# The Great Computer Challenge, 2025
# *Artificial Intelligence (A.I.), Level 4*
# *Problem 2*

## Background

Early and accurate identification of breast cancer is crucial for improving patient outcomes and guiding effective treatment strategies. High-dimensional medical data can introduce challenges such as increased computational cost, redundancy, and overfitting, which may affect the accuracy of classification models. Dimensionality reduction techniques help address these challenges by simplifying the dataset while retaining essential information and improving model efficiency and interpretability. By reducing complexity, these techniques enhance visualization, speed up computations, and improve the generalization of machine learning models for breast cancer diagnosis.

## Guidelines & Requirements

Participants will work on the Breast Cancer Wisconsin (Diagnostic) dataset, a widely used benchmark dataset for binary classification in medical diagnostics. It contains **569** samples, each representing a tumor, with **30 numerical features** derived from digitized images of fine needle aspirates (FNAs). Each sample is classified into one of two classes: **malignant (212 samples)** and **benign (357 samples)**, labeled 0 and 1, respectively. Features describe tumor characteristics such as radius, texture, smoothness, and symmetry. The data can be downloaded at https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html.

## Challenge 1

1. Perform **exploratory data analysis (EDA)** to understand the statistical distribution of features and their relationships.

## Challenge 2

2. Apply one or both dimensionality reduction techniques:
    - Principal Component Analysis (PCA) to reduce the features to two components and visualize them (hint: this is dimensionality reduction, not feature selection).
    - Linear Discriminant Analysis (LDA) to reduce the features to one component and analyze the class separation (hint: this is dimensionality reduction, not feature selection).

# Challenge 3

3. Develop and evaluate a classification model on the original features, PCA-reduced features, and LDA-reduced data using any method such as logistic regression, decision trees, random forests, or support vector machines.

# Challenge 4

4. Analyze and compare the classification performance across the original features, PCA-reduced features, and LDA-reduced features to evaluate the impact of dimensionality reduction on model accuracy and effectiveness.

# Judging Criteria

The exploratory data analysis evaluation will focus on the ability to generate meaningful visualizations and statistical insights that
(1) Capture the dataset's overall structure (10 points) and,
(2) Reveal relationships between key features (10 points).

The classification model will be assessed based on the following criteria:
(1) Proper division of the dataset into training and test sets (10 points),
(2) Application of k-fold cross-validation to enhance reliability (10 points),
(3) evaluation and comparison of performance using classification metrics such as accuracy, precision, recall, and F1-score based on models built on the original features, PCA-reduced features, and LDA-reduced features, (15 points) and
(4) Analysis of key features impacting classification outcomes and their real-world relevance in breast cancer diagnosis (5 points).

*The judging criteria below override the default judging criteria.*

Participating teams are allowed to use multiple computers to distribute the workload. Participants are not allowed to use any forms of large language models such as GPT, Gemini, and Llama during the competition.

Participating teams are allowed to use multiple computers to distribute the workload.

The codes must be runnable on Google Colab and shared with fanchyna@gmail.com. The coding will be reviewed based on its design (5 points), smell (e.g., comments are clear and useful, good naming, etc.) (5 points), modularity (5 points), and reconfigurability (5 points).

The report, which is a single PDF file, should be sent to fanchyna@gmail.com. The report should describe the model(s), external data (if any), results (e.g., text, tables, figures, visualization, screenshots), the process to reproduce the results, discussion (if any), conclusion, references (if any). No page limit. No changes can be made after the competition ends. The report will be scored based on the model's performance (metrics in the evaluation section above), comprehensiveness, clarity, and accuracy. (report quality overall score: 20 points)

The judging committee reserves the right of post-competition validation. Awards will be revoked if plagiarism or fabrications are found.

A team's final score is determined by the consensus of all judges.

## SOL Correlation

Not applicable.

***Have fun and thanks for participating in the Great Computer Challenge, 2025!***