

# **Contextual Text Extraction From PDF Files Using Machine Learning Practices**

Project report submitted for  
Industrial Project Based Learning in Data Science and Machine Learning

Under the Supervision of

**P.Mohan**  
**Director, Gyan Astra IT Solutions**

**&**  
**Y.V.N Phani Kishore**  
**Director, Gyan Astra IT Solutions**

By

## **Team Members Name (Roll Number)**

P Pranavi	21R11A6742
C Manognasri	21R11A6729
A. Sai Srujana	21R11A05A8
P Moulika	21R11A05E0

Department of Computer Science & Engineering

Nov 2023 - May 2024



**Geethanjali College of Engineering & Technology**

**Accredited by NBA (UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi, NAAC - A+)

# **Geethanjali College of Engineering & Technology**

(Affiliated to J.N.T.U.H, Approved by AICTE, NEWDELHI.)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

## **CERTIFICATE**

This is to certify that Mr./Ms. C.Manognasri, A.Sai Srujana, P Moulika , P Pranavi bearing roll no 21R11A6729,21R11A05A8,21R11A05E0, 21R11A6742 has successfully completed Industrial Project Based Learning in Data Science and Machine Learning held during November 2023 to May 2024 at Geethanjali college of Engineering and Technology.

**Coordinator**

**Mr. E. Mahender**

**Assistant Professor**

**Dept of C.S.E**

**H.O.D - C.S.E**

**Dr. A. Sree Lakshmi**

**Professor & Head**

**Dept of C.S.E**

# ACKNOWLEDGEMENT

We are greatly indebted to the authorities of Geethanjali College of Engineering and Technology, Cheeryal, R.R Dist, for providing us the necessary facilities to successfully carry out this Industrial Project Based Learning titled “Contextual Text Extraction From PDF Files Using MachineLearning Practices” under Value added course.

Firstly, we would like to express our sincere gratitude to our **Principal Dr. S.Udaya Kumar** for providing the necessary infrastructure to complete our Value added course.

Secondly, we thank and express our solicit gratitude to **Dr. A. Sreelakshmi**, Prof & HOD, CSE department, Geethanjali College of Engineering, and technology, for her invaluable help and support which helped us a lot in successfully completing of our course.

We express our gratitude to **P.Mohan**, Director, Gyan Astra IT Solutions, **Y.V.N Phani Kishore** Director, Gyan Astra IT Solutions, **Dr Kamakshiah Kolli**, Associate Professor, Dept of CSE, GCET, **Mr. E.Mahendra**, Dept of CSE, GCET and **N.Madhavi, Assistant Professor**, Dept of CSE-Data Science, GCET for their valuable suggestions and encouragement which helped us in the successful completion of our course.

Finally, we would like to express our heartfelt thanks to our parents who were very supportive both financially and mentally and for their encouragement to achieve our set goals.

## Team Names and roll Numbers.

P Pranavi	21R11A6742
C Manognasri	21R11A6729
A. Sai Srujana	21R11A05A8
P Moulika	21R11A05E0

## INDEX TABLE

<b>S.NO</b>	<b>LIST OF CONTENTS</b>	<b>PAGE NO</b>
1.	ABSTRACT	5
2.	INTRODUCTION	6
3.	PROBLEM STATEMENT	7
4.	OBJECTIVES	8
5.	METHODOLOGY	9
	5.1 Data Source	
	5.2 Efficient Director Details Extraction from PDFs	
	5.2.1 Checking for Details	
6.	Observations	11
7.	Algorithm	12
8.	Implementation	13
9.	Final implementation of model Using Flask Website	15
10.	Final Observation	17

## **1. ABSTRACT**

In this study, we delve into the application of machine learning methods for the extraction of pertinent data from PDF files across different domains. While PDFs hold abundant information, their unorganized format poses challenges for specific detail extraction. Through the integration of machine learning techniques, our objective is to craft algorithms capable of autonomously discerning and extracting relevant textual content from PDF documents. Our methodology entails the training of machine learning models on a curated dataset of annotated PDFs, wherein the pertinent text has been meticulously labeled. These models undergo training to discern intricate patterns and structures within the PDFs, thereby facilitating accurate extraction of desired information.

Through rigorous experimentation and thorough evaluation, we gauge the efficacy of our machine learning models in extracting contextual text from PDF files. Additionally, we investigate methodologies aimed at enhancing both the precision and speed of the extraction process. These techniques encompass feature engineering and model optimization, all geared towards furnishing a dependable and scalable solution for harvesting valuable information from PDFs through machine learning methodologies. Our ultimate aim is to offer a versatile solution applicable across diverse domains and applications.

## 2.INTRODUCTION

The project is dedicated to harnessing machine learning techniques to tackle the common challenge of efficiently extracting specific information from a range of PDF documents, such as director names, their type (independent or executive), and their Director Identification Number (DIN). PDFs are prevalent across various domains, and extracting information from them is pivotal for tasks like legal analysis, data processing, and research. However, manual extraction is time-consuming and prone to errors. By leveraging machine learning, the project aims to automate this process, enabling swift and accurate extraction of pertinent information from PDF files.

Machine learning models excel at discerning patterns and structures within PDF documents, including text layout and positioning, thereby facilitating efficient extraction while taking neighboring text into account. This approach offers a scalable solution capable of handling diverse PDF formats and structures. The project's objective is to devise a solution that seamlessly extracts the necessary information, such as the director's name, type, and DIN, with minimal manual intervention. Such a solution would significantly enhance efficiency, particularly in scenarios involving the processing of large volumes of PDFs, such as analyzing bank statements or tax forms.

By overcoming the challenges associated with unstructured data in PDFs, the project aims to streamline the extraction process, ultimately improving accessibility and readability for users with visual impairments or difficulties with small or blurred text. Leveraging machine learning techniques presents a promising solution, enabling accurate text recognition and extraction regardless of the PDF's layout or structure. Machine learning models can learn to interpret text layout and positioning within PDFs, leveraging contextual clues from neighboring text to enhance their ability to understand document structures effectively.

Through the training of machine learning algorithms on annotated datasets containing labeled information, such as director names and their associated attributes, the project seeks to develop robust extraction models capable of accurately identifying and extracting the requisite information. Additionally, the project aims to explore various machine learning applications tailored to text extraction from PDFs, including natural language processing (NLP) and optical character recognition (OCR), to preprocess PDF files and optimize extraction performance.

Through meticulous exploration of text PDFs and iterative development of extraction solutions, the project endeavors to streamline the process of retrieving critical information, ultimately enhancing accessibility, readability, and efficiency across diverse domains such as legal documentation.

### **3.PROBLEM STATEMENT**

Frequently, the necessity arises to extract data from PDFs, especially considering the importance of accessibility and readability for individuals with vision impairments or difficulty reading small or blurred text. This requirement is crucial across various contexts, including legal proceedings, data analysis, and research. Extraction becomes necessary when utilizing text or image content from PDFs in other documents, aiming to save time and prevent errors.

Machine Learning (ML) techniques stand out as optimal methods for PDF extraction due to their ability to achieve highly accurate text recognition and extraction, regardless of the file structure. These models effectively capture information regarding both the layout and position of text, taking neighboring text into account, which enhances their ability to generalize and learn document structure efficiently.

The challenge of efficiently extracting specific information from a multitude of PDFs is a common obstacle encountered by numerous applications and industries. Extracting information from bank statements or tax forms, for example, poses significant difficulties. The conventional method of manually sifting through numerous PDFs is time-consuming and prone to producing inaccurate or inconsistent data. Furthermore, the presence of unstructured data in PDFs complicates the task for automated systems attempting to extract the necessary information.

Designed solution should be able to extract the Name of the Director.

Director type : Independent / executive

DIN

## **4.OBJECTIVES**

1. Devise a machine learning-based solution tailored to extract precise details from PDF documents, with a particular focus on retrieving the Director's Name, their classification (Independent or Executive), and their unique Director Identification Number (DIN).
2. Automate the extraction process to enhance the accessibility and legibility of PDF files, catering to individuals with vision impairments, while also optimizing efficiency across legal, research, and data analysis endeavors.
3. Address the hurdles presented by unstructured data within PDFs by employing advanced ML techniques adept at accurately discerning and extracting text, irrespective of the file's layout or organization.
4. Offer a scalable and efficient solution capable of widespread application,



## 5. METHODOLOGY

### 5.1 Data Source

#### ➤ Brief description of the data source

The provided data source outlines the necessity and challenges associated with extracting data from PDF files, highlighting the importance of accessibility and readability, particularly for individuals with vision impairments or difficulties reading small text. It emphasizes the significance of accurate text extraction for various purposes such as legal matters, data analysis, and research, noting instances where extraction is required, such as retrieving information from bank statements or tax forms.

Furthermore, the text underscores the effectiveness of machine learning (ML) techniques in addressing these challenges by enabling highly accurate text recognition and extraction from PDF files, regardless of their structure. ML models are noted for their ability to understand the layout and position of text within PDFs, as well as the contextual information surrounding it, which enhances their capability to generalize and learn document structures efficiently.

The challenges associated with manual extraction processes are also highlighted, including the time-consuming nature and potential for producing inaccurate or inconsistent data. The unstructured nature of data within PDFs further complicates automated extraction efforts.

### 5.2 Efficient Director Details Extraction from PDFs

This Python script utilizes the SpaCy library for named entity recognition (NER) to extract person names from PDF files. Subsequently, it filters these names based on their occurrence in sentences containing the term "director". For each extracted director name, the script identifies their classification (Independent/Executive) and retrieves associated Director Identification Numbers (DINs). DINs are detected using regular expressions and their proximity to the director's name within the text. Finally, the script outputs the extracted director names along with their classifications and corresponding DINs. This script serves as an automated tool for extracting director-related information from PDF documents, thereby facilitating streamlined processes in diverse domains such as financial analysis, legal documentation, and corporate governance.

#### 5.2.1 Checking for Details:

equitas.pdf

```
Director Name: Anil Kumar
Director Type: Executive
DIN Numbers:
-----
Director Name: Arun Ramanathan
Director Type: Executive
DIN Numbers:
- 00308848
```

polyplex.pdf

```
Director Name: Yogesh Kapur
Director Type: Executive
DIN Numbers:
-----
```

Triveni1.pdf

```
... Director Name: Amrita
Gangotra
Director Type: Independent
DIN Numbers:
- 08333492
-----
Director Name: Sonu Halan Bhasin
Director Type: Independent
DIN Numbers:
-----
Director Name: Dhruv M. Sawhney
Director Type: Executive
DIN Numbers:
-----
Director Name: Mahindra First Choice Wheels
Director Type: Executive
DIN Numbers:
-----
Director Name: Amrita Gangotra
Director Type: Independent
DIN Numbers:
-----
Director Name: Nikhil Sawhney
Director Type: Independent
DIN Numbers:
...
Director Name: Hosier Complex
Director Type: Executive
DIN Numbers:
```

Vinati.pdf

```

Director Name: Prashant Barve
Director Type: Executive
DIN Numbers:
-----
Director Name: M. Lakshmi
Director Type: Executive
DIN Numbers:
-----
Director Name: J. C. Laddha
Director Type: Executive
DIN Numbers:
-----
Director Name: Item Nos
Director Type: Independent
DIN Numbers:
-----
Director Name: Lagnam Spintex Limited
Director Type: Independent
DIN Numbers:
-----

```

## Varun.pdf

```

... Director Name: Raj Gandhi
Director Type: Executive
DIN Numbers:
- 00003649
-----
Director Name: Varun Jaipuria
Director Type: Executive
DIN Numbers:
-----
Director Name: Rajinder Jeet Singh Bagga
Director Type: Executive
DIN Numbers:
- 08440479
-----
Director Name: Naresh Trehan
Director Type: Independent
DIN Numbers:
-----
Director Name: Bagga Ravi Batra
Director Type: Executive
DIN Numbers:
-----

```

## 6. Observations:

- Names undergo filtration based on their occurrence within sentences containing the term "director," signifying their relevance to director-related data.
- Upon extraction of each director name, the script employs a straightforward rule-based methodology to ascertain their director type (Independent/Executive) by examining the contextual cues in the surrounding text.
- DIN numbers linked to each director are extracted utilizing regular expressions, assuming a standard format of 8-digit numbers. The script identifies DINs in close proximity to the director's name within the text.
- The script systematically prints the extracted director names, their respective types, and associated DIN numbers separately, offering a concise summary of the director-related information gleaned from the PDF document.

## 7. ALGORITHM

NER (Named Entity Recognition) model:

Leveraging SpaCy, a renowned natural language processing (NLP) library, for named entity recognition (NER) involves the following steps:

- 1. Data Preparation:** Text data extracted from PDF files undergoes preparation for NER processing.
- 2. Named Entity Recognition (NER):** Utilizing SpaCy's pre-trained model, such as "en\_core\_web\_sm," entities like persons, organizations, and locations are identified in the text. Although SpaCy doesn't rely on traditional machine learning algorithms for NER, it internally utilizes deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to train its models.
- 3. Filtering:** Extracted entities are filtered based on specific criteria, like being labeled as a "PERSON" and having a length greater than one, effectively eliminating single-word entities such as "Mr." or "Dr."
- 4. Additional Processing:** Further refinement of results is achieved through processes like checking for uppercase initials and filtering based on the presence of the word "director" in the text.

Overall, while traditional machine learning algorithms aren't explicitly employed in this code snippet, SpaCy's NER capabilities leverage deep learning techniques behind the scenes to extract person names from text data.

## 8.Implementation

```
tractor > views.py > ...
from flask import Flask, render_template, request
import pdfplumber
import re
import spacy
app = Flask(__name__)
nlp = spacy.load("en_core_web_sm")

def extract_text_from_pdf(pdf_file):
    with pdfplumber.open(pdf_file) as pdf:
        text = ""
        for page in pdf.pages:
            text += page.extract_text()
        return text

def extract_director_detail(text):
    director_details = []
    director_pattern = r"(?:Dr\.|Mr\.|Prof\.|Mr\.|Ms\.)(\w+)[DIN: (\d{8})"
    director_info = re.findall(director_pattern, text)
    director_details = []

    # Define patterns for DIN and director types
    din_pattern = re.compile(r'\b\d{8}\b') # Assuming DIN is an 8-digit number
    director_type_pattern = re.compile(r'independent|executive', re.IGNORECASE)

    # Use sets to store unique director names and DIN numbers
    unique_director_names = set()
    unique_din_numbers = set()

    for ent in nlp(text).ents:
        if ent.label_ == "PERSON":
            director_name = ent.text
            director_text = text[text.find(director_name) - 100:text.find(director_name) + len(director_name) + 100]

            # Classify director type
            if "independent" in director_text.lower():
                director_type = "Independent"
            else:
                director_type = "Executive"

            # Extract DIN for director
            din_matches = din_pattern.findall(director_text)
            for din in din_matches:
                # Check if director's name is in the text near the DIN number
                if director_name.lower() in director_text[director_text.find(din) - 50:director_text.find(din) + 50].lower():
                    # Check if the DIN number is unique
                    if din not in unique_din_numbers:
                        director_details.append({
                            'director': director_name,
                            'director_type': director_type,
                            'din': din
                        })
                        # Add the DIN number to the set of unique DIN numbers
                        unique_din_numbers.add(din)

            # Check if the director name is unique
            if director_name not in unique_director_names:
                # Add the director name to the set of unique director names
                unique_director_names.add(director_name)

    return director_details

@app.route('/')
def index():
    return render_template('index.html')

@app.route('/upload', methods=['POST'])
def upload_file():
    if 'file' not in request.files:
        return render_template('results.html', error='No file part')
    file = request.files['file']
    if file.filename == '':
        return render_template('results.html', error='No selected file')
    if file:
```

```
def upload_file():
    if file:
        text = extract_text_from_pdf(file)
        director_details = extract_director_detail(text)
        return render_template('results.html', director_details=director_details)

if __name__ == '__main__':
    app.run(debug=True)
```

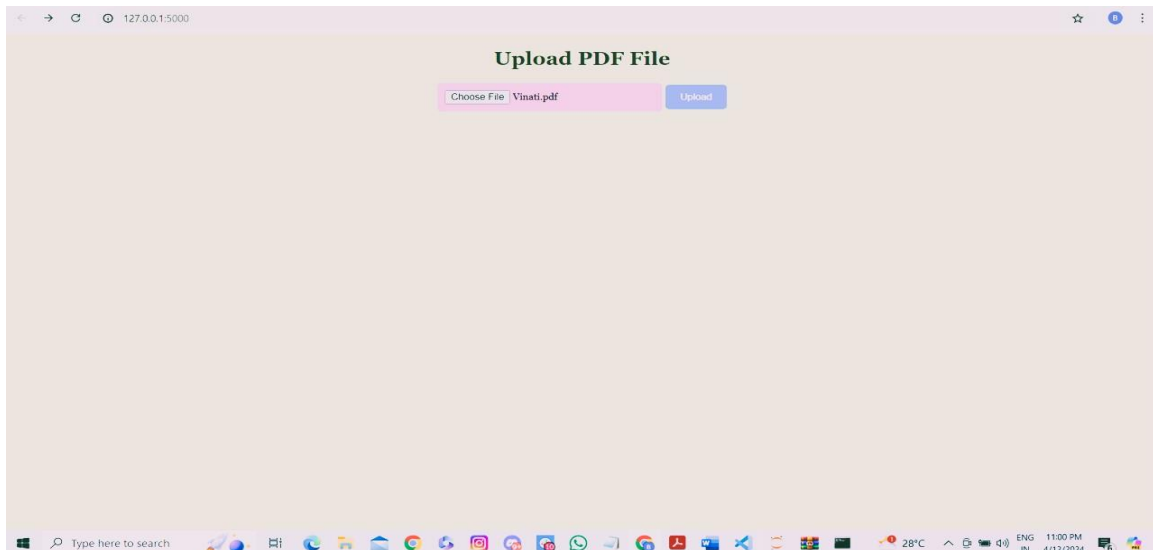
### Observations:

- The Flask web application is designed to extract director details from uploaded PDF files and display the results using HTML templates.
- It utilizes the Flask framework to create a web server, with routes defined for the homepage ("/") and file upload ("/upload").
- The application imports the required libraries, including Flask, pdfplumber, regular expressions (re), and SpaCy for natural language processing (NLP).
- The extract\_text\_from\_pdf function extracts text from a PDF file using pdfplumber library.
- The extract\_director\_detail function processes the extracted text to identify director details, including names, types (Independent/Executive), and Director Identification Numbers (DINs).
- Director names are identified using SpaCy's NER capabilities for PERSON entities. The extracted director details are stored in a list of dictionaries.
- Upon file upload, the Flask application processes the uploaded PDF file, extracts director details, and renders the results using an HTML template.
- Error handling is implemented to manage scenarios where no file is uploaded or no file name is selected.

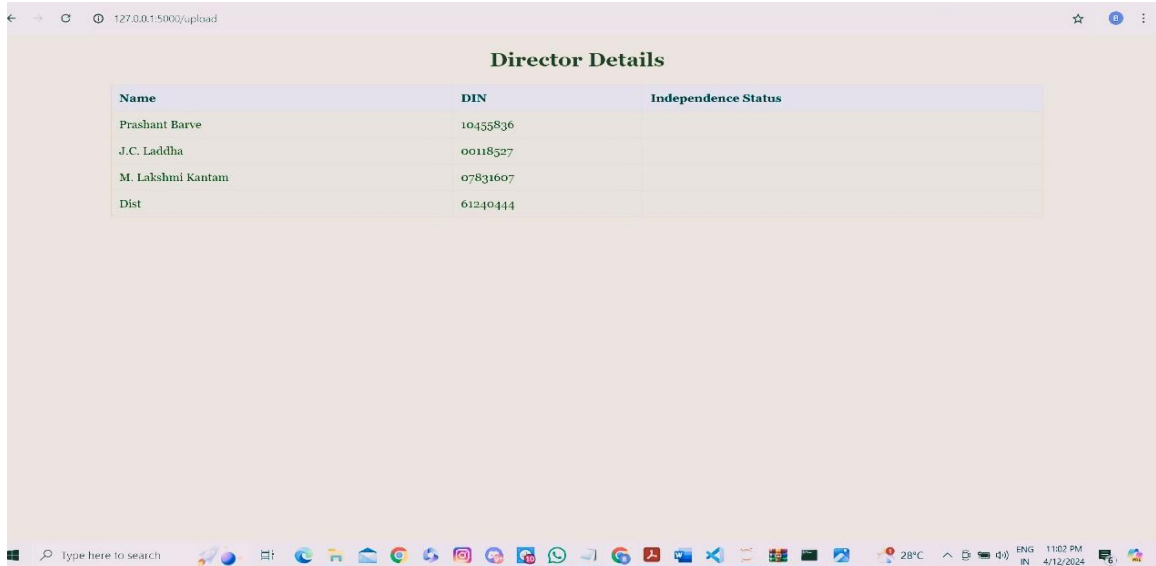
The application is run using the `app.run(debug=True)` command, enabling debugging mode for development purposes.

## 9. Final implementation of model Using Flask Website

We need to upload a pdf file,such that we can get Director's details:



Vinati.pdf  
(output)



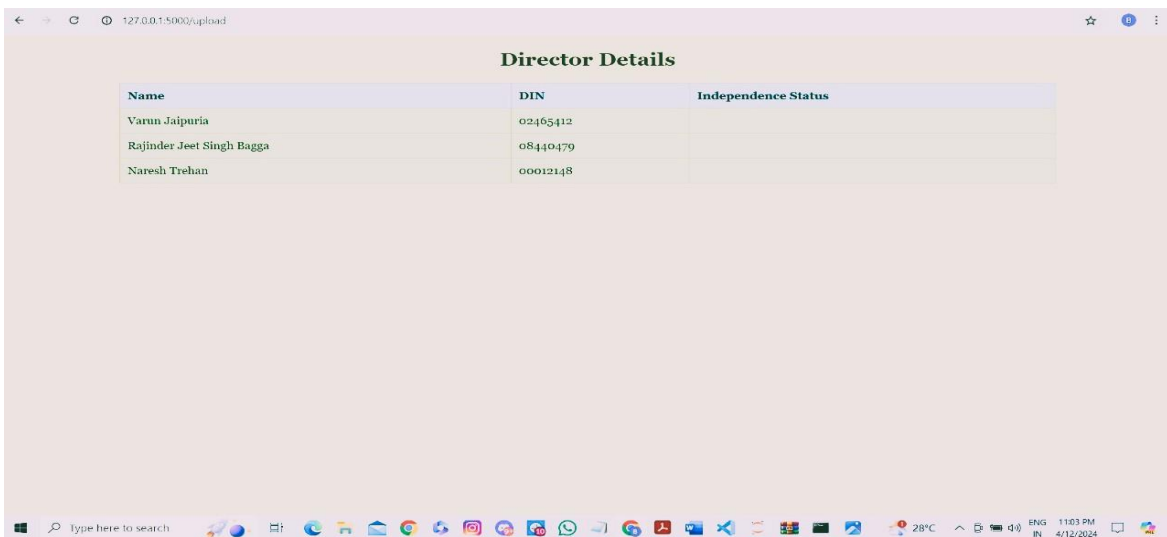
127.0.0.1:5000/upload

Name	DIN	Independence Status
Prashant Barve	10455836	
J.C. Laddha	00118527	
M. Lakshmi Kantam	07831607	
Dist	61240444	

Type here to search

28°C 11:03 PM 4/12/2024

Varun.pdf  
(output)



127.0.0.1:5000/upload

Name	DIN	Independence Status
Varun Jaipuria	02465412	
Rajinder Jeet Singh Bagga	08440479	
Naresh Trehan	00012148	

Type here to search

28°C 11:03 PM 4/12/2024

triveni2.pdf  
(output)



Name	DIN	Independence Status
Sonu Halan	02872234	
Amrita Gangotra	08333492	
Asabov Copyto	01961162	
Dhruv M. Sawhney	00102999	
Homai A. Daruwalla	00365880	

## 10.Final Observation

When employing machine learning methods to extract contextual text from PDF files, the resulting output usually comprises structured data encompassing the extracted details. For each director identified within the PDF files, the output typically presents:

- **Director Name:** The full name of the director extracted from the PDF.
- **Director Type:** The type of directorship, categorized as either "Independent" or "Executive," based on the context in which the director's name appears.
- **DIN (Director Identification Number):** The unique identification number associated with the director, extracted from the text.