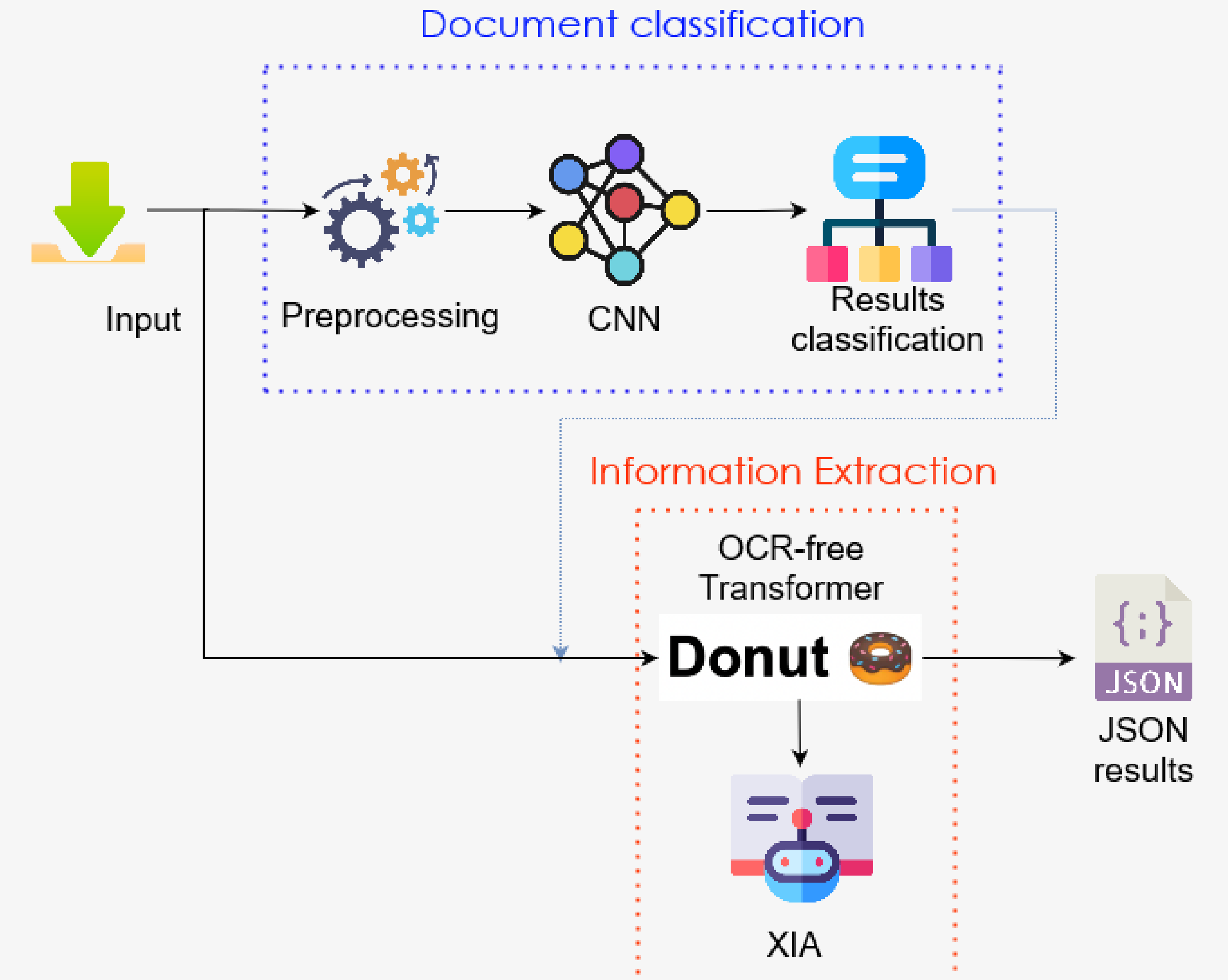
EXTRACTING INFORMATION FROM LEGAL DOCUMENTS

Unlock the full potential of your business with Information Extraction - the key to digital transformation

ARCHITECTURE



PROBLEM STATEMENT

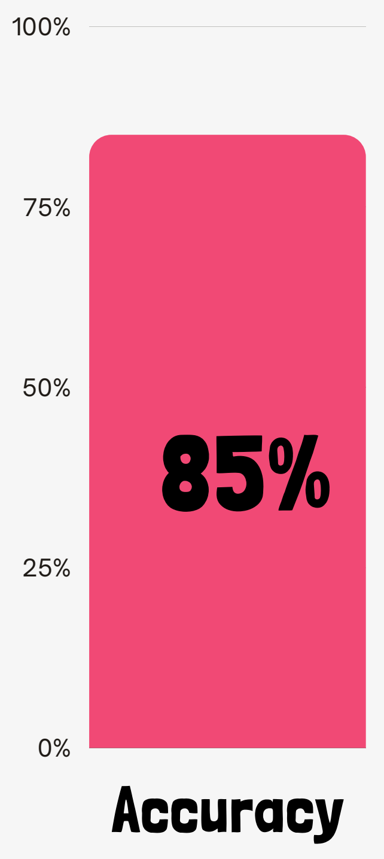
Extracting structured information from different unstructured data sources is a critical step for digitalization. This will improve business and organizations' productivity, cost savings, and decision-making. It leads to streamlining operations and improving overall effectiveness.

CLASSIFICATION

Transfer learning approach using InceptionResNetv2 as a base.

Added custom 'tail' to the model architecture.

Accuracy score: 85%



OBJECTIVE

This project aims to use machine learning approaches to extract relevant and useful information from images of documents and use Explainable AI to understand how the information was extracted.

TYPE OF DOCUMENTS

Invoices

Receipts

TWO AREAS OF DEVELOPMENT

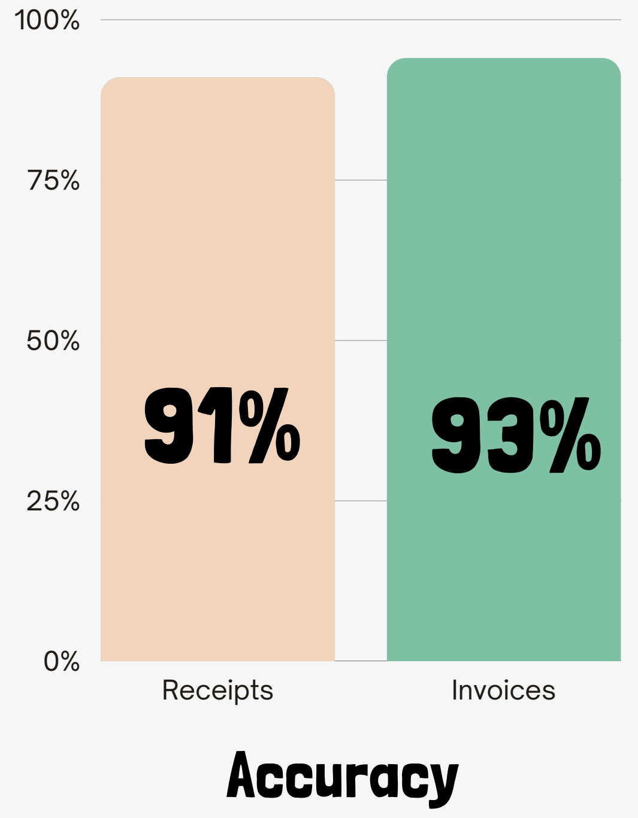
Document classification

Information extraction

INFO EXTRACTION

Receipts: hight quality realistic data.

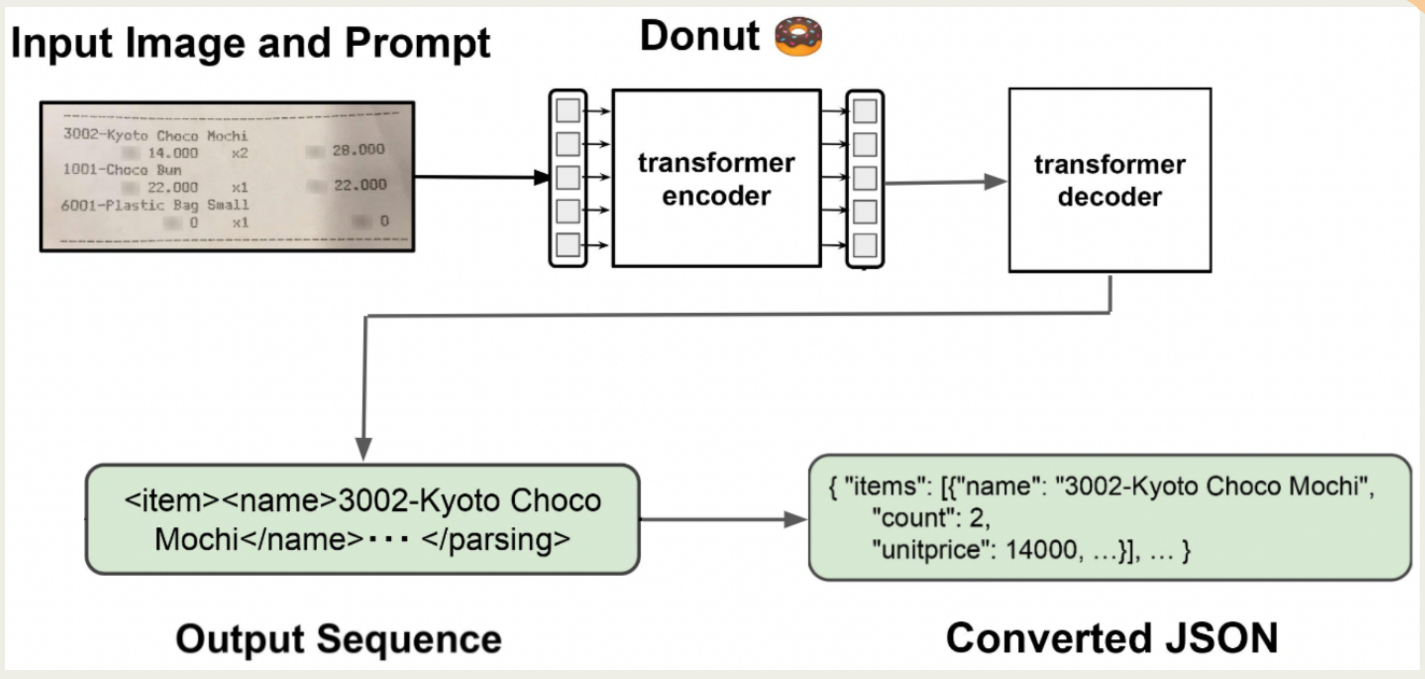
Invoices: Synthetic data, lack of samples, some overfitting.



DONUT

Document understanding transformer, is a new method of document understanding.
OCR-free Transformer model.

Donut does not require off-the-shelf OCR engines/APIs, yet it shows state-of-the-art performances on various visual document understanding tasks, such as information extraction (a.k.a. document parsing).



```
<item><name>3002-Kyoto Choco Mochi</name>··· </parsing>
```

```
{ "items": [{"name": "3002-Kyoto Choco Mochi",
   "count": 2,
   "unitprice": 14000, ...}], ... }
```

NEW DATASET INVOICES

280 New Images
More than 500 invoices in process
32 Different labels of information

Images + Annotated data

JSON

PROCESS OF CREATION

- Search the data
- Labeling the data with Nanonets
- Manually checking

Upload to:

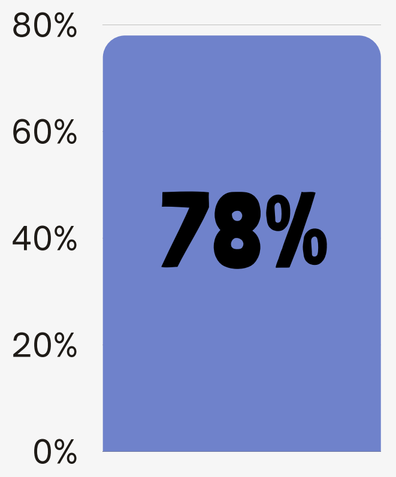
zenodo

EXPERIMENT

Fine tune the base of the Donut extraction model

Using only our own dataset

Accuracy score: 78%



Team 3 & Iconicchain

Debayan Bhattacharya, Juan Carlos Pichardo, Sofiia Charnota

EDISS
MASTER'S PROGRAMME