# Final Project Report: Proteomic Signatures for Breast Cancer Prediction

## Abstract

This project explores the use of machine learning to predict breast cancer stage using proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). We developed a pipeline for data preprocessing, including handling class imbalance with SMOTE. A baseline RandomForestClassifier achieved a weighted F1-score of 0.61. We further explored an FT-Transformer model to improve predictive performance. The results demonstrate the potential of using proteomic data for cancer stage prediction, while also highlighting challenges such as class imbalance.

## Background and Motivation

Breast cancer is a heterogeneous disease and a leading cause of cancer-related mortality in women worldwide. Early and accurate diagnosis is critical for improving patient outcomes. While traditional methods rely on histopathology, molecular subtyping offers a more precise characterization of tumors. Proteomics, the large-scale study of proteins, provides a direct readout of the functional state of a cell. By analyzing the proteomic landscape of breast tumors, we can potentially identify molecular signatures associated with different cancer stages and subtypes. This project was motivated by the opportunity to apply advanced machine learning techniques to the rich CPTAC dataset to explore the predictive power of proteomics for breast cancer staging.

## Dataset Summary

The data for this project was sourced from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). The dataset includes:

- \*\*Proteomics Data:\*\* `BRCA\_proteomics\_gene\_abundance\_log2\_reference\_intensity\_normalized\_Tumor.txt`

- \*\*Clinical Metadata:\*\* `BRCA\_meta.txt`

- \*\*Phenotype Data:\*\* `BRCA\_phenotype.txt`

- \*\*Survival Data:\*\* `BRCA\_survival.txt`

The preprocessing pipeline involved loading and merging these files, followed by cleaning and feature scaling. To address the significant class imbalance in the tumor stage labels, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. The final dataset was split into training, validation, and test sets, which were saved for model training and evaluation.

A screen shot of a computer

AI-generated content may be incorrect.

A graph of a patient age

AI-generated content may be incorrect.

A graph with blue squares

AI-generated content may be incorrect.

## Method Description

Our workflow proceeded as follows:

1. \*\*Data Preprocessing:\*\* A standardized preprocessing pipeline was created to load, clean, scale, and split the data. SMOTE was a key component in addressing class imbalance.

2. \*\*Baseline Model:\*\* We first trained a `RandomForestClassifier` as a baseline model to establish initial performance benchmarks.

3. \*\*Advanced Model:\*\* We then implemented and trained an `FT-Transformer`, a deep learning model designed for tabular data, to potentially capture more complex patterns in the data.

4. \*\*Evaluation:\*\* Both models were evaluated on the held-out test set. The primary evaluation metrics were precision, recall, and F1-score, computed for each class and as a weighted average.

## Results

The performance of the baseline `RandomForestClassifier` is summarized in the table below:

| Stage | Precision | Recall | F1-Score | Support |

| :-------- | :-------- | :----- | :------- | :------ |

| Stage I | 0.00 | 0.00 | 0.00 | 1 |

| Stage II | 0.70 | 0.93 | 0.80 | 15 |

| Stage III | 0.50 | 0.17 | 0.25 | 6 |

| \*\*Accuracy\*\* | | | \*\*0.68\*\* | \*\*22\*\* |

| \*\*Weighted Avg\*\* | \*\*0.61\*\* | \*\*0.68\*\* | \*\*0.61\*\* | \*\*22\*\* |

## Conclusion & Discussion

This project successfully demonstrated a machine learning workflow for predicting breast cancer stage from proteomic data. The baseline `RandomForestClassifier` showed moderate performance, but struggled with the underrepresented "Stage I" and "Stage III" classes, as evidenced by the low F1-scores. This highlights the significant challenge of class imbalance in this dataset.

The exploration of the FT-Transformer model represents a step towards more sophisticated modeling. The primary limitation of this study is the small size of the dataset, especially the test set, which makes it difficult to draw robust conclusions. Future work should focus on acquiring larger datasets, exploring more advanced deep learning architectures, and integrating other omics data (e.g., genomics, transcriptomics) for a more comprehensive multi-modal analysis.

## Data and Code Availability

The CPTAC dataset is publicly available through the CPTAC data portal. The code and documentation for this project are available at the following GitHub repository: [https://github.com/your-username/proteobrca](https://github.com/your-username/proteobrca).

## Acknowledgments

We gratefully acknowledge the CPTAC for providing the high-quality dataset used in this study. The data analysis, preprocessing, and model construction were conducted with guidance and suggestions from Google’s Gemini, a large-scale language model that assisted in developing the machine learning pipeline and preparing this report.

## References

1. Clinical Proteomic Tumor Analysis Consortium (CPTAC). [https://proteomics.cancer.gov/programs/cptac](https://proteomics.cancer.gov/programs/cptac)

2. Carrasco-Zanini, J. et al. Proteomic signatures improve risk prediction for common and rare diseases. Nat Med 30, 2489–2498 (2024).

3. You, J. et al. Plasma proteomic profiles predict individual future health risk. Nat Commun 14, 7817 (2023).

4. Enroth, S. et al. High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. Commun Biol 2, 221 (2019).

## Appendices

Most chat history has been recorded in the following \*json files.

Desktop/Courses/DDLS/final\_project/checkpoint-final\_chat.json

Desktop/Courses/DDLS/final\_project/checkpoint-computer-lab-phase2.json

Desktop/Courses/DDLS/final\_project/checkpoint-computer-lab-model\_refinement.json