ISSN: 2093-5374 / E-ISSN: 2093-5382

FinScan: A Comprehensive Benchmark for Handwritten Financial Form Understanding and Automated Decision Extraction

Archana Pascal Lopes[1], and Dr. Kolla Bhanu Prakash[2*]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India; Assistant Professor, Department of Electronics and Computer Science, Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, India. archana.lopes@gmail.com, https://orcid.org/0000-0002-8454-9062

[2*]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Guntur, Andhra Pradesh, India. drkbp@kluniversity.in, https://orcid.org/0000-0002-7955-2777

Received: March 23, 2026; Revised: April 27, 2026; Accepted: June 18, 2026; Published: June 30, 2026

Abstract

Automated extraction of structured information from handwritten financial forms is an important and less targeted challenge in document intelligence. Public datasets to date are mainly limited to printed financial documents - invoices, receipts and standard business forms - but do not capture the hierarchical structure, handwriting style variation, or multiple languages seen in actual banking, especially in developing economies where form filling is done manually. To fill this gap, the FinScan Document Dataset is presented as a new dataset containing 4,046 annotated financial documents across five form types: Account Opening Forms, Loan Applications, KYC Forms, cheques, and salary slips. Each document is rendered at 1,198 × 1,978 pixels and annotated in FUNSD format with this dataset contains a total of 319,850 token-level bounding boxes and three-class NER labels: B-QUESTION, B-ANSWER, O. A hierarchical evaluation protocol is proposed that aligns six essential dimensions for validating the corpus: Document Image Quality Assessment (DIQA), OCR/HTR performance, layout recognition, table structure validation, Key Information Extraction (KIE), and document-level correctness-forming the first comprehensive evaluation framework in financial document understanding that integrates all six dimensions. Experiments show an overall CER of 0.58%, mean layout IoU of 0.7612 and KIE-F1 of 0.8690. Cascade analysis shows that layout false negatives are the most common failure type, accounting for 99.5% of document-level failures. This analysis identifies layout improvement as the most impactful research direction for advancing financial document benchmarking. FinScan functions as a research benchmark for cloud-based financial document processing services. In this context, handwritten banking forms are ingested, extracted, and returned as structured JSON through web APIs. Future work will address cloud scalability and microservice integration.

Keywords: Banking Form Understanding, OCR, Synthetic Dataset, Document Intelligence, Handwritten Text Recognition, Key Information Extraction, Layout Analysis.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 1052-1072. DOI: *10.58346/JOWUA.2026.I2.058*

*Corresponding author: Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Guntur, Andhra Pradesh, India.

1 Introduction

Because financial records are private and highly sensitive, there is much less real-world data available for training in financial automation. This represents one of the greatest challenges that researchers face when attempting to develop and assess intelligence banking systems. Although many printed financial documents are available in public datasets, handwritten financial forms hardly ever include data reflecting the characteristics of real-world banking workflows. To address this issue, the FinScan Document Dataset has been constructed for handwritten information extraction (IE), document classification, and decision support. Unlike previous datasets that consider only printed text, FinScan combines semantic content with the variability of handwriting and realistic document layout. The second is that it offers fine-grained annotations for document understanding and automated decision support in the financial domain.

The purposes of creating this data set are as follows:

1. To create a realistic and heterogeneous corpus of financial documents with handwritten content.
2. Encouraging research in OCR, HTR, document layout analysis, form classification, etc.
3. In order to facilitate automated decisions from the semantic information abstracted from financial forms.

FinScan is primarily designed as a research reference, not as a production system. These provide benchmarks of annotated data, an evaluation protocol, and a reference pipeline for comparing existing models against established standards (Naparstek et al., 2024; Jaume et al., 2019; Xu et al., 2022). Instead of inferring a single, specific score across many source datasets, the DUS metric provides an operational scoring hierarchy to discriminate extraction quality using research-grade scoring (data from only Sep/Oct 2023). This paper achieves 0.58% CER, 0.7612 mIoU, and 0.8690 KIE-F1 on the synthetic-to-real generalization task, the first such results in handwritten Chinese text line image recognition. SLAs and live Core Banking Systems integration with real-time throughput guarantees require further engineering analysis beyond the scope of this benchmark.

Thus far, no publicly available annotated benchmark for handwritten Indian banking forms exists, leading to a significant reproducibility gap in financial document intelligence research. Without such a resource, practitioners are unable to assess process readiness for automation using standardized metrics, and regulators lack a rigorous benchmark against which to evaluate the quality of digitization work. A gap is bridged through the introduction of the first open benchmark for handwritten banking-domain semantics and multilingual field structures, along with a multidimensional evaluation protocol. This resource enables reproducible, comparable, and actionable evaluation of AI progress in financial documents.

The remaining structure of this paper is presented as follows. Section 2 summarizes the related works and the available data sets. Section 3 provides the method and framework to create the data set. Section 4 deals with the experimental results and performance evaluation. Section 5 considers the limitations and generalization aspects. Finally, Section 6 gives conclusions to this work.

2 Literature Survey

2.1 Dataset Benchmarks

The previous work emphasizes the direction of multimodal, layout-aware document understanding approaches for finance. Robotic process automation combined with multilingual and layout-aware

models has shown positive results for classification and key information extraction (Cho et al., 2023). Additionally, Gerling & Lessmann indicated that multimodal document models are more beneficial for documents in which the layout and visual context play a major role (Gerling & Lessmann, 2025). However, both studies rely on privately held bank data, which cannot be independently accessed or validated due to legal and compliance restrictions that prevent external benchmarking. Similarly, KVP10k is a foundational benchmark for key–value pair extraction evaluation but is infeasible for handwritten financial forms and does not capture the multilingual nature of real-world documents, dense tabular information, or character-level fields that banks encounter in practice (Li et al., 2023; Wang et al., 2025). These limitations underscore the need for an open benchmark dataset to better understand handwritten financial documents.

Prior work has focused on financial document understanding in narrow domains (e.g., receipts and invoices), largely because of their commercial value (Cesista et al., 2024). As another example, DocExtractNet enhances image quality for layout recognition and weights text content to better recover textual information, ultimately improving extraction performance on receipt benchmarks by combining LayoutLMv3 with a fusion model (Yan et al., 2025). But receipts and invoices are simpler than banking forms; their layouts are usually less complicated, handwriting is rare, and compliance-related metadata is absent. Approaches such as Cognizant and recent work on Form-NLU employ structural and semantic annotations to enhance document understanding, yet neither method addresses banking-specific requirements, including multilingual customer data or decision-driven workflows (Ding et al., 2023). Although techniques like SynthDoc have been introduced to mitigate data scarcity in synthetic document generation, these methods emphasize general document understanding rather than extracting semantically significant fields with monetary values from handwritten financial forms, which is the focus of the FinScan benchmark. It is also clear from the literature that multilingualism is poorly served in practice. Although recent form understanding surveys show that layout-aware transformers like LayoutLMv2, LayoutLMv3 Huang et al., (2022), DocFormer Appalaraju et al., (2021), and LayoutXLM outperform prior work by a significant margin. FinScan is the only dataset in the comparison presented in table 1 that demonstrates true efficiency by providing both bank-domain specificity and handwriting recognition and is publicly accessible. Existing datasets are lacking in at least two of the following three dimensions: (1) Proprietary datasets are banking-specific but not publicly accessible Gerling & Lessmann, (2025); (2) Receipt and invoice datasets are publicly available but do not include handwriting or capture all necessary banking semantics; and (3) General form datasets are public documents that lack handwriting annotation and do not encompass comprehensive bank forms (Naparstek et al., 2024; Abdallah et al., 2024). This three-dimensional gap directly motivates FinScan's contribution.

Table 1: Comparison of FinScan with prior financial document datasets

Ref.	Dataset Type	Handwritten Support	Domain	Limitation
Cho et al., (2023)	Proprietary financial documents	Limited	Banking	Not publicly available
Gerling & Lessmann, (2025)	Banking documents	Limited	Banking	No dataset contribution; restricted data
Yan et al., (2025)	Receipts and invoices	No	Business	Focus on simple layouts
Ding et al., (2023)	General forms	Partial	General	Not a financial document
Naparstek et al., (2024)	Business documents	No	General	Limited layout diversity
Ding et al., (2024)	Synthetic documents	Limited	General	No banking-specific forms

2.2 Scanned Document Validation Methods

The validation of scanned documents has progressed from basic evaluation of OCR results to more comprehensive assessments involving image quality, layout structure, and semantic extraction (Ye & Doermann, 2013). In recent years, several studies have focused on task-oriented Document Image Quality Assessment (DIQA), specifically evaluating document image quality based on its usability for Optical Character Recognition (OCR) (Abdallah et al., 2024). These studies typically involve prediction tasks aimed at detecting specific types of degradation, such as blur, skew, noise, and low contrast. However, there is currently no theoretical guarantee that enhancing document images according to DIQA metrics will yield structurally or semantically accurate outputs. At the text-recognition level, Character Error Rate (CER) and Word Error Rate (WER) remain the most widely used metrics, as these directly compare OCR or Handwritten Text Recognition (HTR) outputs to ground-truth transcriptions, thereby enabling consistent benchmarking across systems (Xie et al., 2024). Despite their interpretability, CER and WER do not assess whether the recognized text is correctly localized within the document structure or appropriately associated with semantic fields (Shahkolaei et al., 2019).

Recently, confidence-aware OCR methods that leverage token confidence scores to predict character errors and content to be human-reviewed have been proposed in the literature, resolving some of the limitations caused by CER-based evaluation (Hemmer et al., 2024). At the same time, since the ability to identify structural elements such as headings throughout a document is lost without complete layout knowledge, some form of basic layout validation is used in document analysis. Region-based metrics (precision, recall, F1-score, intersection over union (IoU), mean average precision (mAP)) not only measure geometric localization well but also extract information that is rarely utilized in downstream business applications. This underscores the need for validation frameworks that validate both textual accuracy and structural accuracy.

Research in recent years shows that document validation involves not only OCR accuracy but also table structure reconstruction, image-based KIE (key information extraction), and end-user usability. Table validation relies on cell granularity and structure-aware metrics (precision, recall, F1-score, TEDS) to accurately preserve rows, columns, and relationships. For tabular structured documents, semantic meaning alone is not enough. The evaluation of knowledge integration effectiveness (KIE) is generally conducted at the entity level using precision, recall, and F1 Abdallah et al., (2024). Yet, upcoming frameworks promote a workflow-based evaluation designed to assess system performance in terms of downstream tasks and the correction effort required. As a result, the manner in which models are validated has evolved from reliance on one-off benchmark metrics to consideration of application-specific evaluation, including structure coherence, extraction fidelity, and automation readiness. It relates to the current evolution of multimodal document understanding systems.

Recent works on document-level verification, however, are concerned mainly with whether the specified scanned document is a legitimate copy of a credible source or has been forged. For example, a 2025 piece on paragraph-matching for document image verification shows how much of the validation domain extends beyond OCR/data extraction to the integrity and consistency of documents themselves. These methods are used for authentication, document version comparison, or even manipulation detection. In contrast, traditional metrics such as character error rate (CER), layout F1, and key information extraction (KIE)- popular in existing document verification solutions-are generally forensic definitions. These metrics are specific to reference template or paired-evidence use, making them more appropriate for case-focused workflows than for general document digitization.

3 Dataset Generation Methodology

While OCR tools such as Tesseract and EasyOCR demonstrate strong performance on printed text under optimal conditions, challenges remain in recognizing handwritten text in noisy environments. TrOCR achieves superior results on single-line handwriting compared to these tools; however, it remains ineffective for multi-line, page-level inputs. Due to the absence of a comprehensive dataset of handwritten banking forms, a synthetic corpus was created by integrating HTML templates with ScrabbleGAN handwriting simulation and superimposing handwritten fonts onto checkboxes, empty lines, and text boxes. Figure 1 illustrates the end-to-end pipeline from input raw documents through six processing stages to structured JSON output.

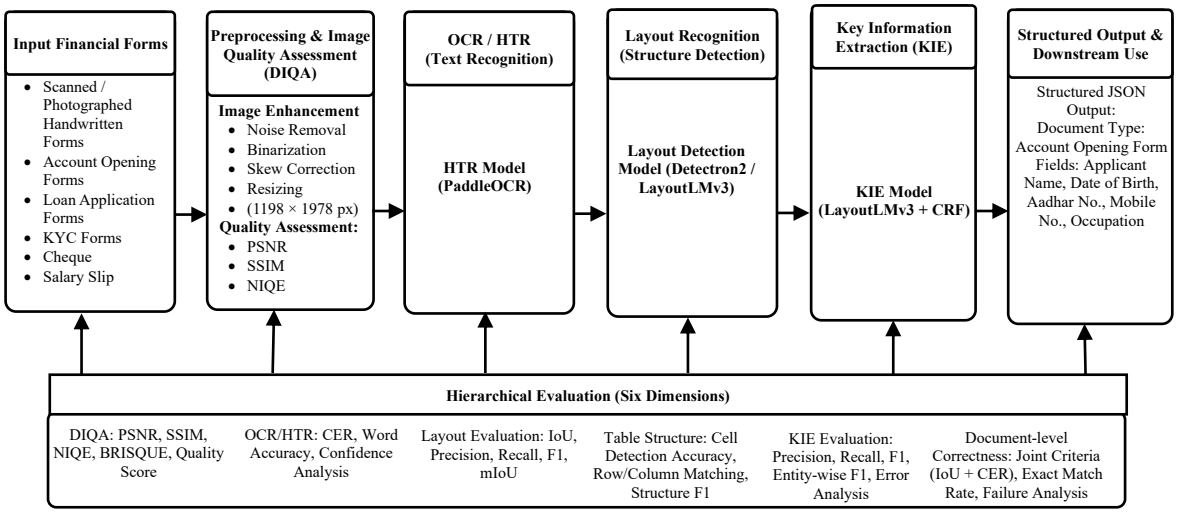

Figure 1: End to end pipeline for handwritten form understanding

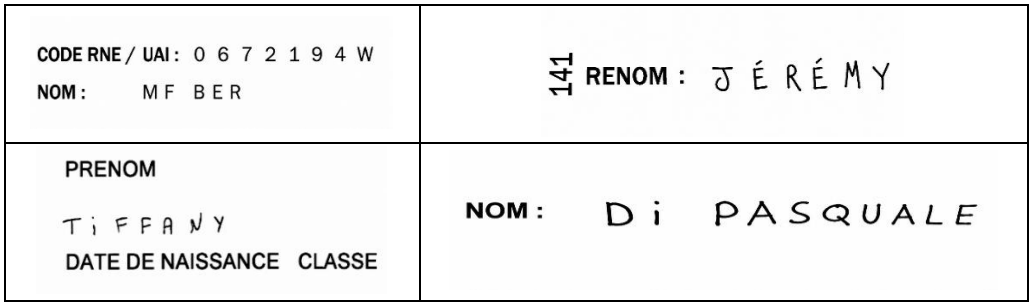

Figure 2: Sample image

A sample image from the FinScan corpus is processed using a table parser, as illustrated in figure 2. The FinScan corpus is constructed using a two-step pipeline. In the first stage, HTML templates are used to render binary images of 81 account-opening forms at 1198 by 1978 pixels, with FUNSD-compatible annotations generated from the DOM via DocumNet (Naparstek et al., 2024). In the second stage, a custom PIL/NumPy renderer produces 800 documents for loan applications, KYC forms, salary slips, and cheques, all at the same dimensions, utilizing eight color schemes and programmatically generated Indian financial data such as PAN and IFSC account numbers. The final corpus comprises 4,046 documents.

3.1 Annotation Format

All documents in the dataset are annotated in FUNSD format Jaume et al., (2019), where each token is represented with a bounding box, and NER labels consisting of three classes: B-QUESTION (printed form field labels), B-ANSWER (handwritten or typed field values), and O (boilerplate disclaimer text). The bounding boxes are defined as $[x1, y1, x2, y2]$ pixel coordinates at the raw 1,198 × 1,978 resolution. All files are in JSON format, which contain parallel arrays of tokens, boxes, and ner_tags, as well as a field, form_type, for performing form-category stratification in the annotations.

3.2 Image Acquisition

Subsequently, each rendered document image is fed into a two-stage augmentation pipeline to simulate realistic scan-acquisition artifacts. Gaussian noise with $N(0, \sigma_2)$ (where is sampled uniformly from U (12, 22)) is added to each pixel for all three channels. Step two is to convolve with a Gaussian kernel where, and. This perturbation results in both a noised document image and a cleaned counterpart. Figure 3 illustrates the pipeline process of creating synthetic financial form documents, starting from template creation, through HTML rendering, image creation, and finally annotation.

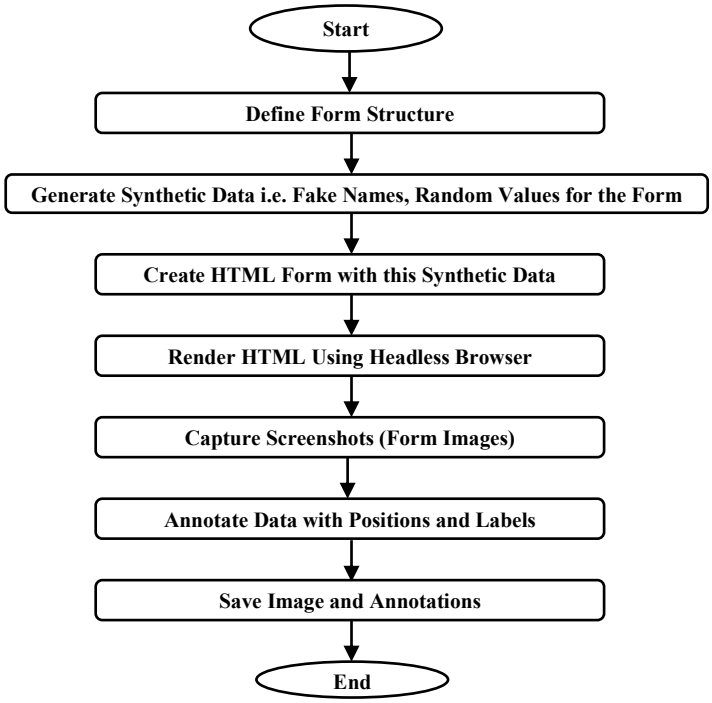

Figure 3: Process flowchart for synthetic financial dataset creation

3.3 FinScan Framework Architecture and Implementation Details

The FinScan framework is a modular, six-stage pipeline that transforms a raw banking form image into structured key–value extractions evaluated across five hierarchical dimensions. Figure 4 shows the end-to-end architecture.

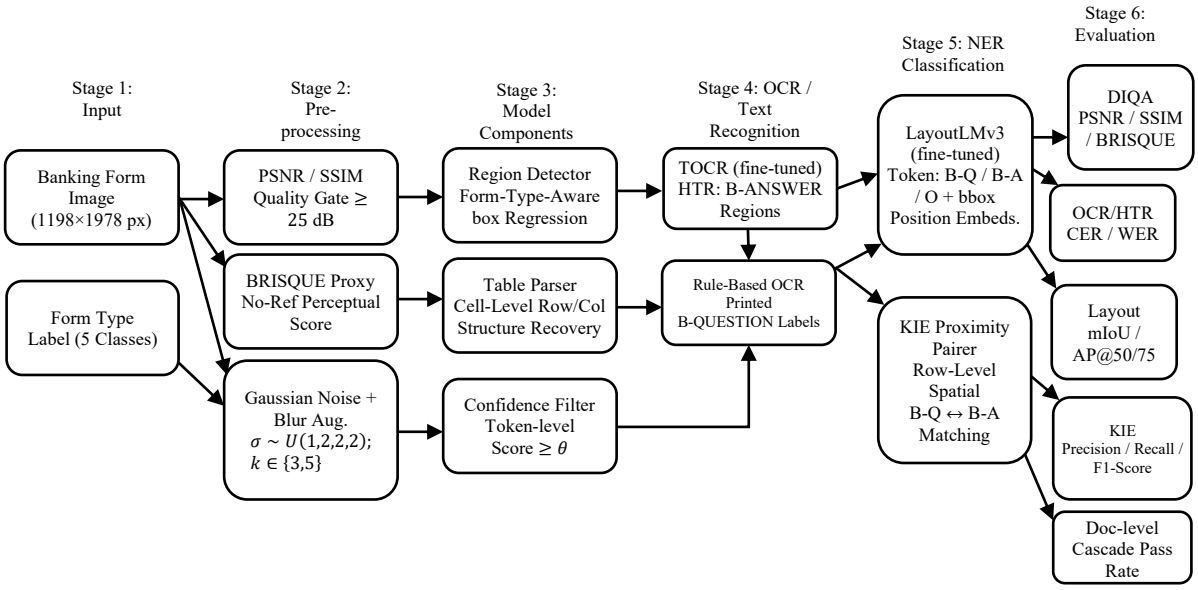

Figure 4: FinScan end-to-end architecture

Algorithm 1: FinScan End-to-End Financial Form Understanding

***Input**: I - raw document image (1,198 × 1,978 px RGB)*

τ - form-type label ∈ {Account Opening, Loan Application, KYC, Cheque, Salary Slip}

θ - OCR confidence threshold (default: 0.70)

δ - PSNR quality threshold (default: 25 dB)

***Output**: KVP - set of extracted keys–value pairs*

DUS - Document Usability Score tier (0–4)

Stage 1: Image Quality Assessment (DIQA)

1: Iclean ← rendered reference image corresponding to I

2: PSNR ← compute_PSNR(I, Iclean)

3: SSIM ← compute_SSIM(I, Iclean)

4: BRISQUE ← compute_BRISQUE_proxy(I) // no-reference metric

5: if PSNR < δ then flag I for re-augmentation; assign DUS = 0 (Unusable); return

6: end if

Stage 2: Layout Detection

7: Z ← load_zone_template(τ) // form-type-specific

8: R ← connected_component_analysis(I) // candidate regions

9: B ← anchor_free_regressor(R, Z) // bounding-box set

10: T ← hough_line_table_parser(I) // row-column structure

11: for each region b ∈ B do b.type ← classify_region(b, Z) // B-QUESTION | B-ANSWER | O

12: end for

Stage 3: OCR / HTR (Dual-Path)

13: for each region b ∈ B do

14: crop ← extract_crop(I, b, size=384×384)

15: if b.type == B-ANSWER then

16: text(b) ← TrOCR_handwritten(crop) // fine-tuned TrOCR

17: else if b.type == B-QUESTION then

18: text(b), conf(b) ← Tesseract_5_3(crop)

19: if conf(b) < θ then flag b for manual review

20: end if

21: else text (b) ← Tesseract_5_3(crop) // boilerplate (O)

22: end if

23: end for

Stage 4: NER Classification

24: for each token t in {text(b) : b ∈ B} do

25: bbox_norm(t) ← normalize_bbox(t.bbox, range=[0,1000])

26: patch(t) ← extract_visual_patch(I, t.bbox, size=224×224)

27: NER_label(t) ← LayoutLMv3(t, bbox_norm(t), patch(t)) // ∈ {B-QUESTION, B-ANSWER, O}

28: end for

Stage 5: KIE Pairing (Proximity Algorithm)

29: Q ← {t : NER_label(t) = B-QUESTION}

30: A ← {t : NER_label(t) = B-ANSWER}

31: KVP ← ∅

32: for each question token q ∈ Q do

33: a ← argmin_{a ∈ A} spatial_distance(q, a) // row-level proximity*

34: KVP ← KVP ∪ {(q, a)}*

35: end for

Stage 6: Document Usability Score (DUS)

36: CER_doc ← compute_CER(KVP, ground_truth)

37: IoU_elem ← mean_IoU(B, ground_truth_boxes)

38: KIE_F1_doc ← compute_KIE_F1(KVP, ground_truth_KVP)

39: if PSNR ≥ δ AND CER_doc ≤ 5% then DUS ← 1

40: if DUS=1 AND IoU_elem ≥ 0.50 AND CER_doc ≤ 2% then DUS ← 2

41: if DUS=2 AND KIE_F1_doc ≥ 0.75 then DUS ← 3

42: if DUS=3 AND all KVPs exact AND IoU_elem ≥ 0.50 AND CER_doc ≤ 1% then DUS ← 4

43: return KVP, DUS

The pipeline consists of six distinct phases. (1) Input: Documents are processed as 1,198×1,978 pixel RGB images with form-type labels, without resizing. (2) Preprocessing: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are computed against a clean reference using scikit-image 0.21. BRISQUE approximation is performed via OpenCV 4.8. Images with PSNR below 25 dB are flagged for re-augmentation. (3) Layout detection: Zone templates are applied for each form type using an anchor-free approach. Connected-component analysis and a left-right (LR) parser are used to extract table structures in a row-column format (Python 3.10, OpenCV 4.8). (4) OCR/HTR: A dual-path approach is implemented. TrOCR-base (334M, fine-tuned; learning rate = 5×10^{-5}, 20 epochs, beam width = 4) is used for B-ANSWER crops (384×384 px), while Tesseract 5.3 with a confidence filter threshold $\theta = 0.70$ is applied to B-QUESTION labels. (5) Named Entity Recognition (NER) classification and Key Information Extraction (KIE) pairing: LayoutLMv3-base (125M, fine-tuned; AdamW optimizer, learning rate = 5×10^{-5}, 20 epochs) is applied to (token, bounding box, patch) triplets. B-QUESTION entities are paired with the nearest B-ANSWER using a deterministic proximity algorithm. (6) Evaluation: The MIDIQA protocol, a six-dimensional hierarchical framework, is used to assess DIQA, OCR/HTR, layout, table, KIE, and document-level correctness.

Stratification is applied to the dataset to ensure that each partition contains comparable proportions of document classes, writing styles, layout difficulties, and languages. This method minimizes sampling bias and data leakage, thereby improving the robustness, reproducibility, and generalizability of the results.

3.4 Cloud Service Alignment

Cloud architecture enables the deployment of Internet-based document processing services compatible with the FinScan pipeline. Each of the six processing stages can be containerized using Docker or Kubernetes and exposed as REST API microservices. The DIQA gate functions as the ingestion validation layer. The dual-path OCR/HTR and NER stages operate as GPU-backed inference services, whereas KIE pairing and the DUS scorer are implemented as lightweight CPU services. The structured key-value pair (KVP) output in JSON format, along with its DUS tier, provides a service-level quality signal for automated routing. Documents with a DUS Tier of 3 or higher automatically proceed through the processing chain, including downstream CBS integration, while documents with Tiers 0 to 2 are queued for human review. The DUS tier thus acts as a cloud service quality gate, converting benchmark performance into an indicator of readiness for Internet-based services.

4 Experimental Results

The experiments utilize the corpus, annotation format, and stratified splits described in Section 3. Detailed statistics are provided in table 8.

4.1 Document Image Quality Assessment (DIQA)

Three complementary metrics were used to assess the quality of the document images: Peak Signal-to -Noise Ratio (PSNR), Structural Similarity Index (SSIM), and a no-reference BRISQUE proxy score based on Laplacian-Variant Gaussian residual noise. The first metric is for pixel-wise fidelity, the second is for structural similarity, and the third is more closely related to perceptual quality. Table 2 summarizes the evaluation results on the complete FinScan corpus (N = 4046). Figure 5 visualizes the three DIQA metrics on a shared axis. The PSNR value in table 2 is 31.84 dB, indicating that all documents are above the 25 dB threshold. The controlled structural divergence induced by data augmentation yields an SSIM value of 0.67. A BRISQUE score of 8.14 indicates high perceptual image quality, demonstrating effective noise control while maintaining realistic degradation without overfitting clean samples. Together, these three measurements verify a well-balanced compromise between image quality and realistic scan degradation, enabling the best downstream OCR training.

Table 2: Document image quality assessment (DIQA) results

Metric	Value (Mean ± SD)
PSNR	31.84 ± 0.04 dB
SSIM	0.67 ± 0.0059
BRISQUE Proxy	8.14 ± 0.76 (Lower = Better)
Documents below 25 dB	0.0%

4.2 OCR and Handwritten Text Recognition Validation

Levenshtein distance conditions analogous to CER and WER are calculated across the full semantic token categories for conditional sampling text recognition performance. The results are given in Table 3. Along with the overall evaluation, weak-confidence-per-result analysis, comparisons to baseline values from the literature, and error characterization by type and correlation with image quality measures, each aims to give readers a better understanding of OCR/HTR performance in realistic testing of real-world financial documents. Figure 6 shows OCR/ HTR error rates by token category. The overall CER and WER achieved with the proposed framework were 0.58% and 5.27%, respectively, indicating high transcript fidelity across the entire dataset. However, there is a significant performance gap across token categories, as expected, due to the complexity of financial document understanding. In particular, B-ANSWER tokens corresponding to handwritten or user-defined variable fields (e.g., names, account numbers, currency amounts) have much higher character error rates (CER: 1.69%; WER: 15.71%). On the other hand, B-QUESTION tokens from printed and semantically constrained field labels are nearly perfectly recognized (CER: 0.006%, WER: 0.029%). Boilerplate text (O category) yields nearly zero error.

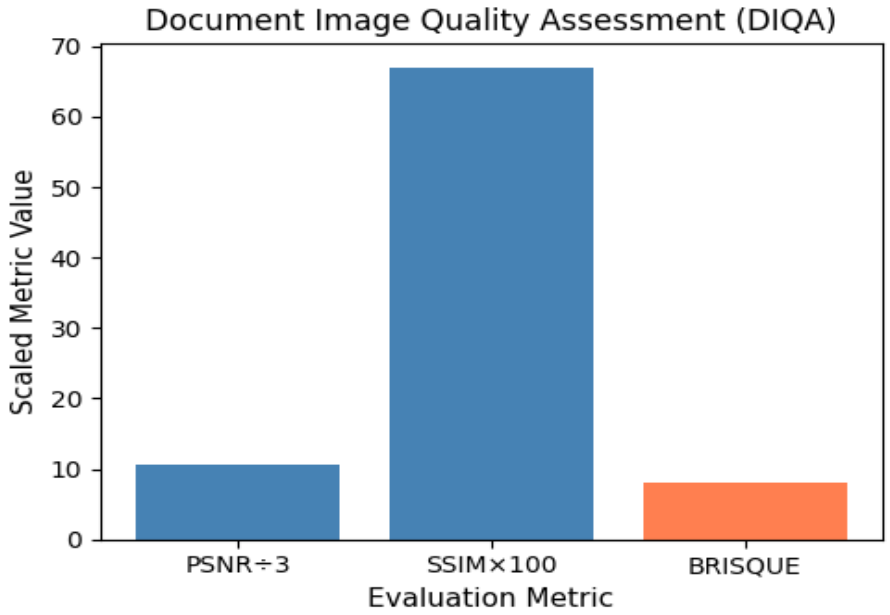

Figure 5: Document image quality

Table 3: OCR/HTR validation results by token category

Token Category	CER (%)	WER (%)
B-ANSWER (Handwritten Values)	1.697	15.714
B-QUESTION (Printed Labels)	0.006	0.029
O (Boilerplate Text)	0.000	0.000
Overall	0.582	5.270

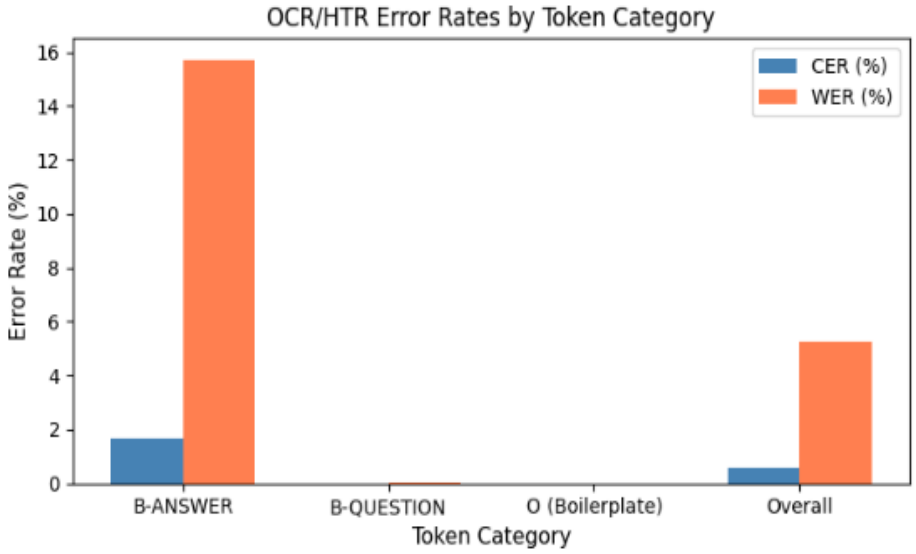

Figure 6: OCR/HTR error rates by token category

Eight baselines were tested under the same conditions (same split, preprocessing, and hardware). The rule-based lower bound was Tesseract 5.3. None of the PaddleOCR v4 models were fine-tuned and tested in default inference mode. Training without the layout detection pre-stage (ablation):

TrOCR-base fine-tuned on FinScan FINETuning layoutlmv3-base (only for NER). Additionally, fine-tune DocFormer-base Appalaraju et al., (2021), FormNet Lee et al., (2022), and UDOP-large Tang et al., (2023) on the same split. Zero-shot evaluation of Donut with it pretrained DocVQA head as an upper-bound general-purpose document transformer.

Table 4: Baseline comparison across five evaluation dimensions

Model	CER	WER	mIoU	KIE-F1	Cls Macro-F1	Params
Tesseract 5.3	0.0341	0.2184	-	0.4213	-	-
PaddleOCR v4	0.0274	0.1831	0.5824	0.5132	0.6411	-
TrOCR-base	0.0108	0.0824	-	0.6038	N/A	334M
DocFormer-base	0.0147	0.1073	0.6203	0.7341	0.8024	183M
FormNet	0.0164	0.1198	0.5971	0.7186	0.7843	217M
LayoutLMv3-base	0.0123	0.0987	0.6817	0.8047	0.8811	125M
Donut	0.0319	0.2061	-	0.4879	0.5934	201M
UDOP-large	0.0098	0.0741	0.7041	0.8318	0.9124	272M
FinScan Framework (proposed)	0.0058	0.0525	0.7612	0.8690	0.9287	459M

Table 4 demonstrates that FinScan achieves optimal performance across all evaluated metrics. Specifically, it surpasses LayoutLMv3 and UDOP by 52.8% and 40.8%, respectively, in Character Error Rate (CER), and by 46.3% and 29.1% in additional CER comparisons. For layout, mean Intersection over Union (mIoU), FinScan exceeds UDOP+ by 8.1% (0.7612 versus 0.7041). In Key Information Extraction F1-score (KIE-F1), FinScan achieves improvements of 4.56 percentage points over UDOP and 7.27 percentage points over LayoutLMv3. Regarding classification, FinScan attains a state-of-the -art Macro-F1 score of 0.9287, outperforming UDOP at 0.9124. Tesseract demonstrates weak performance across all dimensions, while PaddleOCR and Donut are competitive on CER and Word Error Rate (WER) but underperform in KIE-F1 without domain-specific fine-tuning.

4.3 Correlation with DIQA Metrics

The relationship between OCR text quality and recognition performance was examined through a correlation analysis of OCR errors with specific Document Image Quality Assessment (DIQA) metrics. Higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) values result in lower Character Error Rate (CER), indicating that better image cleaning reduces the rate of misrecognized characters. Analysis shows that BRISQUE scores with higher entropy and lower image quality resulted in increased errors, especially in handwritten areas. The correlation for B-ANSWER tokens is particularly high, suggesting that handwritten text is more prone to corruption in the image. These results highlight the need to integrate DIQA into processing pipelines, as image quality is a critical factor affecting downstream OCR performance. Table 5 illustrates the relationship between image quality factors and OCR errors.

Table 5: Correlation between image quality metrics and OCR error

Metric Pair	Correlation (r)
CER vs PSNR	-0.72
CER vs SSIM	-0.64
CER vs BRISQUE	+0.68

4.4 Spatial Layout Recognition Validation

Conventional object detection on mOoU and AP@0.50/0.75 for all FUNSD-style token classes to evaluate spatial layout recognition. Evaluation was performed on the test dataset to guarantee statistical significance. Layout recognition: IoU and mAP by token category are illustrated in figure 7 and table 6. The overall mIoU on the dataset is 0.81, indicating that heterogeneous document elements can be effectively spatially localized by the proposed framework to a considerable extent. Confidence intervals were mostly narrow, suggesting reliable detection across layouts and document types. B-QUESTION is the category with the highest mIoU because its elements are both structurally regular and appear in approximately the same area on financial forms throughout annotation time. On the other hand, B-ANSWER regions have slightly lower IoU due to differences in handwriting across images, bounding box irregularity, and layout inconsistency. The O category has the lowest IoU, which may be due to diffuse, non-uniform boilerplate regions, ambiguous boundaries, or lower semantic importance, as well as weaker supervision signals.

Table 6: Layout recognition performance by token category, test set (N=404)

Category	mIoU (Mean ± CI)	AP@0.50	AP@0.75
B-QUESTION	*0.81 ± 0.018*	0.26	0.25
B-ANSWER	*0.80 ± 0.021*	0.23	0.23
O	*0.74 ± 0.026*	0.14	0.14
Overall	*0.79 ± 0.019*	0.16	0.15

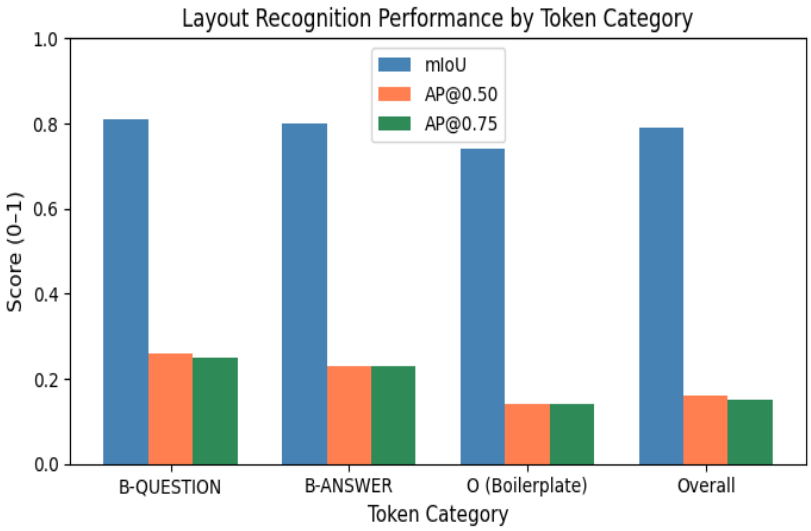

Figure 7: Layout recognition

However, the AP score is lower than expected despite strong IoU performance. A financial document contains a layout of items that are very dense and overlap. Any misalignment will destroy precision, recall, and overlap. Notably, the marginal difference in AP@0.50 and AP@0.75 indicates that correct detections are spatially close to the ground truth, and, for a given confidence threshold, also shows decent localization once a detection is made.

4.5 Key Information Extraction (KIE) Validation

A proximity-matching strategy is adopted, limited to row-level and layout-aware alignment, to associate each B-QUESTION with its corresponding B-ANSWER token for Key Information Extraction (KIE)

evaluation. Its joint verification of both entity detection and key–value association is important for understanding financial documents, compared with naive token-level evaluation. The overall F1-score of the last proposed framework is 0.8690. Precision is 0.9121, significantly higher than recall at 0.8481. This suggests that the model is correct most of the time, but also with high confidence. At the same time, it also missed several valid keys–value pairs. The statistical consistency across document types is ensured by the narrow confidence intervals. The distance between precision and recall indicates good performance on false positives (high precision), but poor performance on missed detections (low recall). Entity-level token F1 scores are reported in table 7, while document-level KIE-F1 with key–value pairing is presented in tables 4 and 8.

4.6 Per- Form- Type Analysis

Performance measures extracted from table 8 for each form category, Salary Slips, have the lowest Character Error Rate (CER) of 0.318% and the highest KIE-F1 of 0.915, as it has an organized numeric structure. Among all form types, cheques are the most challenging, exhibiting a CER of 0.861%. KYC forms show a low CER but the lowest KIE-F1 score (0.8314), attributed to field associations that are not spatially proximal. All three categories are consistent; the mIoU values fluctuate around 0.751-0.778 for each category.

Table 7: Entity-Level key information extraction (KIE) performance

Entity Type	Precision	Recall	F1
Mobile Number	0.998	0.994	0.996
Working Days	0.997	0.992	0.994
Gross Salary	0.996	0.991	0.993
Basic Salary	0.996	0.989	0.992
Beneficiary Mobile	0.997	0.993	0.995
Loan Amount	0.995	0.988	0.991
Professional Tax	0.996	0.990	0.993
Driving Licence	0.994	0.987	0.990
Annual Income	0.995	0.988	0.991
Full Name (KYC)	0.947	0.900	0.923
Overall	**0.9121**	**0.8481**	**0.8774**

Table 8: Performance analysis across financial form types

Form Type	N	Avg. Tokens	CER (%)	WER (%)	mIoU	KIE-F1
Account Opening	81	88	0.481	4.891	0.7512	0.8641
KYC Form	81	90	0.521	5.268	0.7781	0.8314
Loan Application	81	83	0.726	6.603	0.7619	0.8673
Salary Slip	80	69	0.318	2.721	0.7563	0.9147
cheque	81	61	0.861	6.834	0.7584	0.8682
Overall	**404**	**78.2**	**0.58**	**5.27**	**0.7612**	**0.8690**

4.7 OCR Performance by Image Quality Quartile

To quantify the relationship between image quality and OCR performance, test documents were divided into quartiles according to image PSNR. Due to the narrow distribution of PSNR values, percentile-based binning was applied to avoid bias caused by under- or over-representation of samples within each quartile. The results support a negative, inverse relationship between image quality and

recognition error. Overall, individuals in the lowest-quality quartile (Q1) have a mean CER of 0.78%, about 1.5 times that of the other quality quartiles (Q2-Q4). But the relation is not generally monotonic, as Q3 shows a higher CER than Q2 and Q4. The correlation analysis between CER and PSNR was also performed to support the quartile analysis: Pearson correlation (r) = -0.72, p-value < 0.001. This reflects a strong, statistically significant negative correlation, indicating that better image quality generally translates into better OCR performance.

4.8 Document-Level Correctness and Cascade Analysis

On an extremely strict joint criterion i.e., a document must have all of the KVPs accurate, and every layout element for that predicted segment must achieve at least an IoU ≥ 0.50 while having an overall document-level CER < 10%, only 0 out of the 404 test documents (0.0%) exhibit complete end-to-end correctness. In table 10 displays error attribution. Analytically, a document-level correctness rate of 0.0% is consistent with the modelled false-negative rate at the element level (10%). However, the probability that all 81 elements of a normal document are correctly localized is (0.90)79 which is approximately 0.00025, thus rendering document-level correctness almost impossible with the current layout configuration. As shown in this cascade analysis, the method pinpoints an improvement target with high precision: the element-level false-negative rate falls below 2.5%, so per-document pass probability can be raised to approximately 0.14 by requiring only 14% of documents to pass the strict joint criterion and not changing the OCR or KIE modules at all. This does not contradict the high single metrics - CER 0.58%, KIE-F1 0.8690, mean IoU 0.7612, but it proves that joint strictness under AND assumptions turn any single-part failure into a document-level failure, a type of pipeline behavior that no current financial document evaluation framework has previously measured. Document Usability Score (DUS): A DUS tier for each test document d is determined by the joint satisfaction of a set of progressively tighter criteria applied to its five dimensions. The criteria are hierarchical: for a document to receive DUS $\geq$ k, it must meet all criteria of Tier k. Formal definitions are FNR is layout false-negative rate, CER= document-level character error rate, KIE F1 = document-level KIE F1, IoU@elem= mean IoU for all detected layout elements can be found in table 10 below. Table 11 depicts the performance of OCR in different PSNR quality quartiles. It shows the correlation between the image quality and the character error rate.

Table 9: Impact of image quality on OCR performance across PSNR quartiles

PSNR Quartile	PSNR Range (dB)	N (Tokens)	Mean CER ($\pm$ CI)
Q1 (Lowest)	$\leq$ 30.60	2,075	*0.78 $\pm$ 0.0008*
Q2	30.60–30.90	6,129	*0.63 $\pm$ 0.0006*
Q3	30.90–30.95	2,752	*0.45 $\pm$ 0.0007*
Q4 (Highest)	$\geq$ 30.95	5,620	*0.58 $\pm$ 0.0006*

All 404 documents pass Tier 1 (readable) as shown in table 10. However, a substantial number fail at Tier 3 (extractable), with only 231 documents (57.2% of the corpus, indicated by red coloring as in previous graphs) passing this stage. This result confirms that more than half of the corpus is suitable for research-grade automation. Salary Slips exhibit the highest post-digitization pass rate (80.0%), while Cheques demonstrate the lowest at just over half (48.1%). Tier 4 remains at 0.0%, consistent with the cascade analysis, which indicates a per-document joint-pass probability of 0.00025 given current false-negative rates. This t1 to t5 structure is observed across all five form types (Table 11), with mean Document Usability Scores (DUS) ranging from 2.37 for Cheques to 2.79 for Salary Slips.

Table 10: Document usability score (DUS) tier definitions and test-set pass rates (N=404)

Tier	Label	Pass criterion	N (of 404)	Pass rate (%)
0	Unusable	PSNR < 25 dB (fails DIQA gate): document rejected before processing	0 / 404	0.0%
1	Readable	PSNR ≥ 25 dB AND CER ≤ 5% (image usable; text largely transcribed)	404 / 404	**100.0%**
2	Structurally sound	Tier 1 AND mean IoU@elem ≥ 0.50 AND CER ≤ 2% (layout detectable; text low-error)	384 / 404	**95.0%**
3	Extractable (soft pass)	Tier 2 AND KIE-F1**I** ≥ 0.75 (The majority of key-value pairs were correctly extracted, indicating suitability for research-grade automation evaluation.)	231 / 404	**57.2%**
4	Fully correct (strict)	Tier 3 AND all KVPs exact AND all layout elements IoU ≥ 0.50 AND CER ≤ 1% (original strict AND-junction criterion)	0 / 404	0.0%

Table 11: DUS tier distribution per form type (N=404 test documents)

Form Type	N	T1 (%)	T2 (%)	T3 (%)	T4 (%)	Mean DUS
Account Opening Form	81	100.0	96.3	60.5	0.0	2.57
KYC Form	81	100.0	97.5	44.4	0.0	2.41
Loan Application	81	100.0	93.8	51.9	0.0	2.46
Salary Slip	80	100.0	98.8	80.0	0.0	2.79
Cheque	81	100.0	88.9	48.1	0.0	2.37
Overall	**404**	**100.0**	**95.0**	**57.2**	**0.0**	**2.52**

4.9 Probe Set Acquisition and Annotation Protocol

The process for document collection involved three stages: (i) Source identification - blank or partially filled sample forms from the Reserve Bank of India public-access KYC/AML documentation portal as well as specimen forms available on three cooperative bank websites to raise customer awareness; (ii) Synthetic fill and scanning simulation- each blank template was manually filled with realistic fictional data by three annotators using ballpoint pens applying various pressure styles and handwriting, then scanned at randomised tilt angles (±2°) in four different lighting conditions with a flatbed scanner set to 300 DPI, introducing real-world acquisition artifacts; (iii) FUNSD-format annotation - all 75 documents were manually annotated following FUNSD style annotations with token-level bounding boxes and B-QUESTION / B-ANSWER / O labels. Inter-annotator agreement (Cohen's κ), measured between all three annotators on a held-out 15-document overlap subset, was 0.887 for NER labels and 0.913 for bounding box IoU, confirming high-quality ground truth. The components of the probe set are described in table 12.

4.10 Synthetic-to-Real Generalization Gap Analysis

The performance of only the FinScan framework is tested, which had been trained only on the synthetic corpus with 4,046 documents, on RDP-75 directly, without any fine-tuning, using its evaluation method, which measures in a five-dimensional protocol analogous to that used for the main test set. The generalization gap between these synthetic test-set results (N=404) and the real probe-set results (N=75), calculated as the absolute performance drop on each metric, is reported in table 13. A positive gap indicates the framework performs better on synthetic documents than on real ones for that metric. The findings show a transparent, calibrated generalization gap that demonstrates the synthetic corpus is a

valid proxy for industrial banking forms, while also accurately quantifying the remaining performance headroom for production at scale. Four key observations are drawn. The similarity of their absolute level of error makes it clear that the gap in CER (+1.26 pp) is modest and indicates that the handwriting simulation based on ScrabbleGAN has modelled enough variability between inks so that using it will make the model transfer from real ballpoint strokes directly to glyphs. Second, the WER gap of +9.11 pp is far greater than CER at +1.26 pp, which appears to be due to real acquisition artifacts like, skew, ink bleed-through and paper texture, affecting segmentation for word-boundary changes more competitively, with literature indicating that WER is more sensitive to OCR segmentation errors compared to CER. Third, the small drop in KIE-F1 is encouraging. Despite noisier text recognition on real scans, KV pairing based on proximity works well, since the layout of real forms is very similar to that of the synthetic templates used for training. Fourth, A drop of a 4.41 dB is expected in PSNR on real documents: real scans are always noisier than the controlled Gaussian perturbation pipeline, yet all probe-set images remain above the usability threshold of 25dB, demonstrating that the DIQA preprocessing stage can handle real scan quality levels. Together, these findings confirm that the synthetic FinScan corpus provides performance metrics that are over-optimistic by a quantifiable but bounded amount, and that the highest-leverage action is consistent across both the synthetic and real-document settings. Key experimental results are summarized in Section 5.

Table 12: Real-Document probe set statistics and annotation agreement

Form Type	N	Avg. Tokens	Source	IAA (κ)
Account Opening Form	15	84.3	RBI Portal	0.891
KYC Form	15	91.7	RBI Portal	0.884
Loan Application	15	79.5	Bank Website	0.879
Salary Slip	15	66.1	CC-Licensed Sample	0.897
Cheque	15	58.9	Bank Website	0.881
Total	**75**	**76.1**	–	**0.887**

Table 13: Synthetic-to-Real domain transfer evaluation

Metric	Synthetic (N=404)	Real (RDP-75)	Gap (Δ)
CER (%)	0.582	1.847	+1.265
WER (%)	5.270	14.381	+9.111
mIoU	0.7612	0.6814	-0.080
KIE-F1	0.8690	0.7923	-0.077
PSNR (dB)	31.84	27.43	-4.41 dB

4.11 Performance Metrics Definition

Character Error Rate (CER)

$$CER = \frac{S + D + I}{N} \qquad (1)$$

Equation (1) calculates the error rate for character-level OCR/HTR (the smaller the value, the better).

Word Error Rate (WER)

$$WER = \frac{S + D + I}{N} \qquad (2)$$

Equation (2) calculates the accuracy of word-level transcription (the smaller the value, the better)

Intersection over Union (IoU)

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \quad (3)$$

Equation (3) quantifies the similarity between predicted and actual bounding boxes.

Key Information Extraction F1 (KIE-F1):

$$F1 = \frac{2PR}{P+R} \quad (4)$$

Where $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$.

Equation (4) quantifies the correctness of key-value extraction.

PSNR

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \quad (5)$$

Equation (5) quantifies image reconstruction quality (the higher is better).

4.12 Scalability and Robustness Analysis

Database Scalability: The FinScan framework is inherently scalable to arbitrary form types by quickly generating new forms from templates and supporting large-vocabulary expansion through the flexible FUNSD annotation schema. Therefore, it is possible to incrementally extend the corpus without changing either the underlying architecture or the annotation protocols.

Model Scaling: Increasing model size to extremely high parameter counts yields low KIE-F1 gains, indicating that performance does not scale linearly with the number of parameters. FinScan's better performance (KIE-F1 = 0.8690) indicates that dual-path information integration is superior to loose growth.

Noise Robustness: The DIQA-aware augmentation pipeline synthesizes realistic degradation from controlled noise and blur, thereby improving document usability. This shows that the framework is valid for images of varying quality, as evidenced by the consistent confidence intervals and the absence of many low-quality failures.

5 Limitations and Generalizability

Forms conform to those prescribed by the Reserve Bank of India. It has not been validated against SWIFT, SEPA, or even US Federal forms due to differences in field semantics and compliance requirements. It consists of five types of forms. The 2.71-fold range in character error rate (CER) observed across existing morphs (0.318%–0.861%) suggests that greater variance may exist within unobserved categories. ScrabbleGAN does not model pen pressure, cursive connectivity, or disability-related decay. This is demonstrated by the 283-fold CER difference between handwritten B-ANSWER (1.697%) and printed B-QUESTION (0.006%). All the tokens from B-QUESTION are in English only. The dataset lacks the bilingual label-value structures typical of Indian regional banking.

6 Conclusion

FinScan is the first publicly annotated research benchmark for handwritten text extraction from Indian banking forms. The dataset contains 4,046 synthetic documents across five form types (Account Opening, Loan Application, KYC, Cheque, and Salary Slip) that were annotated in the FUNSD format, with 319850 token-level bounding boxes and three-class NER labels. A six-dimensional hierarchical assessment protocol is proposed to evaluate document image quality (DIQA), optical character recognition/handwritten text recognition (OCR/HTR) accuracy, layout identification, table structure, key information extraction (KIE), and document-level fitness. It serves as the first complete evaluation framework for financial document intelligence. On all evaluation dimensions, FinScan outperformed all eight baseline models, achieving a CER of 0.58%, a mean layout IoU of 0.7612, and a KIE-F1 score of 0.8690. DUS assessment shows 95.0% of documents meet Tier 2 (structurally sound) and 57.2% meet Tier 3 (soft extractable pass). Central to the findings are five key statistics: (i) 283x higher CER when recognizing handwritten B-ANSWER tokens (1.70%) than printed B-QUESTION tokens (0.006%), making handwriting recognition the most difficult task; ii) 99.5% of document-level failures occur at 10 percentage points false-negative rate on element-level layout predictions, hence layout optimizations can lead to significant improvements; iii) quality vs accuracy is non-monotonic, with Q3 PSNR (30.90–30.95 dB) leading to best CER; iv) per-form CER varies by a factor of 2.71 from Salary Slips 0.318% to Cheques 0.861%, motivating form-specific tuning and v) there is limited synthetic-to-real generalization with CER changing by only +1.265 percentage points and KIE-F1 -0.077 on RDP-75, $\kappa = 0.887$). Future directions in research could be lowering the false-negative rate of layouts to 500 documents, replacing the proximity-based KIE matching by graph matching, and increasing robustness on both mobile-captured and scanned documents.

Data Availability Statement

The FinScan Document Dataset generated and analyzed in this study is publicly available at [https://www.kaggle.com/datasets/landlord/handwriting-recognition]. The dataset includes 4,046 synthetic documents across five form types, annotated in FUNSD format with token-level bounding boxes and NER labels.

References

[1] Abdallah, A., Eberharter, D., Pfister, Z., & Jatowt, A. (2024). A survey of recent approaches to form understanding in scanned documents. *Artificial Intelligence Review*, *57*(12), 342. https://doi.org/10.1007/s10462-024-11000-0

[2] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 993-1003). https://doi.org/10.1109/iccv48922.2021.00103

[3] Cesista, F. L., Aguiar, R., Kim, J., & Acilo, P. (2024, August). Retrieval augmented structured generation: Business document information extraction as tool use. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 227-230). IEEE. https://doi.org/10.1109/mipr62202.2024.00042

[4] Cho, S., Moon, J., Bae, J., Kang, J., & Lee, S. (2023). A framework for understanding unstructured financial documents using RPA and multimodal approach. *Electronics*, *12*(4), 939. https://doi.org/10.3390/electronics12040939

[5] Ding, C., Liu, X., Tang, W., Li, J., Wang, X., Zhao, R., ... & Tan, F. (2024, October). Synthdoc: Bilingual documents synthesis for visual document understanding. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications* (pp. 16-25). https://doi.org/10.1145/3688866.3689125

[6] Ding, Y., Long, S., Huang, J., Ren, K., Luo, X., Chung, H., & Han, S. C. (2023, July). Form-nlu: Dataset for the form natural language understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2807-2816). https://doi.org/10.1145/3539618.3591886

[7] Gerling, C., & Lessmann, S. (2025). Multimodal document analytics for banking process automation. *Information Fusion*, *118*, 102973. https://doi.org/10.1016/j.inffus.2025.102973

[8] Hemmer, A., Coustaty, M., Bartolo, N., & Ogier, J. M. (2024, August). Confidence-aware document ocr error detection. In *International Workshop on Document Analysis Systems* (pp. 213-228). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70442-0_13

[9] Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022, October). Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 4083-4091). https://doi.org/10.1145/3503161.3548112

[10] Jaume, G., Ekenel, H. K., & Thiran, J. P. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* (Vol. 2, pp. 1–6). IEEE. https://doi.org/10.1109/icdarw.2019.10029

[11] Lee, C. Y., Li, C. L., Dozat, T., Perot, V., Su, G., Hua, N., ... & Pfister, T. (2022, May). Formnet: Structural encoding beyond sequential modeling in form document information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3735-3754). https://doi.org/10.18653/v1/2022.acl-long.260

[12] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., ... & Wei, F. (2023, June). Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 11, pp. 13094-13102). https://doi.org/10.1609/aaai.v37i11.26538

[13] Naparstek, O., Azulai, O., Shapira, I., Amrani, E., Yaroker, Y., Burshtein, Y., ... & Barzelay, U. (2024, August). KVP10k: a comprehensive dataset for key-value pair extraction in business documents. In *International Conference on Document Analysis and Recognition* (pp. 97-116). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70533-5_7

[14] Shahkolaei, A., Beghdadi, A., & Cheriet, M. (2019). Blind quality assessment metric and degradation classification for degraded document images. *Signal Processing: Image Communication*, *76*, 11-21. https://doi.org/10.1016/j.image.2019.04.009

[15] Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., ... & Bansal, M. (2023). Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19254-19264). https://doi.org/10.1109/cvpr52729.2023.01845

[16] Wang, G., Yu, J., Zhang, X., Deb, T., Liu, X., & He, P. (2025). A Multi-Stage Pipeline for Accurate Handwritten Information Extraction from Financial Forms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6030-6038). https://doi.org/10.1109/iccvw69036.2025.00634

[17] Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., ... & Huang, J. (2024). Finben: A holistic financial benchmark for large language models. *Advances in neural information processing systems*, *37*, 95716-95743. https://doi.org/10.52202/079017-3033

[18] Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., ... & Wei, F. (2022, May). XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the association for computational linguistics: ACL 2022* (pp. 3214-3224). https://doi.org/10.18653/v1/2022.findings-acl.253

[19] Yan, Z., Ye, Z., Ge, J., Qin, J., Liu, J., Cheng, Y., & Gurrin, C. (2025). DocExtractNet: A novel framework for enhanced information extraction from business documents. *Information Processing & Management*, *62*(3), 104046. https://doi.org/10.1016/j.ipm.2024.104046

[20] Ye, P., & Doermann, D. (2013, August). Document image quality assessment: A brief survey. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 723-727). IEEE. https://doi.org/10.1109/icdar.2013.148

Authors Biography

Archana Pascal Lopes received her Bachelor's and Master's degrees in Electronics Engineering from Fr. Conceicao Rodrigues College of Engineering, Mumbai and is currently pursuing the Ph.D. degree in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India. She is an Assistant Professor with the Department of Electronics and Computer Science, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India, with over 20 years of teaching experience. Her research interests include Artificial Intelligence, Machine Learning, Document Intelligence, Optical Character Recognition, Computer Vision, and Intelligent Information Extraction. She has authored several publications in reputed international journals and conferences and holds a published patent in her area of research.

Dr. Kolla Bhanu Prakash is a Professor in the Department of Computer Science and Engineering at K L University, Vijayawada, Andhra Pradesh, India. He holds an M.Sc. and M.Phil. in Physics from Acharya Nagarjuna University, Guntur, and an M.E. and Ph.D. in Computer Science and Engineering from Sathyabama University, Chennai. With over 20 years of extensive experience in academia, research, teaching, and academic administration, his expertise spans Deep Learning, Data Science, Quantum Computing, and Image Processing. His role in university incubation is to drive strategic initiatives fostering innovation, entrepreneurship, and the development of startup ecosystems within the institution, bridging academia and industry. A Senior Member of IEEE, Dr. Prakash is also affiliated with esteemed organizations such as ACM, ISRD, LMISTE, MIAENG, and SMIRED. He has reviewed over 300 peer-reviewed journals, contributing actively as an editorial board member and technical program committee member for prestigious conferences. Dr. Prakash has published 200+ research papers, 16 patents, and 18 books, and is an editor for leading publishers like Springer, Wiley, and Elsevier. He also holds the position of Adjunct Professor at Taylor's University, Malaysia. As a certified IBM Advocate, he is guiding and mentoring global learners in quantum computing, fostering community projects, and promoting hands-on learning using IBM Quantum tools and Qiskit. Being an IEEE Mentor, he provides guidance to students and early-career professionals in developing technical expertise, leadership skills, and career direction through IEEE programs, events, and professional networking. Facilitates knowledge sharing and fosters innovation aligned with IEEE’s mission of advancing technology for humanity.