Handwritten Digit Recognition with Machine learning

Participant:

Kavya Konakati

Mentor:

Dr. Karthikeyan E

ABSTRACT:-

- Handwritten Digit Recognition is the capacity of a computer to interpret the manually written digits from various sources like messages, bank cheques, papers, and pictures, and in various situations for web-based handwriting recognition on PC tablets, identifying number plates of vehicles, handling bank cheques, digits entered in any forms.
- We performed various classifications on MNIST dataset to determine which is the most accurate for real-time applications .
- Handwritten digit recognition is the process to provide the ability to machines to recognize human handwritten digits. It is not an easy task for the machine because handwritten digits are not perfect, vary from person-to-person, and can be made with many different fonts.
- The main objective of this project is to ensure effective and reliable approaches for recognition of handwritten digits.

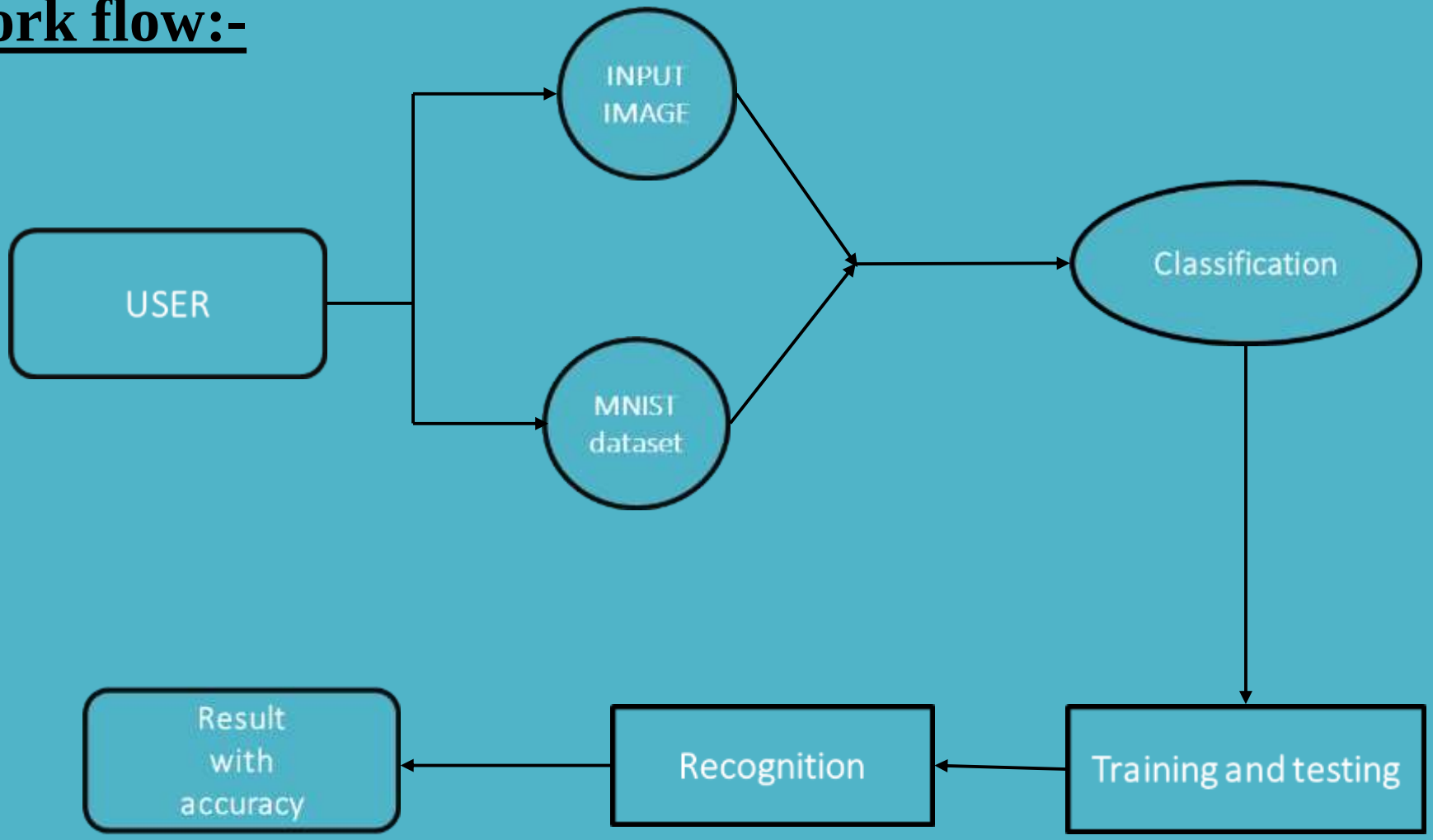
INTRODUCTION :-

- Machine Learning is the art of programming computers to make them learn from data, Systems that can learn new concepts and tasks, systems that can learn by experience and systems that can understand the data
- Machine Learning provides us various methods through which human efforts can be reduced in recognizing manually written digits.
- Handwriting recognition has been one of the most fascinating and challenging research areas in field of image processing and pattern recognition. The Handwritten Digit Recognition system is being developed and tested with MNIST dataset.

MNIST DATASET :-

- The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems
- We are using the MNIST Data Set, which consists of 28 by 28 pixels per image, with 10,000 rows and 785 columns.
- MNIST dataset has a training set of 60,000 examples and test set of 10,000 examples.
- https://www.kaggle.com/datasets/oddrationale/mnist-in-csv

Work flow:-


INPUT IMAGE
USER
MNIST dataset
Classification
Result with accuracy
Recognition
Training and testing

CHALLENGES IN HANDWRITING RECOGNIZATION

1.Huge variability and ambiguity of strokes from person to person.

2.Handwriting style of an individual person also varies time to time, and it is inconsistent.

3.Poor quality of the source document/image due to degradation over time.

4.Cursive handwriting makes separation and recognition of characters challenging.

USE CASES

- Postal mail sorting
- Banking (bank check processing)
- Form data entry
- We can use these algorithms in hospitals application for detailed medical diagnosis, treatment and monitoring the patients
- Autonomous Cars
- License Plate readers for parking structures/security cameras

HANDWRITTEN RECOGNITION METHODS

Online Methods :- Online methods involve a digital pen/stylus and have access to the stroke information, pen location. they tend to have a lot of information with regards to the flow of text being written they can be classified.

Offline Methods :- Offline methods involve recognizing text once it's written down, hence won't have information to the strokes/directions involved during writing.

In real world it's not always possible/scalable to carry a digital pen with sensors to capture stroke information and hence the task of recognizing text offline is a much more relevant problem.

Thus, we use various methods of classifications to solve the problem.

METHODS USED

1. Logistic Regression
2. Naive Bayes
3. K-Nearest Neighbors
4. Decision Tree
5. Support Vector Machines
6. Random Forest

LOGISTIC REGRESSION:

- Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
- It is a form of regression where the target variable is binary. What Logistic Regression is great for is an initial benchmark model on a binary classification problem with well-behaved data. It is one of the more transparent algorithms and doesn’t work well with really messy data.
- The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.

Naive Bayes:

- Naive — Bayes is a classifier which uses Bayes Theorem. It calculates the probability for membership of a data-point to each class and assigns the label of the class with the highest probability.
- Naive Bayes is one of the fastest and simple classification algorithms and is usually used as a baseline for classification problems.
- It performs well with clean and noisy data.

K- Nearest Neighbour:

- KNN is an instance-based learning algorithm. There are two main benefits of using KNN algorithm, that is, it is robust to complex training data, and it is very efficient if the data is very large in size.
- It is a supervised learning algorithm for multiclass classification.
- The algorithm considers a new data point as its input and performs classification by calculating distance between new and labeled data points using the Euclidean distance formulas.

Decision Tree:

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome It is used to categorize or make predictions based on how a previous set of questions were answered.

Advantages:

- Works for numerical or categorical data and variables
- Model problems with multiple outputs
- Tests the reliability of the tree

Support Vector Machine:

- In this type of algorithm, there are data items which are considered as points in an n-dimensional space.
- This classifier finds the hyper plane by performing classifications between the two classes. One of the main advantages of this algorithm is that it provides a regularization parameter which avoids the over fitting problems.

Random Forest:

- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- A Random Forest merges a collection of independent decision trees to get a more accurate and stable prediction. It says that there is an immediate connection between the total number of trees and the result it gets. This classifier can deal with the missing quantities.

Report :-

CLASSIFIER	ACCURACY
Logistic Regression	87.35%
Naive Bayes	54.75%
K-Nearest Neighbour	95.1%
Decision Tree	81.75%
Support Vector Machine	85.15%
Random Forest	95.35%

Conclusion:-

- We used here 6 different approaches for supervised Machine learning Classification. By these, we found the accuracy in the data set in each method.

- The K-nearest neighbor algorithm is fast to train the data but is slow to compute the results.

- On the other hand, the Random Forest is faster to classify the data. The results obtained with Linear SVC were less good, but this could probably be improved by using better parameters.

- We can increase the accuracy by Taking huge datasets • Adopting many suitable algorithms • Hyper-parameter tuning • Compile the model with a greater number of epochs