Research Proposal: Problem solving
Michael Cornelisse s1059020, Nienke Helmers s1016904

Our project will revolve around gene expression data for different types of cancer cells: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (link below). The data consists of a total of 72 samples of 47 AML and 25 ALL patients. For all patients, the expression of 6817 genes was measured. Our end goal is to write an algorithm that can classify a cell to the correct type of cancer cell. **For this we will be using neural networks: there is the possibility of a Multilayer Perceptron Neural Network (MLPNN) and a Convolutional Neural Network (CNN) and we will determine which of the two will yield a better result in classifying the data.** To write our algorithms, we will be using pytorch and potentially supplement it with an API.

There have already been many attempts at creating a model that can classify cancer cells, mostly based on pictures of the cells and their surrounding tissues. The researchers that assembled this dataset have done the same, albeit not using data mining concepts like decision trees, but by sorting the genes based on correlation and making predictions per gene, which were then assigned a weight and summed. As an accuracy test they used cross validation. It will be interesting to see how their results compare to ours and their methods might also give us some ideas on how to tackle this data set.

Data:
https://www.kaggle.com/crawford/gene-expression?select=data_set_ALL_AML_independent.csv

Possible useful sources:

- Molecular Classification ofCancer: Class Discovery andClass Prediction by GeneExpression Monitoring: https://www.researchgate.net/publication/12779876_Molecular_classification_of_cancer_class_discovery_and_class_prediction_by_gene_monitoring **Original paper**
- Microarray data classified by artificial neural networks https://pubmed.ncbi.nlm.nih.gov/18220242/
- Microarray Data Analysis Using Neural Network Classifiers and Gene Selection Methods https://link.springer.com/chapter/10.1007/0-387-23077-7_16
    - https://www.researchgate.net/publication/227064821_Microarray_Data_Analysis_Using_Neural_Network_Classifiers_and_Gene_Selection_Methods
- Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665029/
- https://www.simplilearn.com/deep-learning-algorithms-article
- Artificial Neural Networks Methods for the Identification of the Most Relevant Genes from Gene Expression Array Data https://www.researchgate.net/publication/4030466_Artificial_Neural_Networks_Methods_for_Identification_of_the_Most_Relevant_Genes_from_Gene_Expression_Array_Data