**Executive Project Summary: Protein Profiles Team**

**Project Title**: Detecting Cancer from Blood Tests

**Team Members**: Parinaz Fathi, Simeiyun (May) Liu, Nihan Akis Man, Cerise Chen

**Project Overview:** Early cancer detection is crucial for improving survival rates, but traditional diagnostic methods like imaging are often costly and may not be practical unless someone is already suspected of having cancer. This project explores the potential of using machine learning to identify patients with cancer using their **blood protein profiles.** We aimed to distinguish between cancer and non-cancer patients, as well as to predict specific cancer types.

We combined data from **four published datasets** using **NPX** values to quantify protein levels in blood plasma samples. All datasets utilized Olink technology for profiling the plasma proteome of the subjects. The datasets include: 1) **Pancancer dataset** with levels of 1463 proteins from 1477 patients with 12 different types of cancer (4 of which were condensed into a "blood cancer" category for this project). 2) **Esophageal Cancer dataset** with levels of 92 proteins from 91 patients with Esophageal cancer. 3) **Hodgkin's Lymphoma dataset** with levels of 92 proteins in 167 samples collected at different times from 54 patients with Hodgkin's Lymphoma. 4) **Southern German Population-based cohort** with levels of 728 proteins from 173 individuals in a population-representative cohort, which we are using as the control group in our analyses.

**The goal of this project** was to predict disease status (Cancer or Control; binary classification) and disease type (which type of cancer; multiclass classification) based on the protein profiles. The **49 proteins in common** across these 4 datasets were used for our exploratory data analysis and our machine learning models that predicted cancer status (Cancer or Control). Our exploratory data analysis included principal component analysis, correlation matrices, interactive heatmaps, bar plots, swarm plots, and pair plots. The **1463 proteins in the Pancancer dataset** were used in our machine learning models that predicted cancer type (out of 9 types of cancer).

**Stakeholders:** **Healthcare Providers**- to improve diagnostic accuracy and offer quicker, less invasive diagnostic tools. **Patients**- to benefit from faster detection methods that enable earlier treatment and improved survival rates. **Biotech and Pharmaceutical Companies**- To develop potential commercial diagnostic tools for early cancer detection.

**Key Performance Indicators (KPIs):** **Accuracy**- the percentage of correct predictions for both binary cancer detection (Cancer vs. Control) and multiclass cancer type classification. **Precision, Recall, and F1-Score**- to assess the balance between sensitivity and specificity in predictions. **ROC-AUC**- to measure the ability of models to discriminate classes. **Geometric Mean Score**- to evaluate balanced performance across all cancer types.

**Preprocessing Steps**: Within cross-validation loops, missing values were imputed using K-Nearest Neighbors (KNN) imputation to ensure no loss of data. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied. For feature selection, SelectKBest was used to prioritize the most predictive features.

**Workflow**: 5-fold cross-validation was used to assess the performance of each model, ensuring a robust evaluation. GridSearchCV was used to optimize model hyperparameters, particularly for models like Random Forest to avoid overfitting. Performance was evaluated using Accuracy, F1-Score (macro average), ROC-AUC, Recall, Precision, and Geometric Mean Score (G-Mean).

**Model Results:** To predict cancer presence and type, we experimented with machine learning models to address three objectives.

**Objective 1. Binary Classification (Cancer vs. Healthy)**: We applied two models to the combined dataset that included the levels of 49 proteins from all 4 datasets.

- **Logistic Regression**: Distinguished cancer from healthy samples effectively with a training accuracy of 100 %.
- **K-Nearest Neighbors (KNN)**: Using both 5 and 13 neighbors, this model was also able to distinguish between cancer and healthy samples with a training accuracy of 100 %.

We selected **Logistic Regression** as the final model for this objective, and it had a testing accuracy of 100 %.

**Objective 2. Multiclass Classification (Cancer Type Prediction)**: The 49 proteins alone were not effective in distinguishing between the different types of cancer. To maximize the ability to distinguish between the different cancer types, we focused on using all 1463 proteins to distinguish between the 9 different types of cancer in the pancancer dataset. We tested 6 models on this dataset (1463 proteins for 1477 patients, each of whom was categorized as having one of nine types of cancer).

- **Logistic Regression**: Was able to distinguish between 9 different types of cancer with the best accuracy (Accuracy: 0.7663) among the 6 models.
- **K-Nearest Neighbors (KNN)**: Performed poorly in distinguishing between 9 different types of cancer (Accuracy: 0.3937, F-1 score: 0.4030, ROC AUC: 0.8184).
- **Random Forest**: Was able to distinguish between 9 different types of cancer with an accuracy of 0.6528, F-1 score of 0.6536, ROC AUC of 0.9189. We also visually compared Random Forest and Logistic Regression. Additionally, despite GridSearchCV and hyperparameter optimization, Random Forest exhibited significant overfitting (with high variance between training and testing sets), and the model ran too slowly with GridSearchCV.
- **Extra Trees**: Was able to distinguish between 9 different types of cancer with an accuracy of 0.6096, F-1 score of 0.6119, and ROC AUC of 0.9029. This model also struggled with overfitting, similar to Random Forest, and faced computational challenges when dealing with large datasets.
- **XGBoost**: Was able to distinguish between the 9 different cancer types with an accuracy of 0.7163, F-1 score of 0.7160, and ROC AUC of 0.9498.
- **Multinomial regression:** Tested for distinguishing nine different types of cancer (accuracy: 0.5932, F-1 score: 0.5920, ROC AUC: 0.9122).

**Objective 3**. **Multiclass Classification with Fewer Features (Cancer Type Prediction):** Although the 1463 proteins could be used to distinguish between the 9 different cancer types with a high ROC AUC, routinely measuring such a high number of proteins may not be feasible. We used SelectKBest to determine the top 200 proteins and compared the results of a Logistic Regression Model trained with all 1463 features with a Logistic Regression model trained with the top 200 features.

- **1463 Features**: A Logistic Regression model incorporating a L1 penalty and a maximum of 500 iterations resulted in a training accuracy of 0.7654, ROC AUC of 0.9625, and F-1 score of 0.7662.
- **200 KBest Features:** A Logistic Regression model incorporating a L1 penalty and a maximum of 500 iterations resulted in a training accuracy of 0.7197, ROC AUC of 0.9461, and F-1 score of 0.7208.

Although a loss of model performance was observed with a reduction in features, the final model trained with the 200 KBest features exhibited a testing accuracy of 0.7331, ROC AUC of 0.9486, and F-1 score of 0.7324.

**Strengths**: 1) **Comprehensive Model Comparison**: We tested multiple models across both binary and multiclass tasks, providing a thorough evaluation of their performance. 2) **Feature Selection**: The use of **SelectKBest** helped identify the most important features, improving model performance and reducing dimensionality. 3) **Cross-validation**: Ensured reliable, unbiased performance metrics for each model.

**Limitations**: 1) **Overfitting**: Despite the use of GridSearchCV and hyperparameter tuning, Random Forest and Extra Trees exhibited overfitting. These models performed well on the training set but had high variance when tested on unseen data, resulting in reduced test accuracy. 2) **Computational Cost**: The application of bootstrapping and GridSearchCV made the models computationally expensive. These models were slow to train, especially when dealing with large datasets and high-dimensional protein features (1463 proteins). 3) **Class Imbalance**: Despite using SMOTE for balancing classes, some rare cancer types were still underrepresented, which impacted model performance, especially in multiclass classification tasks. 4) **Data Quality**: The performance of models could have been impacted by the quality of the data and any imputation of missing values.

**Future Work:** Incorporating **additional datasets** from more cancer types or patient populations could improve generalizability and model robustness. Control datasets with detailed health profiles of the participants would enable additional validation of our models. While combining datasets offers numerous advantages, challenges such as batch effects, variability in sample collection, and differences in assay platforms must be addressed. Although we used Olink technology in this study, many other datasets employ different methods for protein measurement. Testing our model's performance on datasets collected using alternative methods could provide valuable insights into its robustness and generalizability.

**Conclusions:** This project highlights the potential of blood protein profiling for machine learning-based cancer detection. The models demonstrated promising results for binary cancer detection and multiclass cancer type classification. With further optimization, expansion of datasets, and improvements in computational efficiency, this work could serve as a foundation for a practical diagnostic tool in clinical applications.