Kidney cancer classification using ML on gene expressions

Jonas Balke • Maximilian Greß • Tim Lukas Nolte • Xiang Zhou • Xinyue Gong

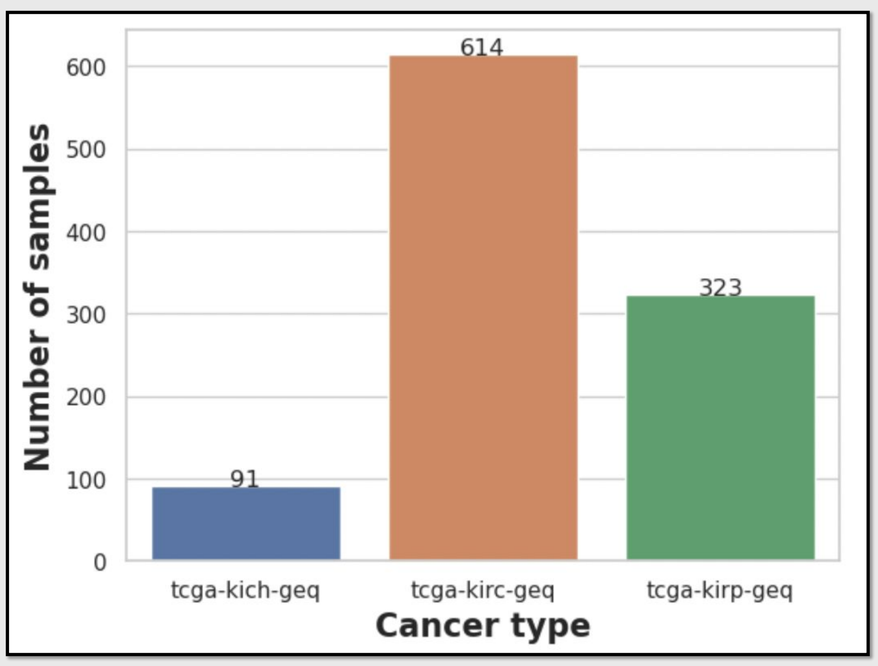
Introduction

In this project, we use different machine learning approaches to predict the specific subtype of kidney cancers given the transcriptome profile data.

Dataset Analysis

Our models are trained with a subset of the TCGA dataset which contains:

- 60660 gene expressions (given as both TPM and FPKM for each gene),
- 1024 samples,
- and an imbalanced class distribution (as shown in the figure).



Methods

In 2018, Muhamed Ali et al.[1] used miRNA to classify kidney cancer subtypes. They adopted a neural network (LSTM) for classification. Both accuracy and Matthews Correlation Coefficient (MCC) were measured with 5-fold cross-validation. In 2020, Hamzeh, O., Alkhateeb, A., Zheng, J. et al.[2] tried to predict the tumor location in prostate cancer tissue with TPM values. Besides other methods, they used the Synthetic Minority Oversampling Technique (SMOTE) to handle imbalanced data and information gain (IG) as a feature selection method. In their study, support vector machine (SVM) achieved the best results.

Inspired by these works, we also **use SMOTE to generate new synthetical samples** for the smallest class and designed the following two approaches:

A. **Traditional ML algorithms**
- Data normalized in range (0,1)
- Feature selection with **select-k-best** (100, 1000, 10000) algorithm and IG as score function
- Grid Search with **5-fold cross-validation**

(For **random forest**: number of estimators, tree depth)
(For **SVM**: C, gamma, kernel)

B. **Deep Learning**
- Data normalized in range (0,1)
- No feature selection
- **A neural network with 3 hidden layers** (1024, 512, 256 units per layer)

Results

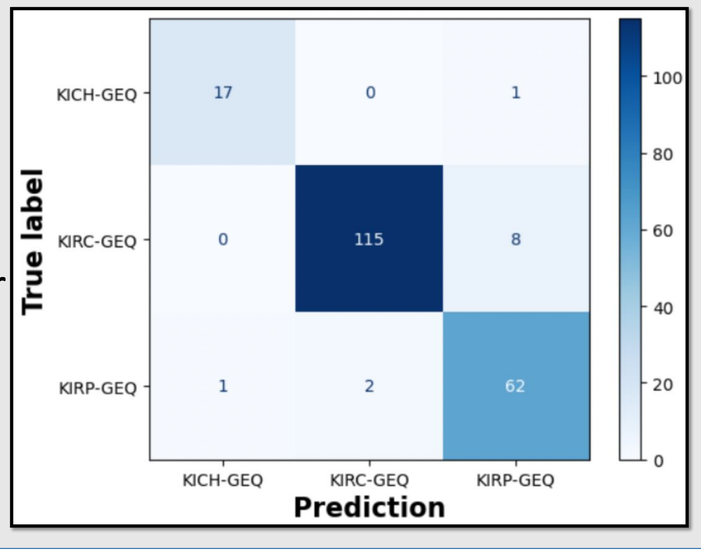
Experiments were conducted on both **TPM and FPKM** values.
The traditional ML approach preferred TPM while **FPKM worked slightly better for our neural network**. The main metric used for comparison is the **F1 score** as shown in the table below.

gene expression	model	balance	number of features	macro f1-score	f1-kich	f1-kirc	f1-kirp
tpm	random forest	Balanced	10000	0.914	0.872	0.963	0.906
tpm	SVM	Imbalanced	10000	0.938	0.944	0.958	0.912
tpm	Neural Network	Balanced	all	0.933	0.895	0.971	0.932

TABLE I
BEST TEST SCORES OF EACH APPROACH

To compare with the study of Muhamed Ali et al., who reached an MCC score of 0.92, we also measured the **MCC score**.

With the traditional ML approach, our highest MCC score was only 0.89 with the best SVM model. However, our neural network was able to achieve a better MCC score of 0.92. The confusion matrix shows the prediction result of our best neural network on the test set.



[1] Muhamed Ali, A.; Zhuang, H.; Ibrahim, A.; Rehman, O.; Huang, M.; Wu, A. A Machine Learning Approach for the Classification of Kidney Cancer Subtypes Using miRNA Genome Data. Appl. Sci. 2018, 8, 2422. https://doi.org/10.3390/app8122422
[2] Hamzeh, O., Alkhateeb, A., Zheng, J. et al. Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. BMC Bioinformatics 21 (Suppl 2), 78 (2020). https://doi.org/10.1186/s12859-020-3345-9

Institut für Informationsverarbeitung
Applied Machine Learning in Genomic Data Science

Leibniz Universität Hannover