GSTR Filing Data Extraction

Aditya Tekriwal
Artificial Intelligence and Machine Learning
R.V. College of Engineering
Bengaluru, Karnataka
adityatekriwal.ai22@rvce.edu.in

P Shreyas
Artificial Intelligence and Machine Learning
R.V.College of Engineering
Bengaluru, Karntaka
pshreyas.ai22@rvce.edu.in

Abhinav
Artificial Intelligence and Machine Learning
R.V.College of Engineering
Bengaluru, Karntaka
abhinav.ai22@rvce.edu.in

***Abstract*—Goods and Services Tax (GST) filing is a crucial yet time consuming process for businesses, needing accurate extraction of tax-related details from invoices. This paper presents an ai-driven approach to automate GSTR filing by extracting relevant data from invoice images using deep learning algorithms and Optical Character Recognition (OCR). This approach reduces manual effort, minimizes errors, and accelerates tax compliances, making it a valuable solution for businesses handling large volumes of invoices. By automating the GSTR filing process, this system reduces manual intervention, enhances efficiency, minimizes tax compliances errors, and streamlines financial reporting for enterprises.**

Natural Language Processing (NLP) algorithms further enhance entity recognition to ensure accuracy in tax fields such as GSTIN, invoice number, taxable amount, and tax percentages. The extracted information id validated against GST regulations before being formatted for direct integration into the GSTR filing system/

The system employs a two stage pipeline : first, a computer vision model that detects and annotates key fields using labeled data, and second, an OCR engine that extracts text from these regions for structured data entry. Maintaining records for taxation purposes can be a hassle, especially when dealing with small and mid-cap businesses, where most of the invoices are hand-written and need to be stored in good condition for references. Therefore, a system that automates the process can be a very useful way of saving time, energy and manual costs.

I. Introduction

The Goods and Services Tax (GST) is a compliance indirect tax levied on the manufacture, sale, and consumption of goods and services in India. Implemented in 2017, it replaced a complex web of central and state taxes, aiming to create a unified national market. GST is a dual tax, with both the Central Government (CGST) and the State Governments (SGST) levying taxes on intra-state supplies. For inter-state supplies, the iNtegrated Goods and Services Tax (IGST) is levied by the Central Government.

GST is a multi-stage tax, meaning it's levied at every stage of the supply chain, from manufacturing to final consumption. However, it's designed to avoid the cascading effect of taxes, where taxes are levied on taxes. This is achieved through input tax credit, allowing businesses to claim credit for taxes paid on inputs, reducing the overall tax burden.

GST has simplified the tax system, reduced compliance costs, and improved tax collection. It has also led to a reduction in prices of many goods and services, benefiting consumers. However, GST has also faced challenges, including initial teething problems, complexities in implementation, and concerns about its impact on small businesses.

II. Literature Review

Automated data extraction for Goods and Services Tax Return (GSTR) filing is a rapidly evolving field that combines Optical Character Recognition (OCR), deep learning, and Natural Language Processing (NLP) to streamline tax compliance. Various studies have explored invoice digitization, document processing, and AI-driven automation for financial tasks. This survey presents an overview of existing approaches, techniques, and advancements in automated tax data extraction.OCR plays a vital role in extracting textual data from scanned invoices and digital receipts. Tesseract OCR is widely used due to its open-source nature and adaptability for multiple languages [1]. Studies have compared Tesseract, EasyOCR, and PaddleOCR, highlighting their performance in structured and semi-structured documents [2]. Advanced techniques, such as Long Short-Term Memory (LSTM) models and Transformer-based OCR, have improved text recognition accuracy [3].To enhance OCR accuracy, researchers have explored image preprocessing techniques, including binarization, noise reduction, and contrast enhancement [4]. The use of Convolutional Neural Networks (CNNs) for handwritten and printed text recognition has also demonstrated significant improvements [5].Detecting and segmenting key invoice fields is essential for structured data extraction. YOLO (You Only Look Once) and Faster R-CNN have been widely adopted for object detection-based invoice processing [6].

These models identify regions of interest (ROIs) such as invoice numbers, tax fields, and GSTIN. Studies comparing YOLOv3, YOLOv5, and Faster R-CNN suggest that YOLOv5 provides a better trade-off between accuracy and speed for real-time applications [7].Hybrid approaches, combining OCR with deep learning-based Named Entity Recognition (NER), have been proposed to improve structured text extraction [8]. Bidirectional LSTM (BiLSTM) models and Transformer-based architectures (BERT, LayoutLM) have also been employed to enhance invoice field recognition [9].After text extraction, NLP techniques are used for field classification, validation, and contextual understanding. Research has shown that Named Entity Recognition (NER) models trained on financial documents improve tax-related information retrieval [10].Rule-based approaches, combined with machine learning models like Support Vector Machines (SVM) and Random Forests, have been explored for validating tax calculations and detecting missing fields [11]. Transformer-based models (BERT, RoBERTa) have been fine-tuned on invoice datasets for improved classification accuracy [12].Automated tax filing systems must comply with GST regulations to ensure accuracy and legal validity. Research has focused on building frameworks that integrate AI-based data extraction with GST filing APIs [13]. Studies have proposed rule-based engines for tax validation and fraud detection using machine learning [14].Blockchain-based smart contracts have also been explored to ensure transparent and tamper-proof GST transactions [15].

These technologies aim to minimize errors and automate tax submissions directly to government portals [16].Several comparative studies have benchmarked OCR models, deep learning architectures, and NLP techniques for invoice data extraction [17]. Datasets such as SROIE (Scanned Receipt OCR and Information Extraction), FUNSD, and InvoiceNet have been used to evaluate model performance [18].Research suggests that hybrid models combining CNNs, Transformers, and OCR outperform standalone techniques [19]. The introduction of self-supervised learning and semi-supervised learning has further improved model generalization for diverse invoice formats [20].

The process of Goods and Services Tax Return (GSTR) filing is an essential but highly tedious task for businesses, requiring accurate extraction of tax-related details from invoices. Manually entering data from invoices into tax filing systems is time-consuming and prone to errors. Invoices come in various formats, and businesses often deal with a high volume of transactions, making manual data extraction an inefficient approach. Additionally, errors in GST filing, such as incorrect tax amounts or missing GSTINs, can lead to penalties, compliance issues, and potential legal disputes.

Automating this process with Artificial Intelligence (AI)-driven solutions can significantly reduce errors and improve efficiency. By using Optical Character Recognition (OCR) for text extraction, deep learning-based models for field detection, and Natural Language Processing (NLP) for structured information retrieval, this system can streamline the extraction of key invoice fields such as invoice number, GSTIN, taxable amount, CGST, SGST, IGST, and total tax payable. The proposed system aims to develop a robust AI-powered invoice processing tool that automatically extracts tax-relevant details, validates the extracted data, and formats it for seamless GSTR filing, thereby ensuring compliance with taxation laws and improving business efficiency.

III. Dataset Description

The dataset used for this project is the FATURA dataset, designed to contain invoice images from 13 companies, annotated. It contains the image file as .jpg and the annotation file as .xml structure, which can be used directly to train an object detection model using tools like roboflow, YOLOv8, YOLOv11, etc. The dataset contains 10,000 images of invoices annotated to specify fields like invoice number, client address, client GSTIN, seller GSTIN, Invoice Type, etc.

This dataset is not commonly known, and was developed by the fatura group, in 2020, aiming to train a model that was able to detect and classify these annotation classes.

Dataset Description:

Table 1: Dataset description

Image Resolution	2480X3508 Pixels
Class	**Total Images**
Invoice No	10013
date	10103
seller address	1000
seller tax id	1130
iban	10300
client address	254
client tax id	4540
vat	2011
Total Images	**10300**

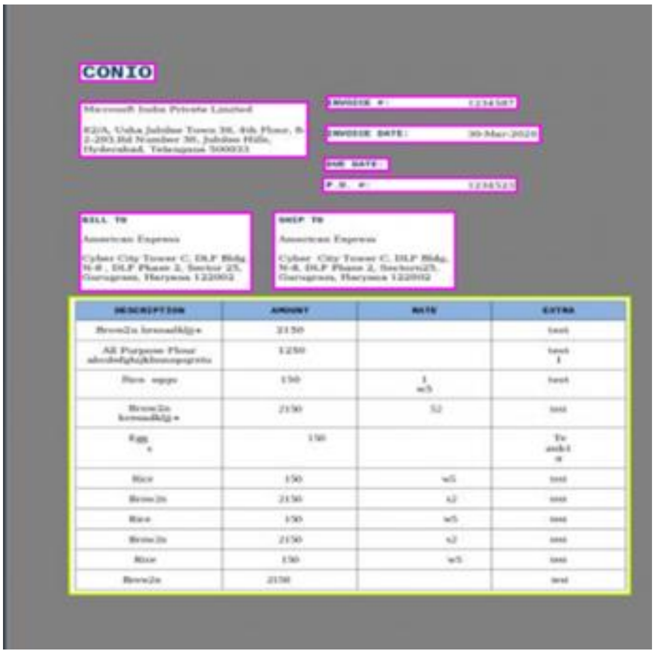


Figure 1: labelled invoices

IV. Methodology And Architecture Diagram

The architecture of the model is structured into three key components: the Object Detection for bounding and labeling our images, the OCR for extracting the text from these bounded and labelled boxes and the data storing part where the data is stored into a temporary file in the computer which can then be used to upload into the GSTR portal through plugins.

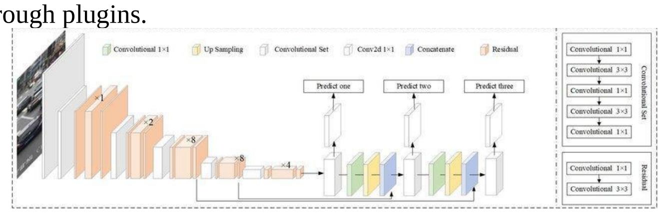


Figure 2: Yolo architecture

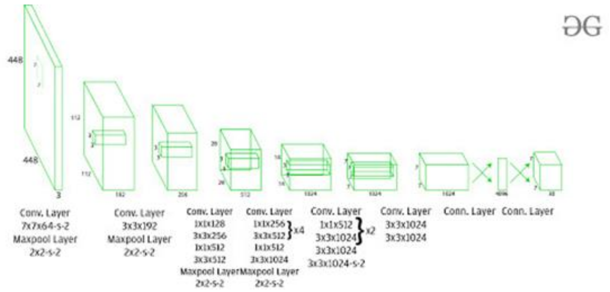


Figure 3: Model architecture

1. Image Preprocessing Module

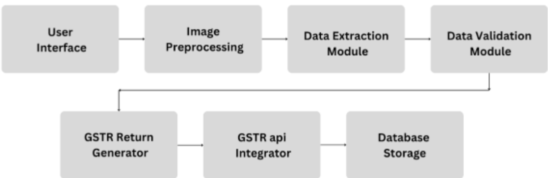
To prepare the invoices images for OCR by improving size and reducing noise. Image resizing to standard dimensions, conversion of colored images to greyscale, noise removal and image enhancement.

2. OCR Processing Module

To extract text from the invoice images using optical character recognition. OCR using libraries like Tesseract or EastOCR, Text extraction and initial formatting.

3. Data Extraction module

To extract relevant data fields from OCR text using deep learning models. Trained deep learning models to detect key fields, extract data such as GSTIN, invoice number, tax amounts, etc.



ACKNOWLEDGMENT *(Heading 5)*

V. RESULTS AND DISCUSSION

EFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.