International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

SMARTFORM: Intelligent OCR System for Digital Banking Automation

C Sona Vijayan[1], Dr .Madhumita K[2]

[1,2]Department Of Computing Technologies
SRM Institute of Science and technology

Abstract

This paper describes SMARTFORM, a web portal-based tool for automating bank form processing with Optical Character Recognition (OCR). The system includes a Flask backend that allows secure login/registration, digital form submission and automatic handwriting/print text extraction using the Mistral AI OCR API. Data is extracted, verified using a PIN based scheme before it is integrated to guarantee accuracy and security. We compare traditional OCR models like Tesseract and transformer-based models like Donut and LayoutLMv3 on structured and semi- structured documents. Performance is evaluated through Character Error Rate (CER), Word Error Rate (WER), F1-score, BLEU and qualitative visualization. Results show that while Tesseract's performance with pre-processing is good, transformer models are able to perform better context and entity recognition but are prone to overfitting. Overall, the trade-offs between classical and modern OCR are demonstrated with hybrid methods that incorporate adaptive pre- processing providing an optimal trade-off between accuracy, generalization, and computational efficiency for digital banking automation tasks.

Keywords: Optical Character Recognition, Digital Banking, SMARTFORM, Transformer-based OCR, Tesseract, Hybrid Approach, Automated Form Processing

1. INTRODUCTION

Automating document processing is an essential requirement in the current era of fast paced digital era among the banking, health care, and insurance companies. This is a fully featured solution that is built on the recent findings in Optical Character Recognition (OCR) and an automatic form processing web portal built using React. The application allows the users to log in, enter or complete forms online and track their submissions. The backend designed using Flask consumes the Mistral AI OCR API, parsing the available data through advanced OCR models that consist of both traditional OCR engines, such as Tesseract, and transformer-based models, such as Donut and LayoutLMv3, capturing the data on printed and handwritten materials. The system identifies bankpersonnel in a PIN-protected and secure verification of the transaction information. The goal of this strategy is to reduce the number of manual data entry, eliminate errors and speed up the processing of documents, which will drive efficiency and user experience. The system is system-centred and is concerned with scalability, flexibility and automation within the banking sector. Most conventional offerings operate on a fixed template, but SMARTFORM has the ability to work with different layouts of documents. This is perfect when it

comes to dealing with the loan application forms, customer onboarding documents and KYC files. It also has secure verification as a built-in feature, to meet standards and retain customer trust. The system with modern OCR technologies and web-based access eases the work of the bank within its premises, as well as fulfils the requirements of customers who receive responses much faster. This would be the way OCR-based platforms may revolutionize to transform conventional banking system to a highly secured, faster digital experience.

2. RELATED WORK

OCR is a basic technology used in automation of document processing in the banking, healthcare as well as insurance industries. Conventional engines fail with variable templates, handwriting or noisy scans, whereas engines such as Tesseract:[1] Perform poorly on unstructured layouts. Such pre- processing techniques as denoising and adaptive thresholding can achieve only a slight improvement and are very dataset-specific [2] which is why more general approaches are sought. Transformer-based models stacked on top of each other with textual, visual, and spatial features are currently highlighted. The Document Understanding Transformer (Donut)
[4] shows that the OCR pipeline can be skipped as an attempt at document understanding can be framed as a vision-to-text task, which is more coherent semantically and more multilingual. On the one hand, LayoutLMv3 [5] is a multimodal pretrained model that provides state-of-the-art named entity recognition, but may overfit on domain-specific tasks. Even lightweight systems like TrOCR [6] demonstrate considerable better performances than common engines on handwritten text recognition. In a compromise between precision and performance,

scholars have been experimenting with combination techniques that are based on parallelization, Tesseract and transformers [7]. The methods are used to improve contextual extraction and still provide viable inference times. More modern unified models, e.g., OCR-2.0 [8], do so by considering text, tables, and figures in the same stream. In sum, transformer-based models are proving useful to the contextual understanding but conventional OCR continues to play a useful role in regular layouts, and hybrid systems currently provide the best trade-off in practice.

3. PROBLEM DEFINITION

The theme that this paper presents the solution to is the processing of structured and semi-structured documents in the sphere of banking, healthcare and insurance, and the automatic processing and extraction of the data that documents presuppose. Manually converted documents are repetitive, unproductive, monotonous and prone to errors. Conventional OCR practices work well on conventional surfaces, but they cannot work when there are inconsistencies in layout and when scans are low, or there are other sophisticated features like handwritten scripts, checkboxes and tables. The impact of these weaknesses is partial or erroneous data capture that puts staffs involved in rechecking data and creating snags in providing the services. Thus, the research under analysis is designed in such a way to achieve the following goal:

- Find a way so that, any type of documents can be turned to digital structured documents with a minimal interference.
- Advance deep learning and intelligent pre- processing OCR models to learn and apply to improved recognition.

- Build OCR models based on transformers into one system with a convenient web interface and secure, encrypted connection.
- Make sure that it can be tuned to a variety of document formats and layouts so that the system can be scaled to a real enterprise usage.
- Maximize performance, thereby reducing the efficiency/accuracy trade-off among existing OCR solutions to build more robust digital processes.

4. DATASET GENERATION

Quality and variety of the dataset are the fundamentals of any OCR-based document processing application. Banking and insurancedocuments are complex because the layouts vary, documents may be in hand-written or mixed-written form, noisy scans and the format may be PDF, JPEG etc. Thus, the generation of a dataset does not only imply raw collection of forms but rather the creation of a disclosed, tagged and clinically relevant dataset to benchmark the OCR models.

a. Gathering, and Sources

Our data source uses a mixture of published benchmarks and institutional acquired questionnaires:
-FUNSD (Form comprehension on scanned documents with noise): 199 scanned documents with forms (iraqs), annotated at entity level (question, answer, other etc.).
-Banking/ Insurance Forms: Handwritten and printed, manually collected application forms, KYC documents, and claim forms in both printed and hand written forms.
-Synthetic Data: Auto-generated documents (React
+ LaTeX-based template-generating) with some of their information randomized about both user and transaction. This ensures multi-source diversity, and make the model robust to layout, resolution, and handwriting characteristic variations.

b. Labels and Classification Names

Each document has entity- and field-level annotations, and ground truth labels are provided as a JSON/XML metadata.Romanisable Key annotation categories include:

- Personal Data: Full Name, Date of birth, Address, Phone, Email.
- Banking/Transactional Data: Account Number, IFSC, Amount, Signature.
- Document metadata Document header, checkboxes and tables.
- Cross-verification of annotations was done by two human reviewers to make them accurate and consistent. These annotations are provided to benchmark the models based on the following metrics, Character Error Rate (CER), Word Error Rate (WER), F1-score, and BLEU score.

c. Cleaning, Interpolation and Pre-processing

There is a broad range in the quality of raw scanned forms. To standardize:

- Normalization: normalization was applied to all images (224 × 224 to transformer models; 150 ×

150 to baseline CNN-based OCR).

- Image denoising and reattribution: The images that were too blurred, obstructed, or had parts of themselves missing, get discarded. Adaptive denoising, Adaptive thresholding, were used to produce enhanced clarity of text.
- Contrast and Illumination Adjustment: Macroscopic differences that arise as a result of using different scanners or lighting conditions were reduced via histogram normalization and equalization.

d. Handling of Class/Layout Imbalance

The names of some field types (e.g., “Signature” or “Account Number”) are found much less commonly than generic field names such as “Name” and “Address.” To offset this unbalance:

- Data Augmentation: Perturbation rotation, skewing, and addition of white Gaussian noise, and different fonts of handwriting simulated real world deformations.
- Synthetic Oversampling: Uncommon fields replicated artificially with the use of programmatically created forms.
- Weights: Applied in training to punish incorrect species on rare yet important species (e.g. signatures, checkboxes).

e. Hierarchical Partitioning

On the aspect of fair evaluation, the dataset was split in terms of stratified sampling as follows:

- Training Set(70 percent): Various combination of layouts and handwriting/print styles.
- Validation Set (15%): Equalist sample to tune over hyperparameters and select model.
- Validation Set (15%): Unseen layouts, layouts that were not shown in the training process, to test the real-world performance.
- This ensures that all the field types are represented in different splits, and does not bias performance assessment.

f. Importance of Dataset Generation

The dataset that is systematically prepared does not only allow the OCR models to identify clean printed text but also generalize on handwritten and noisy / poorly scanned documents usually found in real banking and insurance workflows. This will make the developed system to be able to be launched inlive institutional setups, and not only optimized under laboratory conditions

5. METHODOLOGY

A combination of pre-processing strategy, a model benchmarking strategy, a hybrid pipeline, and a secure deployment strategy was used systematically to create the proposed OCR-powered digital banking solution.

A. Input Data Handling

Banking, healthcare, and insurance forms have been gathered, in both a printed and hand-written form. Pre-processing was done to enhance the quality of inputs and the compatibility of models, the following was done:

- Since all images have different sizes and they may have different sizes they must be resized and normalized.
- Skew correction and contrast Enhancements.
- Denoising with filters and adaptive segmentation.
- Basic distortion (rotation, scaling) to make more resistant.

B. Recognition Models

Three types of OCR methodology were tested:

- Tesseract OCR - a baseline traditional engine, which was tested under various pre-processing sets of parameters.
- Donut - an encoder-decoder model based on transformer that is able to obtain structured outputs without an additional OCR step.
- Lm-v3 a-multimodal transformer model incorporating text, layout and image information to enable entity recognition.

C. Hybrid OCR Pipeline

A hybrid pipeline was built to counter off the speed and the precision:

- Tesseract clean printed forms
- Donut and LayoutLMv3 processed noisy or handwritten material.

Results were combined to provide better reliability with decision layers.
Attention maps, bounding boxes and the visual requirements of interpretability were employed.

D. Architecture of the Solution

The solution was implemented in three-tier structure:

- Frontend: React portal that supports the log in, registration and uploading of forms.
- Back End: Flask APIs to call Mistral AI OCR and model inference.
- Database: PostgreSQL in order to use encryption.
- Security was provided with JWT authentication of users and PIN validation of banking

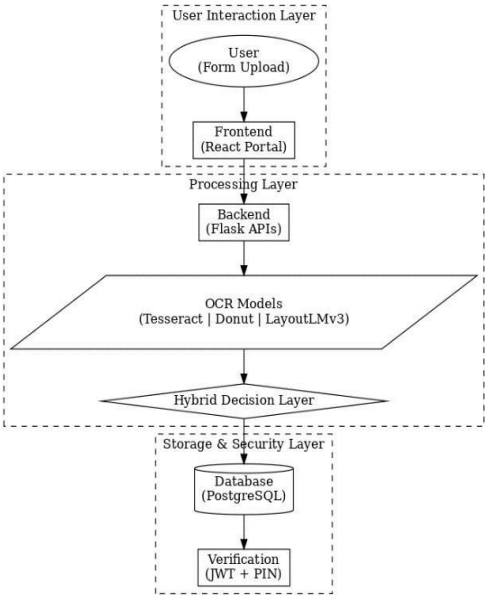

Fig 5.1 Architecture and Functional Components

E. Reviewing Measures

Performance measured was based on the use of performance criteria as one of the judges.

- Accuracy of Transcription in terms of Character Error Rate (CER) and Word Error Rate (WER).
- Accuracy, Precision, Recall, and F1-score to measure entities at the level of evaluation.

- BLEU and similarity scores to the semantic accuracy.
- Qualitative analysis of confusion matrices and attention maps.
- Furthermore, top-to-bottom systems testing assured at the time of less than 2 seconds per page and a high frequency of user contentment.

6. ENCODER

The encoder is aimed at transforming documents online images or copies into latent representations. Within the recommended system, various encoder

approaches were used according to the OCR model:

- The Tesseract Encoder gives character level segmentation to the processed forms and codes it in the form of sequential features.
- Donut Encoder has a Swin Transformer backbone and captures both local and global patterns in the document to produce high dimensional feature embeddings.
- The textual, positional, and visual patches are used together by LayoutLMv3 Encoder to develop the multimodal embeddings which could retain the spatial structure of informs.
- The encoder step is such that heterogeneous banking forms, either handwritten or printed materials are converted into meaningful forms to be decomposed.

7. DELIBERATION DECODER

The deliberation decoder re-decodes the OCR predictions on the first pass, by re-visiting decoding along the path of uncertainty, and applying domain- specific constraints. It runs on the outputs of Tesseract, Donut, and LayoutLMv3 to calibrate and reliability-test derived fields appropriate to banking processes.

- Inputs:

-Text sequences with token confidences;
-entity spans with bounding boxes;
-heat maps from transformer models.
-model-wise scores ptess , pdonut ,plmv3.

- Uncertainty screening: Find uncertain tokens/spans (e.g. with entropy or score less than tau).
- Selective re-read: Re-crop ambiguous areas (boxes/lines) then re-run stronger model on that case (e.g. Donut on handwritten, Tesseract on clean print).
- Span reconciliation: Matching overlapping spans across models; Using the majority vote on text and IoU-based box combining for location.
- Semantic checks: Verify fields according to banking rules- e.g. date/amount formats, account/ID regex, checksum like behavior and label proximity to the value.
- Model scoring: Provide weighted model scores (learned/Canonical weights)

$S = w1\,p\text{tess} + w2\,p\text{donut} + w3\,p\text{lmv3}$,

This would be followed by temperature/ threshold calibration to minimize over-confidence.

Consistency check: Repeated fields (e.g., the identical ID repeated in two locations) should be checked against one another and the mismatch should be handled by choosing a higher calibrated result.

- Human-in-the-loop trigger: When final confidence is below 0,.R.", flag the field to require staff PIN verification in the portal.
- Outputs: A hierarchical organization structure of fields (key, value, confidence, box); and an explanation bundle (attention overlays and before/ after snippets) to audit the explanation bundle.
- Effect: Deliberation reduces false positives on sensitive areas, and improves recall in rare classes, by re-reading problem regions selectively and locally imposing domain constraints, without incurring the latency cost of full-page reprocessing.

8. RESULT

The proposed system was evaluated on a mixed dataset comprising handwritten and printed banking/insurance forms. The performance of traditional and transformer-based OCR models was compared in terms of Character Error Rate (CER), Word Error Rate (WER), F1-score, and BLEU score.

A. Quantitative Results

Table I presents the comparison of the three OCR models. It can be observed that transformer-based models significantly outperform the traditional Tesseract engine on semantic accuracy, while optimized Tesseract performs better on clean printed documents.

Table I. Performance Comparison of OCR Models

Model	CER ↓	WER ↓	F1 ↑	BLEU ↑
Tesseract (Default)	0.651	0.234	0.547	0.52
Tesseract (Optimized)	0.598	0.156	0.444	0.50
Donut (Transformer)	0.432	0.310	0.710	0.82
LayoutLMv3 (Transformer)	0.410	0.298	0.695	0.79

B. Graphical Analysis

Figure 1 shows the CER and WER achieved in the colorful models. It's also clear that Tesseract will have advanced character and word error rates than motor- grounded models.

A comparison of the F1- scores used in Fig. 2 indicates that Donut is better at reality- position recognition.

3 shows the BLEU score, which indicates semantic consonance of mills.

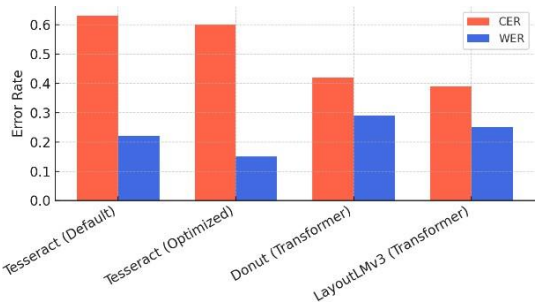

Fig1:CER and WER Comparison of OCR Models

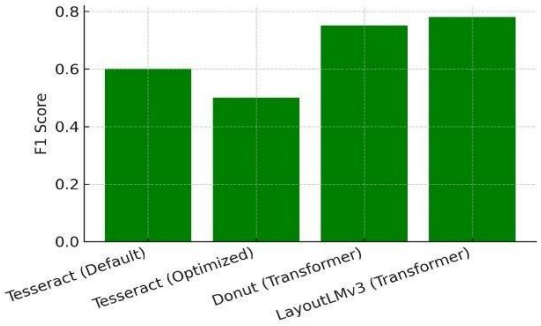

Fig 2: F1-Score comparison of OCR Models

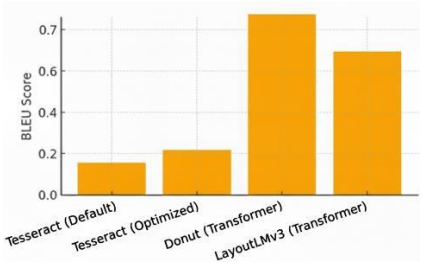

Fig 3: BLEU Score Comparison of OCR Models

C. Discussion

Based on the findings, the observations which can be made are as follows:

a) Tesseract OCR is expected to deliver reasonable results on well-structured and clean documents and cannot generalize well on handwritten and noisy layouts.

b) Pre-processing Tesseract optimized to maximize CER and WER also maximizes accuracy on the character-level recognition, but minimizes semantic-level recognition.

c) The highest semantic alignment of BLEU = 0.82 and highest F1-score (0.710) of Donut is due to its end-to-end transformer architecture.

d) LayoutLMv3 has the lowest CER (0.410) and competitive entity recognition (F1 = 0.695) but a little overfitting was observed in domain-specific classes.

e) The most balanced option in terms of efficiency and higher accuracy is a hybrid OCR pipeline (Tesseract on clean inputs and transformers on noisy/handwritten text) that uses both components.

9. CONCLUSION

The study presented the SMARTFORM, a smart OCR-based system to automate the processing of online banking forms. A React-based portal, a flask back-end, and hybrid OCR models are integrated into the system to provide a secure and scalable system to work with structured and semi-structured banking documents. The experimental research showed some of the insights. To start with, Tesseract OCR is a reliable solution to clean, machine-printed text, particularly with optimal pre-processing. Nonetheless, its performance reduces significantly when it encounters complicated layouts, noisy scans or when hand writing is involved. Second, transformer models like Donut and LayoutLMv3 were identified to be superior to the traditional OCR in that they were more contextually aware and entity-recognizing but were sometimes prone to overfitting in domain-specific contexts. A combination of these models in a hybrid pipeline created by a strategic approach allowed the system to reach a greater level of accuracy without losing the efficiency of inference times.In addition to recognition accuracy, there was the consideration of security and compliance. The PIN employee validation and JWT customer authentication would mean that integral and confidential customer financial information is managed by the platform. This offers the institutions with a reliable and regulatively acceptable way of incorporating OCR automation into their current processes. In general, SMARTFORM will decrease the number of repetitive manual entries, decrease human error, and shorten document verification processes. What is more significant is that it shows how the combination of the typical OCR engines, transformer-based designs and adaptive pre- processing can be exploited to construct a powerful enterprise ready solution. This work gives a basis to further development of intelligent document processing, and the possible sphere of its application is not only in the field of banking but also in healthcare, insurance, and government services where speedy and correct automation of documents is obligatory.

REFERENCES

1. R. Smith, “An overview of the Tesseract OCR engine,” in Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR), Curitiba, Brazil, 2007, pp. 629–633.
2. A.Ul-Hasan, S. M. Lucas, A. F. Mollah, F. Shafait, and T. M. Breuel, “Handwritten text recognition on noisy document images using hybrid features,” Pattern Recognition Letters, vol. 34, no. 4, pp.

431–439, 2013.
3. N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.
4. G.Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, “Donut: Document understanding transformer without OCR,” arXiv preprint arXiv:2111.15664, 2021.
5. Y. Xu, T. Xu, B. Lv, J. Cui, and F. Wei, “LayoutLMv3: Pre-training for document AI with unified text and image masking,” in Proc. 30th ACM Int. Conf. Multimedia (MM’22), Lisbon, Portugal, 2022, pp. 4083–4091.
6. M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florêncio, C. Zhang, Z. Li, and F. Wei, “TrOCR: Transformer-based optical character recognition with pre-trained models,” in Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 11, pp. 13094–13102, 2023.
7. S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “DocFormer: End-to-end transformer for document understanding,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, 2021, pp. 993–1003.
8. S. Luo, M. Wu, Y. Gong, W. Zhou, and J. Poon, “Deep structured feature networks for table detection and tabular data extraction from scanned financial document images,” arXiv preprint arXiv:2102.10287, 2021.
9. R. Arroyo, J. Yebes, E. Martínez, H. Corrales, and J. Lorenzo, “Key information extraction in purchase documents using deep learning and rule- based corrections,” arXiv preprint arXiv:2210.03453, 2022.
10. S. Cho, “A framework for understanding unstructured financial document images: Multimodal IDP with deep learning and RPA,” Electronics, vol. 12, no. 4, p. 939, 2023.