# Comparative Analysis of Machine Learning Models for Image Classification on the EMNIST Dataset

## Introduction

Machine learning is currently one of the most demanding fields in the everyday world and image classification is at the centre of it all as it is utilised in many areas, from helping to identify medical issues in patients to helping autonomous driving systems identify obstacles. As image classification has become so vital due to the various ways in being implemented to help people it is important to recognise the variation in how well the various machine learning models perform as they can have significant impacts on how applications are developed from small to larger scales. This project aims to compare the performance of 4 different machine learning models and analyse how accurate they are. The aim of the project will be achieved by training and testing machine learning models on a subset of the EMMNIST dataset to recognise handwritten letters and evaluate them at the end.

## Data and Preparation

To compare and analyse the 4 machine learning models with the most accuracy a large enough dataset is required for the training and testing purposes of this study and so the dataset being that has been chosen for this is the EMNIST dataset which consists of 26,000 images which are handwritten letters along with their corresponding labels. Each image in this dataset is of size 28x28 pixels and stored as a reshaped 1x784 vector while being grey scaled for easier manageability as we are specifically testing machine learning models to recognise the handwritten letters. Figure 1 displays an example of how the dataset looks like with their corresponding labels. The dataset is imported into MATLAB and prepared by splitting the data with their corresponding labels in a 50:50 ratio for training and testing by randomisation.

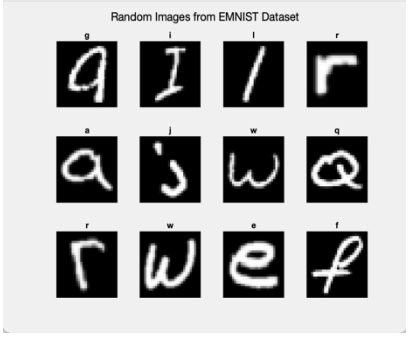


*Figure 1 consists of randomised subset of data with their corresponding labels.*

## Methodology

The overall methodology of this study includes the training and testing of the 4 machine learning models to recognise handwritten letters and classify them to their appropriate subsets and evaluate them. The models being used to train and test data are self-implemented versions of K-Nearest Neighbour (KNN) for Euclidean and cosine distances algorithms and the rest being from MATLAB's implementations of multi-class level vectors and decision tree algorithms. KNN is an effective algorithm for supervised image classification due to its simplicity and effectiveness in finding similarities between data points as it classifies datapoints based on the majority of its KNN. The SVM model is a powerful supervised algorithm that works best on smaller datasets complex datasets making it a good fit for this study. The decision tree algorithm works by building a tree that splits the data into subsets based on its features and then traversing through the tree to find where new datapoints belong, so it is easy for visualisation and was the reason of choice to be added.

## Results

| Algorithm | Accuracy | Elapsed Time |
|---|---|---|
| Self-Implemented KNN Euclidian distance | 78.0615% | 142.6144 seconds |
| Self-Implemented KNN Cosine distance | 80.5615% | 298.8428 seconds |
| SVM for Multiclass (with default parameters) | 74.1% | 2.5157 seconds |
| Decision Tree (with default parameters) | 57.5231% | 0.027471 seconds |

The results above show that KNN Models had achieved the highest accuracies to complete while Cosine version of it being 2.5% higher while SVM falls short of both KNN to an accuracy of 74.61% and decision tree having the worst accuracy. It can also be inferred that Cosine distance took the longest to run while decision tree being the shortest.

## Conclusion

Based on the results the algorithm that I would recommend implementing the SVM model as based on the results section of this report it had a 74.1% accuracy followed up with a 2.5157s elapsed time to complete. The benefit of this model is a high level of accuracy that is relative to the time it takes to assess images and would be most suitable in image classification while the limitations of it not having accuracy closer to a 100% as this is a desired result.