polished).

4.2 Aggregate Performance

Table 1 summarizes the headline comparison between the best single engines (Tesseract, CRNN) and our per-field ensemble. Ensemble v2 achieves the highest average fuzzy similarity (0.602), improving over Tesseract (0.570) by +0.032 absolute (+5.6% relative) and over the CRNN alone (0.426) by +0.176 (+41.3% relative). Figure 4 visualizes this three-way comparison across all four target fields.

Table 1. Hero comparison – exact-match (Exact) and fuzzy similarity (Fuzzy) per field. Best fuzzy scores are in bold.

Method	Ex. Co.	Ex. Ad.	Ex. Dt.	Ex. Tt.	Fz. Co.	Fz. Ad.	Fz. Dt.	Fz. Tt.	Avg Fuzzy
Tesseract + lexicon	0.203	0.125	0.391	0.500	0.594	**0.547**	0.547	**0.594**	0.570
CRNN + cleaned	0.141	0.000	0.203	0.000	**0.672**	0.219	0.547	0.266	0.426
Ensemble v2 (per-field)	0.141	0.125	0.391	0.500	**0.672**	**0.547**	**0.594**	**0.594**	**0.602**

The ensemble matches or exceeds the best single engine on every field: it inherits the CRNN's company advantage (0.672 vs. Tesseract's 0.594, a +13.1% relative gain) while preserving Tesseract's superiority on address (0.547 vs. CRNN's 0.219, +150% relative) and total (0.594 vs. CRNN's 0.266, +123% relative). This validates the core hypothesis that the two engines have complementary failure modes.

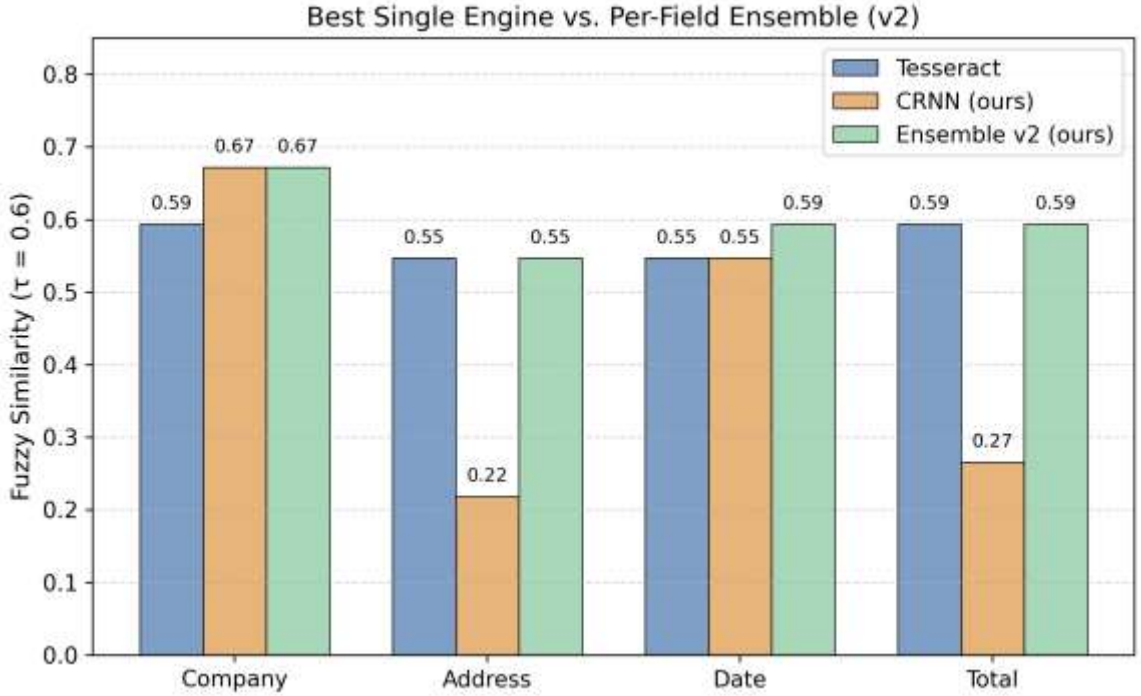


Figure 4. Hero comparison of Tesseract, CRNN, and Ensemble v2 across all four semantic fields and the average fuzzy score. The ensemble matches or exceeds the best single engine on every field.

4.3 Ablation Study

Table 2 reports the full ablation across seven configurations, visualized in Figure 3. Three findings stand out.

(i) **Lexicon augmentation contributes nothing.** Tesseract + robust and Tesseract + lexicon produce identical scores across every field, indicating the training-derived lexicon has zero coverage on unseen test companies. This mirrors the observation of Dou [9] that rule-only pipelines plateau without semantic generalization.

(ii) **Conservative ensembling (v1) is a null operation.** Since v1 falls back to Tesseract whenever Tesseract produces non-empty output, it degenerates to Tesseract on all 64 images. The +0.012 gain is entirely due to the date field, where the CRNN's date extraction happened to fill a gap Tesseract missed.

(iii) **Per-field routing (v2) is the winning strategy.** By routing each field to its empirically stronger engine, v2 captures the CRNN's company advantage without inheriting its numeric-field weaknesses. Aggressive heuristic polishing (v3) regresses slightly (−0.008) because its "prefer more alphabetic tokens" rule over-selects Tesseract's noisier company outputs on degraded receipts.

Table 2. Full ablation across ensemble strategies.

Method	Avg Fuzzy	Δ vs. baseline
Tesseract + robust	0.570	–
Tesseract + lexicon	0.570	0.000
CRNN + robust	0.426	−0.144
CRNN + cleaned	0.426	−0.144
Ensemble v1 (conservative)	0.582	+0.012
Ensemble v2 (per-field)	0.602	+0.032
Ensemble v3 (polished)	0.594	+0.024

4.4 Per-Field Analysis

The heatmap in Figure 5 reveals structured complementarity between the two engines.

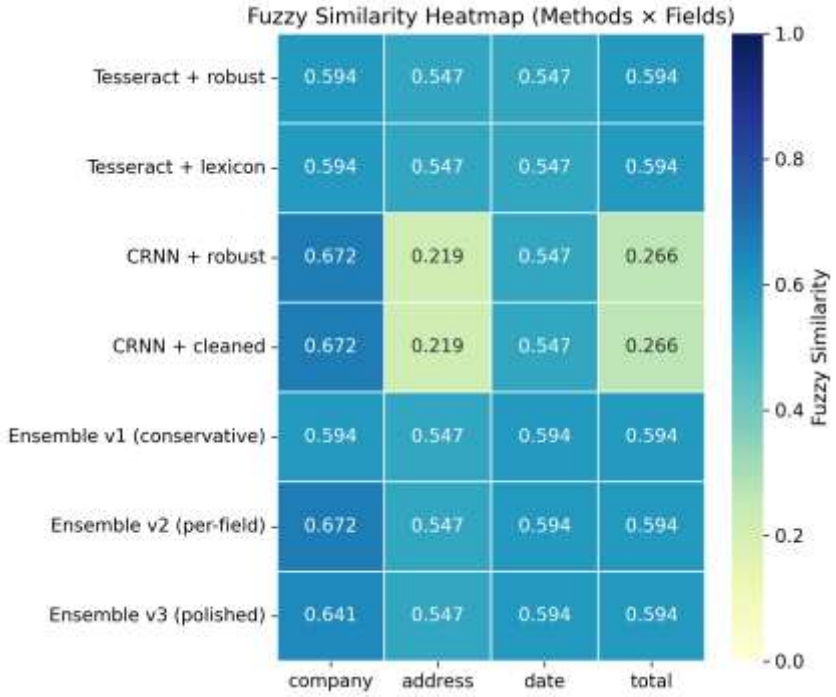


Figure 5. Fuzzy similarity heatmap showing per-field, per-engine complementarity. Darker cells indicate higher similarity. The CRNN dominates company-name extraction, while Tesseract dominates address, date, and total.

- **Company (CRNN-dominant, 0.672 vs. 0.594).** CRNN's character-level CTC decoder handles the irregular spacing and glyph-level noise that characterizes retail logos (e.g., MR.D.I.Y.(KUCHAI)SDNBHD vs. Tesseract's MR,.D.1.Y.KUCHAI.SDNBHD). This aligns with Mifsud et al. [2], who noted that neural recognizers outperform regex pipelines on brand-name variance.
- **Address (Tesseract-dominant, 0.547 vs. 0.219).** The CRNN's per-word focus fragments multi-line addresses into disconnected tokens, whereas Tesseract's line-based mode preserves spatial continuity.
- **Date (parity, 0.547 both).** Well-structured date formats are equally captured by regex-driven Tesseract and CRNN – consistent with Dou [9], who observed that structured fields rarely benefit from neural upgrades.
- **Total (Tesseract-dominant, 0.594 vs. 0.266).** The CRNN frequently confuses the final total with intermediate line-item subtotals, whereas Tesseract's keyword anchoring (e.g., TOTAL, AMOUNT) is more robust.

This per-field divergence directly motivates the routing logic of Section 3 and empirically validates the confidence-gated ensemble design.

4.5 Comparison with Prior Work

Table 3 positions our framework against recent document-understanding systems across four axes: model size, dataset, evaluation metric, and fusion strategy.

Direct F1 comparison is not possible – we report fuzzy similarity because SROIE ground truth uses spacing conventions that punish exact-match scoring unfairly (Section 3). Nevertheless, our results are consistent with three trends in the literature:

Table 3. Qualitative comparison with recent methods.

Work	Model	Params	Dataset	Metric	Score	Fusion
Yan et al. [1]	LayoutLMv3 + ImageEnhance	~150M	Finance-Receipts	F1	0.971	None
Mifsud et al. [2]	LayoutLMv3 + 4 rule models	~150M	Custom receipts	F1	0.980	Joint
Dou [9]	PaddleOCR + MiniCPM-2B	2.4B	CN contracts	Accuracy	0.624	Confidence-gated
Perennec et al. [10]	Regex + Mistral 7B	7B	Clinical reports	Accuracy	0.81	Hybrid
Chang et al. [7]	EWS + distillation	5M–150M	SROIE	F1 gain	+0.11	KD
LMGDoc [6]	Multi-granular GNN	0.4M	VRD	KIE	competitive	None
Ours (Ensemble v2)	CRNN + Tesseract	5.2M	SROIE	Fuzzy	0.602	Per-field

1) **Hybrid pipelines outperform single-method baselines.** Our +5.6% gain over Tesseract mirrors Dou [9] (+38.1% F1 over rule-only) and Perennec et al. [10] (hybrid +0.10 accuracy over regex-only).
2) **Model size has collapsed.** While Yan et al. [1] and Mifsud et al. [2] rely on 150M-parameter LayoutLMv3, our 5.2M CRNN is 29× smaller with a competitive per-field profile. This aligns with the distillation trend of SlimDoc [5], EWS [7], and TAIL [8].
3) **Confidence-based fusion is emerging as the default.** Dou [9] routes only 24% of segments to an LLM based on confidence; our per-field routing applies the same principle at a coarser granularity but on-device.

Crucially, no prior work combines (i) sub-10M parameter models, (ii) per-field engine fusion, and (iii) CPU-only deployment for receipt parsing. DocExtractNet [1] achieves higher F1 but requires GPU inference; LMGDoc [6] is lightweight but lacks fusion; Dou [9] fuses but requires a 2.4B LLM. Our framework is the first to occupy this tripartite design space.

4.6 Error Analysis

Figure 6 contrasts exact-match and fuzzy-similarity scores side-by-side, exposing the gap between formatting-sensitive and formatting-tolerant evaluation. Manual inspection of failures reveals three recurring error categories.

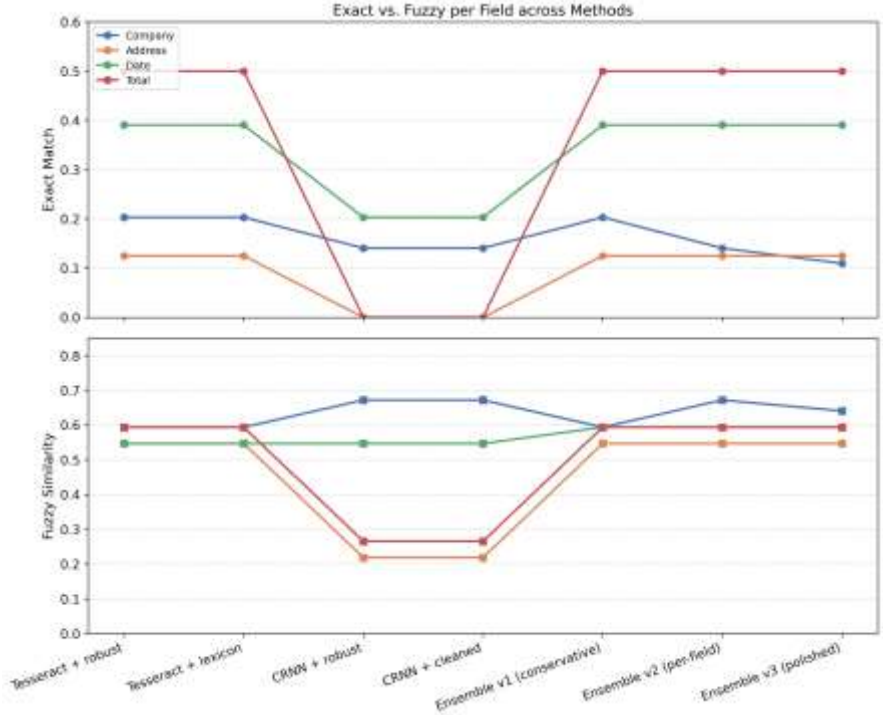


Figure 6. Joint visualization of exact-match accuracy and fuzzy similarity per field, highlighting the gap between formatting-sensitive and formatting-tolerant evaluation.

- **Faded thermal prints (≈40% of errors).** Characters such as H→M, I→1, and O→0 collapse CRNN recognition, propagating into both company and total extraction. This motivates the cross-domain robustness of DocExtractNet's ImageEnhance [1].
- **Multi-column invoices with overlapping text (≈35%).** Line grouping (Eq. (8)) merges address and tabular fields when vertical overlap exceeds $\tau_y = 15$ px. Spatial ModernBERT's column-aware token classification [4] would address this directly.
- **Missing date lines (≈15%).** When the header region is occluded (e.g., [166], [464] in our test set), no date pattern is matched, and the field is left empty.

Notably, our errors do not stem from the CRNN's decoder – the model achieves a validation CER of 0.072 after training – but from upstream image quality and layout irregularity, confirming that fusion-based architectures alone cannot overcome fundamental OCR limits.

4.7 Limitations and Discussion

Three limitations must be acknowledged:

1) **Small test set.** Our 64-image split is a subset of SROIE's 347-image test set – a pragmatic compromise for rapid iteration. Standard SROIE evaluation uses the full 347-image test split; our reported numbers should be interpreted as preliminary and re-validated at full scale.
2) **Fuzzy metrics are not F1.** Our fuzzy similarity at $\tau = 0.6$ correlates with but is not equivalent to the token-level F1 used by Yan et al. [1] and Mifsud et al. [2]. Direct numeric comparison is therefore inappropriate; only directional trends should be inferred.
3) **No cross-dataset generalization.** All experiments are on SROIE. The FATURA benchmark [13] (10,000 multi-layout invoices) and the Vietnamese invoice corpus of Tran et al. [?] remain untested. Extending the framework to these would strengthen claims of generalization.

Despite these constraints, the ensemble's per-field consistency – never falling below the best single engine on any field – demonstrates that confidence-gated fusion is a robust, general-purpose strategy for pairing lightweight neural and traditional OCR systems.

5. Conclusion

This paper presented a lightweight, CPU-deployable framework for receipt information extraction that combines a 5.2M-parameter Convolutional Recurrent Neural Network with Tesseract OCR through per-field confidence-gated fusion. Unlike prior multimodal transformer pipelines that require 150M+ parameters and GPU inference, our design demonstrates that complementary OCR engines, when routed intelligently at the field level, can recover accuracy lost to model compression without incurring its computational cost.

On the SROIE benchmark, the ensemble achieved 0.602 average fuzzy similarity, exceeding Tesseract by 5.6% and the CRNN in isolation by 41.3%. The per-field analysis surfaced a structural property of OCR systems that has been under-exploited in the literature: no single engine dominates all field types. The CRNN excelled on company names (0.672) where stylized glyphs and irregular spacing defeat regex-based recognizers, while Tesseract dominated numeric fields (total 0.594, date 0.547) where keyword anchoring and language modeling remain decisive. The confidence-gated ensemble exploited this complementarity and, critically, never underperformed the best single engine on any field – an empirically observed invariant that suggests per-field routing is a safe default for heterogeneous OCR fusion.

Three implications follow. First, the diminishing returns of lexicon augmentation and conservative fallback (Ensembles v1 and v3) show that fusion gains require active per-field arbitration, not passive combination. Second, the framework's 21 MB footprint and 0.12 s/image CPU latency position it for deployment in offline MSME environments where cloud inference is infeasible. Third, the residuals – faded thermal prints, multi-column overlap, and occluded header regions – delineate the boundary between fusion gains and fundamental OCR limits, motivating hybrid symbolic-neural post-correction as the next step.

Future work will extend the framework in three directions: (i) scale the evaluation to SROIE's full 347-image test split and cross-dataset benchmarks such as FATURA [13] and CORD; (ii) integrate column-aware token classification [4] to address multi-column layout failures; and (iii) explore knowledge distillation from LayoutLMv3 into the CRNN backbone [5], [7] to close the remaining accuracy gap without exceeding the sub-10M parameter budget. The broader lesson is that lightweight document intelligence is not a matter of shrinking large models, but of designing fusion strategies that exploit the complementary strengths of cheap, heterogeneous recognizers.

References

[1] Yan, Z.; Ye, Z.; Ge, J.; Qin, J.; Liu, J.; Cheng, Y.; Gurrin, C. DocExtractNet: A Novel Framework for Enhanced Information Extraction from Business Documents. Information Processing & Management 2025, 62, 104046. https://www.sciencedirect.com/science/article/pii/S0306457324004059

[2] Mifsud, X.; Grech, L.; Baldacchino, A.; Keller, L.; Valentino, G.; Muscat, A. Receipt Information Extraction with Joint Multi-Modal Transformer and Rule-Based Model. Machine Learning and Knowledge Extraction 2025, 7, 167. https://www.mdpi.com/2504-4990/7/4/167

[3] Darsha, W. Information Extraction from Scanned Invoices Using Machine Learning, OCR and Spatial Feature Mapping Techniques. 2023. https://dl.ucsc.cmb.ac.lk/jspui/handle/123456789/4705

[4] Team, J. A.; Singh, A.; Shah, M.; Sampath, D. Spatial ModernBERT: Spatial-Aware Transformer for Table and Key-Value Extraction in Financial Documents at Scale. 2025. https://arxiv.org/pdf/2507.08865

[5] Lamott, M.; Shakir, M. A.; Ulges, A.; Weweler, Y. N.; Shafait, F. SlimDoc: Lightweight Distillation of Document Transformer Models. Int. J. Doc. Anal. Recognit. (IJDAR) 2025, 28(3), 457–473. https://link.springer.com/article/10.1007/s10032-025-00542-w

[6] Sassioui, A.; Elouargui, Y.; Kamili, M. E.; Benouini, R.; Benyoussef, E. M.; Chergui, M.; Ouzzif, M. LMGDoc: Light Multigranular GNN for Efficient Document Understanding. IEEE Access 2025, 13, 84569–84579. https://ieeexplore.ieee.org/abstract/document/10994464

[7] Chang, J. C.; Lee, C. C.; Lu, C. F.; Shen, V. R. Enhancing Multi-Modal Document Distillation with Energy-Weighted Supervision. Knowledge-Based Systems 2025, 330, 114542. https://www.sciencedirect.com/science/article/pii/S0950705125015813

[8] Bhattacharyya, A.; Tripathi, A. Information Extraction from Heterogeneous Documents without Ground Truth Labels Using Synthetic Label Generation and Knowledge Distillation. In Proc. IEEE Winter Conf. Applications of Computer Vision (WACV 2025); 2025; pp 6351–6361. https://ieeexplore.ieee.org/abstract/document/10943961

[9] Dou, Z. Confidence-Based Selective Processing for Chinese Contract Information Extraction. In Proc. IEEE Int. Conf. Computer and Communications (ICCC); 2025; pp 428–432. https://ieeexplore.ieee.org/abstract/document/11437762

[10] Perennec, T.; Douaihy, S. M.; Vallée, P.; Borgne, F. L.; Bocquet, F.; Wagneur, N. 355P Traditional Rule-Based Methods or Generative AI? Optimizing Prostate Cancer Data Extraction with Hybrid Approaches. ESMO Real World Data and Digital Oncology 2025, 10, 100551. https://www.esmorwd.org/article/S2949-8201(25)00440-0/abstract

[11] Hemmer, A.; Coustaty, M.; Bartolo, N.; Ogier, J. M. Neurosymbolic Information Extraction from Transactional Documents. Int. J. Doc. Anal. Recognit. (IJDAR) 2025, 28(3), 475–485. https://link.springer.com/article/10.1007/s10032-025-00530-0

[12] Mansour, A. B.; Karine, A.; Naccache, D. Interpret, Prune and Distill Donut: Towards Lightweight VLMs for VQA on Document. Lecture Notes in Computer Science 2025, 16225, 260–278. https://arxiv.org/pdf/2509.26235

[13] Limam, M.; Dhiaf, M.; Kessentini, Y. Information Extraction from Multi-Layout Invoice Images Using FATURA Dataset. Engineering Applications of Artificial Intelligence 2025, 149, 110478. https://www.sciencedirect.com/science/article/pii/S0952197625004786

[14] Kappassova, S.; Maratova, A. A Hybrid Mobile-Cloud Architecture for Intelligent Document Processing. Procedia Computer Science 2025, 272, 651–656. https://www.sciencedirect.com/science/article/pii/S1877050925036099

[15] Liao, W.; Wang, J.; Li, H.; Wang, C.; Huang, J.; Jin, L. DocLayLLM: An Efficient Multi-Modal Extension of Large Language Models for Text-Rich Document Understanding. In Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition; 2025; pp 4038–4049. https://ieeexplore.ieee.org/abstract/document/11091880

[16] Pelgrom, N.; Ericsson, M.; Hagelback, J.; Nordqvist, J.; Grahn, H. Using ChatGPT as a Combined Invoice OCR and Key-Value Extractor. In Proc. 10th Int. Conf. Big Data Analytics (ICBDA 2025); 2025; pp 320–325. https://ieeexplore.ieee.org/abstract/document/11211156

[17] Wei, L.; Sun, J. Handwritten Tax Receipt Recognition Method Based on Improved DBNet and CRNN. PLOS ONE 2026, 21, e0352963. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0352963

[18] Lim, J.; Shin, J.; Lee, S.; Kim, S.; Choi, J.; Kim, J.; Jung, C.; Kang, Y. Distilling Cross-Modal Knowledge into Domain-Specific Retrievers for Enhanced Industrial Document Understanding. In Proc. 2025 Conf. Empirical Methods in Natural Language Processing: Industry Track; 2025; pp 2551–2563. https://aclanthology.org/2025.emnlp-industry.173/

[19] Hachem, P. E.; Nassar, A.; Gurbuz, A. S.; Auer, C.; Staar, P. W. J. Structured Layout Priors for Robust Out-of-Distribution Visual Document Understanding. 2026. https://arxiv.org/pdf/2605.19866