Introduction

Early detection of cancer is crucial for successful treatment and improved patient outcomes. As cancer cells rapidly proliferate, they often outgrow their blood supply, resulting in regions of hypoxia (low oxygen) within the tumor microenvironment. Hypoxic conditions promote tumor aggressiveness, resistance to therapy, and the generation of genetic alterations that drive cancer progression.

The aim of this project is to develop a robust method for detecting hypoxia and normoxia in cells using gene expression analysis and classification AI models. Furthermore one special objective is to investigate whether clusters of cells that do not conform to hypoxia/normoxia classifications can be explained by variations in the expression levels of genes associated with different stages of the cell life cycle. By exploring this possibility, we aim to gain a comprehensive understanding of the complex interplay between hypoxia, cell cycle progression, and cancer development.

Ultimately, the successful implementation of such tools holds great potential for early cancer detection and the development of targeted therapeutic strategies, thus improving patient outcomes in the battle against cancer.

This report is divided in the following sections

- Data Preprocessing: cleaning and formatting the dataframes, obtaining a unique dataframe for both cell lines of SmartSeq
- Exploratory Data Analysis: visualizing some graphs and statistics of genes and cells, and an analysis of the differential expression
- From Eda to Modelling: feature selection with LASSO penalty, PCA and autoencoders
- Unsupervised Learning: clustering algorithms (K-means and Gaussian mixture), and domain-specific analysis of the cell cycles with Scanpy
- Supervised Learning: regressions, SVM, boosting algorithms, ensemble methods and neural networks

Import libraries