Devnagari-Handwritten-Chars-Classification

Sheetal Munjewar

Bellevue University

Course Number: DSC 680 Applied Data Science

Amirfarrokh Iranitalab

02/23/2025

https://smunjewar.github.io/Portfolio/

Abstract

Devanagari script, used by over 500 million people worldwide, is a widely adopted writing system for languages such as Hindi, Marathi, and Nepali. Despite its significance, limited research has been conducted on handwritten Devanagari script recognition compared to English. This paper explores the use of machine learning and deep learning techniques for recognizing handwritten Devanagari characters, highlighting the challenges, methods, and potential applications in real-world scenarios.

Business Problem

The recognition of handwritten Devanagari script can aid in digitizing historical documents, automating postal address reading, processing bank checks, and creating digital libraries. Unlike English, where automated recognition is well-developed, Devanagari script presents unique challenges due to its complex character set and variations in handwriting styles.

Background/History

In countries like the United States, mobile applications facilitate check deposits through image-based recognition. Similar technologies are lacking for Devanagari script, despite its extensive use. The lack of research and available datasets necessitates the development of robust recognition systems leveraging machine learning and deep learning.

Data Explanation

The dataset used for this project is obtained from the Machine Learning Repository. It contains 92,000 images of handwritten Devanagari characters across 46 classes, split into 85% training data (78,200 images) and 15% testing data (13,800 images). Each image is 32x32 pixels with a centered character and

added padding. The dataset is structured into train and test folders, with each character stored in separate directories.

Methods

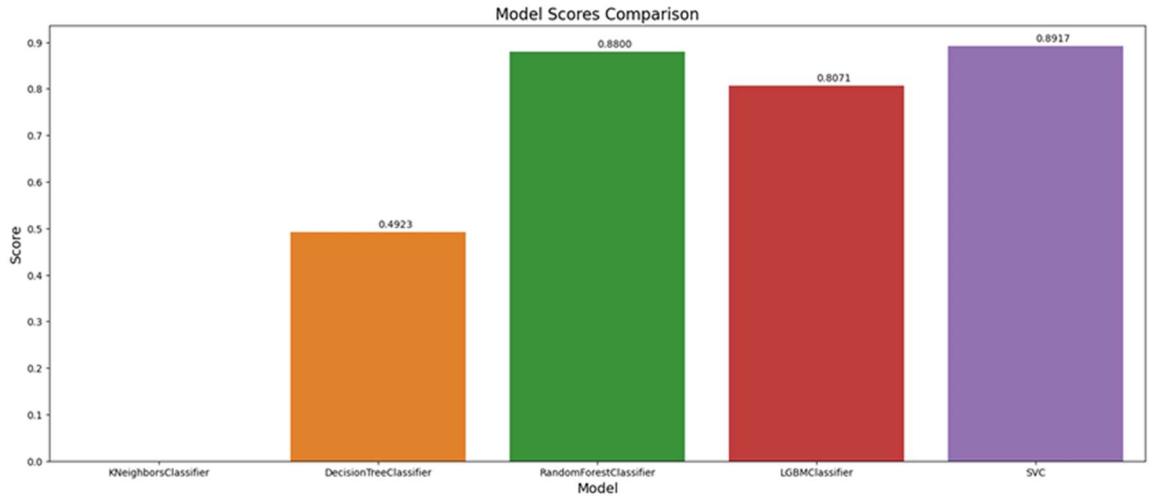
Two approaches were implemented

Machine Learning Models: Used scikit-learn classifiers, including KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, LGBMClassifier, and SVM. Data preprocessing involved normalization and label encoding. Grid search was performed to optimize hyperparameters.

Deep Learning Model: Utilized TensorFlow and a Convolutional Neural Network (CNN) with 11 layers. Image augmentation was applied to enhance training data. The model was trained with batch size 32 and optimized with dropout layers to prevent overfitting.

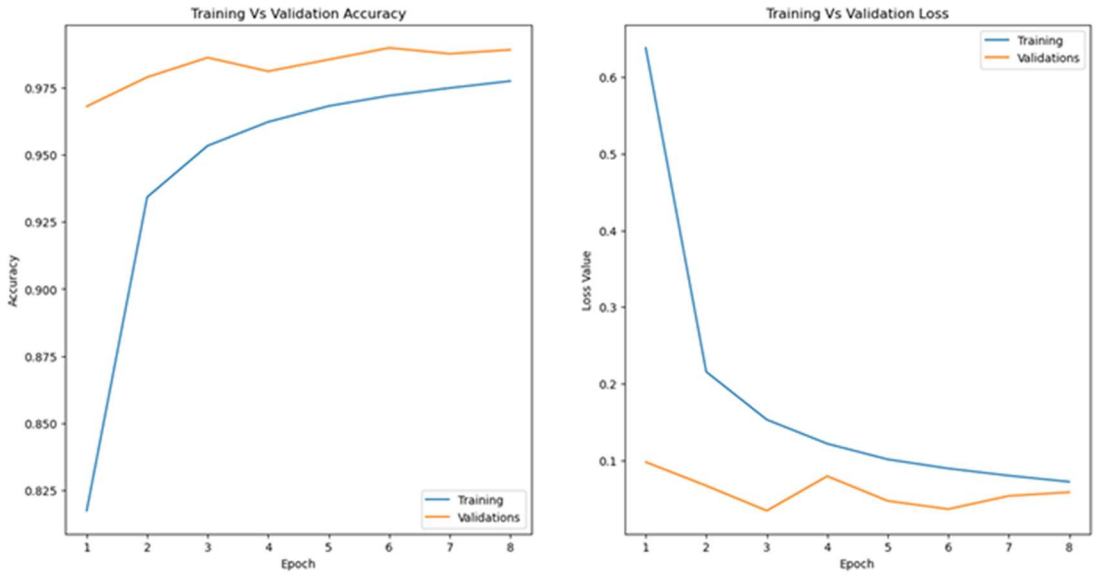
Analysis

Machine Learning models: The first approach with machine learning models yields accuracy around 49% to 89% on the validation dataset while training. The Decision Tree is being the worst with 49% while the SVM is the best with 89% accuracy on the training dataset.



Then SVM model is finally trained with full training data and best parameters found in grid search as it yields the best results. When this model was tested with a test dataset it shows an accuracy of 60%. This shows that the model is over-fitted and needs more work.

Deep Learning Model: With the second approach of the deep learning model, the model consistently produced good results on the validation dataset while training. The validation dataset accuracy was around 96.83% in the first epoch and around 98.91% on the last epoch. The following plots show the accuracy and loss comparison of the model on the training and validation dataset. When this model is tested with the test dataset, it yields an accuracy of ~99%. This shows that the model is not over-fitted and it is working very well on unseen data.



Discussion/conclusion – Next steps

The model is doing quite well on unseen data, so we can say that this model is ready for the real world. On Devanagari Handwritten Character Dataset Data Set[1] it is mentioned that they have created a model that can yield accuracy up to 98.91%. That model is explained in the paper._ I am happy that the model can surpass the accuracy(97.83%) achieved by that model accuracy(98.91%). I think this model could do that because of augmented images addition in the training dataset.

Assumptions

The dataset represents diverse handwriting styles.

Augmentation improves model performance.

CNN architecture is effective for character recognition.

Limitations

Requires computational resources for deep learning training.

Limited dataset diversity might not capture all handwriting variations.

Hardware constraints prevented testing more complex models like XGBoost.

Challenges

Variability in handwritten Devanagari script.

Data augmentation needed to improve accuracy.

Computational requirements for deep learning models.

Future Uses/Additional Applications

Extending recognition to other scripts and languages.

Real-time handwriting recognition for smart devices.

Integration into postal systems, banking, and document digitization.

Recommendations

Enhance dataset with more handwriting variations.

Experiment with advanced deep learning models.

Optimize CNN architectures for efficiency.

Implementation Plan

Phase 1: Data collection and preprocessing.

Phase 2: Model training and evaluation.

Phase 3: Optimization and augmentation.

Phase 4: Deployment and real-world testing.

Ethical Assessment

Ensuring privacy and data protection is crucial in handwriting recognition applications. Misuse of automated handwriting recognition could lead to identity theft or unauthorized data access. Transparent policies should govern data collection and application deployment.

Acknowledgments

The author of this project has referred to the Center for Machine Learning and Intelligent Systems[1] for the data set. The author referred to some notebooks available on Kaggle and GitHub. There is a significant number of solutions available on the Internet for the problem. The author has referred to some of those available solutions and has designed my solution. The author has also referred to multiple websites(not limited to the below references) including data science articles from various authors, machine learning websites, mediam.com, towardsdatascience.com, and many others for basic machine learning and statistics concepts and practical examples.

References

1. Devanagari Handwritten Character Dataset Data Set - Center for Machine Learning and Intelligent Systems
   https://archive.ics.uci.edu/ml/datasets/Devanagari+Handwritten+Character+Dataset
2. Classifier comparison on Devanagari recognition - Prashanth Aditya Susarla -
   https://www.kaggle.com/spadix/classifier-comparison-on-devanagari-recognition
3. Handwritten Devanagari character identification with ResNet - Akshay Chougule -
   https://medium.com/@akshaychougule/handwritten-devanagari-character-identification-using-resnet-b90894b42c4d
4. Devanagari - https://en.wikipedia.org/wiki/Devanagari
5. Deep Learning based Character Classification using Synthetic Dataset - Krutika Bapat -
   https://learnopencv.com/deep-learning-character-classification-using-synthetic-dataset/
6. Devanagari Character Set - Rishi Anand -
   https://www.kaggle.com/rishianand/devanagari-character-set
7. Intro to Computer Vision-Indian National Language - Shruti Iyyer -
   https://www.kaggle.com/shrutimechlearn/intro-to-computer-vision-indian-national-language
8. Devanagari Character Dataset - Ashok Pant -
   https://www.kaggle.com/ashokpant/devanagari-character-dataset

9. Image classification - https://www.tensorflow.org/tutorials/images/classification

10. Recognizing Handwritten Digits with Scikit-learn -

11. Manthan Bhikadiya - https://medium.com/codex/recognizing-handwritten-digits-with-scikit-learn-90ca6e2471ed

12. Handwriting recognition - https://en.wikipedia.org/wiki/Handwriting_recognition

13. An Analysis of Irregularities in Devanagari Script Writing – A Machine Recognition Perspective - Satish Kumar - http://www.enggjournals.com/ijcse/doc/IJCSE10-02-02-30.pdf