



# Chars74K dataset

## Character Recognition in Natural Images



Villem Lassmann & Eva-Maria Pedosk

### Introduction

Since character recognition is an ongoing problem in data science, we decided to take on the challenge with the Chars74K dataset.

The aim of the project was to train a proficient model to achieve the highest possible character recognition precision.

The goal was set to have **95% of accuracy**.

The original dataset and experiments were done by Microsoft Research India.



Test sample



Training sample

Data samples

### Working with data

The dataset contained images that were categorized into two groups - computer generated and natural images.

The models in our project used computer generated images as a training dataset and natural images as a testing dataset. The validation dataset was 10% of the training dataset.

Data preprocessing involved:

- Making the images black and white
- Normalization
- Data transformation

Since CNN (*Convolutional neural network*) is widely used and performs well in image recognition, it was applied to the project.

The most efficient and used activation function was ReLu.

### Results and conclusion

The best model achieved an accuracy of 96% and exceeded the goal that was set for the project.

In conclusion, the model performed well on the training dataset and the accuracy could not have been much higher. The best accuracy on the testing data was 18%. It is unspecified if the model was composed poorly or if the quality of the images provided in the dataset was too low. Nevertheless, the achieved accuracy was better than random and the model should still be considered successful.

### Potential applications

A model trained on the Chars74K dataset could be highly useful for identifying data from pictures.

This project could be directly beneficial e.g for the post and logistics company Omniva. Best model could be used for identifying addresses and names on international parcels. It could also help to lower future research costs and quicken the researches.

Visit also our project [GitHub](#) and [Chars74K website!](#)

