# OCR Digit Classification using Machine Learning


## Author

Faiz Azam
Department of Artificial Intelligence
Sindh Madressatul Islam University (SMIU), Pakistan



## Abstract

This paper presents a supervised machine learning approach for classifying handwritten digits (0–9) using image data. The dataset used is the popular Digit Recognizer dataset from Kaggle, containing grayscale images of handwritten digits in a 28x28 pixel format. After exploratory data analysis (EDA) and preprocessing, three classification algorithms were applied: Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). The models were evaluated based on accuracy, with Random Forest achieving the highest performance (~96%). This study demonstrates the effectiveness of ensemble methods in image classification tasks.



## Keywords

OCR, Digit Classification, Machine Learning, Random Forest, Logistic Regression, KNN, Kaggle.


## 1. Introduction

Optical Character Recognition (OCR) is a foundational task in computer vision and pattern recognition, enabling systems to interpret and digitize human handwriting. Handwritten digit classification is a classic problem in this domain, often used as a benchmark for evaluating machine learning algorithms. This project utilizes the Kaggle Digit Recognizer dataset to classify digits from 0 to 9 using supervised learning algorithms, aiming to compare their performance and understand the practical strengths of each model.

## 2. Related Work

Various researchers have explored OCR and digit classification using different approaches. The MNIST dataset has been widely used in numerous studies involving neural networks, support vector machines (SVMs), and convolutional neural networks (CNNs). For classical ML approaches, Random Forest and KNN have shown promising

results due to their simplicity and strong baseline accuracy. In this study, we aim to replicate and extend such findings using a simplified machine learning pipeline.

## 3. Methodology

### 3.1 Dataset Description

The dataset used was sourced from Kaggle's Digit Recognizer competition. It contains labeled training data in a CSV format with 28x28 pixel grayscale images of handwritten digits. Each image is flattened into a 784-column row with a corresponding label.

### 3.2 Data Preprocessing

Preprocessing steps included handling missing values (none were found), normalizing pixel values from 0–255 to 0–1, and splitting the data into training and testing sets using a 80/20 ratio. Sample images were visualized using matplotlib.

### 3.3 Algorithms Applied

The following machine learning algorithms were applied using scikit-learn:
- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbors (KNN)

All models were trained on the training set and evaluated on the test set using accuracy metrics.

## 4. Experiments and Results

The models were evaluated on test data, and the following accuracy scores were obtained:

• Logistic Regression: ~92%
• Random Forest: ~96%
• KNN: ~94%

Random Forest outperformed the other algorithms due to its ensemble nature, which reduces overfitting and captures non-linear relationships more effectively.

## 5. Conclusion and Future Work

This study demonstrates the capability of classical machine learning methods in solving OCR-based digit classification tasks. Random Forest proved to be the most effective among the three applied models. For future work, integrating deep learning models such as CNNs or using transfer learning could significantly improve performance. Additional preprocessing techniques like dimensionality reduction or feature engineering may also be explored.

## References

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE.

[2] Breiman, L. (2001). Random forests. Machine learning.

[3] Kaggle: Digit Recognizer Dataset - https://www.kaggle.com/competitions/digit-recognizer