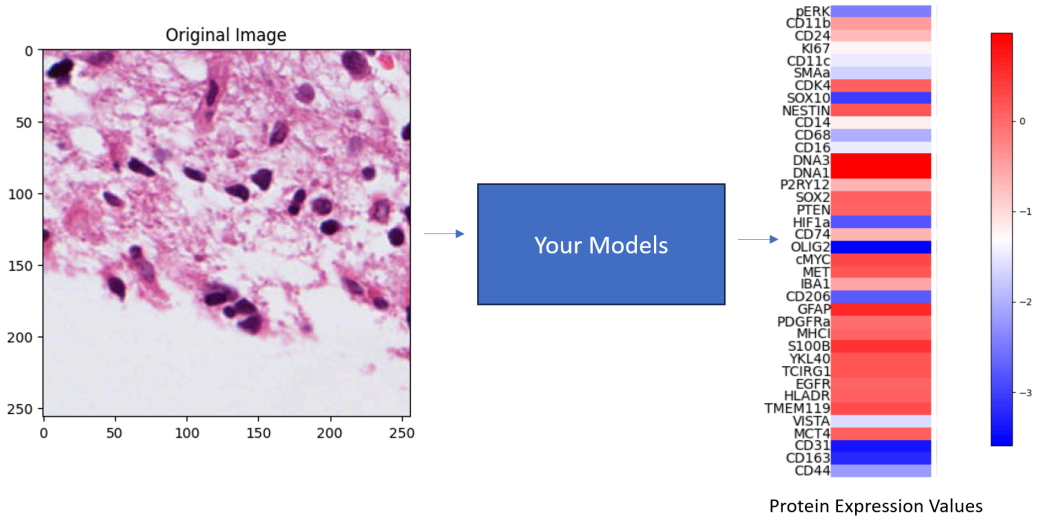
Data Mining 2025 Assignment 2: Prediction of Protein Expression

(by Fayyaz Minhas)

In this assignment, the objective is to develop regression models for predicting the level of expression of different proteins in a given biological tissue image. You do not need to know any biology for solving this machine learning exercise. **Your task is to develop machine learning models that use training data (images with known protein expression values) to predict protein expression in test images.**



Data Availability, Format and Reading:

The data comes from 4 biological tissue specimens (labelled A1, B1, C1 and D1) and each specimen contains multiple “spots”. Each spot corresponds to a spatial location in a specimen. For each spot, we have a Red-Green-Blue (RGB) image at that location in the specimen in the form of a png file and the corresponding expression values of 38 different proteins.

Instructions on how to download the data and view it (along with other helpful hints, e.g., on how to convert the image to a Hematoxylin-Eosin-DAB (DAB) space which is useful for analysis of such images) are given in the notebook:
https://github.com/foxtrotmike/CS909/blob/master/DataMining2024_assignment_2.ipynb

Training and Testing:

Unless otherwise specified, use (all or some) data from specimens **B1, C1 and D1 for training and validation** (it is entirely up to you how much data you use for training and how much data you use for validation) and the data for specimen **A1 for testing**. Do not perform testing on specimen A1 until you have developed your model fully. Also do not use any protein expression data from A1 in training or model selection or hyperparameter optimization. Wherever applicable, performance metrics for the test data are to be reported unless otherwise specified. **Note** that Q3ii asks you to report “Leave one specimen out” cross-validation results.

Submission: You are expected to submit a **single Python Notebook** containing all answers and code. Include all prediction metrics in a presentable form within your notebook and include the output of the execution of all cells in the notebook as well so that the markers can verify your output. **Also submit a consolidated table of your performance metrics within the notebook to indicate which model performs the best (MANDATORY).**

Use of additional libraries: You can use other libraries where needed. But please include the installation instructions of those in the notebook along with a reason why you needed to use them.

Use of additional data: You can use other datasets if you want. Please explain any such uses clearly in your notebook.You are free to do any augmentations or any other strategies to improve prediction performance as long as you do not use target variable information directly or indirectly in doing so.

Restrictions:

Students are restricted from sharing the data files or the assignment solutions. Each student needs to submit a single solution which should be developed by the student without assistance from other sources.

Question No. 1: (Data Analysis) [15 Marks]

Using training data, answer the following questions:

i. Counting Examples: Determine the number of "examples" or spots in each specimen. **[1 mark]**

ii. Protein Expression Histograms: For each specimen, generate histograms to visualize the expression values of 'CD11b' and discuss your observations. **[2 marks]**

iii. Image Pre-processing: Convert a selection of images from RGB to HED color space, focusing on the Hematoxylin channel (H) to highlight cellular nuclei. Provide visual examples following the hints in the notebook linked above. **[2 marks]**

iv. Calculate the average intensity of the H-channel for each image. Create a scatter plot comparing these averages against the expression levels of CD11b for each image. Assess the

correlation between H-channel intensity and CD11b expression. Discuss the potential of H-channel average as a predictive feature for CD11b expression. **[3 marks]**

iv. Calculate the average intensity of the blue channel for each image. Create a scatter plot comparing these averages against the expression levels of CD11b for each image. Assess the correlation between E-channel intensity and CD11b expression. Discuss the potential of E-channel average as a predictive feature for CD11b expression. **[2 marks]**

v. Visualize (as a scatter plot) and quantify the degree of correlation or dependence between average H and average blue channel intensities across images. What are the implications of this? **[2 marks]**

v. Is there association between protein expression levels of different proteins? If so, how can we quantify this association? **[3 marks]**

Question No. 2: (Feature Extraction and Classical Regression) [30 Marks]

<u>For the following questions, use the expression of CD11b as the output prediction target variable.</u>

i) [15 Marks]

Extract “informative” features from an image. For example you can use one or more of the following types of features:

a. Average and variance for each of the 'H', ‘E’, ‘D’ (from HED) and the original red, green, and blue channels
b. PCA (Principal Component Analysis): Applying PCA, such as randomized PCA or incremental PCA, can significantly reduce dimensionality while preserving the variance in the image data, making it easier to identify patterns. Refer to sklearn.decomposition.PCA for implementation details. You might choose to reduce the dataset size or image dimensions for PCA to manage computational complexity.
c. Any other features of your choice but you do need to give justification of those features in terms of their appropriateness for this problem.

Note: You can also consider resizing images or selecting specific image regions or reducing the number of training images if necessary to manage computational load.

ii) [15 Marks]

Apply the following regression models using the features from Q2(i):

- Ordinary Least Squares (OLS) Regression
- Random Forest or Support Vector Regression (Pick one or more – your choice!)

What is the appropriate metric for this prediction problem? Provide a clear justification. [3 Marks]

For each model of choice, create scatter plots to compare the true and predicted values on the test data. Additionally, evaluate and report your models' performance using the following metrics: RMSE, Pearson Correlation Coefficient, Spearman Correlation Coefficient, and R2 score. Reference for metrics: sklearn.metrics.

It's your responsibility to select appropriate hyperparameters.

Deliverables:

Scatter plots for true vs. predicted values for each model type.

Performance metrics (RMSE, Pearson, Spearman, R2 score) on the test data.

Question No. 3 (Using Neural Networks) [55 Marks]

(i) [20 Marks] Develop a Neural Network using PyTorch to predict the expression level of CD11b from input images, following the approach outlined in part (ii) of Question (2). Design the architecture of the model to take an image as input and output a single value representing the CD11b expression level. You have the freedom to select the structure of the network and the loss functions to be used. You can use pre-trained models and perform transfer learning if needed.

Evaluate your model's performance on the test dataset by creating a scatter plot that compares the true vs. predicted CD11b expression values. Additionally, quantify your model's predictive performance using the following metrics:

- RMSE (Root Mean Square Error)
- Pearson Correlation Coefficient
- Spearman Correlation Coefficient
- R2 score

Your model will be assessed based on its architecture design and the achieved performance metrics. Aim for the best possible performance on the test set, ensuring that the test data is not used during training. Include in your submission convergence plots that illustrate the change in loss across training epochs, demonstrating how your model's performance improves over time.

(ii) [20 Marks] Create a neural network using PyTorch to simultaneously predict the expression levels of all proteins from given image patches. You have the flexibility to choose the architecture of the neural network and the loss functions you deem appropriate for this task.

For model validation, employ a "leave one specimen out cross-validation" strategy. This approach involves sequentially using data from one specimen as the test set and the combined

data from the remaining specimens as the training set. This method is similar to a 4-fold cross-validation but specifically tailored to ensure that each specimen is used as a test set exactly once. This validation technique ensures that your model's performance is evaluated on entirely unseen data, mimicking a scenario where the model is tested on data from a new specimen. For a practical understanding of how this is implemented, you can refer to the GroupKFold method in scikit-learn.

Finally, quantify the performance of your optimal model for each protein with the following statistical metrics in the form of a single table.

- RMSE (Root Mean Square Error): Measures the model's prediction error.
- Pearson Correlation Coefficient: Assesses the linear relationship between predicted and actual values.
- Spearman Correlation Coefficient: Evaluates the monotonic relationship between predicted and actual values.
- R2 Score: Indicates the proportion of variance in the dependent variable predictable from the independent variable(s).

For each metric, report both the average and standard deviation across the specimens for every target protein. This comprehensive evaluation will help in understanding the model's predictive accuracy, reliability, and the nature of its errors or biases.

Also, report the number of proteins for which the average spearman correlation coefficient is above 0.7.

iii) **[10 Marks]** For the questions below, you will be graded on the feasibility and practicality of your ideas and you can get bonus marks depending upon whether you show any preliminary or pilot results.

A. How can we utilize the location data of each spot or any other information available in the dataset for improving the prediction? [5 marks]
B. How can you determine whether your predictor is truly predicting CD11b expression independently, rather than just reflecting the combined influence of multiple correlated proteins? What are the potential implications if your predictor is not independent, and how can you address this limitation to improve the reliability of your findings? [5]