Handwriting to Digital Text Conversion (H2DTC)

JV Aditya (12140840),C Nikhil(12140530), B Sri Bhargav Ram (12140460)

---

1. Data Pre-processing : We have utilized the MNIST and EMNIST datasets, which together comprise over four hundred thousand handwritten words
Character Recognition utilizes image processing technologies to convert characters on scanned documents into digital forms. It typically performs well in machine-printed fonts.
There are 206,799 words in total. The data was divided into a training set (331,059), testing set (41,382), and validation set (41,382) respectively.

```
# Extract features from images
n_samples = len(train_X)
n_features = img_size * img_size
train_X = np.array(train_X).reshape(n_samples, -1)

n_samples_val = len(val_X)
val_X = np.array(val_X).reshape(n_samples_val, -1)

# Scale the data
n_components = 64  # Adjust the number of components as needed
pca = PCA(n_components=n_components, svd_solver='randomized', whiten=True).fit(train_X)

train_X = pca.transform(train_X)
val_X = pca.transform(val_X)
```

The provided code segment prepares image to flaten into 1D array.Subsequently, the Principal Component Analysis (PCA) for dimensionality reduction. It aims to reduce the data's complexity by setting `n_components` to 64, meaning it wants to represent the data using only 64 principal components.This dimensionality reduction aids in reducing computational complexity while preserving essential information, making the data more suitable for machine learning tasks involving high-dimensional image data.

2. Training with the basic model, validation, and completion of the data pipeline Which models did you use, what training/validation accuracy have you achieved? Is your data pipeline completed?
In the process of developing and training our machine learning models, we employed Support Vector Machines (SVM) and K-Means clustering algorithms. The performance of these models was evaluated through training and validation, resulting in the following accuracy scores: the SVM model achieved an accuracy of 79%, while the K-Means clustering model achieved an accuracy of 15%. These accuracy scores indicate the models' ability to correctly classify and cluster the data.
Additionally, it's worth noting that our data pipeline is indeed completed.In our case, the pipeline involves collecting and preprocessing the handwritten name data, flattening and scaling the images using PCA, and then feeding this processed data into the SVM and K-Means models for training and validation. This completed pipeline ensures that the data flows smoothly from its raw form to being ready for model training, making it a critical

component of our overall machine learning workflow.

3. Identification of the exact tasks you want to complete for the final submission. What challenges you are facing and how you plan to address them. What will be your final deliverables?
For our final submission, our primary objectives are to enhance the accuracy of our existing machine learning models. To achieve this, we plan to incorporate Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in the final phase of our project. These advanced neural network architectures are well-suited for image and sequence data, and we expect them to improve our model's performance significantly.
One of the main challenges we've encountered during our project is related to segmentation. To address this challenge, we will explore alternative segmentation models and techniques to improve the accuracy of this crucial step in our data processing pipeline.
Our final deliverables will include the improved machine learning models with higher accuracy, as well as a detailed report documenting our methodology, findings, and any novel techniques we develop to overcome the segmentation challenge.

C. Each group member needs to identify what they have done so far.
As we have already mentioned in the proposal, **a) Data Collection:** All group members contributed to the data collection process.
b) Data Preprocessing, Feature Extraction, Character Testing, Word Testing: Nikhil and Aditya were responsible for data preprocessing, feature extraction, character testing, and word testing.
c) SVM, Grapheme Segmentation, K-Means Clustering: Aditya and Bhargav worked on implementing the Support Vector Machine (SVM) model, grapheme segmentation, and K-Means clustering.