

# E2DIEVS: End-to-End Document Information Extraction and Verification System

## Final Project Report

Yuvraj Bhadoriya  
202411002

Priyanshi Patel  
202411009

Manish Prajapati  
202411012

### 1 Abstract

As we witnessed rapid digitalization in our country, India, there grows the need for automatic document verification systems. Earlier, the KYC (know your customer) methods were done manually, as opposed to today, we can now perform KYC anywhere on Earth. Traditionally, OCR-based pipelines are getting used to extract text, and after applying several rule-based methods, we extract information from it. These methods failed miserably on Indian Identity documents such as Aadhar cards, and Pan cards. LayoutLM is a pre-trained model, known for extracting textual information from scanned documents. We employ that model, fine-tune it on the datasets which only consist of scanned form images and receipts with primary information, that eventually enhanced the performance. But, it faces limitations on scanning and understanding textual information from the Aadhar cards and Pan cards due to unrecognized formats. We overcome this limitation by using YoloV11, pre-trained on the dataset of aadhar cards and pan cards, which gives us text along with its category.

### 2 Introduction

Extracting information from scanned documents such as forms, invoices, receipts and identity cards has been emerging critically in AI. Microsoft Research introduced LayoutLM (Xu et al., 2019) which brought revolution in this domain. It incorporates spatial awareness into the transformer which led to the state-of-the-art results on benchmarks such as FUNSD (Jaume et al., 2019) and SROIE (Huang et al., 2019). But it has a drawback of heavily relying on the accurate OCR and bounding box inputs during inference. That makes its multi-modal pre-training computationally expensive

and less effective.

To address limitations faced by the LayoutLM, Donut (Kim et al., 2022b) proposed an OCR-free approach. It uses a vision-transformer encoder which is combined with a text decoder to directly generate structured JSON output from raw document images without intermediate text recognition. The drawbacks of Donut are: training on diverse document types, and underperforms on densely formatted or multi-lingual documents, such as Aadhar cards, and Pan cards.

The accurate field extraction from Aadhar cards and Pan cards remains particularly challenging due to multi-lingual content and inconsistent layouts across states. We came across an advanced object detection model, such as YoloV11 (Logasanjeev, 2024), which is fine-tuned on Indian ID datasets. It is more capable than LayoutLM and Donut for extracting the information from Indian IDs along with their categories. Our project contains the following novel ideas:

- LayoutLM fine-tuned on FUNSD and SROIE datasets showing improved performance.
- A novel pipeline that can extract information from the printed digital forms as well as Aadhar cards or Pan cards, and verify with the ground truth data.
- Used a hybrid similarity scoring method of 60% character and 40% token-based similarity jaccard method.
- Employed YOLOv11 for classification of documents to dynamically load weights as well as bounding boxes.

### 3 Related Work

Earlier, the systems relied on rule-based approaches and machine learning techniques such as SVMs and CRFs were trained on OCR outputs (Mao and Kanungo, 2003; Russell and Norvig, 2006), which achieved acceptable performance only on fixed-layout forms, that later on failed for variable or noisy documents. Hybrids of CNN-RNN such as CharGrid (Katti et al., 2018) and CRAFT-based pipelines (Baek et al., 2019) introduced visual feature learning by considering documents as images or character grids, which improved robustness over pure-text based methods. These approaches still remain highly sensitive to the quality of images. Some OCR-dependent variants of transformers such as BERTgrid (Denk and Dietrich, 2019), and Graph-based models such as GraphIE (Qian et al., 2019) incorporate 2D position or relational information, which still inherits OCR inaccuracies due to low-quality images. These challenges often occur in extracting information from Indian Identity documents such as Aadhar cards and Pan cards.

### 4 Data Modeling and Abstraction

In this project, we adopted the data abstraction method similar to LayoutLM. It jointly represents text tokens and their corresponding 2D spatial coordinates within the document image. Following (Xu et al., 2019), each input sequence consisting of words obtained from OCR together with their bounding boxes  $[x_1, y_1, x_2, y_2]$ . These coordinates are normalized to the range  $[0, 1000]$  relative to the image dimensions, and then injected into the model through four dedicated positional encoding embedding layers. We used AUTO tokenizer with the maximum sequence length of 512. Special tokens such as [CLS] and [SEP] were assigned zero bounding boxes  $[0, 0, 0, 0]$  as they do not carry any spatial meaning. This modeling choice ensures that both textual and spatial information are used simultaneously. Following are the details about modeling the data for different encodings:

#### 1. Text Encoding:

- Input: List of words (from OCR: Tesseract for SROIE, ground-truth for FUNSD).

- Tokenizer: *AutoTokenizer* (WordPiece, uncased) with *is\_split\_into\_words = True*.
- Sub-words of the first token for each word inherits the BIO label.

#### 2. Layout (2D Position) Encoding

- Bounding boxes are normalized for every document, making them the size of  $[0, 1000] \times [0, 1000]$ .
- Formula:

$$x' = \left\lfloor \frac{x \cdot 1000}{\max(x)} \right\rfloor, \quad y' = \left\lfloor \frac{y \cdot 1000}{\max(y)} \right\rfloor$$

where  $x$  and  $y$  are the dimensions of the images.

- Four coordinates are fed into the learned 2D position embeddings as in (Xu et al., 2019).

#### 3. Input representation

- Each token receives:  
 $[CLS] + \text{Token Embedding} + \text{Segment Embedding} + 2\text{-D Position Embedding } (x_1, y_1, x_2, y_2) + [SEP]$ .

### 5 Methodology

#### 5.1 Dataset Configuration

1. FUNSD (Jaume et al., 2019): A dataset for understanding the forms in noisy scanned documents. There are 149 datapoints in train-set and 50 datapoints in test-set. The language in forms is majorly English. More details can be seen in Table 1.

Table 1: FUNSD dataset configuration.

Attribute	Details
Total Images	199 scanned forms
Word-level annotations	30,000
Entities	10,000
Entity Types	Header, Question, Answer & other
Annotation Format	BIO tagging + bounding boxes + key-value links
Image Resolution	100 DPI grayscale

2. SROIE (Huang et al., 2019): The full-form is Scanned Receipts OCR and Information Extraction. This is used for OCR and key

information extraction from scanned receipts. It consists of 626 images in train-set and 347 images in test-set. The language in receipts is English. More details can be seen in Table 2.

Table 2: SROIE dataset configuration.

Attribute	Details
Total Images	1,000 scanned receipts
Key Entities	4
Image resolution	Varied
Annotations	OCR bounding boxes and key-value pairs

3. Aadhar and Pan cards: The YoloV11 model that we employed, is fine-tuned on a custom dataset comprising of real-world mobile-captured images of Indian identity documents. There are 7 different types of document classes: aadhar-front, aadhar-back, pan-card-front, driving-license-front, driving-license-back, passport and voter-id-card. It has also been incorporated with 45 distinct entity classes covering key information such as Name, Date of Birth, Aadhar Number, PAN number, and many.

## 5.2 Baseline Models

For complex form layouts, we use LayoutLM (Xu et al., 2019), specifically *layoutlm-base-uncased*. It is pre-trained multi-modal transformer that learns textual and spatial layout information from images of the documents. It is pre-trained on IIT-CDIP dataset (Lewis et al., 2006), which consists of a variety of documents. A snippet of dataset it was pre-trained on can be seen in Figure 1. There are 400,000 grayscale images in 16 classes, each class accounting 25,000 images. After train, test and validation split, there are 320,000 images in the training set, 40,000 images in the validation set and 40,000 images in the test set. The images are resized such that the largest dimension do not exceed 1000 pixels.

The model specification of the pre-trained LayoutLM can be seen in Table 3. The embeddings of the input modalities consists of the following information:

1. Adding 2D spatial embeddings to each token:
  - x-position-embeddings

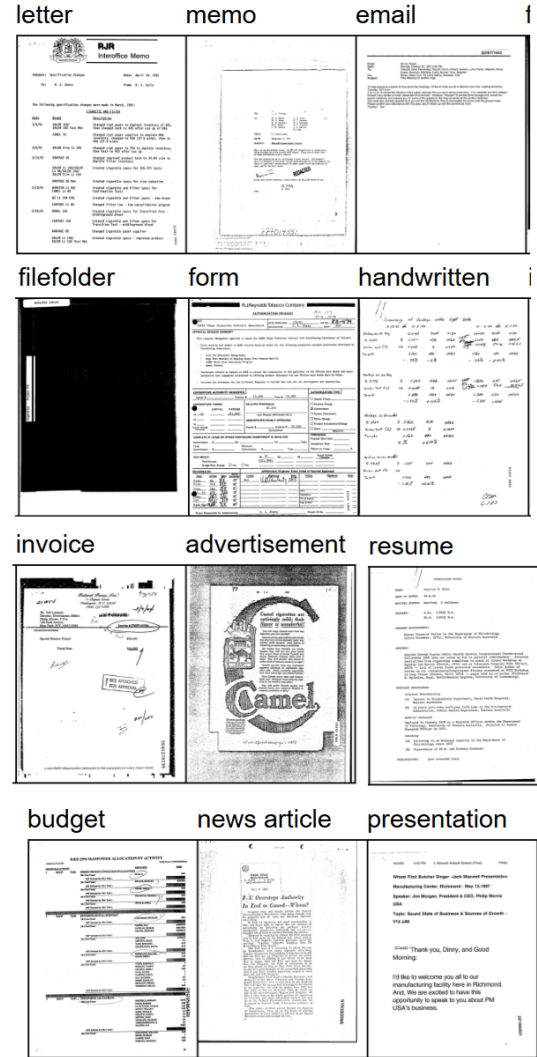


Figure 1: IIT-CDIP train dataset snippet. There 16 different types of scanned documents in the dataset.

- y-position-embeddings
- h-position-embeddings (height)
- w-position-embeddings (width)

2. Pre-training on real scanned documents (not synthetic text), preserving:

- Font styles
- Table structures
- Form fields
- Spatial alignments

For simpler OCR-free extraction, we also employed a pre-trained Donut model (Kim et al., 2022a). It is vision-transformer that directly generates structured JSON from raw images without intermediate tokenization. Donut excels in end-to-end tasks such as receipt parsing on CORD (Park et al., 2019), reported with

Table 3: Model Specifications of LayoutLM.

Component	Details
Architecture	BERT-like Transformer
Input Modalities	1. Text (words) 2. 2D Layout
Pre-training Corpus	IIT-CDIP
Tokenizer	WordPiece (uncased)
Max Sequence Length	512
Parameters	~110M

more than 90% accuracy on structured outputs, hence, no fine-tuning of Donut is required in our pipeline.

We have used pre-trained YOLOv11 model from (Logasanjeev, 2024), which achieved good results on custom Indian ID datasets. We employ this model to extract information from Aadhar card and Pan card.

### 5.3 Proposed Approach

#### 5.3.1 Fine-tuning LayoutLM

LayoutLM introduced a powerful pre-training paradigm for scanned documents. However, downstream evaluation was constrained. Earlier work fine-tuned the model on FUNSD and test-set of the SROIE dataset corpora. We find our enhancement far from being trivial for the following reasons:

1. The objectives of pre-training operate at the representation level and do not expose the model to task-specific BIO schemas.
2. Pre-training data consist primarily of scanned books and articles from IIT-CDIP, exhibiting fundamentally different spatial statistics.
3. Character-level fidelity as a Proxy for deployment success.

Thus, our work including the fine-tuning, completes the experiment from the original paper. The model configuration can be seen in the Table 4.

As described in Table 4, we choose 15 number of labels, which represent the BIO-schema heads: HEADER, QUESTION, ANSWER, COMPANY, DATE, ADDRESS, TOTAL, etc. We kept the value of the maximum sequence

Table 4: Model configuration for fine-tuning on the dataset corpora of FUNSD and SROIE.

Parameter	Value
Number of labels	15 (BIO-schema)
Max. sequence length	512
Batch Size	8
Learning rate	5e-5
Optimizer	AdamW
Epochs	30

length same as pre-training. To be memory efficient with the bounding box embeddings, we kept the batch-size as 8. Optimizer is AdamW with weight decay of 0.01. Initially, we kept model warmup comprising of 10% of the steps in order to stabilize the training. For collating the data, we use dynamic padding that gives no batch size errors, and ensures no data loss.

#### 5.3.2 Pipeline Layout

Following are the steps of our pipeline:

1. **User registration:** We fetch primary details from user, such as, name, date of birth, aadhar number, etc.
2. **Document upload:** Based on the required details given by the user, it will upload the relevant document for verification. The document can be a scanned copy of digital form, aadhar card, or pan card.
3. **Document identification:** Based on the document uploaded, we will verify whether it is an Indian identity document or scanned copy of digital form. For Indian identity document, we will pass it to YoloV11 and get the information in the form of a structured json. In case of digital copy of the form, we will pass the input to the fine-tuned LayoutLM and Donut. Based on their structured responses whichever being close to the ground truth, we will use that response for further verification.
4. **Information verification:** We used a hybrid similarity scoring mechanism which combines character-level and token-level comparisons. We calculate character-level similarity using Python’s *difflib*.

The *SequenceMatcher* in it combines the Ratcliff-Obershelp algorithm to measure the longest common sub-sequence ratio, which effectively handles OCR-induced typos, spacing inconsistencies, and minor spelling variations. For enhanced robustness on multi-token fields such as names and addresses, we complement this with Jaccard token-level similarity, calculated as the ratio of intersecting to union tokens after whitespace splitting. The final verification score is a weighted fusion – 60% character-level similarity + 40% Jaccard similarity – providing a balanced, noise-tolerant metric.

The flow diagram of our pipeline layout can be seen in the Figure 2. The Figure 3 shows the layout of the report formed after the execution of whole pipeline.

## 6 Evaluation Method

### 6.1 Training and Evaluating Loss

During the fine-tuning, we optimize the standard token classification objective by incorporating cross-entropy loss over the predicted logits and ground-truth labels. For a sequence of length  $T$ , the loss for a particular instance is calculated as:

$$\mathcal{L} = -\frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \log \left( \frac{\exp(\text{logit}_{y_t}[t])}{\sum_{c=1}^C \exp(\text{logit}_c[t])} \right), \quad (1)$$

where:

- $\mathcal{A} = t \mid y_t \neq -100$  is set the of active positions.
- $y_t \in \{0, 1, \dots, C-1\}$  is the ground-truth label at position  $t$ .
- $C = 15$  is the number of classes in our unified BIO schema.
- $\text{logit}_c[t]$  is the model’s un-normalized score for class  $c$  at position  $t$ .
- The positions corresponding to the sub-word continuations, and special tokens ([CLS], [SEP], [PAD]) are assigned label **-100**. This is ignored by the loss function.

Our objective function is similar to that in (Xu et al., 2019).

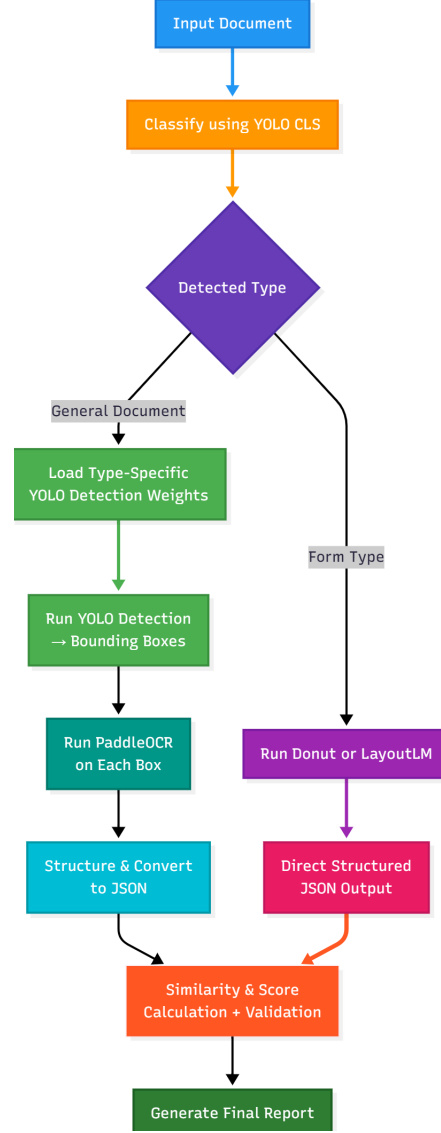


Figure 2: IIT-CDIP train dataset snippet. There 16 different types of scanned documents in the dataset.

#### 6.1.1 Token-level Accuracy

For document-image understanding tasks, the token-level accuracy is defined as:

$$\text{Accuracy} = \frac{\sum_{t \in \mathcal{A}} 1\{\hat{y}_t = y_t\}}{|\mathcal{A}|} \quad (2)$$

#### 6.1.2 Macro F1 Score

This is computed for more than 15 BIO classes:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}, \text{Macro-F1} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (3)$$

where  $P_c$  and  $R_c$  are the precision and recall for class  $c$ .



Figure 3: Final Report after the execution of the pipeline.

## 7 Results

The results for baseline LayoutLM on the dataset of FUNSD can be seen in Table 5, and on the dataset corpora of SROIE can be seen in Table 6.

Table 5: Evaluation of FUNSD dataset on baseline LayoutLM model

Metric	Value
Recall	0.041
Precision	0.254
F1	0.068

Table 6: Evaluation of SROIE dataset on baseline LayoutLM model

Metric	Value
Recall	0.271
Precision	0.249
F1	0.249

The macro-F1 score on FUNSD dataset is 0.068 and, that on SROIE dataset is 0.249. Both the scores are very low, compared to that we obtained upon fine-tuning the model on the FUNSD and SROIE datasets.

Our fine-tuned model achieved the best macro-F1 score of 0.638 at epoch 19, and best evaluation accuracy of 95.169% at epoch 7.

The detailed performance metric of fine-tuned LayoutLM evaluated on FUNSD dataset can be seen in Table 5.

## 8 Conclusion

Our project E2DIEVS demonstrated that combining the layout-aware transformers, such as OCR/vision models, and detection based modules yields a practical pipeline for information extraction and verification across diverse document types. Fine-tuning LayoutLM on the dataset corpora of FUNSD and SROIE, and integrating YOLOv11 for Indian ID cards substantially improved extraction quality compared to baselines, with our best model achieving a large uplift in macro-F1 and high token accuracy after targeted preprocessing and label engineering.

Nevertheless, OCR sensitivity and limited labeled diversity remain the main bottlenecks—especially for multilingual and highly stylized identity documents. Future work should focus on more robust OCR or OCR-free approaches, multilingual/domain adaptation, and active learning to expand coverage; these steps will make the pipeline more reliable for real-world KYC and document-verification deployments.

## References

- Y. Baek and 1 others. 2019. Character region awareness for text detection. In *CVPR*.
- T. I. Denk and C. Dietrich. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. In *NeurIPS Workshop*.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. *ICDAR*. Available at: <https://rrc.cvc.uab.es/?ch=13>.
- Guillaume Jaume, Hazim K Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. *arXiv preprint arXiv:1905.13538*. Available at: <https://guillaumejaume.github.io/FUNSD/>.
- A. R. Katti and 1 others. 2018. Chargrid: Towards understanding 2d documents. In *EMNLP*.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han,

Table 7: Fine-tuned LayoutLM performance on FUNSD dataset.

<b>Class</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
HEADER	0.9479	0.9479	0.9479	96
ANSWER	0.9255	0.9560	0.9405	273
QUESTION	0.9335	0.9425	0.9380	313
<b>Micro Avg</b>	0.9323	0.9487	0.9404	682
<b>Macro Avg</b>	0.9357	0.9488	0.9422	682
<b>Weighted Avg</b>	0.9324	0.9487	0.9404	682

and Seunghyun Park. 2022a. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Geewook Kim and 1 others. 2022b. clovaai/donut: Official implementation of ocr-free document understanding transformer. <https://github.com/clovaai/donut>.

David D. Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Iit-cdip test collection 1.0. <https://ir.nist.gov/cdip/>. Illinois Institute of Technology Complex Document Information Processing (CDIP) test collection.

Logasanjeev. 2024. [indian-id-validator – yolov11 model for aadhaar, pan, and driving license field detection](#). Hugging Face Model Hub.

S. Mao and T. Kanungo. 2003. Text extraction using hidden markov models. In *ICDAR*.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. [{CORD}: A consolidated receipt dataset for post-{ocr} parsing](#). In *Workshop on Document Intelligence at NeurIPS 2019*.

Y. Qian and 1 others. 2019. Graphie: A graph-based framework for information extraction. In *NAACL*.

S. Russell and P. Norvig. 2006. Artificial intelligence: a modern approach. In *Prentice Hall*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. [Layoutlm: Pre-training of text and layout for document image understanding](#). *CoRR*, abs/1912.13318.