OCR MODEL

Team 9

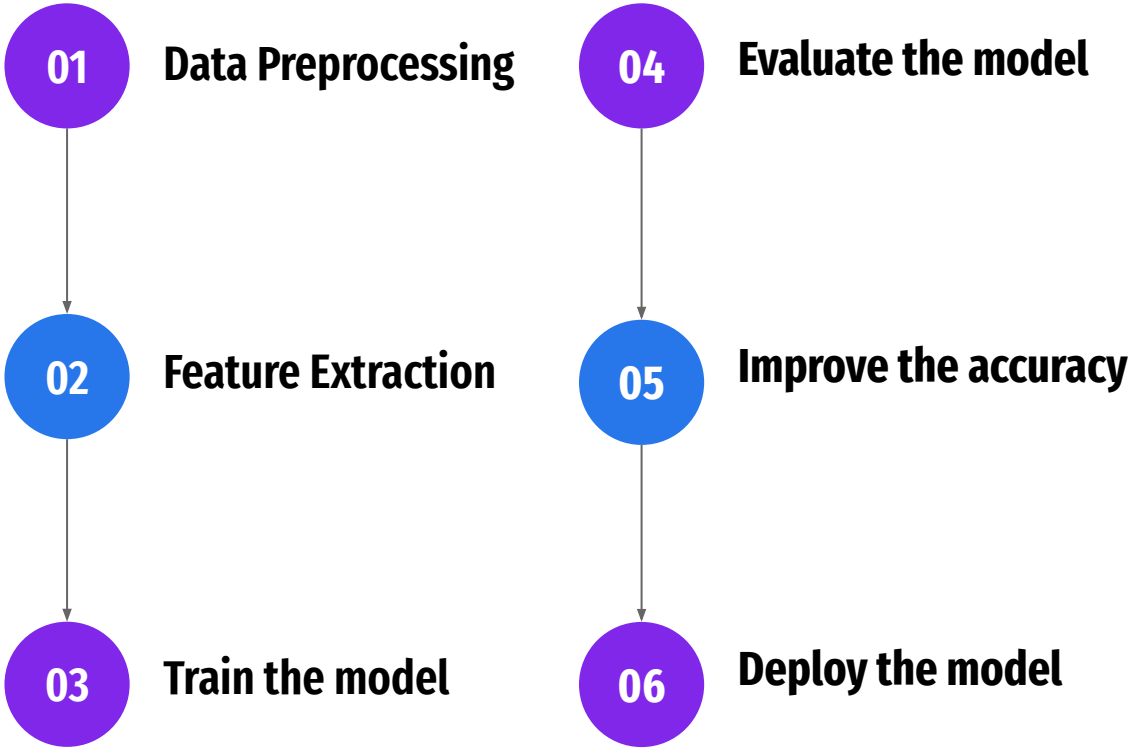
Gunjan Agrawal
Khyati Karia

TABLE OF CONTENTS

INTRODUCTION

- OCR (Optical Character Recognition) utilizes machine learning techniques to extract text from images.
- ML algorithms are applied in image preprocessing, text localization, character segmentation, and character recognition stages.
- The project involves training ML models on a dataset of alphabetic characters to accurately recognize them.
- The system's robustness is evaluated by predicting characters in new images.
- The objective is to achieve accurate character recognition and validate the effectiveness of the implemented ML techniques in OCR.

Flowchart for building OCR model



Comparison

Logistic Regression

- Assumes a linear decision boundary
- Struggle with collinearity between features
- Statistical regression model used for binary classification tasks

Support Vector Machine

- Can model nonlinear decision boundaries by combining multiple decision trees
- Effective for high-dimensional data
- Supervised learning algorithm that performs both classification and regression tasks

Random Forest

- Can model nonlinear decision boundaries by combining multiple decision trees.
- Capture interactions between features.
- Ensemble learning algorithm that uses a collection of decision trees for classification or regression

Challenges in the problem

01 High image noise levels

02 Limited training dataset availability

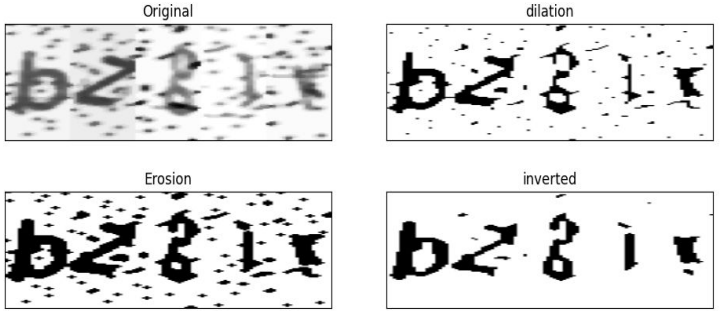
03 Generalizing to new images

04 Handling character appearance variability

Accuracy from Different Models

Models	
Logistic Regression	62.50%
Random Forest	82.40%
Support-vector machines	83.00%
KNN	74.30%
CNN	**87.70%**
Ensemble	80.07%

Methods To Improve Accuracy



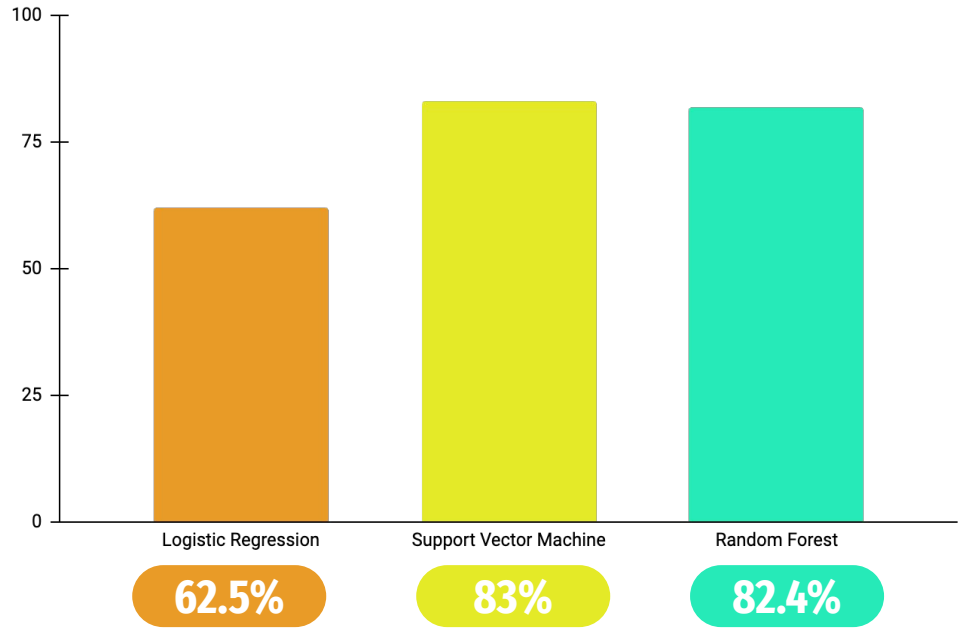
Pre-Processing

- **Median Filter** — Increased accuracy almost by 2%
- **Image contrasting** — Filtered out minute noise and increased the accuracy
- **Kernel** — Transforms data into higher dimensional feature space
- **Gaussian Filter** — Provides balance between noise reduction and feature preservation

Best Model

Support Vector Machine

It performed the best with respect to accuracy



Parameters for models

Support Vector Machine

C = 100 and gamma = 0.01

- Regularization parameter (C)
- Kernel coefficient (gamma)
- Kernel (kernel parameter)

Random Forest

Number of estimators = 200

- Number of estimators
- Maximum Depth

Logistic Regression

Penalty = Ridge

- Penalty
- Solver

Other Algorithms

KNN

It utilizes nearest neighbor search to classify data based on their similarity.

CNN

High accuracy image classifiers, leveraging hierarchical feature learning.

Ensemble Learning

It combines multiple models to improve prediction accuracy and generalization.

Thank You !