CHAIR FOR CLINICAL BIOINFORMATICS
SAARLAND UNIVERSITY

IDENTIFICATION OF TUMOR PATIENTS WITH DEEP LEARNING APPROACHES BASED ON MICRORNAS MEASURED IN BLOOD SAMPLES

Mustafa Kahraman
muka.uni@gmail.com

Identification of tumor patients with Deep Learning approaches based on microRNAs measured in blood samples

In the last 10 years the number of studies and efforts in the area of body fluid diagnostics and research of diverse diseases has been increased. Especially the interest is focused on early identification of tumor patients since the current applied diagnostic methods work only for diseases at advanced stage which is in most cases too late for a successful curative treatment.
While the majority of the studies deals with messengerRNAs as biomarkers, the community researching with microRNAs becomes more and more established. The current research shows that there is a huge diagnostic potential in the regulatory behavior of miRNAs. This is usually analyzed in bioinformatics by classical machine learning methods and hypothesis tests. In the last years Deep Learning is one of the newest trends in data science applied in different fields. However, there are only a few groups using the new methods for researching with miRNAs.

The following project has the aim to find if Deep Learning approaches can find sets of biomarkers that can distinguish good enough between tumor and non-tumor patients, or even outperform classical machine learning methods used on the same dataset. The dataset consists of over 3000 patients of which over 500 have tumors.

miRNAs	The dataset consists of **1,183 microRNAs** (named here as **feature-0001** up to **feature-1183**). MicroRNAs are small non-coding RNA molecules which regulate several biological processes. Studies of the last years shows that some of these miRNAs are involved in the development of diseases like lung cancer.
Samples	The dataset consists of **3,046 samples** (named here as **Sample_0001_disease-state** up to **Sample_3046_disease-state**; disease-state: **LCa** or **Non-LCa**). Each of these samples belongs to an individual who has lung cancer (LCa) or not (Non-LCa). In total, there are **606 LCa-samples** and **2,440 Non-LCa-samples**.

Signature	Your goal is to find one or more suitable sets of **maximum 30 features** (signature) which can **distinguish** well between both groups **LCa (positive class)** and **Non-LCa (negative class)**.
Performance	A **well performing** signature has an **accuracy, sensitivity, specificity** and **AUC-value** of at least 0.85. The performance is based on **cross-validation**.

If you have questions regarding ...

Dataset	**Mustafa Kahraman** muka.uni@gmail.com
Deep Learning Methods	**Pedro Guimarães** pedro.guimaraes@uni-saarland.de

Good luck

for the project

&

your other classes and/or thesis