Amharic Character Recognition: A Machine Learning Approach

Master's Degree in Artificial Intelligence
Department of Computer Science
University of Verona

Academic Year 2024

Author: **Hailemicael Lulseged Yimer**

Contents

1 Introduction

This project aims to recognize Amharic characters using machine learning techniques. Its primary objective is to develop a model capable of accurately classifying Amharic characters from images. The project employs Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimensionality reduction. Additionally, it utilizes various classifiers such as Support Vector Machines (SVM), Logistic Regression, and K-nearest neighbors (KNN) for classification. The performance of each classifier is evaluated before and after applying feature extraction techniques.

2 Motivation

The motivation behind this project is to bridge the gap in existing character recognition systems that primarily focus on Latin characters. Amharic is one of the most widely spoken languages in Ethiopia, and having accurate character recognition systems can facilitate tasks such as text processing, language learning, and document digitization for Amharic speakers.

3 Dataset

- The dataset is prepared by collecting handwriting samples from individuals with diverse characteristics, including different age ranges, levels of education, and both right- and left-handed persons. The dataset contains 14 characters of Amharic, comprising a collection of 4200 images.
- From these, 80 percent are used for training and the remaining for testing.

4 Methodology

4.1 Dataset Preparation

The images are loaded and preprocessed to convert them into a format suitable for training machine learning models. This includes resizing, normalization, and flattening of the image data, as well as shuffling.

- The images are resized to 64x64 and flattened.
- Min-max normalization is used to standardize the data.
- The dataset is also shuffled to ensure randomness.

4.2 Feature Extraction

- PCA and LDA were implemented correctly for feature extraction.
- PCA retained 85

4.3 Model Selection

- After preprocessing and feature extraction, SVM, Logistic Regression, and KNN algorithms were utilized for classification.

4.4 Evaluation and Result

- The classification is conducted in three ways: first, before applying PCA with only normalization; second, after applying PCA on the data; and third, after applying both PCA and LDA.
- SVM, Logistic Regression, and KNN algorithms are tested under each condition.
- Precision, accuracy, and confusion matrix are explained to properly interpret the results.

Table 1: Experimental Results

Model	Accuracy (%)	Precision	F1 Score
SVM (only Normalization)	94.88	95.06	94.90
KNN (only Normalization)	83.10	83.63	83.19
Logistic Regression (only Normalization)	91.55	91.68	91.56
SVM (PCA)	95.36	95.48	95.37
KNN (PCA)	88.21	88.41	88.19
Logistic Regression (PCA)	90.36	90.47	90.37
SVM (LDA)	95.36	95.48	95.37
KNN (LDA)	88.21	88.41	88.19
Logistic Regression (LDA)	90.37	90.47	90.37

5 Conclusions

- The experimental results demonstrate the effectiveness of various machine learning models in recognizing Amharic characters.

- Applying feature extraction techniques such as PCA and LDA improves the classification accuracy of the models.

- Support Vector Machines (SVM) consistently achieved the highest accuracy across different scenarios, closely followed by Logistic Regression and K-nearest neighbors (KNN).

- These findings indicate that SVM could be the preferred choice for Amharic character recognition tasks due to its robust performance.

- Moreover, the results underscore the importance of preprocessing steps such as normalization and feature extraction in enhancing the performance of machine learning models for character recognition tasks.

- By leveraging these techniques, we can effectively address challenges associated with character recognition in languages like Amharic.

6 Recommendation

- Further collection of full-letter data: Expand the dataset by collecting more samples of full-letter handwriting. This can help improve the robustness and generalization of the model.

- Implement advanced analysis techniques: Explore the use of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), such as Long Short-Term Memory (LSTM) networks, for character recognition. These deep learning architectures are well-suited for sequential data like handwriting.

- Experiment with other types of classifiers: Besides SVM, Logistic Regression, and KNN, consider exploring other classification algorithms such as Random Forests, Gradient Boosting Machines, or Neural Networks. Comparing the performance of different classifiers can provide insights into the most effective approaches for Amharic character recognition.