## **Jaypee Institute of Information Technology, Noida**

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AND INFORMATION TECHNOLOGY

## **Minor Project Synopsis**



**Project Title : IndicText OCR**

**Submitted to: Dr Aastha Maheshwari**

## 

| **S.No** | **Name** | **Enroll No.** | **Batch** |
| --- | --- | --- | --- |
| **1.** | **Teesha Kakkar** | **22103098** | **B2** |
| **2.** | **Rishika Aggarwal** | **22103046** | **B2** |
| **3.** | **Mehul Bansal** | **22103048** | **B2** |

**Program: BTech CSE**

**Semester and Year: 5th and 3rd Year**

Develop a powerful Optical Character Recognition (OCR) system for digitizing text in Indian languages. The system will help convert physical documents into digital text, making it easier to access and use.  
  
**Key Features**:

* **Language Support**: Recognizes multiple Indian languages like Hindi, Tamil, Bengali, etc., with models tailored for each language.
* **Data Collection**: Collects a diverse set of documents like aadhar card, voter id, driving license, pan card, etc.
* **Model Development**: Uses advanced deep learning methods for accurate text recognition and image processing.
* **System Integration**: Provides an easy-to-use interface for uploading documents, extracting text, and editing results.

### **Expected Outcomes:**

* **High Accuracy**: Precise OCR for multiple Indian languages.
* **Enhanced Accessibility**: Easier access to content in regional languages.
* **Versatile Performance**: Reliable text recognition across various document types.

**Technologies**:

* **Deep Learning**: Utilizes TensorFlow and PyTorch for training models and improving accuracy.
* **OCR Library**: Uses Tesseract OCR for text recognition.
* **Web Development**: Employs React for the frontend and Node.js for the backend of the web interface.

### **Additional Features & Considerations:**

* **Verification System**: Implement a verification system to ensure the accuracy of the digitized text. This could include unique IDs or checksums for each document processed.
* **Text Validation**: Develop a mechanism to verify the correctness of the input text before processing. This could involve cross-referencing with existing databases or using validation rules.
* **Storage Solutions**: Since the focus is on text extraction, using efficient storage solutions like cloud databases or dedicated document management systems can help handle the volume of data.

**Conclusion:**

The IndicText OCR project aims to significantly advance the digitization of Indian languages, making text extraction from physical documents more accessible and accurate. By leveraging state-of-the-art deep learning techniques and integrating user-friendly features, this project will enhance accessibility to regional content and improve document handling across diverse sectors. The end result will be a robust OCR system that supports multiple languages and delivers reliable performance, contributing to the broader goal of digital inclusion in India.