# **README**

## **Thesis Repository:**

This repository contains the code and data used in the analysis for my Master's thesis

Folder Descriptions:

### **Data/**

* **1\_Breast/** and **2\_Liver/**: Each has:  
  + **raw data/**: Unprocessed clinical or genomic datasets.
  + **[cancer]\_processed/**: Cleaned and feature-engineered datasets used for modeling.

### **Code/**

* **1\_Data Preprocessing and Description**: Scripts for cleaning, transforming, and summarizing raw data.
* **2\_Within-Cancer Initial Regression Models**:  
  + **Breast/** and **Liver/**: Feature selection and preliminary modeling within each cancer type.
* **3\_Data Integration**: Combines results from both cancer types.  
  + **0\_Correlation/**: Visual analyses using correlation plots + csv with overlapping genes selected by liver and breast models
  + **1\_ROC\_AUC/**: Placeholder for ROC-AUC evaluations. This is where the integrated model code can be found (both gridsearch and bayesian optimization methods + sanity check on the code running)
  + **2\_BIC/**: Bayesian Information Criterion comparison across models.
  + **3\_Liver\_Cancer\_Stages\_Validation/**:  
    - **0\_Within\_cancer\_stage\_initial\_models/**: Sub-analysis of stage-wise liver cancer models (stage I II and stage III IV).  
      * **coef\_stage12/**: Coefficients and model outputs for **Stage I & II** patients.
      * **coef\_stage34/**: Coefficients and model outputs for **Stage III & IV** patients.
    - **1\_Correlation/** and **2\_ROC\_AUC/**: Similar evaluations as above, specific to stage-based models.
      * **boxplots/**: Visualizations of distribution differences in selected features between cancer stages.
      * **scatterplots/**: Pairwise feature scatterplots to explore relationships and separability between stages.
      * **csv** with overlapping genes selected by **Stage I & II and Stage III & IV** models
    - **2\_ROC\_AUC/**: Code for the integrated model (gridsearch only) applied to **Stage I & II and Stage III & IV**
    - **Data/**: Processed data used specifically in stage-based liver cancer analysis.
* **4\_Bayesian Statistical Inference**: Scripts and outputs related to Bayesian modeling techniques.  
  + **0\_Two-Step Variational Bayes Logistic Regression/**: Script with implementation of Two-Step Variational Bayes Logistic Regression. Uses gene expression data on patients in stages III/IV and prior information from the elastic net survival model