Cancer Subtype Classification with Machine Learning

Ben Flügge • Meike Liedtke • Bao Tran Nguyen • Hauke Schüle • Tatjana Wehrmann

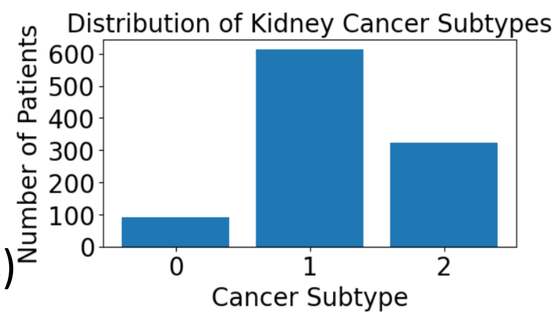
Motivation

Identify Cancer Subtype:

- Kidney Chromophobe (0)
- Kidney Renal Clear Cell Carcinoma (1)
- Kidney Renal Papillary Cell Carcinoma (2)

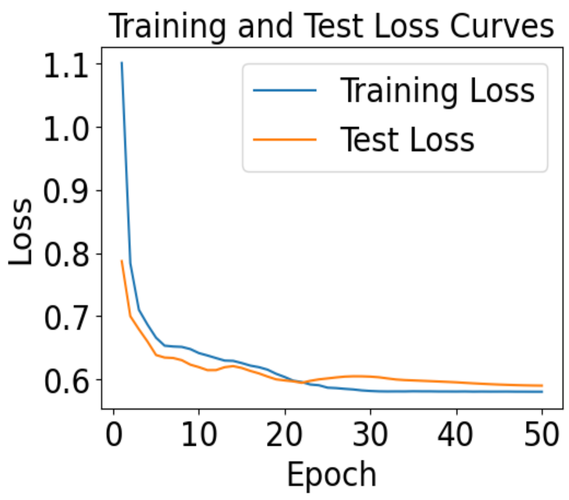
Dataset

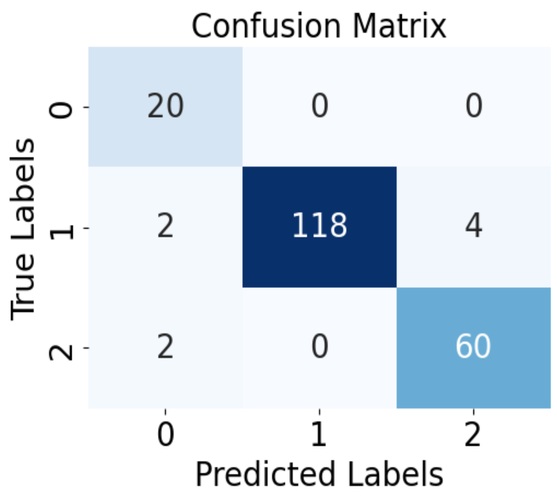
- TCGA Kidney Cancers Dataset
- 1028 cancer patients
  - Amount of data distribution per class (see right)
- 60660 gene expressions (features)
  - Using only TPM



Neural Network with PyTorch

- Neural Network is a more complex supervised learning algorithm
- Layers:
  - Dropout (Regularization)
  - Linear Layer
  - ReLu
  - Linear Layer
  - Softmax
- Adam Optimizer
- Criterion: Cross entropy Loss
- 64 Hidden Layers
- 50 Epochs
- Learning Rate: 0.0005



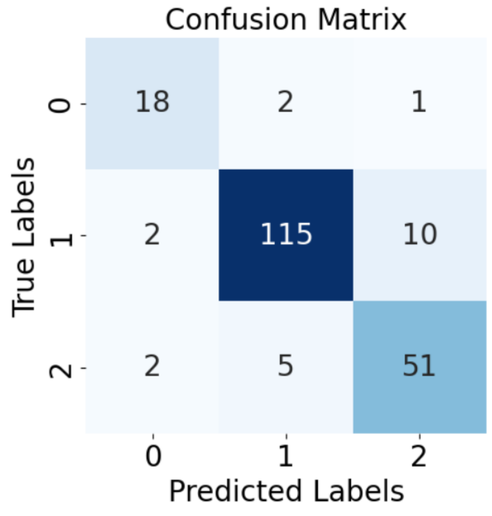


Decision Tree Classifier with Scikit-Learn

- Decision tree is a simple yet efficient supervised learning algorithm
- Criterion: entropy
- Test size: 0.2

Result:

- Tree depth: 7-8
- Root node: gene NDUFA4L2



	F1 Score
(0)	0.87
(1)	0.93
(2)	0.89

Tuning approaches

Decision Tree:
Since some splits resulted in very small result sets, the parameters min-samples-split (= 7), min-samples-leaf (= 7), or min-impurity-decrease (=0.01) were set. The F1 Score and Confusion Matrix showed only slightly better results.

Neural Network:
Learning rate, number of hidden layers and epochs were chosen to allow the model to finish training, without overfitting to the training data. Additionally the dropout layer was added as regularization.

PCA approach:
We used Principal Component Analysis to reduce the number of features. In our first approach we reduced the number of features from 60660 to 1028 with a variance ratio sum of 1.0. This new dataset however proved to have no impact on the results of the decision tree and to be unusable by the neural network. In a second approach we reduced the number of features to 200 with a variance ratio sum of 0.813. This dataset had again no real impact on the results of the decision tree, but produced viable results with the neural network, similar to those using the original data.

01/2024