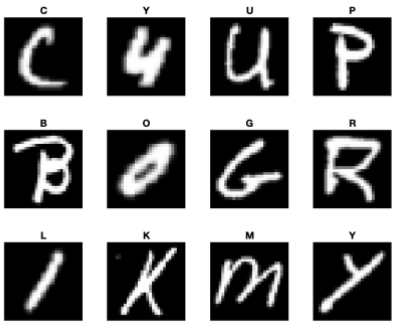
SCC 361 CW Report

Introduction

The objective of this study is to evaluate the effectiveness of various machine learning models in classifying handwritten letters. This assessment is crucial for understanding the impact of different feature extraction techniques and pre-processing steps on the models' ability to interpret handwritten input accurately. Exploring the robustness of these algorithms is vital for advancing human-computer interaction technologies. The study aims to compare the accuracy of these models in recognizing handwritten characters, which could enhance the development of more accessible and efficient digital communication tools.

Data and Preparation



In this study, a subset of the EMNIST dataset comprising 26,000 images of handwritten letters was used, each annotated with its corresponding label. As part of the data preparation process, the dataset was imported and converted to a double floating-point format. The dataset was then divided equally, allocating 50% for training—to learn and identify patterns—and the other 50% for testing to evaluate the models' predictive accuracy. This balanced split aims to prevent overfitting or underfitting.

Methodology

The K-Nearest Neighbours (KNN) algorithm was selected for its straightforward approach to predicting outcomes by looking at similar instances, using different ways to measure similarity. Support Vector Machine (SVM) was included for its capability to handle complex relationships in data, and Decision Trees were used for their ease of interpretation and decision-making transparency. These models were trained on the EMNIST data and was tested to gauge their predictive accuracy. To evaluate the performance of our models, accuracy was used, which simply measures the percentage of predictions our models got right. This approach helped to determine which model was most effective in correctly classifying the data.

Results

Model	Distance Metric	Accuracy (%)	Training Time (s)	Prediction Time(s)
k-nearest neighbours	Euclidean	78.86%	0.16s	33.76s
k-nearest neighbours	Manhattan	76.26%	0.06s	33.00s
Support Vector machine	N/A	73.52%	30.21s	4.04s
Decision Tree	N/A	56.46%	2.33s	0.02s

Conclusion

In assessing the effectiveness of various machine learning models for the classification of handwritten letters, the k-nearest neighbours algorithm utilizing the Euclidean distance metric outperformed others with an accuracy of 78.86%. Despite its longer prediction time of 33.76 seconds, the high accuracy rate positions it as the recommended model for situations where precision is paramount. The k-nearest neighbours with Manhattan distance, while slightly less accurate at 76.26%, offers a negligible improvement in prediction speed, making the Euclidean variant preferable. The Support Vector Machine, with an accuracy of 73.52%, stands out for its quicker prediction time of 4.04 seconds, presenting a viable alternative when faster prediction is required without a substantial sacrifice in accuracy. The Decision Tree model lagged behind with a notably lower accuracy of 56.46%, although it had the fastest prediction time of 0.02 seconds. This trade-off suggests its suitability for applications where speed is critical and some loss in accuracy is tolerable. For future research, it would be beneficial to investigate alternative machine learning algorithms to enhance performance further. The relatively poor performance of the Decision Tree model also warrants a deeper analysis to understand the underlying factors contributing to its lower accuracy, which could reveal insights for improvement.