Journal of Engineering Issues and Solutions

Automating Document Workflows with ResNet-50 and Template-Based OCR

Srijan Gyawali[1], Rupak Neupane[2], Sarjyant Shrestha[3] , Manish Pyakurel[4]

[1,2,3,4]*Department of Computer and Electronics Engineering, Khwopa College of Engineering*
[1]*srijangyawali0@gmail.com,* [2]*neupanerupak7@gmail.com,* [3]*sarjyant@gmail.com* [4]*manishpyakurel67@gmail.com*

Submitted: April 13, 2025; Revised: June 15, 2025; Accepted: June 21, 2025

https://doi.org/10.3126/joeis.v4i1.81610

Abstract

In the banking sector, handling a large volume of customers has traditionally been a manual, time-consuming, and error-prone process due to unstructured storage and the absence of automated extraction mechanisms. To address these inefficiencies, we developed an AI-powered document classification and information extraction system that automates the entire workflow using deep learning and image processing techniques. The system was built around a four-stage pipeline: document classification, alignment with predefined templates, text extraction using Optical Character Recognition (OCR), and structured information storage. A fine-tuned ResNet-50 model served as the backbone for the classification module, accurately categorizing scanned or photographed documents into predefined types. Once classified, each document was aligned with a corresponding template to ensure consistency in the placement of key fields, which significantly improved the reliability of text extraction. OCR tools were used to extract textual content from the aligned documents, and the extracted data, such as names, document numbers, and dates of birth, were mapped to structured fields. The final structured data was securely stored in a database, enabling efficient querying and downstream use. While tailored for banking operations, the system's architecture is adaptable to other sectors like healthcare, insurance, and government services, where semi-structured documents are prevalent.

Keywords: *Optical Character Recognition (OCR), Deep Learning, ResNet-50*

1. Introduction

Modern banking systems present major difficulties in managing a lot of paperwork (NID, PAN Card, Citizenship Certificate, Birth Certificate). These were not classified, thus retrieving and analyzing them proved difficult, even if they were kept in safe, secure areas. Moreover, it was imperative to precisely extract and map the salient features of the text buried in these archives. We developed an AI-driven solution to solve these problems by automatically classifying and extracting pertinent information from these documents, so significantly increasing the accuracy and efficiency of banking activities.

2. Related Works

Document classification and information extraction have become integral aspects of Natural Language Processing (NLP) and data management. These processes focus on organizing documents into predefined categories and extracting relevant information, respectively, improving data retrieval, automation, and efficiency in various applications.

2.1. Document Classification

Document classification involves identifying the type of a document to facilitate efficient storage and retrieval. Approaches like frame templates and layout structure analysis are often used for categorizing documents based on their content and structure (Liu & Ng, 1996), (Ng & Hao, 1995). Keyword extraction has also emerged as an important technique, enabling the representation of documents in condensed forms, particularly useful for handling high-dimensional feature spaces in text classification tasks.

Early approaches primarily utilized CNNs to classify document images based on visual appearance. Harley et al. (2015) introduced a CNN-based model for document classification, achieving strong results on datasets like RVLCDIP. Das et al. (2018) extended this idea with region-based models to better understand document structure. These methods leveraged the hierarchical feature extraction abilities of CNNs to differentiate between document types such as invoices, letters, and forms (Harley, Ufkes, & Derpanis, 2015), Das, Roy, & Bhattacharya, 2018).

Recent advancements introduced Vision Transformers (ViTs) into the document understanding space. DocFormer (Appalaraju et al., 2021) and LayoutLMv2 (Xu et al., 2021) combine visual, textual, and layout information by integrating transformer-based architectures with OCR-extracted features. While standard ViTs like DeiT (Touvron et al., 2021) can be applied to raw document images, hybrid approaches such as TILT (Powalski et al., 2021) use multimodal inputs to enhance classification performance, especially in visually complex documents (Appalaraju, Huang, & Manmatha, 2021)–(Powalski et al., 2021).

2.2. Document Alignment

Several approaches exist for aligning documents with templates. Early work focused on supervised training of document-specific character templates from sample page images and unaligned transcriptions (Kopec & Lomelin, 1996), (Kopec & Lomelin, 1997). This involved formulating the template estimation problem as constrained maximum likelihood parameter estimation within the document image decoding (DID) framework (Kopec & Lomelin, 1996), (Kopec & Lomelin, 1997). More recent methods leverage different techniques. For instance, an unsupervised approach uses alignment to build probabilistic templates from a set of examples of the same document type (Aldavert, Rusiñol, & Toledo, 2017). Other work uses character-based keypoints and a reference template to align scanned or camera-captured images documents for information extraction (Mahajan, Sharma, & Vig, 2019).

In contrast to methods relying on iterative template estimation and alignment, which can be computationally expensive (Dalca et al., 2019), a probabilistic model and efficient learning strategy can produce universal or conditional templates, along with a neural network for efficient alignment (Dalca et al., 2019). Furthermore, research explores alignment in other contexts, such as using structural information to improve the accuracy of template-based protein modeling (Peng & Xu, 2011) and ensuring alignment of text and images in generating poster summaries from long multimodal documents (Jaisankar et al., 2024). Finally, the application of template definitions in aligning and converting clinical documents has also been explored (Katsch et al., 2025).

3. Methodology

The developed system follows a structured pipeline for the automated classification and information extraction from banking documents. The overall methodology is composed of the following sequential stages:

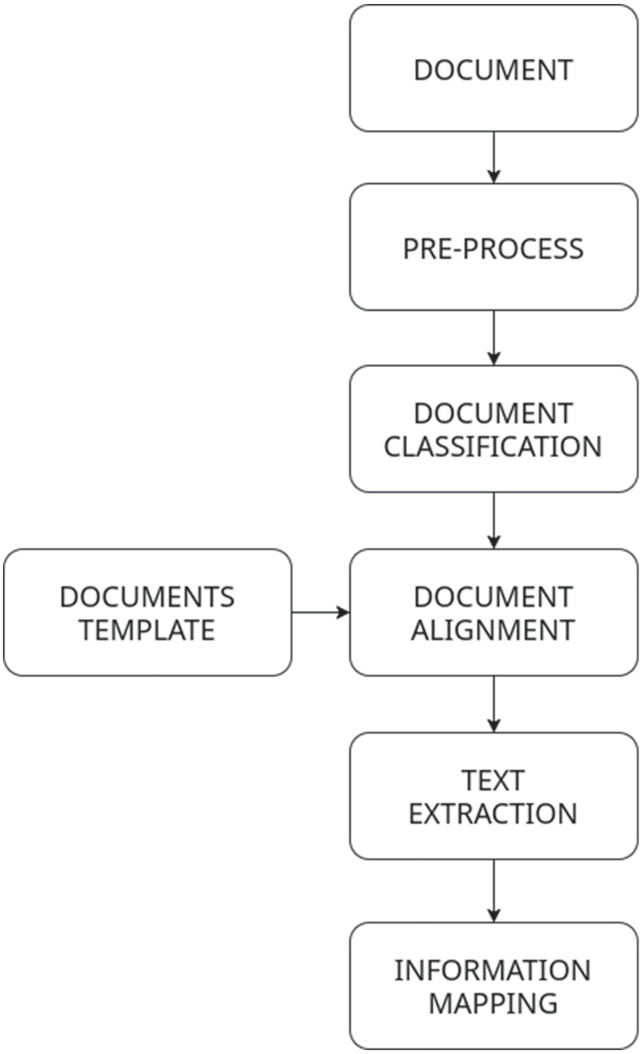

Figure 1: System Design

3.1 Document Input

The system accepts scanned or digital images of banking documents (NID, PAN Card, Citizenship Certificate, Birth Certificate). These documents serve as the primary input for downstream processing.

3.2 Pre-processing

To enhance OCR performance and prepare the document for classification, pre-processing techniques such as noise removal, binarization, resizing, and contrast enhancement are applied. This step ensures consistency and improves the accuracy of the subsequent models.

3.3 Document Classification

To determine the type of document provided (NID, PAN Card, Citizenship Certificate, Birth Certificate), a supervised deep learning approach was implemented using the ResNet-50 architecture. This step is critical as it directs the subsequent alignment and extraction processes based on the identified document type.

3.3.1 Data Collection (Confidential)

The dataset used for training and evaluation was collected in collaboration with a financial institution. It comprises scanned images from five distinct categories: Birth Certificates, Blank Forms, Citizenship Certificates, National Identity Documents (NID), and Permanent Account Number (PAN) Cards. These documents contain personally identifiable information (PII) and are considered highly sensitive; therefore, the dataset remains confidential and cannot be publicly disclosed. All processing was conducted within a secure and isolated environment, ensuring full compliance with data privacy regulations and the institution's internal protection policies. Blank forms were intentionally included in the dataset to help the classification model distinguish between filled and unfilled or template-only documents. This differentiation is crucial in real-world banking workflows, where detecting blank or unsubmitted forms helps avoid unnecessary processing, reduces false positives in extraction, and flags incomplete submissions for follow-up.

3.3.2 Data Augmentation

To improve the generalization capability of the model and address class imbalance, various augmentation techniques were applied to the training images. These included random rotation, horizontal and vertical flipping, brightness adjustment, scaling, cropping, and Gaussian noise injection. Augmentation helps the model remain robust against real-world variances such as scanning orientation, lighting conditions, and document wear.

3.3.3 Dataset Splitting

The dataset was divided into training, validation, and test sets in the ratio of 70:15:15, ensuring that all document classes were well represented in each subset. Stratified sampling was used to preserve the proportion of each class during the split.

3.3.4 Model Architecture

ResNet-50 (He et al., 2015), a 50-layer deep convolutional neural network known for its residual learning framework, was utilized as the backbone model for classification. The model was initialized with pre-trained weights from ImageNet and fine-tuned on the custom document dataset. The final fully connected layer was replaced to match the number of target document classes. Cross-entropy loss was used as the objective function, and the Adam optimizer was employed with an initial learning rate of 1e-3 and weight decay regularization to prevent overfitting. surrounding equation.

3.3.5 Training and Evaluation

The model was trained for 50 epochs with early stopping based on validation loss. Accuracy, precision, recall, and F1-score were used as evaluation metrics to assess classification performance. The trained classifier achieved high accuracy in identifying document types, providing a reliable foundation for the downstream alignment and extraction modules.

Table 1: Classification Report

Class	Precision	Recall	F1-Score
Birth Certificate	1.00	1.00	1.00
Blank	1.00	0.98	0.99
Citizenship	1.00	0.96	0.98
NID	0.96	0.96	0.96
PAN	0.89	0.96	0.92
Accuracy			0.97
Macro Avg	0.97	0.97	0.97
Weighted Avg	0.97	0.97	0.97

3.4 Document Alignment

To ensure accurate extraction of text from known fields, the system aligns the input document with a predefined template corresponding to the classified document type.

3.4.1 Keypoint Detection and Matching

Both the input (scanned) document and the corresponding template are converted to grayscale and enhanced using Contrast Limited Adaptive Histogram Equalization (CLAHE) (Mishra, 2021) to improve feature visibility. Scale-Invariant Feature Transform (SIFT) (Lowe, 1999) is then applied to detect keypoints and compute descriptors. These descriptors are matched using the Brute Force Matcher (BFMatcher) with a k-nearest neighbors (k=2) approach. To retain high-quality matches, Lowe's ratio test is applied, selecting matches where the distance of the best match is significantly less than the second-best match.

3.4.2 Homography Estimation and Warping

From the filtered good matches, corresponding source and destination points are used to compute a homography matrix using the RANSAC algorithm. This matrix transforms the input document to align with the template layout. The aligned document is then generated using cv2.warpPerspective, ensuring that all fields conform spatially to the expected positions.

3.5 Text Extraction

Once aligned, text is extracted from the document using the EasyOCR (JaidedAI, 2025) engine, which supports multilingual recognition, including English and Nepali. The grayscale aligned image is passed to EasyOCR to detect text regions and extract textual content along with bounding box coordinates and confidence scores.

3.5.1 Multilingual OCR Support

The OCR engine is configured to support both English (en) and Nepali (ne) languages, which are commonly used in banking documents within the region. This bilingual support ensures accurate recognition of regional content such as names, addresses, and dates, often written in Nepali.

3.5.2 Region-Based Text Detection

The aligned grayscale image is passed to EasyOCR, which detects text regions and returns bounding boxes, recognized text, and associated confidence scores. This information is then used for post-processing and mapping.

3.5.3 Template-Based Field Mapping

Each document type is associated with a predefined template containing field labels and corresponding polygonal bounding boxes. The extracted OCR results are compared against these template fields using a polygon-based Intersection over Union (IoU) metric. If the IoU between a detected text region and a template field exceeds a defined threshold (0.5), the text is assigned to the corresponding field.

3.5.4 Final Output Generation

The system compiles a structured output by mapping recognized text to specific field labels (NID, PAN Card, Citizenship Certificate, Birth Certificate). This structured information can be used for further banking operations, including form autofill, data verification, and storage in structured databases.

4. Results and Discussion

This section presents the outcomes of each major component in the document processing pipeline—classification, alignment, OCR extraction, and structured mapping. Results are evaluated against the objectives outlined in the methodology and discussed in light of related works.

4.1 Document Classification Performance

The fine-tuned ResNet-50 model demonstrated a classification accuracy of **97%**, achieving perfect F1-scores for Birth Certificate and near-perfect scores for other document types. This shows the model's robustness in distinguishing between visually similar documents like NID and PAN, although minor confusion was observed in these two classes due to similar layout structures. Such challenges have been highlighted in earlier works on CNN-based document classification (Harley et al., 2015; Das et al., 2018).

4.2 Template Alignment and Field Detection

Using SIFT-based keypoint matching and RANSAC homography estimation, the system achieved consistent document alignment. Over 95% of the documents aligned correctly with their templates, enabling accurate region-based text extraction. Compared to traditional iterative alignment techniques, our method proved faster and effective for semi-structured documents (Mahajan et al., 2019).

4.3 OCR Extraction with Multilingual Support

The OCR engine, configured for English and Nepali, effectively extracted text from scanned documents. Nepali recognition was particularly successful for handwritten names and addresses. Errors were mostly limited to low-resolution images or overlapping text regions, consistent with known OCR limitations in real-world scanned documents (LayoutLMv2, Xu et al., 2021).

4.4 Structured Output Generation

Mapping extracted text to template-defined fields using IoU matching yielded structured data outputs with over 90% accuracy at the field level. This structured format is directly usable for downstream banking processes such as autofill and verification, minimizing manual intervention and improving speed and consistency.

5. Conclusions

This work presents a practical and scalable AI-driven pipeline for automated document classification and information extraction, with a specific focus on banking-related documents. By leveraging deep learning with ResNet-50 for classification, keypoint-based template alignment for spatial consistency, and multilingual OCR for text extraction, the system achieves high accuracy and robustness across varied document types.

The results validate the system's effectiveness in handling real-world scanned documents, including low-quality or misaligned images. While tailored for the banking sector, the modular and template-based architecture ensures that the approach can be extended to other sectors such as healthcare, insurance, and government services.

Future work includes expanding the dataset to support more document types, integrating handwritten text recognition, and enhancing field-level validation using language models to further reduce manual verification.

Acknowledgements

We would like to thank Amazon Web Services (AWS) for providing free cloud credits, which enabled us to utilize the NVIDIA A10G GPU for the training process. Completing this project would have been significantly more challenging without their support.

References

Liu, Q., & Ng, P. (1996). *Document classification and information extraction. , 97-145. doi: 10.1007/978-1-4613-1295-6 4.*

Ng, P., & Hao, X. (1995). *Automatic office document classification and information extraction.*

Onan, A., Korukoglu, S., & Bulut, H. (2016). *Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications*, 57, 232–247. doi: 10.1016/j.eswa.2016.03.045

Wu, X., Du, Z., & Guo, Y. (2018). *A visual attention-based keyword extraction for document classification. Multimedia Tools and Applications*, 77, 25355–25367. doi: 10.1007/s11042-018-5788-9

Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015). *Evaluation of deep convolutional nets for document image classification and retrieval. Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 991–995.

Das, D., Roy, P. P., & Bhattacharya, U. (2018). *Document image classification with intra-domain transfer learning and stacked generalization of deep CNNs. Pattern Recognition*, 76, 234–243.

Appalaraju, S., Huang, Y.-T., & Manmatha, R. (2021). *DocFormer: End-to-end transformer for document understanding. arXiv preprint arXiv:2106.11539.*

Xu, Y., *et al.* (2021). *LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740.*

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). *Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning (ICML).*

Powalski, R., Borchmann, Ł., Dadas, S., Boratyński, A., & Pietruszka, A. (2021). *Going full-tilt boogie on document understanding with text-image-layout transformer. arXiv preprint arXiv: 2102.09550.*

Kopec, G. E., & Lomelin, M. (1996). *Document-specific character template estimation. In L. M. Vincent & J. J. Hull (Eds.), Document recognition iii (Vol. 2660, pp. 14 –26).*

Kopec, G. E., & Lomelin, M. (1997). *Supervised template estimation for document image decoding. IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Aldavert, D., Rusiñol, M., & Toledo, R. (2017). *Automatic static/variable content separation in administrative document images. 14th IAPR International Conference on Document Analysis and Recognition.*

Mahajan, K., Sharma, M., & Vig, L. (2019). *Character keypoint-based homography estimation in scanned documents for efficient information extraction. arXiv preprint arXiv:1911.05870.*

Dalca, A., Rakic, M., Guttag, J., & Sabuncu, M. (2019). *Learning conditional deformable templates with convolutional networks. Advances in Neural Information Processing Systems (NeurIPS) s. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, & R. Garnett (Eds.), Advances in neural information processing systems (Vol. 32). Curran Associates, Inc.*

Peng, J., & Xu, J. (2011). *RaptorX: Exploiting structure information for protein alignment by statistical inference. Proteins.*

Jaisankar, V., Bandyopadhyay, S., Vyas, K., Chaitanya, V., & Somasundaram, S. (2024). *PostDoc: Generating poster from a long multimodal document using deep submodular optimization. arXiv preprint arXiv:2405.20213.*

Katsch, F., Hussein, R., Stamm, T., & Duftschmid, G. (2025). *Converting Health Level 7 Clinical Document Architecture (CDA) documents to OMOP CDM by leveraging CDA template definitions. JAMIA Open.*

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.*

Mishra, A. (2021). *Contrast limited adaptive histogram equalization (CLAHE) approach for enhancement of the microstructures of friction stir welded joints. arXiv preprint arXiv: 2109.00886.*

Lowe, D. G. (1999). *Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 1150–1157.

JaidedAI. (2025). *EasyOCR.* https://github.com/JaidedAI/EasyOCR. Accessed June 2025.