



Cancer Detection with Machine Learning – CancerSEEK

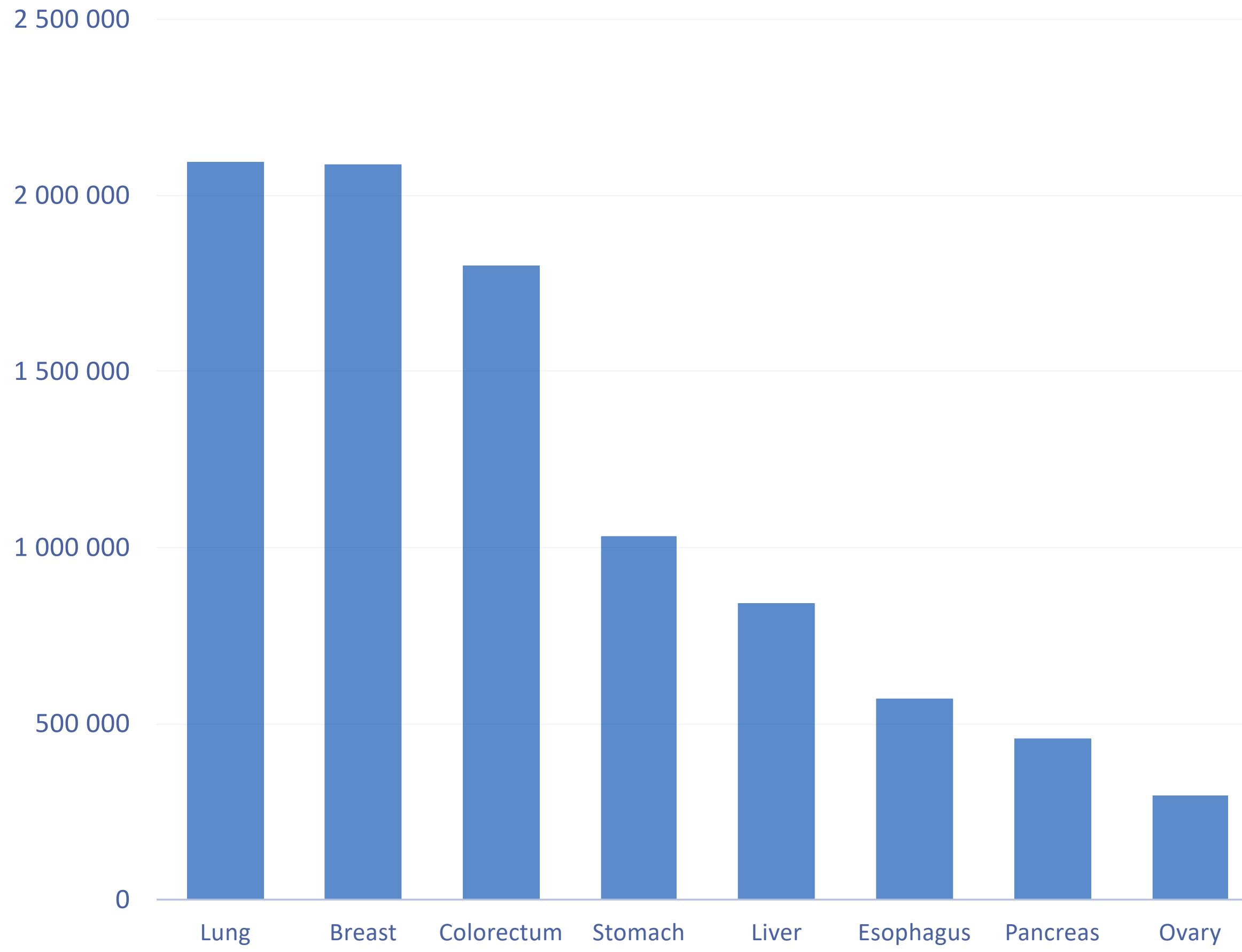
From a Technical Perspective

Background

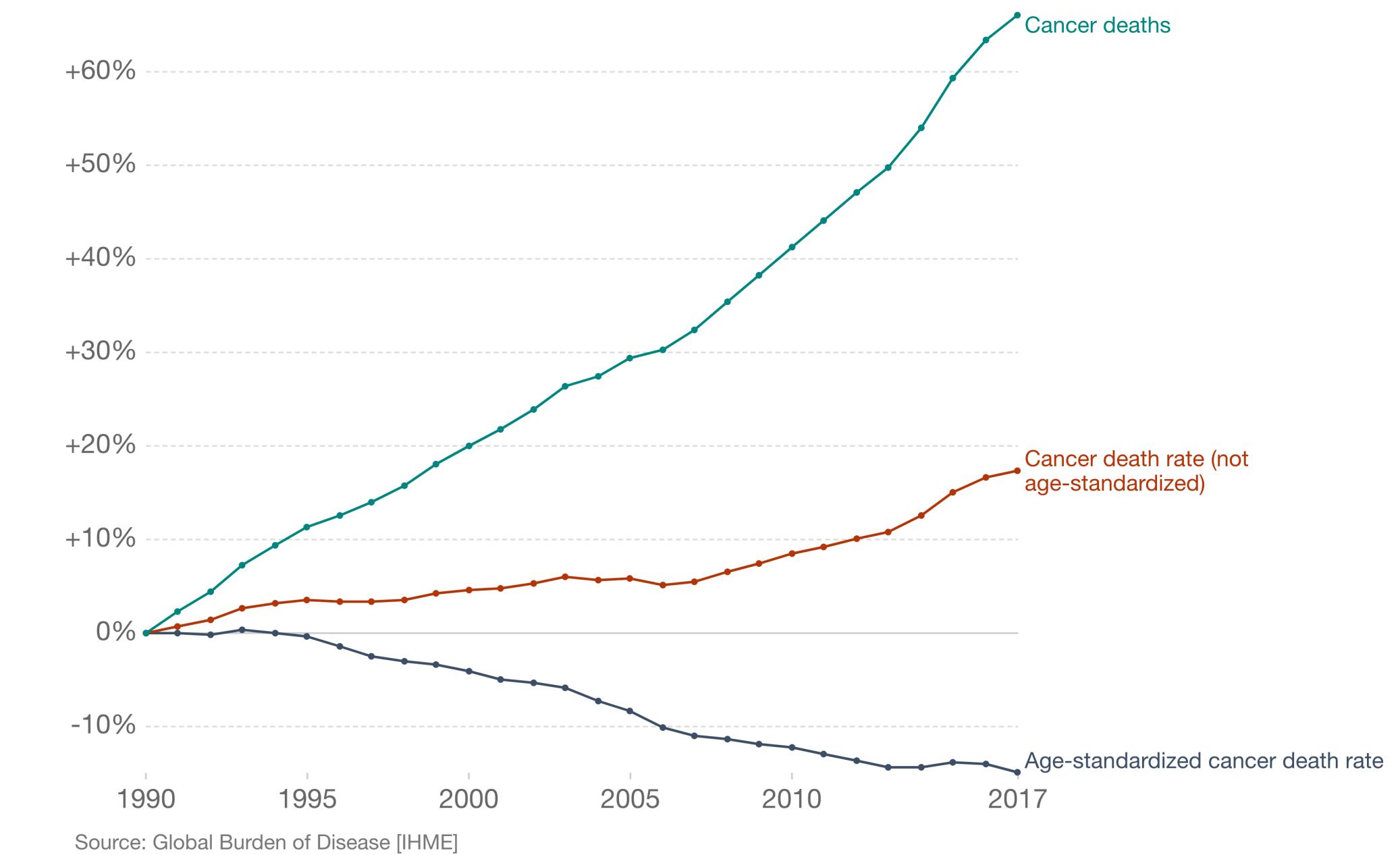
- **Background**
- Different Approaches
- Common Steps
 - Missing Values
 - Feature Transformation
 - Data Visualisation
 - Experimentation
 - Pipeline
- Results
 - Cancer Type Classification (as in publication)
 - Cancer Type Classification (full dataset)
 - Cancer Type Classification (Aneuploidy dataset)
- Conclusions

Background

Cancer Cases 2018 covered by CancerSEEK



Change in three measures of cancer mortality, World, 1990 to 2017



- 18 M new cancer cases and roughly 10M deaths in the world 2018.
- CancerSEEK ≈ 9.2M
- Market size 128 B
- A chance to make a positive impact on the world & people around you!

Different Approaches

- Background
- **Different Approaches**
- Common Steps
 - Missing Values
 - Feature Transformation
 - Data Visualisation
 - Experimentation
 - Pipeline
- Results
 - Cancer Type Classification (as in publication)
 - Cancer Type Classification (full dataset)
 - Cancer Type Classification (Aneuploidy dataset)
- Conclusions

Different Approaches

- Tumor Classification on 626 Cancer Samples (as in publication)
 - Full Feature set
- Tumor Classification on Full Dataset
 - Multiclass Classification
- New Sequencing Technique (follow-up publication)
 - Full 10-Feature Dataset

Common Steps

- Background
- Different Approaches
- **Common Steps**
 - Missing Values
 - Feature Transformation
 - Data Visualisation
 - Experimentation
 - Pipeline
- Results
 - Cancer Type Classification (as in publication)
 - Cancer Type Classification (full dataset)
 - Cancer Type Classification (Aneuploidy dataset)
- Conclusions

Common Steps (1)

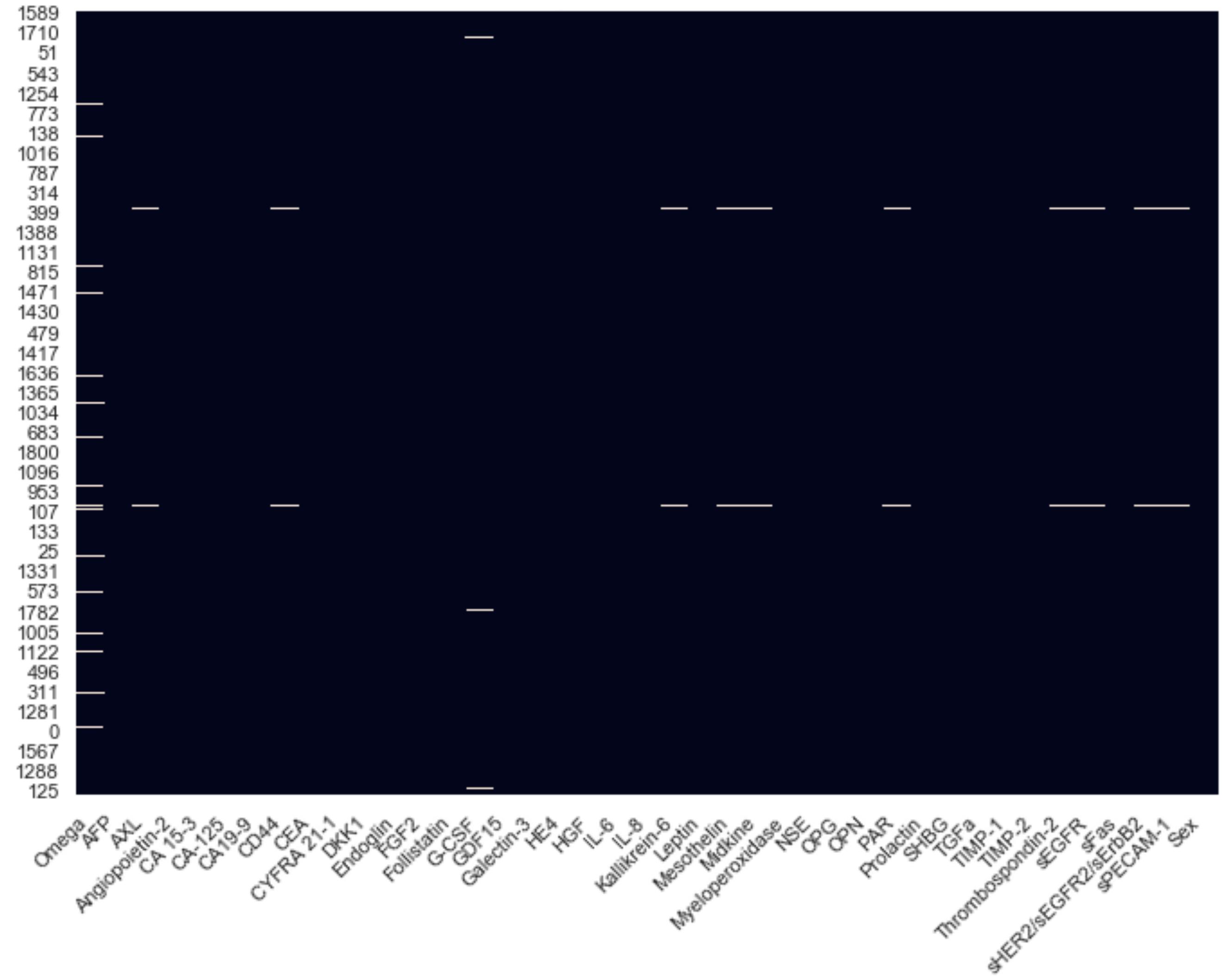


Missing values

Common Steps (1)

- Missing Values

- Combine three datasets
- Remove redundant variables
- Few missing values. Replace with null.
- Dummy variable for Sex



Common Steps (2)



Feature Transformation

Common Steps (2)

- Feature Transformation

```
class PercentileTransformer(BaseEstimator, TransformerMixin):
    ''' Custom transformer that replaces all cancer samples that
        are lower than the healthy 95th percentile with zero.
    '''

    # Class constructor
    def __init__(self, percentile=.95):
        self.percentile = percentile

    # Return self
    def fit(self, X, y):

        # Check if X is DataFrame, if not convert it
        if not isinstance(X, pd.DataFrame):
            X = pd.DataFrame(X)

        # Create copy and fill NaN values with zero
        X = X.fillna(0.0)

        # Calculate thresholds for each column
        thres = X.loc[y == 9, :].quantile(q=self.percentile,
                                         interpolation='linear').to_dict()

        # Zero threshold for Omega
        thres['Omega'] = 0.0

        # Store for later use
        self.thres = thres
        return self
```

```
# Custom transform method to replace cancer values
# that are below the healthy 95th percentile
def transform(self, X, y=None):

    # If X is not DataFrame, convert it to DataFrame
    if not isinstance(X, pd.DataFrame):
        X = pd.DataFrame(X)

    # Create copy and fill NaN values with zero
    X_ = X.copy(deep=True)
    X_ = X_.fillna(0.0)

    # Replace values lower than the (95th) percentile
    for p in self.thres:
        X_[p] = X_[p].apply(lambda x: 0 if x < self.thres[p] else x)

    return X_
```

Common Steps (3)

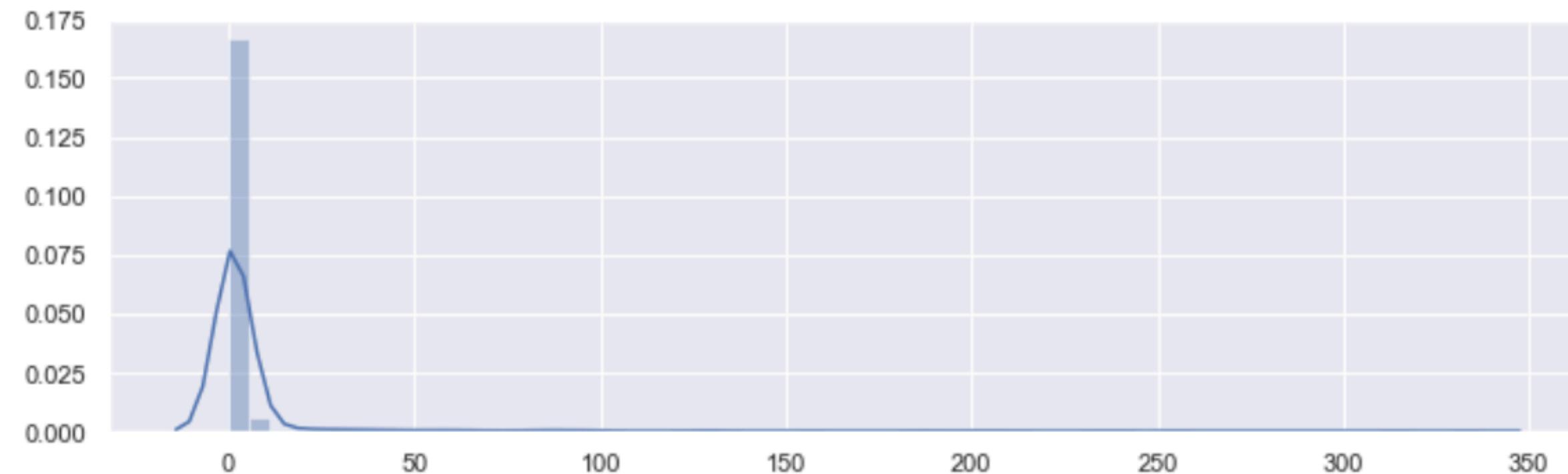


Data Visualisation

Common Steps (3)

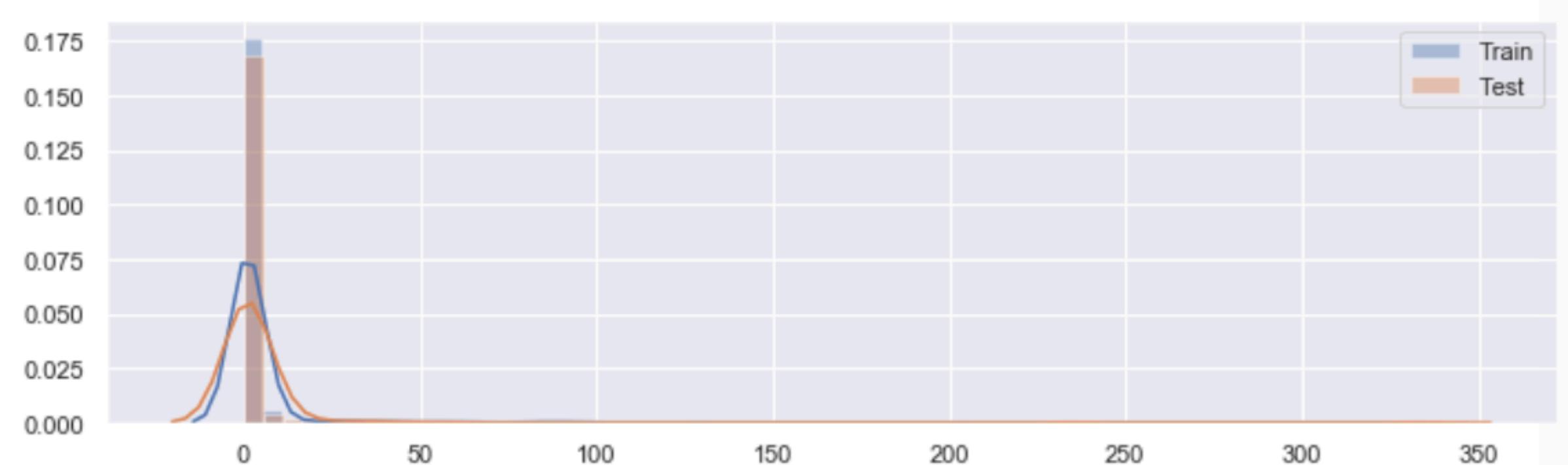
- Data Visualisation – Before & After Custom Transformation

Before

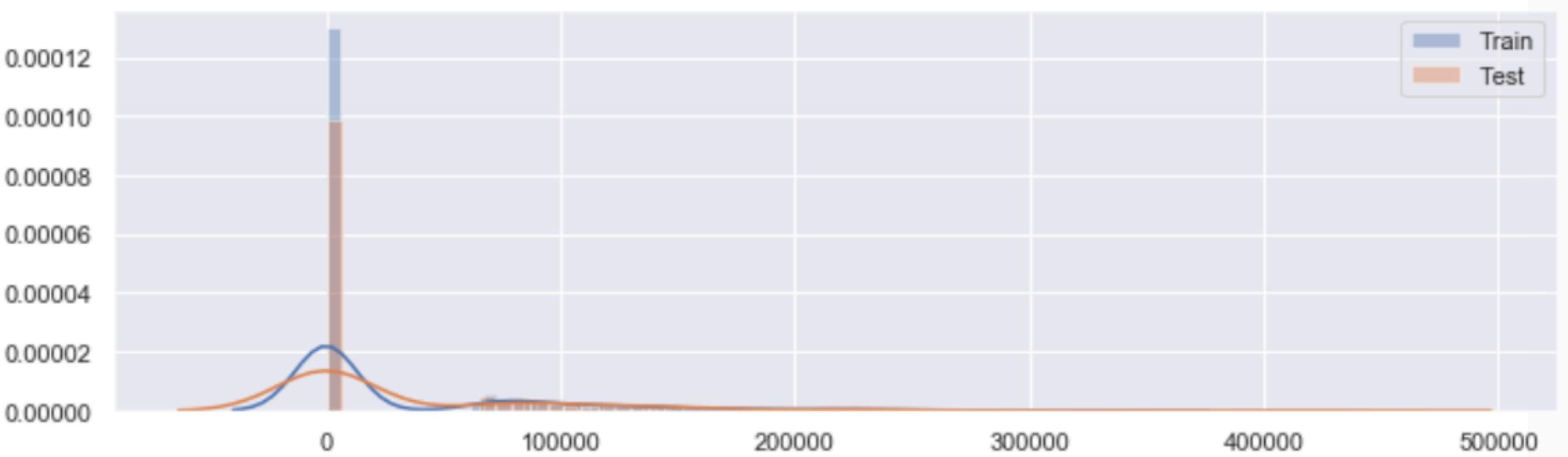
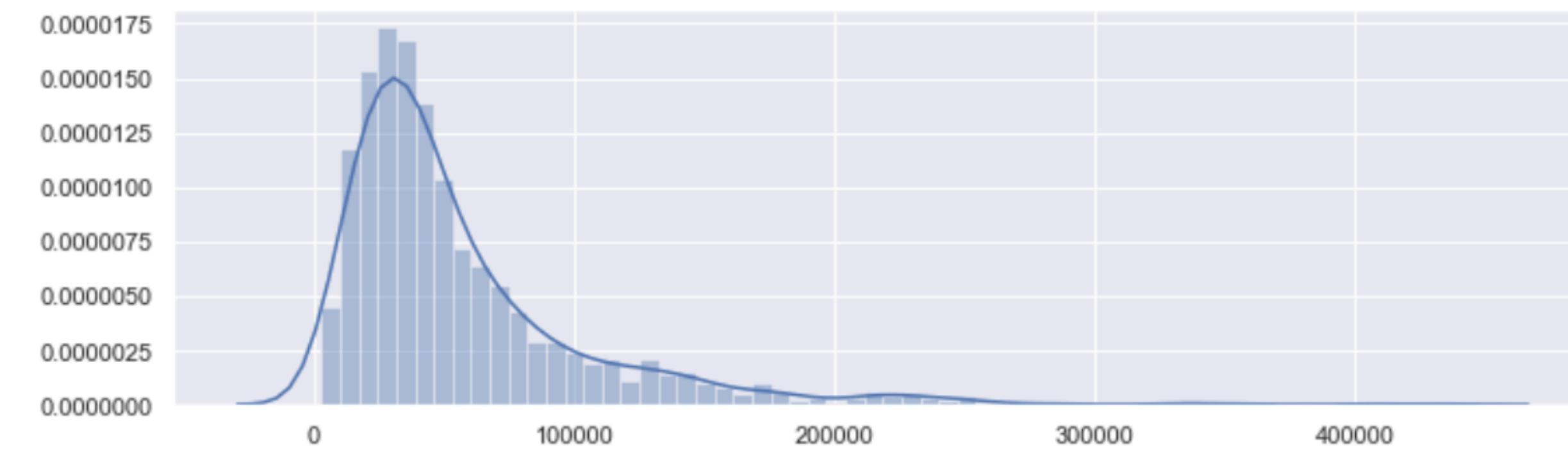


Omega

After

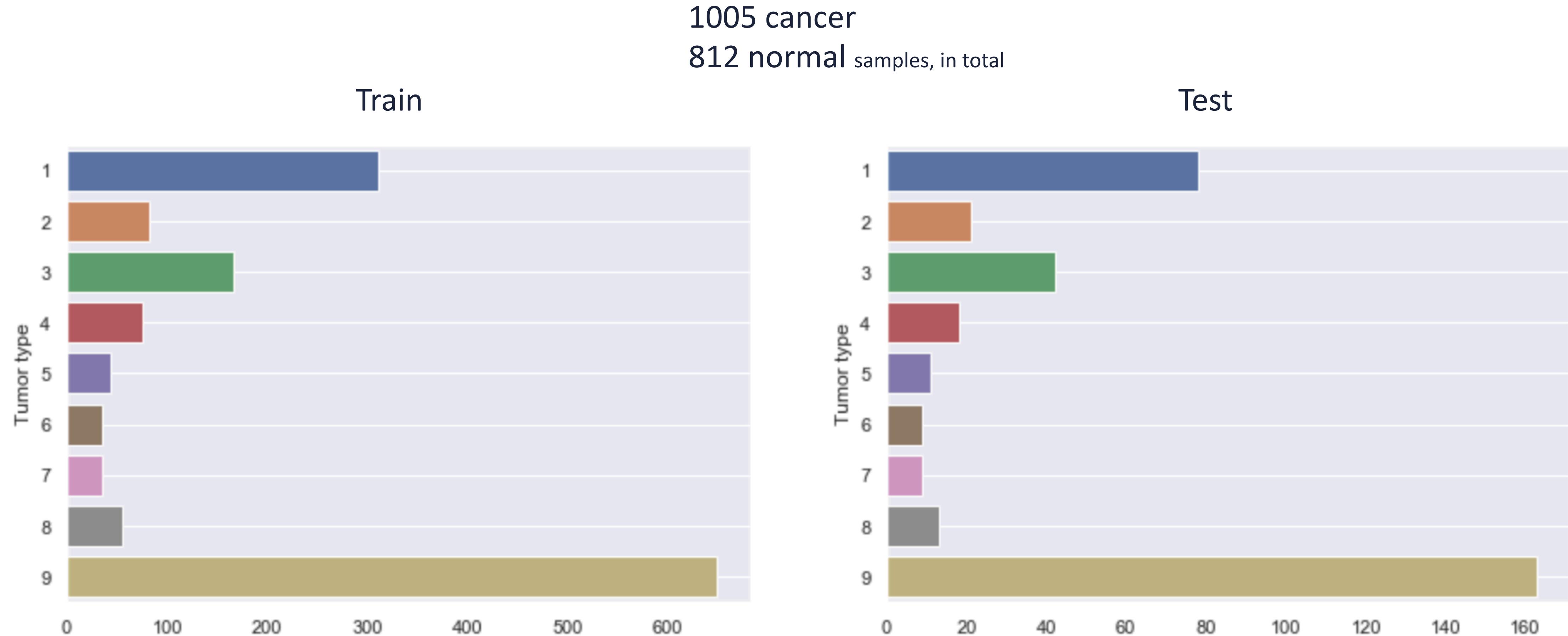


OPN



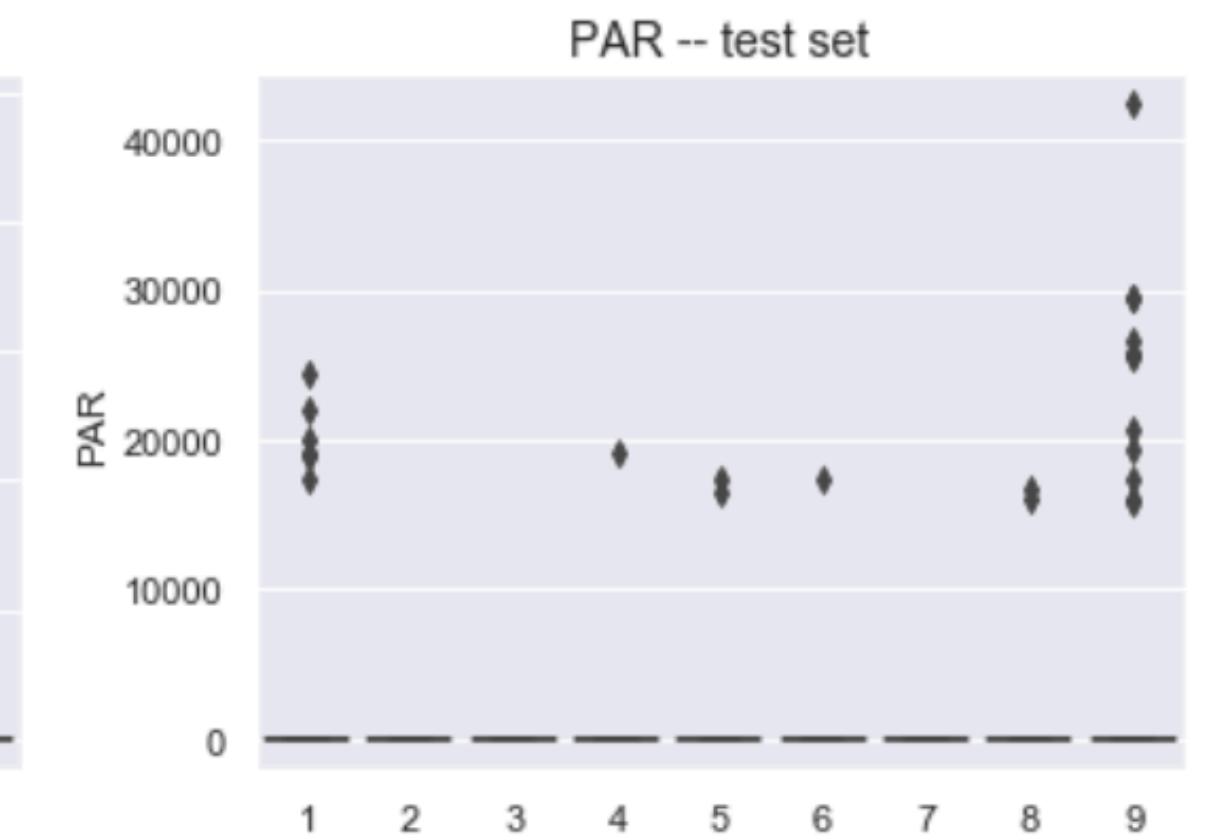
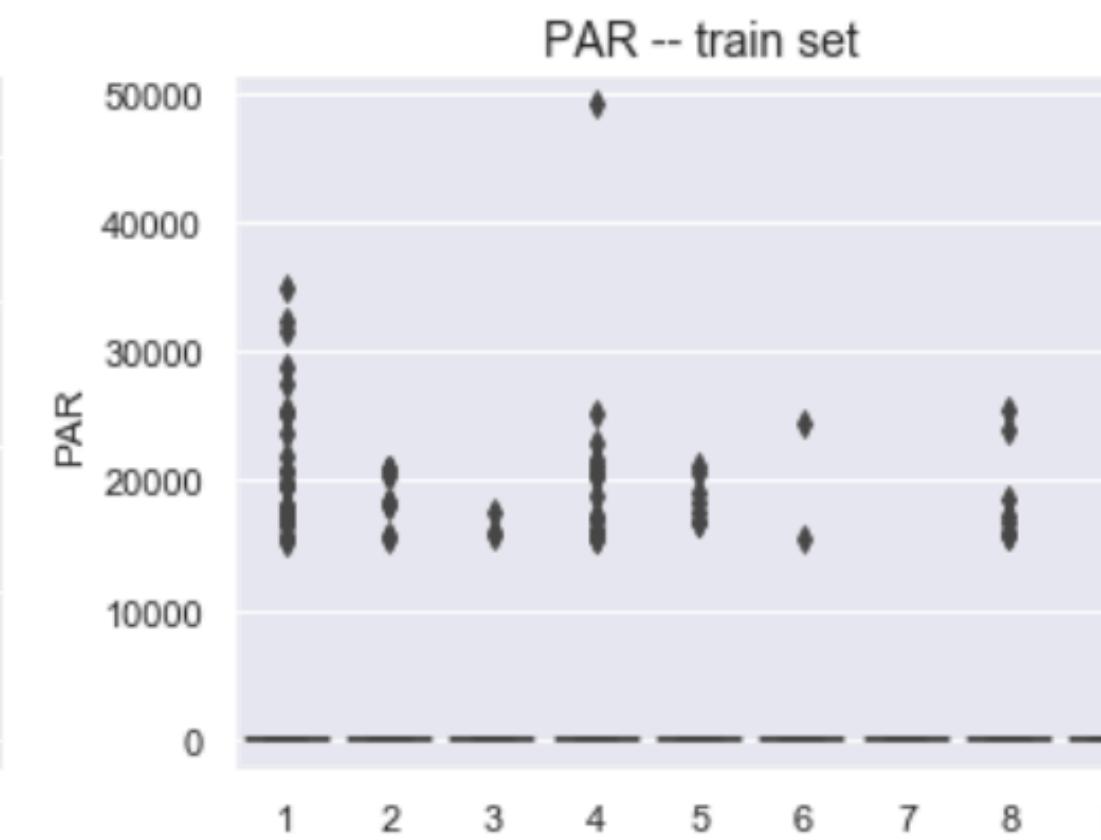
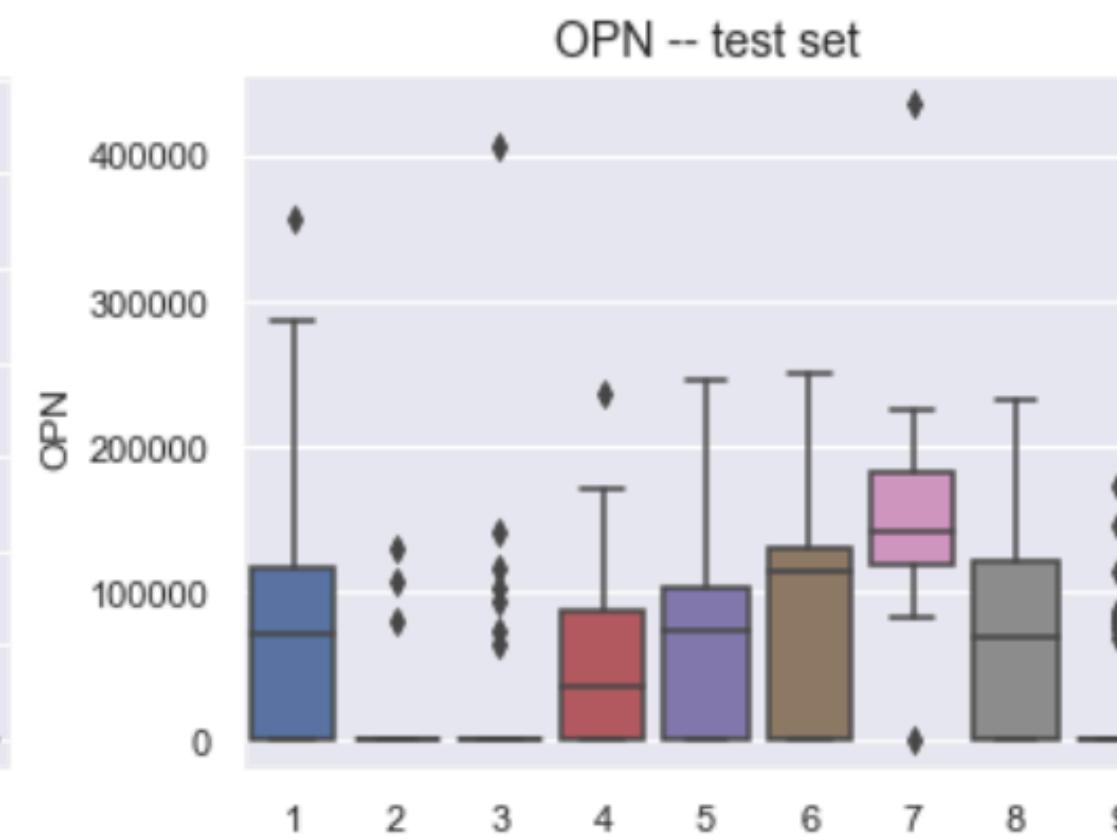
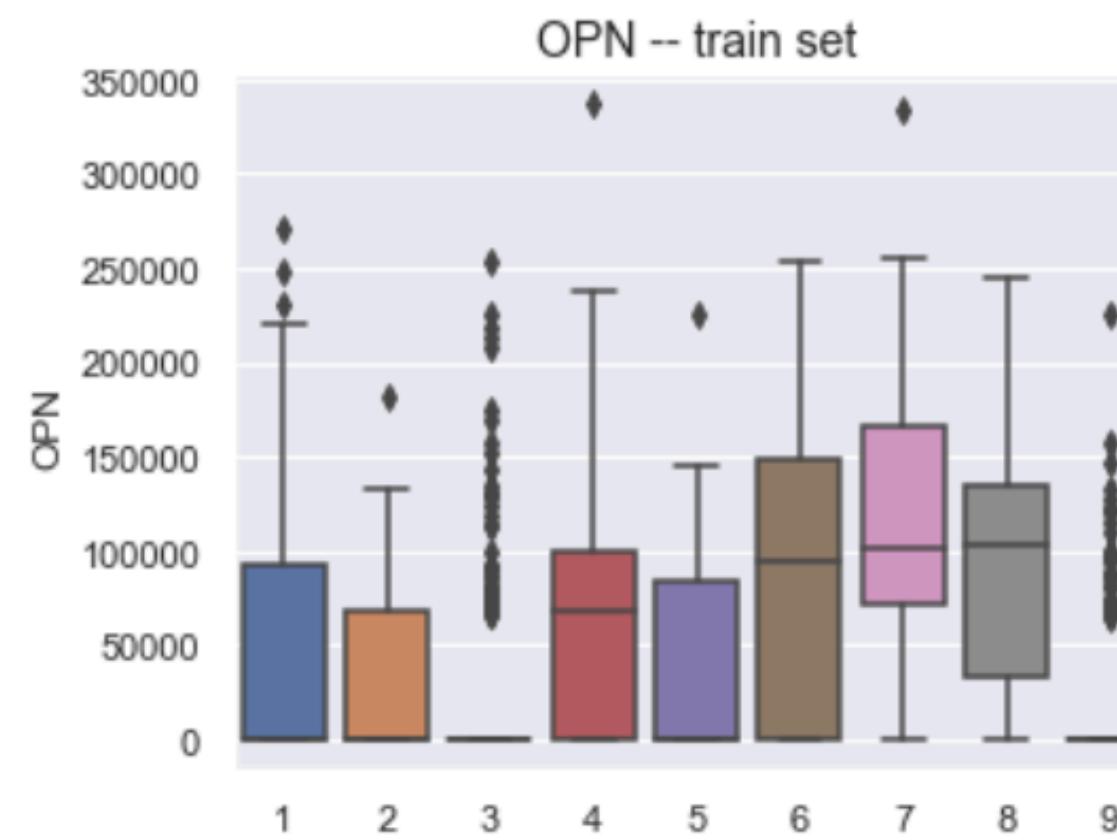
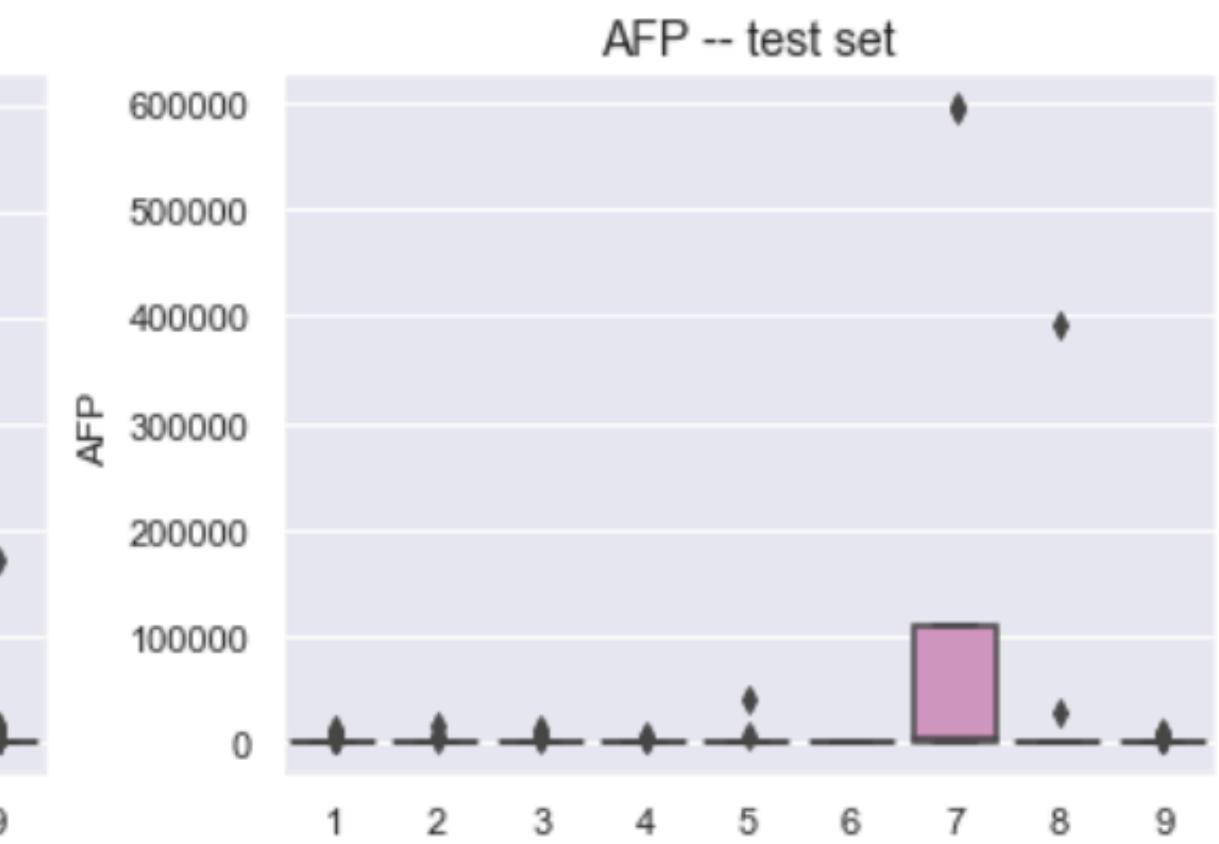
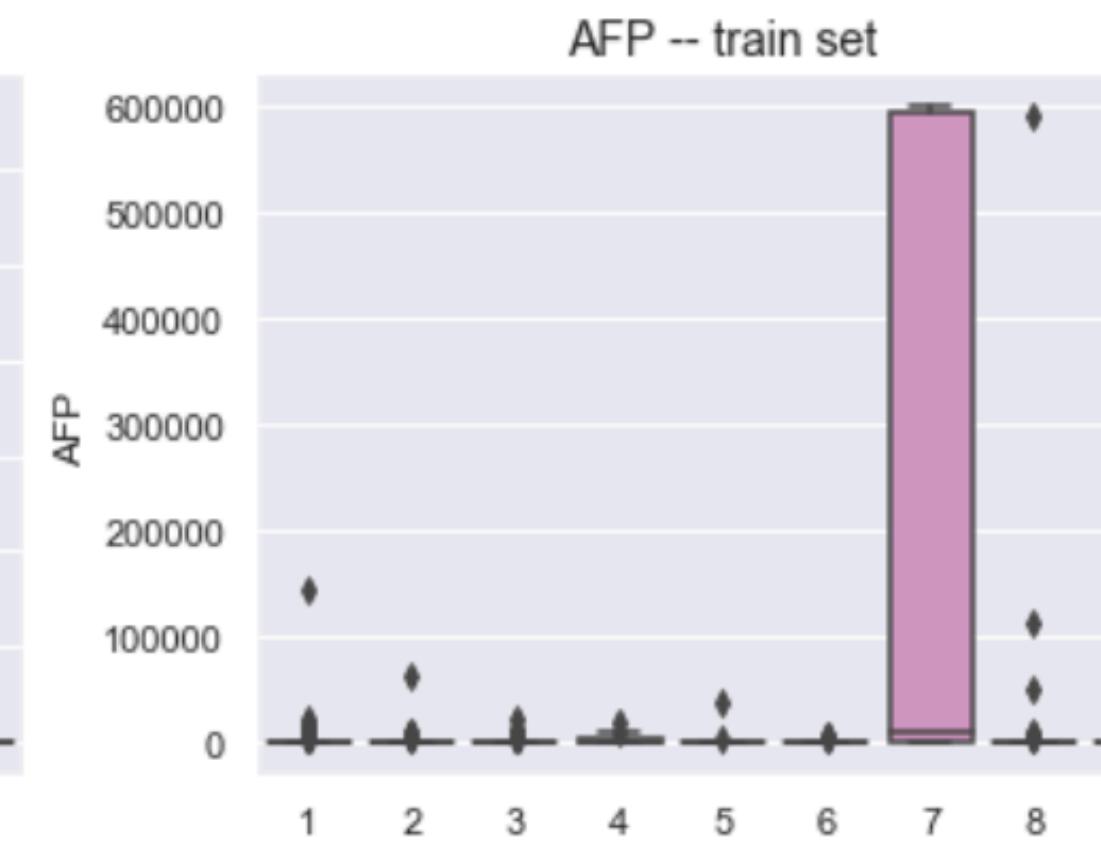
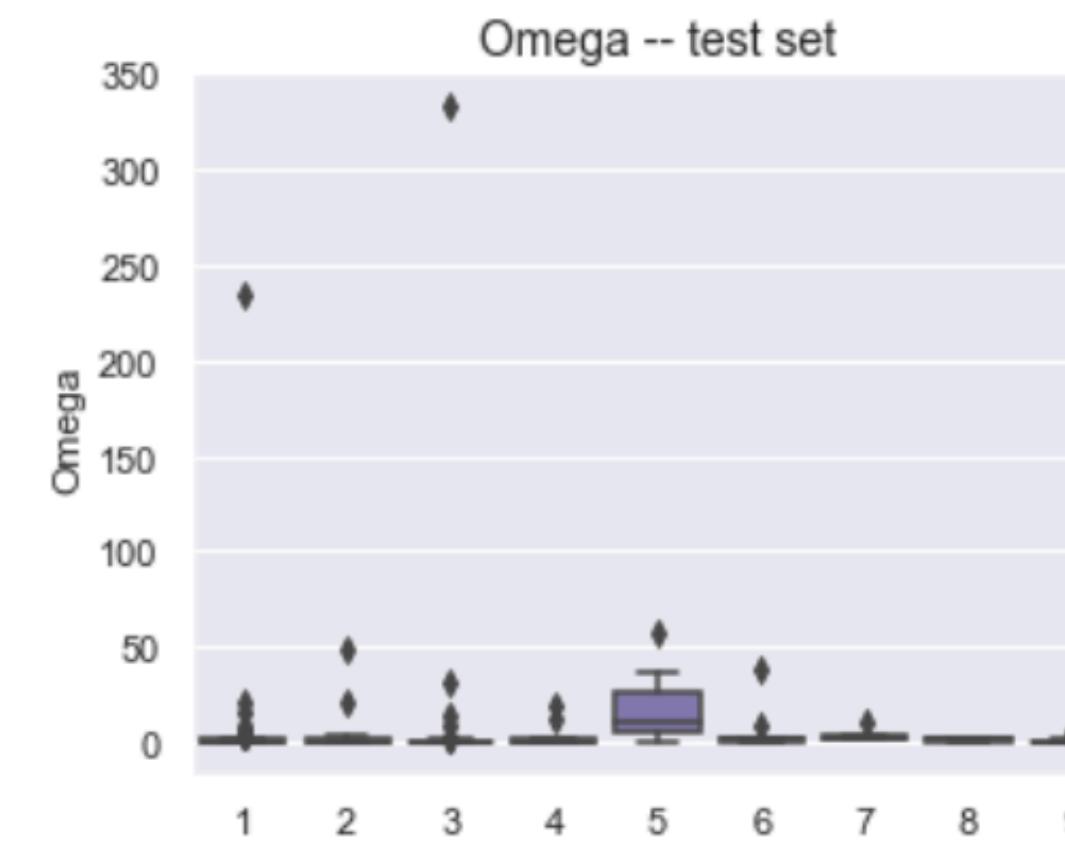
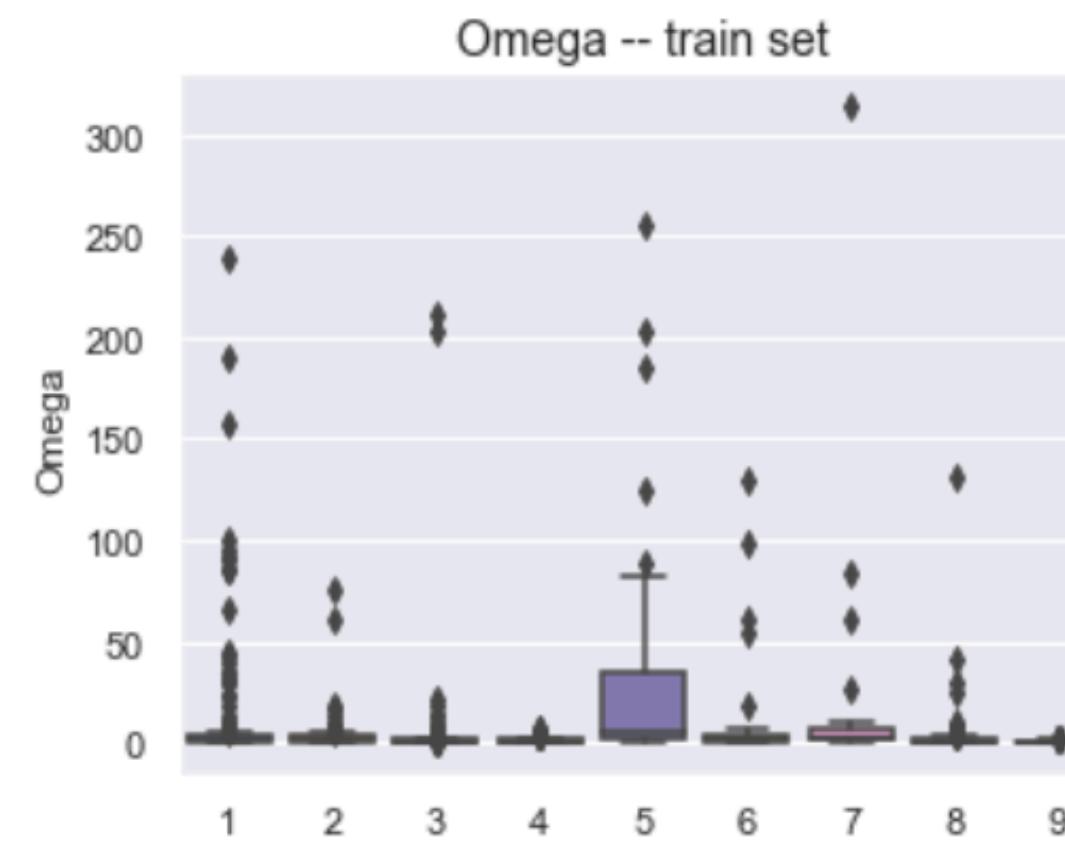
Common Steps (3)

- Data Visualisation – Tumor Counts on Train & Test Sets



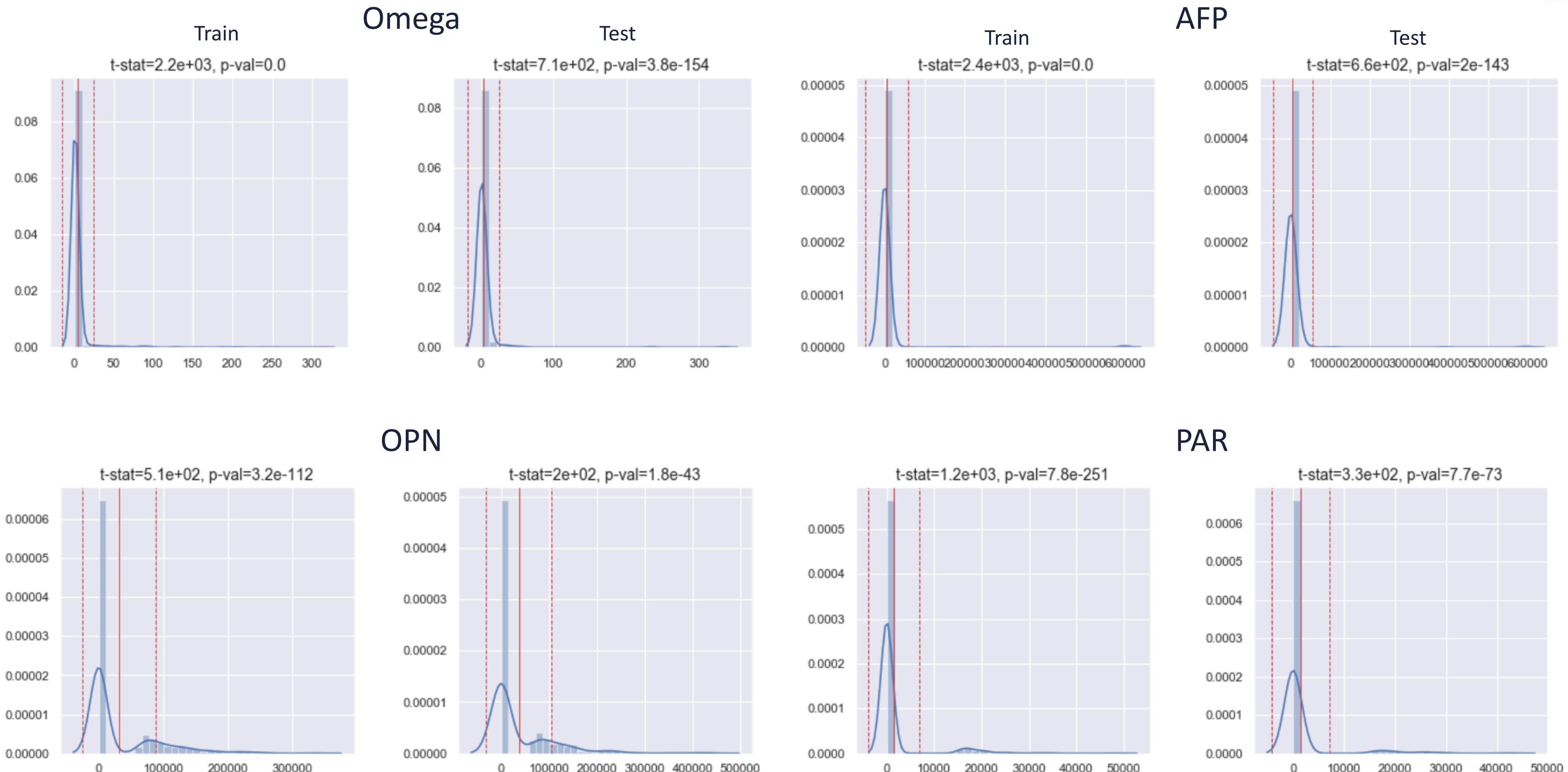
Common Steps (3)

- Data Visualisation – Distribution per Tumor Type after Custom Transformation



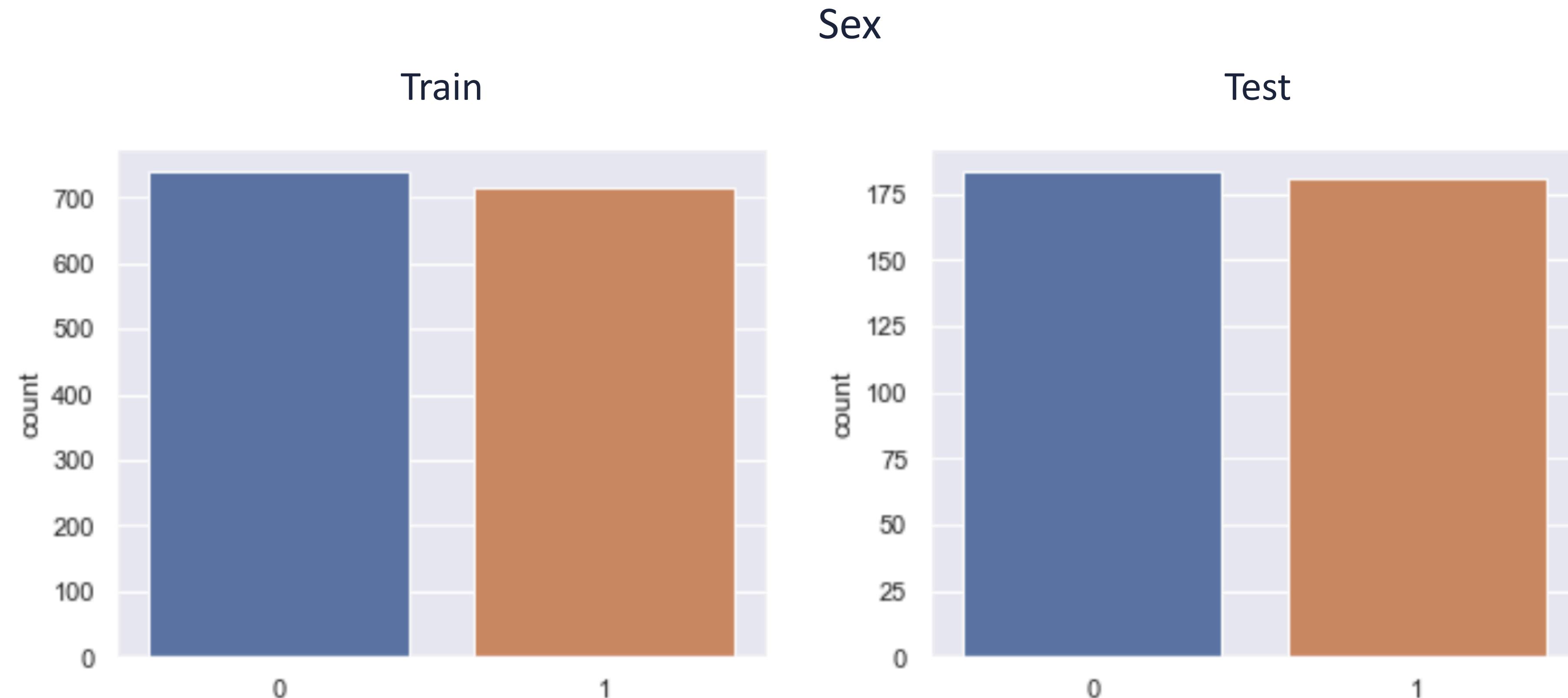
Common Steps (3)

- Data Visualisation – Test for Normality



Common Steps (3)

- Data Visualisation – Categorical Variable's Distribution



Common Steps (3)

- Data Visualisation – Correlation Matrix

Low multicollinearity

Highest correlation with target variable:
Prolactin, OPN, TIMP-1, IL-6, HE4

Tumor type	1	-0.11	0.037	-0.11	0.099	0.047	-0.04	0.018	0.015	0.087	0.039	0.002	0.006	0.050	0.082	0.093	-0.11	-0.11	0.028	-0.12	-0.11	-0.15	0.048	0.019	0.016	0.039	0.087	-0.1	9e-06	0.046	-0.26	0.072	-0.31	0.073	0.003	-0.15	0.004	0.069	0.034	0.007	0.089	0.006	0.087
Omega	-0.11	1	0.026	0.007	0.084	0.2	0.3	0.19	0.098	0.24	0.34	0.073	0.022	0.035	0.075	0.024	0.1	0.028	0.13	0.07	0.01	0.0095	0.04	0.029	0.098	0.0098	0.042	0.0049	0.015	0.16	0.049	0.11	0.007	0.071	0.14	0.031	0.13	0.026	0.005	0.25	0.0028	4e-0	
AFP	0.037	0.026	1	0.062	0.054	0.058	0.006	0.044	0.02	-0.01	0.0024	0.021	0.016	0.063	0.018	0.062	0.084	0.072	0.015	0.19	0.098	0.019	0.066	0.02	0.0038	0.0085	0.24	0.0033	0.011	0.17	-0.022	0.045	-0.009	0.068	0.076	0.023	0.11	0.019	0.02	-0.018	0.01	0.048	
AXL	-0.11	0.007	0.062	1	0.089	0.006	0.04	0.028	0.28	0.026	0.013	0.014	0.095	0.026	0.041	0.019	0.2	-0.007	0.089	0.05	0.029	0.002	0.098	0.019	0.064	0.039	0.0022	0.010	0.004	0.17	0.24	0.011	0.051	0.069	0.26	0.22	0.28	0.13	0.064	0.22	0.21	0.03	
Angiopoietin-2	0.099	0.084	0.054	0.089	1	0.048	0.049	0.014	0.017	0.052	0.069	0.039	0.12	0.017	0.13	0.19	0.32	0.05	0.052	0.2	0.19	0.0004	0.002	0.047	0.041	0.058	0.06	0.009	0.053	0.21	0.12	0.095	0.095	0.044	0.22	0.016	0.28	-0.03	0.022	0.056	0.09	0.029	
CA 15-3	0.047	0.2	-0.0058	0.006	0.048	1	0.41	0.12	0.015	0.092	0.24	0.056	0.0028	0.039	0.017	0.006	0.006	0.003	0.055	0.006	-0.01	0.0023	0.22	-0.0059	0.19	0.0087	0.007	0.017	0.022	0.087	0.003	0.12	0.013	0.0026	0.1	0.0066	0.099	0.003	0.11	0.18	-0.009	0.061	
CA-125	-0.04	0.3	-0.006	0.004	0.049	0.41	1	0.34	0.014	0.17	0.5	0.13	0.033	0.072	0.031	0.015	0.075	0.038	0.15	0.047	0.012	0.069	0.39	-0.02	0.21	0.013	0.037	0.018	0.015	0.14	0.046	0.071	0.017	0.011	0.2	0.046	0.15	0.013	0.025	0.39	-0.015	0.087	
CA19-9	0.018	0.19	0.004	0.028	0.014	0.12	0.34	1	0.04	0.083	0.13	0.23	0.0052	0.12	0.010	0.004	0.081	0.1	0.075	0.089	0.008	0.018	0.005	0.018	0.0039	0.19	0.041	0.012	0.006	0.13	0.066	0.031	0.0029	0.003	0.14	0.016	0.082	0.0097	0.023	0.12	0.01	0.043	
CD44	0.018	0.099	0.02	0.28	-0.017	0.015	0.014	0.04	1	0.013	0.009	0.066	0.088	0.062	0.14	0.032	0.18	0.042	0.12	0.008	0.003	0.007	0.23	0.0092	0.24	0.07	0.028	0.092	0.069	0.14	0.14	0.0059	0.02	0.003	0.19	0.22	0.15	0.14	0.089	0.16	0.11	0.0054	
CEA	0.087	0.24	-0.01	0.026	0.052	0.092	0.17	0.083	0.011	1	0.41	0.046	0.046	0.039	0.006	0.044	0.27	0.3	-0.0079	0.19	0.017	0.033	0.0036	0.042	0.007	0.003	0.13	0.022	0.0086	0.16	0.058	0.026	0.019	0.084	0.12	0.022	0.089	0.026	0.004	70.22	-0.02	0.0062	
CYFRA 21-1	0.039	0.34	0.0024	0.013	0.069	0.24	0.5	0.13	0.009	0.41	1	0.096	0.089	0.06	0.002	0.05	0.34	0.11	0.015	0.082	0.008	0.032	0.018	0.006	0.05	0.0099	0.035	0.071	0.003	0.17	0.045	0.01	0.012	0.069	0.19	0.001	0.25	0.013	0.05	0.71	-0.01	0.045	
DKK1	0.0026	0.073	0.02	0.014	0.039	0.056	0.13	0.23	0.066	0.046	0.096	1	0.066	0.14	0.079	0.018	0.083	0.058	0.093	0.067	0.022	0.008	0.045	0.01	0.046	0.017	0.025	0.35	0.051	0.049	0.19	0.024	0.031	0.1	0.19	0.02	0.062	9e-0	0.19	0.069	0.089	0.022	
Endoglin	0.0065	0.022	0.016	0.095	0.12	0.0028	0.033	0.0052	0.088	0.0046	0.089	0.066	1	0.006	0.085	0.023	0.094	0.014	0.0097	0.046	-0.01	0.0067	0.06	0.032	0.0028	0.026	0.033	0.046	0.029	0.015	0.032	0.004	0.02	0.032	0.1	0.13	0.17	0.086	0.076	0.18	0.12	0.013	
FGF2	0.082	0.035	0.063	0.026	0.017	0.039	0.072	0.12	0.062	0.039	0.06	0.14	0.0061	1	0.073	0.031	0.092	0.019	0.15	0.094	0.014	0.0018	0.006	0.021	0.0028	0.017	0.085	0.11	0.0042	0.12	0.083	0.1	0.013	0.11	0.07	0.086	0.078	0.018	0.07	0.098	0.043	0.052	
Follistatin	0.093	0.075	0.018	0.041	0.13	0.017	0.031	0.014	0.14	0.0066	0.002	0.079	0.085	0.073	1	0.078	0.13	-0.007	0.058	0.046	0.005	0.042	0.067	0.014	0.089	0.069	0.024	0.067	0.07	0.02	0.055	0.09	0.16	0.085	0.057	0.072	0.046	0.015	0.046	0.05	0.048	0.063	0.043
G-CSF	-0.11	0.024	0.062	0.28	0.019	0.006	0.016	0.04	0.032	0.004	0.05	0.018	0.046	0.039	0.004	0.044	0.027	0.09	0.06	0.021	0.007	0.02	0.012	0.069	0.19	0.001	0.25	0.013	0.037	0.011	0.016	0.026	0.044	0.037	0.011	0.012	0.026	0.044					
GDF15	-0.11	0.1	0.084	0.2	0.32	0.006	0.075	0.081	0.18	0.27	0.34	0.083	0.094	0.092	0.13	0.097	1	0.2	0.089	0.23	0.12	0.046	0.086	0.023	0.056	0.068	0.072																

Common Steps (4)

Experimentation



Data Transformation,
Feature Engineering,
Feature Selection
& Algorithms

Common Steps (4)

- Experimentation – Transformations & Algorithms

1st
2nd

	Orig	Orig_sca	Wins_H	Wins_H_sca	Wins_A	Wins_A_sca	Log	Log_sca	BoxC	BoxC_sca	YeoJ	YeoJ_sca	YeoJ_WH	YeoJ_WH_sca
Specificity_NB	0.9509	0.9506	1.0000	1.0000	0.8834	0.9136	0.9568	0.9568	0.9568	0.9568	0.9259	0.9259	0.9816	0.9816
Sensitivity_NB	0.5651	0.5871	0.5971	0.6106	0.5882	0.6195	0.6431	0.6431	0.6243	0.6243	0.6490	0.6490	0.6932	0.6932
Specificity_LR	NaN	0.9753	NaN	1.0000	NaN	0.9753	NaN	0.9753	NaN	0.9815	NaN	0.9815	NaN	1.0000
Sensitivity_LR	NaN	0.6755	NaN	0.6962	NaN	0.6844	NaN	0.6814	NaN	0.6686	NaN	0.6805	NaN	0.7168
Specificity_SGD	0.4451	0.9816	0.5679	1.0000	0.3110	0.9877	0.8704	0.9877	0.9877	0.9877	0.9877	0.9877	1.0000	1.0000
Sensitivity_SGD	0.3284	0.6519	0.3333	0.6844	0.2308	0.6755	0.5723	0.6794	0.6441	0.6785	0.6844	0.6755	0.7041	0.7168
Specificity_KNN	0.8951	0.9571	0.9877	1.0000	0.8704	0.9694	0.9877	0.9877	0.9877	0.9877	1.0000	1.0000	1.0000	1.0000
Sensitivity_KNN	0.5634	0.6598	0.6176	0.6873	0.5457	0.6627	0.6647	0.6647	0.6519	0.6549	0.6844	0.6873	0.6853	0.6873
Specificity_SVC	NaN	0.9755	NaN	1.0000	NaN	0.9877	NaN	0.9877	NaN	0.9815	NaN	0.9877	NaN	1.0000
Sensitivity_SVC	NaN	0.6824	NaN	0.6953	NaN	0.6844	NaN	0.6873	NaN	0.6785	NaN	0.6971	NaN	0.7257
Specificity_DT	0.9693	0.9693	1.0000	1.0000	0.9694	0.9694	0.9694	0.9694	0.9694	0.9632	0.9444	0.9694	1.0000	1.0000
Sensitivity_DT	0.6224	0.6224	0.6657	0.6647	0.6450	0.6450	0.6450	0.6450	0.6450	0.6147	0.6213	0.6450	0.6549	0.6676
Specificity_RF	0.9753	0.9877	1.0000	1.0000	0.9753	0.9753	0.9753	0.9753	0.9691	0.9691	0.9691	0.9691	1.0000	1.0000
Sensitivity_RF	0.7168	0.7176	0.7337	0.7375	0.7071	0.7071	0.7286	0.7265	0.7257	0.7198	0.7249	0.7257	0.7353	0.7367
Specificity_GB	0.9877	0.9816	1.0000	1.0000	0.9816	0.9877	0.9816	0.9877	0.9877	0.9877	0.9816	0.9877	1.0000	1.0000
Sensitivity_GB	0.7670	0.7840	0.7876	0.7876	0.7870	0.7758	0.7751	0.7781	0.7788	0.7699	0.7722	0.7824	0.7817	0.7882
Specificity_XGB	0.9877	0.9877	1.0000	1.0000	0.9815	0.9815	0.9877	0.9877	0.9877	0.9877	0.9877	0.9877	1.0000	1.0000
Sensitivity_XGB	0.7781	0.7758	0.7882	0.7882	0.7817	0.7817	0.7870	0.7870	0.7788	0.7811	0.7847	0.7817	0.7847	0.7929

Common Steps (4)

- Experimentation – Feature Engineering and Selection

Feature Engineering

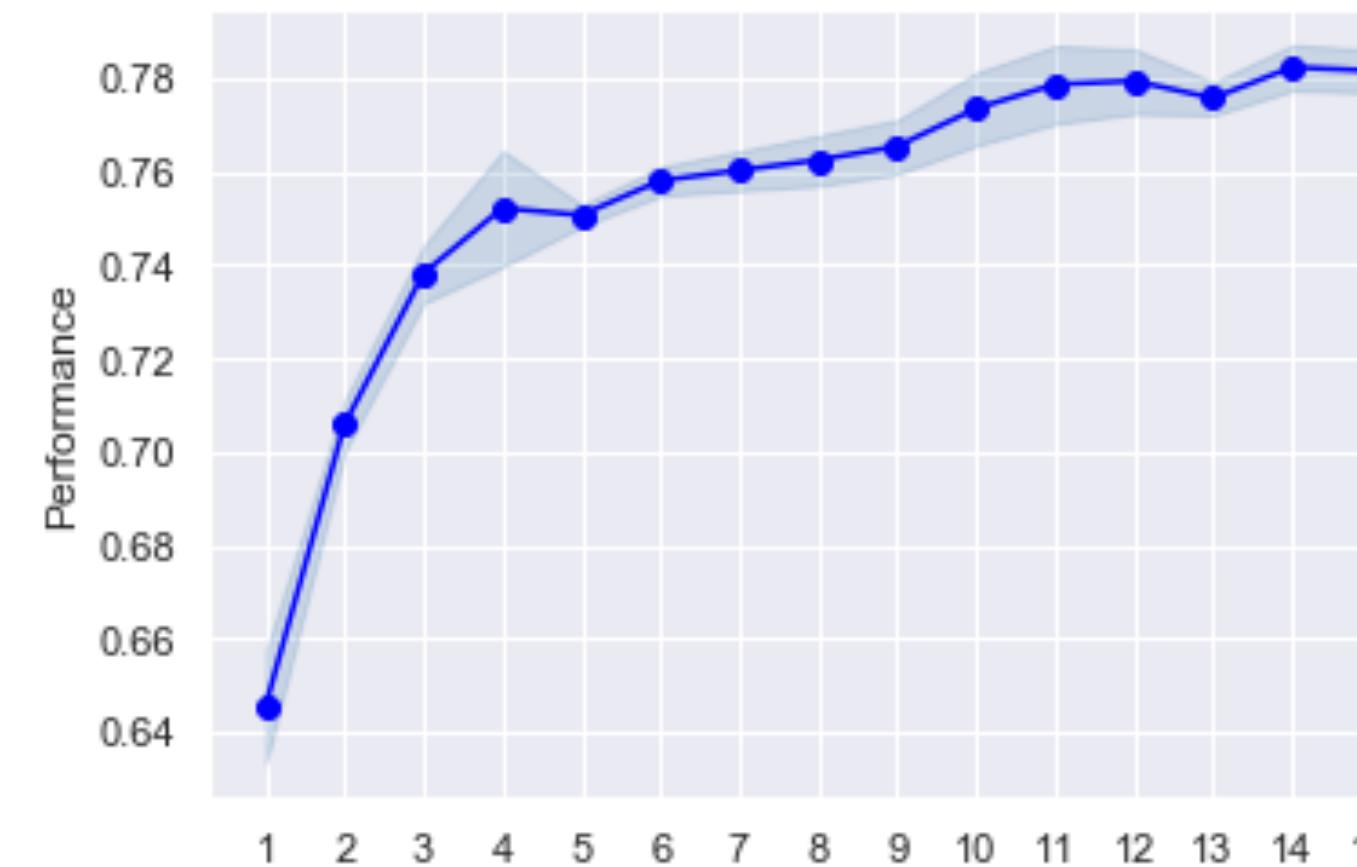
- *Omega, CA19-9, CEA, HGF, OPN*
(as in publication)
- Many other combinations



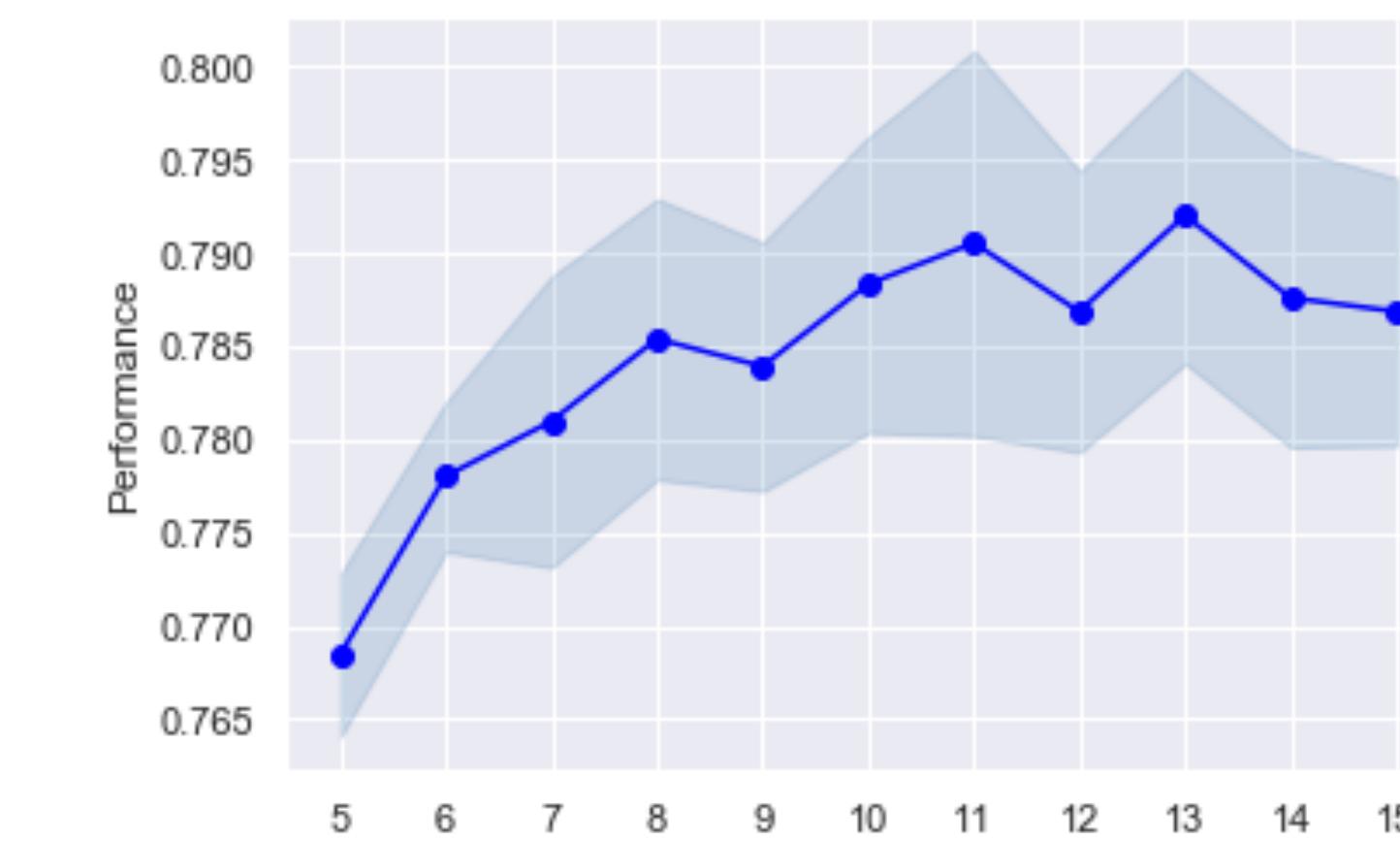
Feature Selection

- Recursive Feature Elimination (RFE)
- Select From Model
- Select K Best
- Sequential Forward Selection (SFS)
- Sequential Forward Floating Selection (SFFS)
- Sequential Backward Selection (SBS)
- Sequential Backward Floating Selection (SBFS)
- Exhaustive Feature Selection (EFS)

SFS CatBoost



SBFS XGBoost



Common Steps (5)



Pipeline

Common Steps (5)

- Pipeline

Naïve Bayes

Logistic Regression (+ SGD)

K-Nearest Neighbors

Support Vector Machine

Decision Trees

Sensitivity: 0.715 (0.714)
AUC: 0.885 (0.885)

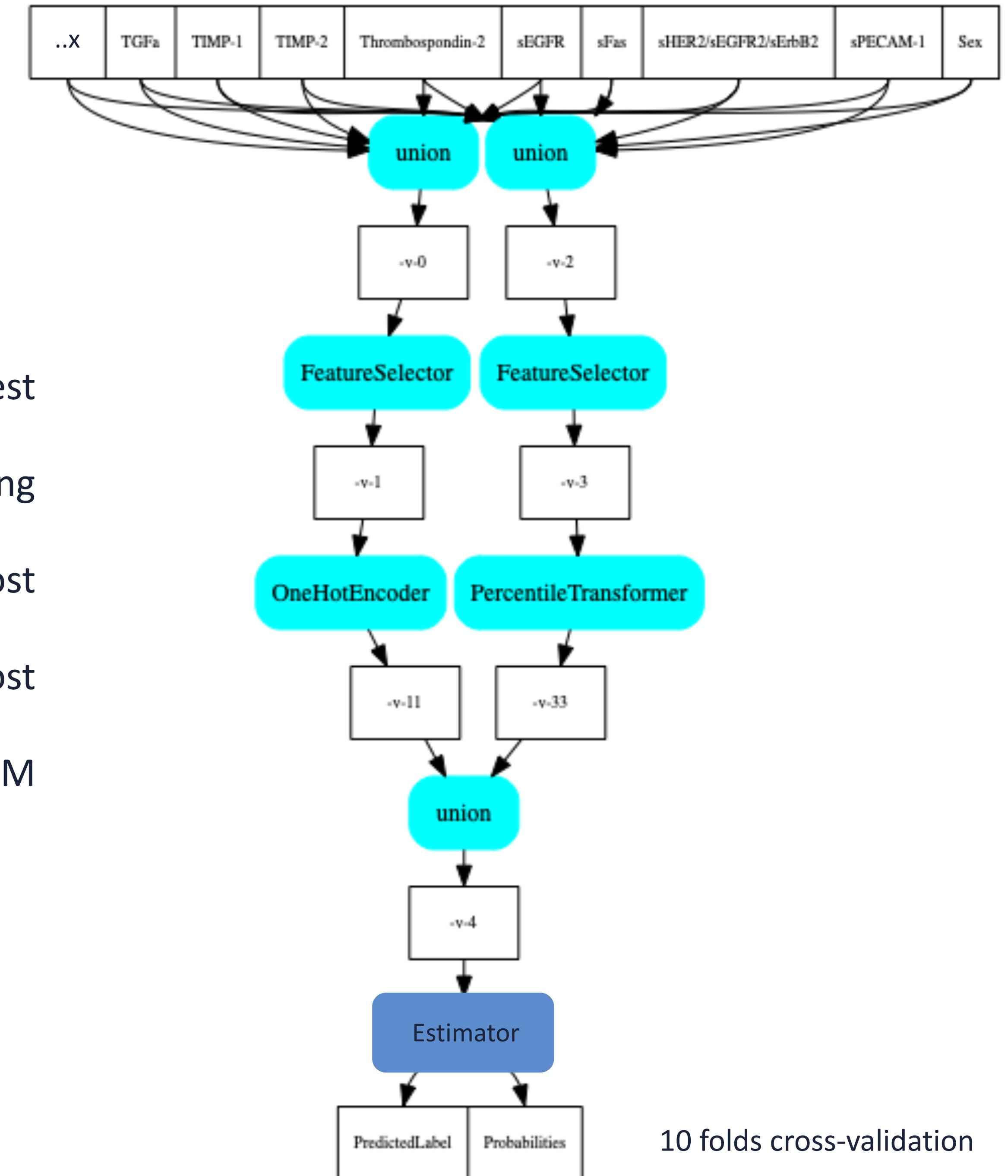
Random Forest

Gradient Boosting

XGBoost

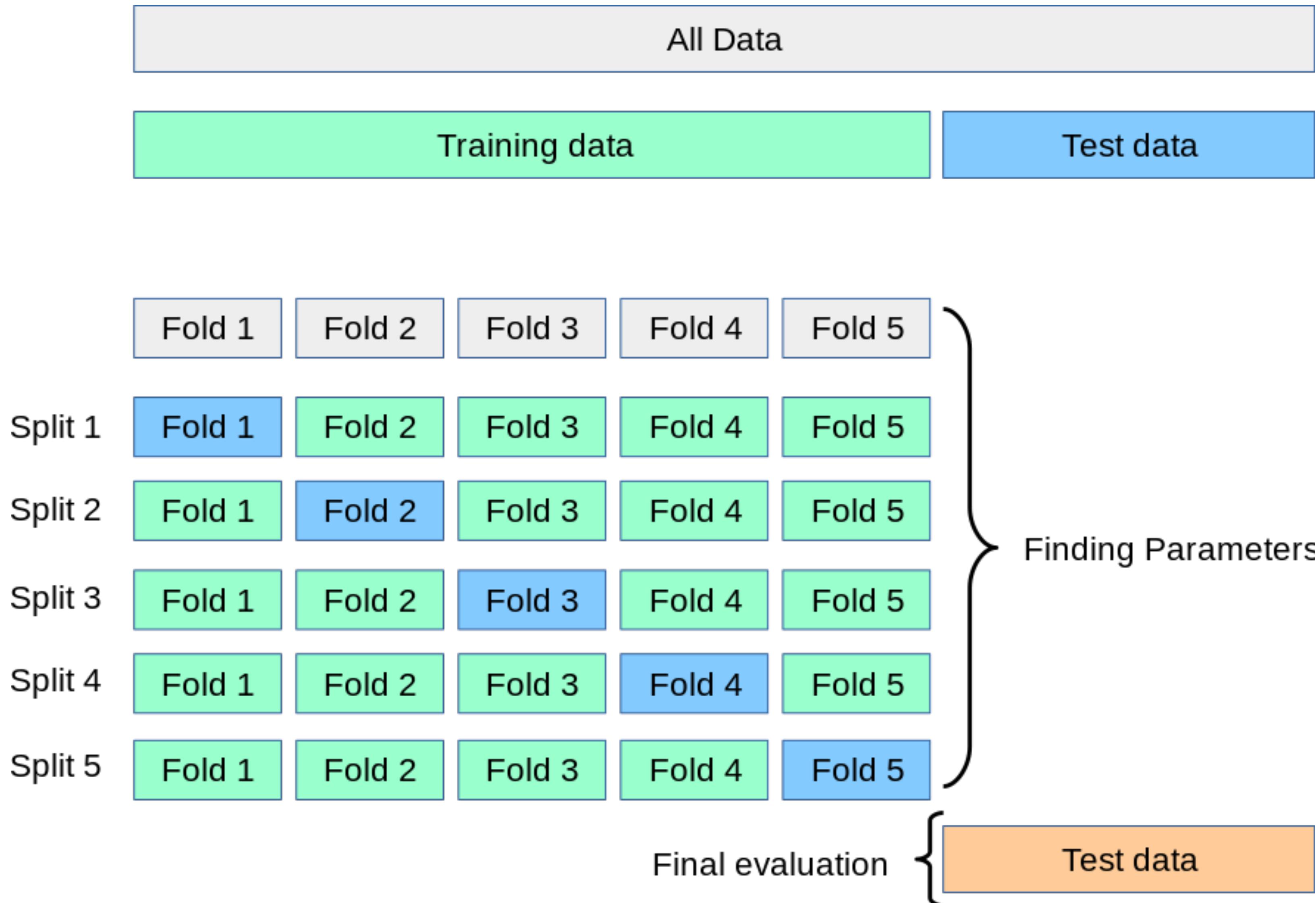
CatBoost

LightGBM



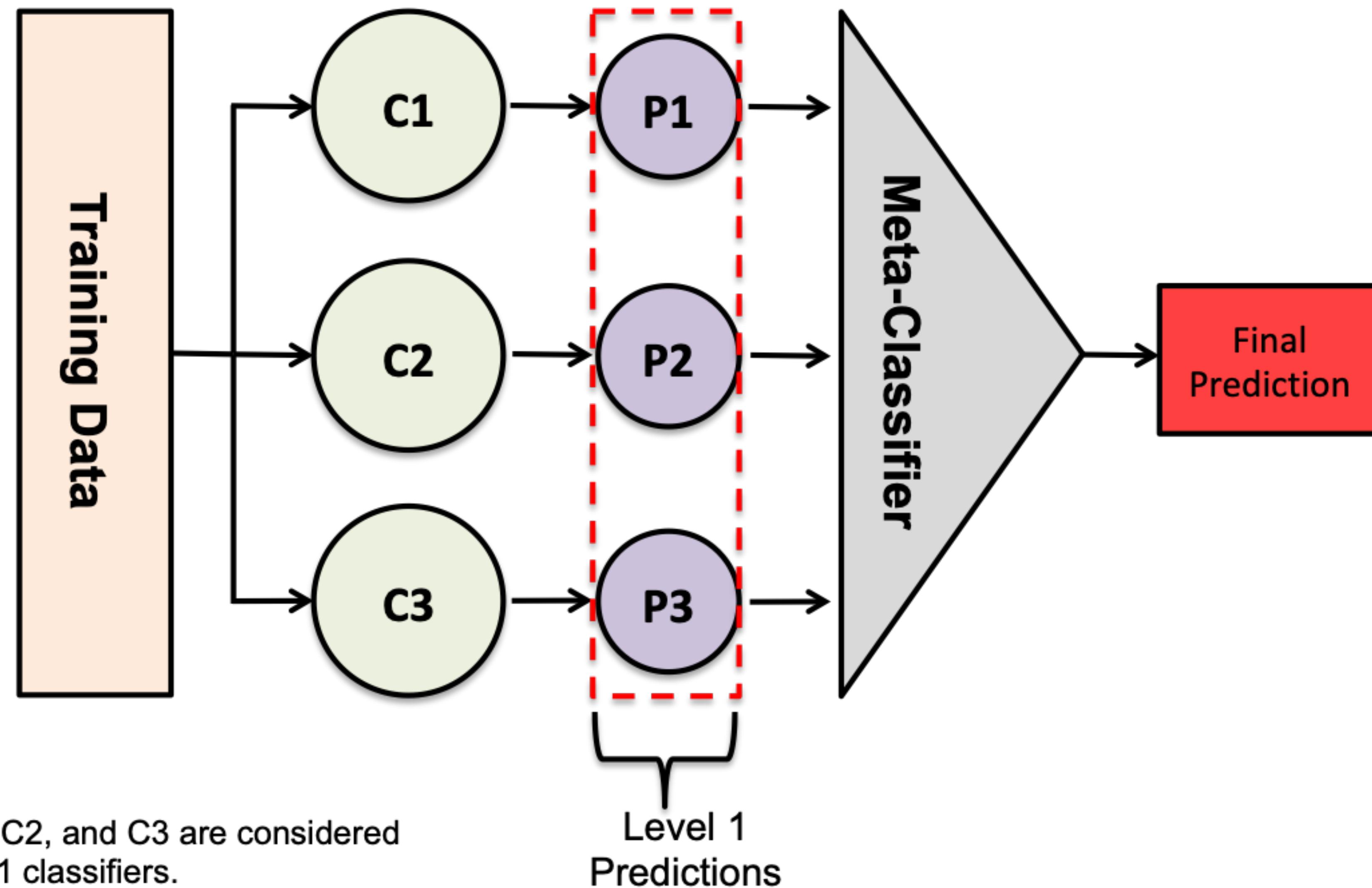
Common Steps (5)

- Grid Search & k-fold Cross-Validation - Concepts



Common Steps (5)

- Stacking Classifier - Concept



Results

- Background
- Different Approaches
- Common Steps
 - Missing Values
 - Feature Transformation
 - Data Visualisation
 - Experimentation
 - Pipeline
- **Results**
 - Cancer Type Classification (as in publication)
 - Cancer Type Classification (full dataset)
 - Cancer Type Classification (Aneuploidy dataset)
- Conclusions

Results (1)

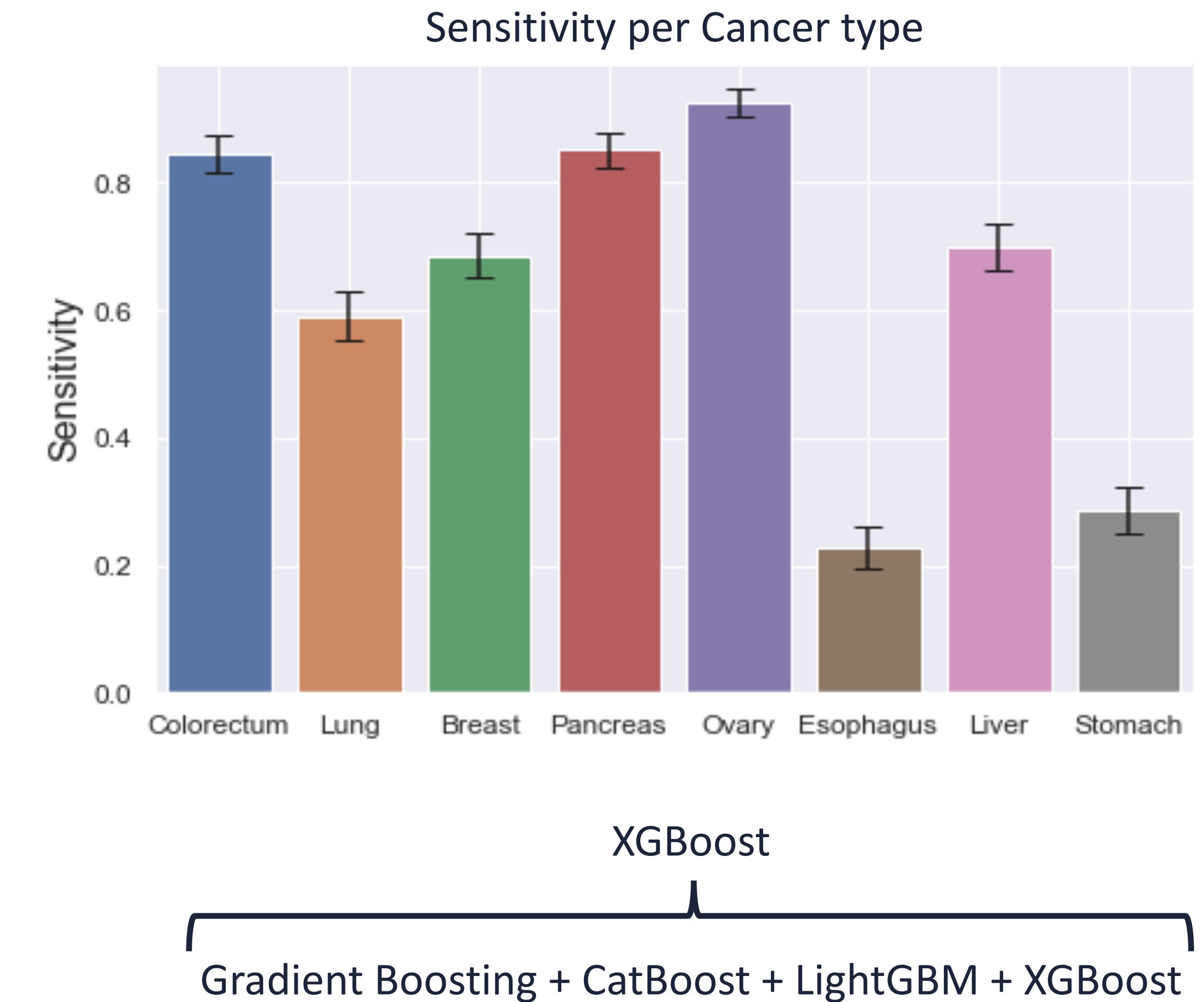
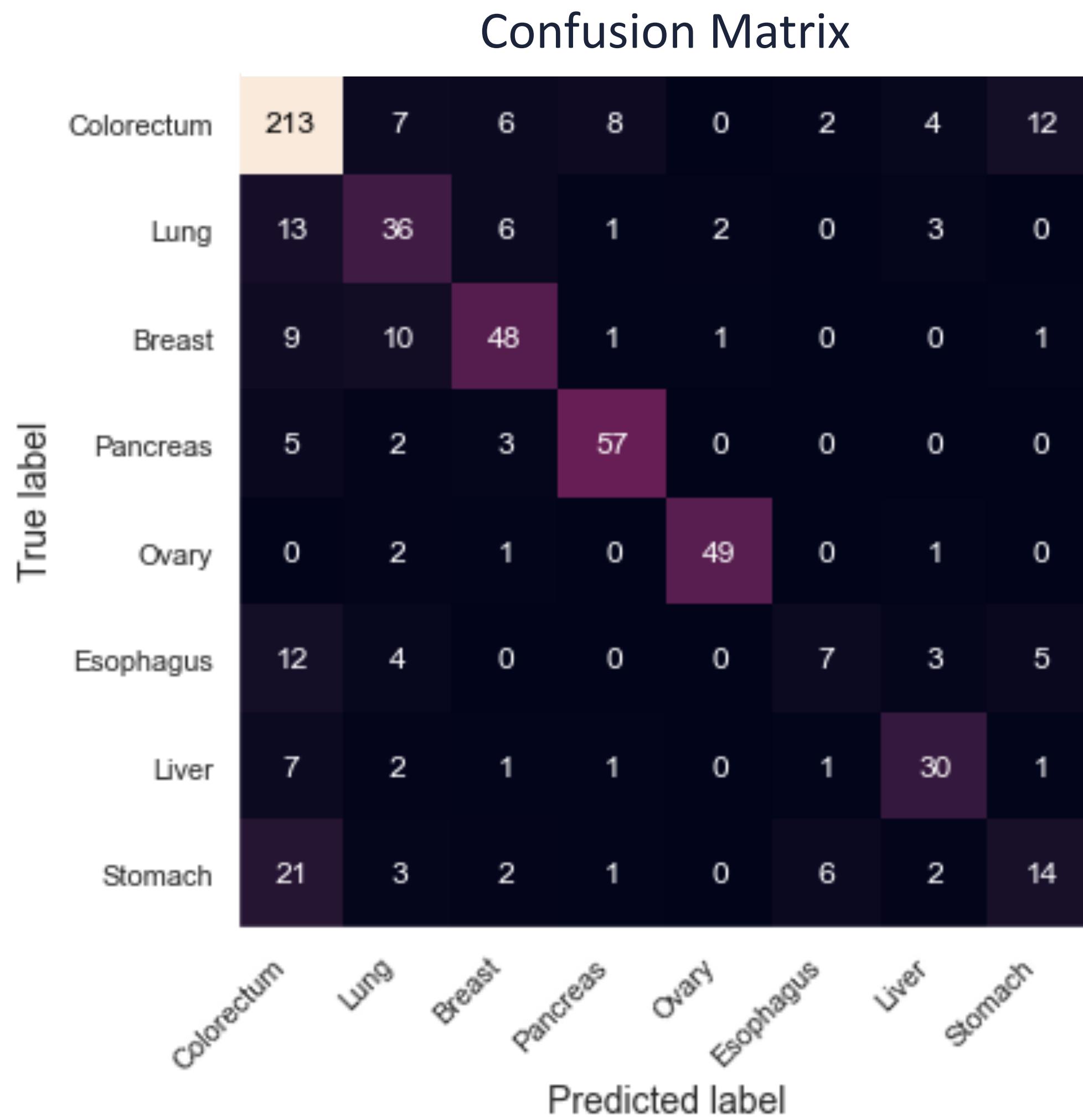


Cancer Type Classification

(as in publication)

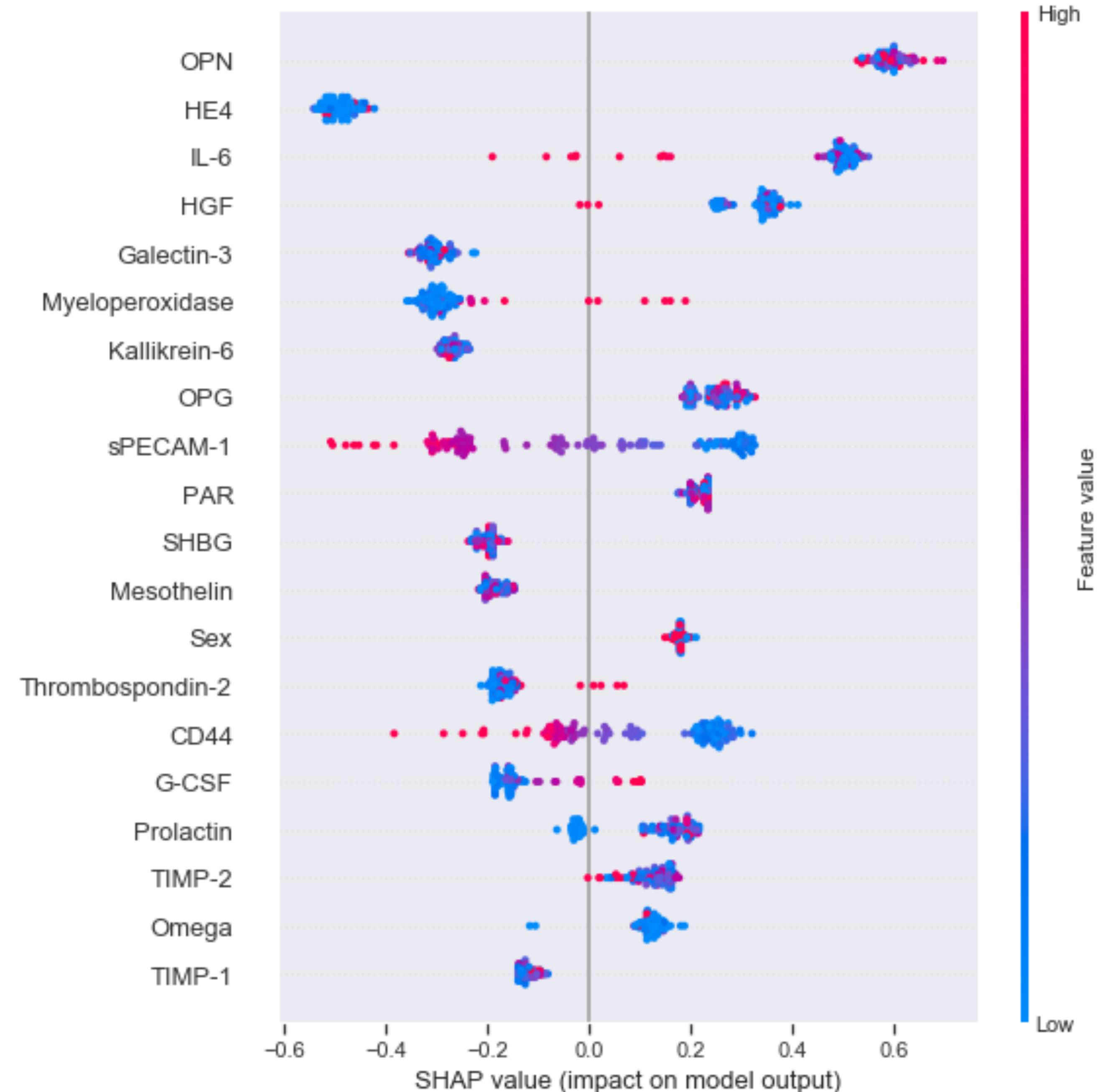
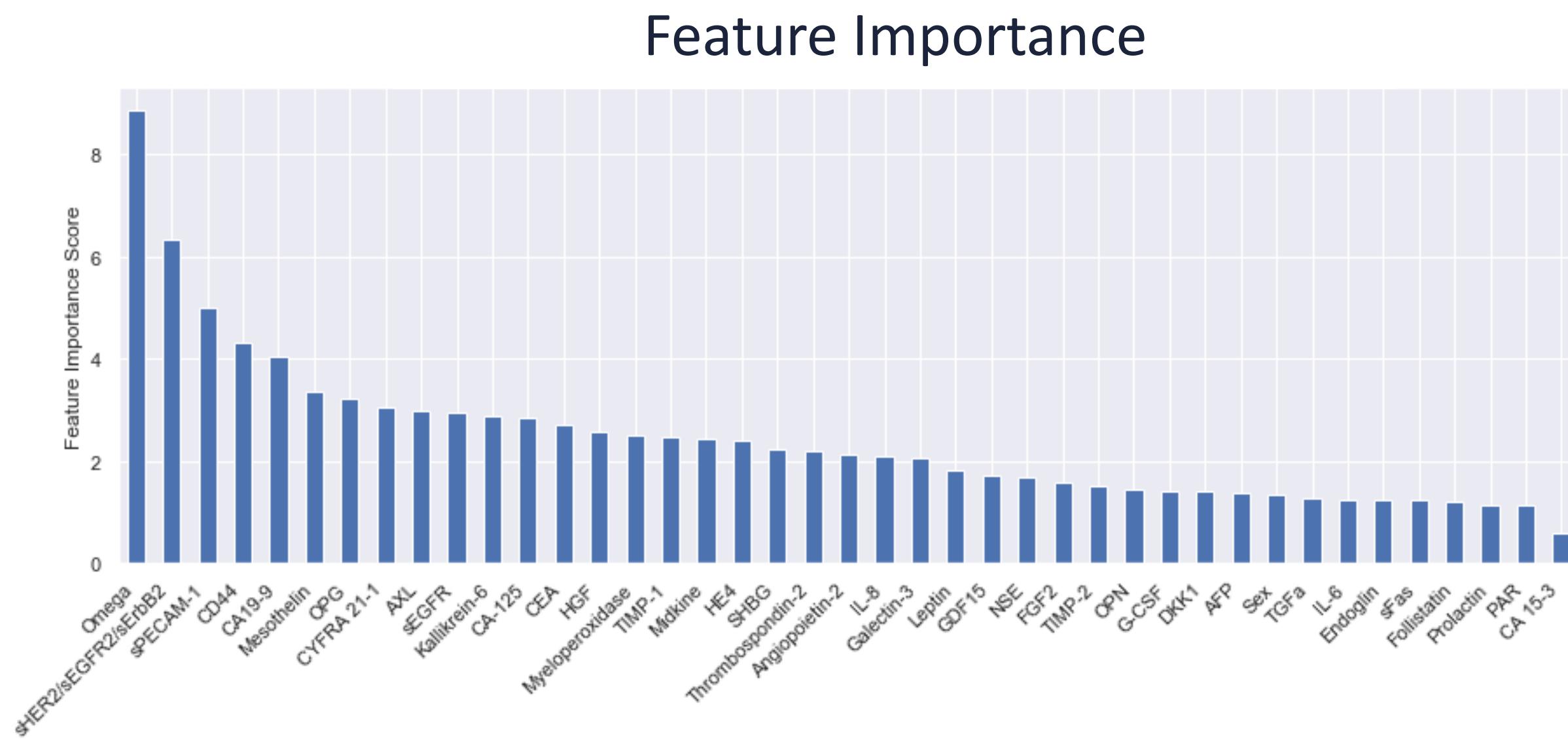
Results (1)

- Cancer Type Classification (as in publication)



Results (1)

- Cancer Type Classification
(as in publication)
- Feature Importance



Results (2)

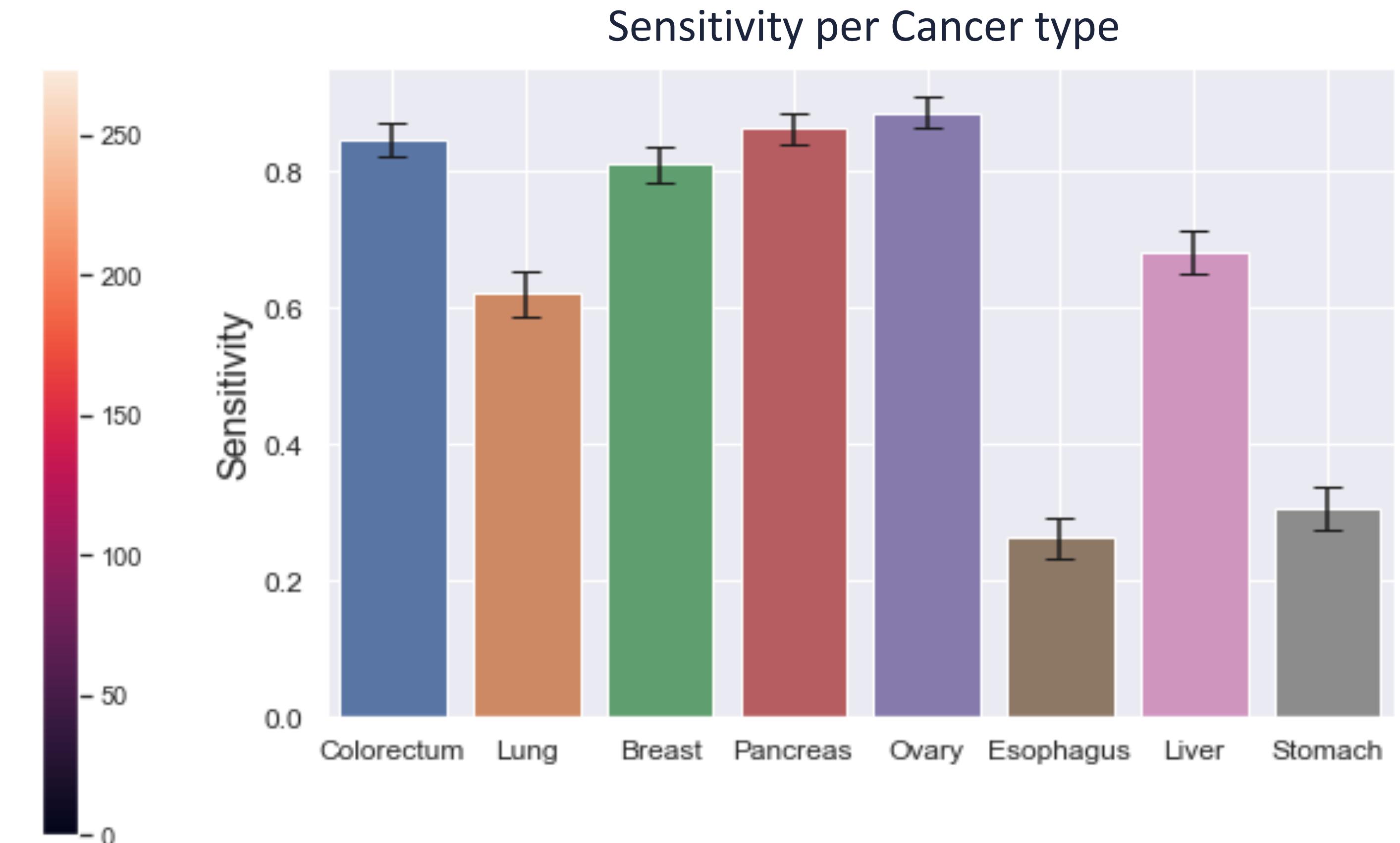
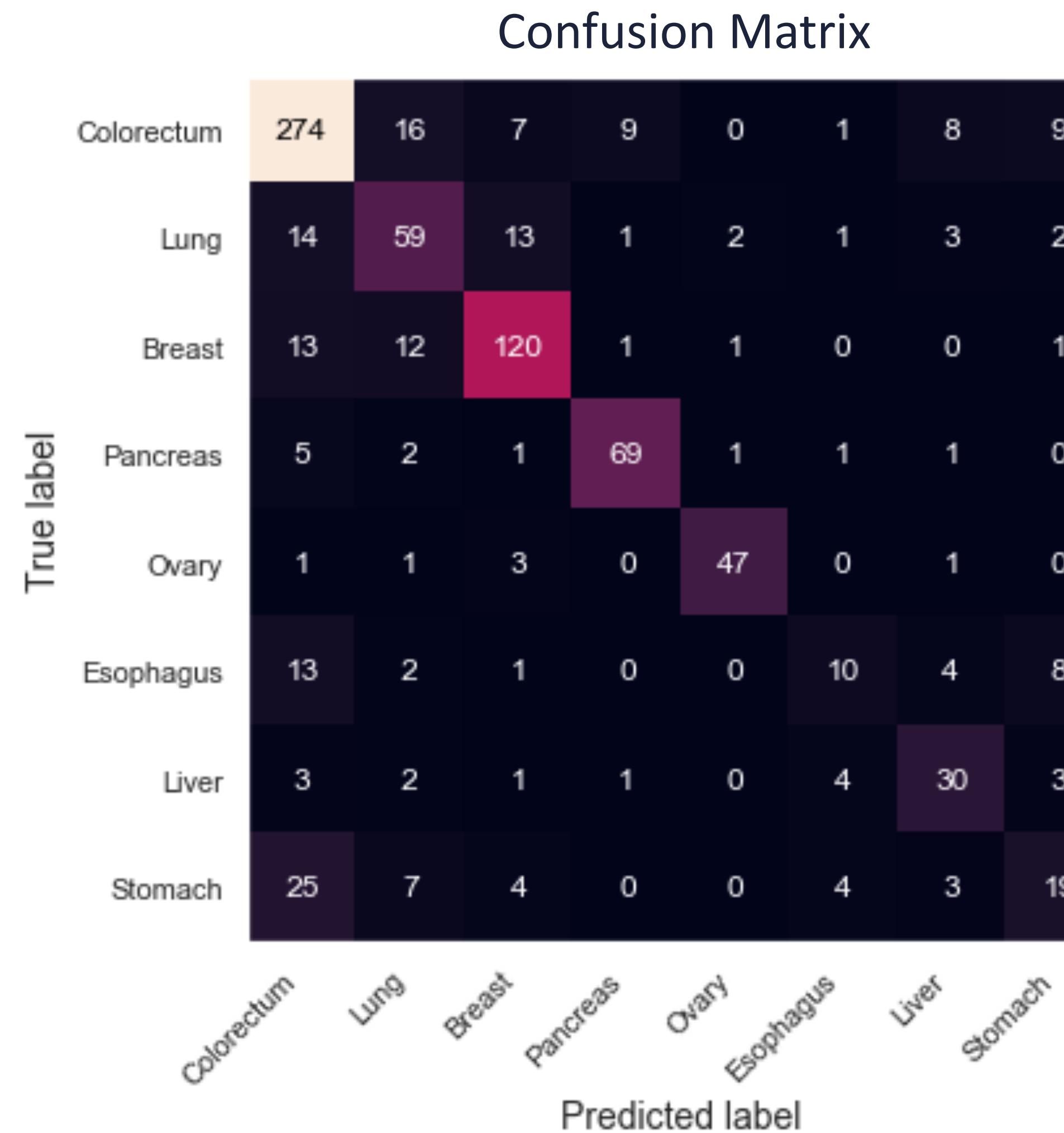


Cancer Type Classification

(On Full Dataset)

Results (2)

- Cancer Type Classification (on Full Dataset)

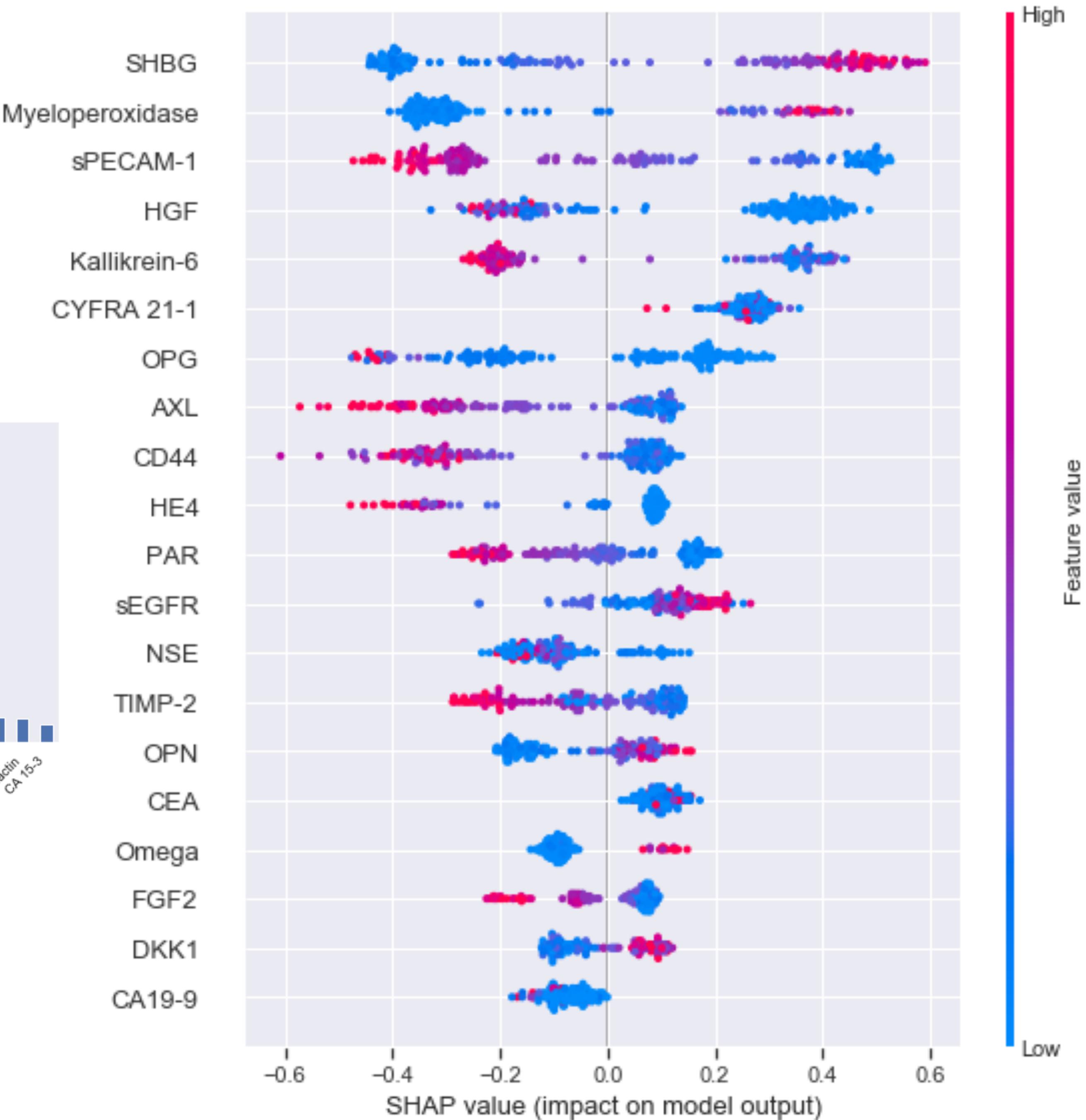
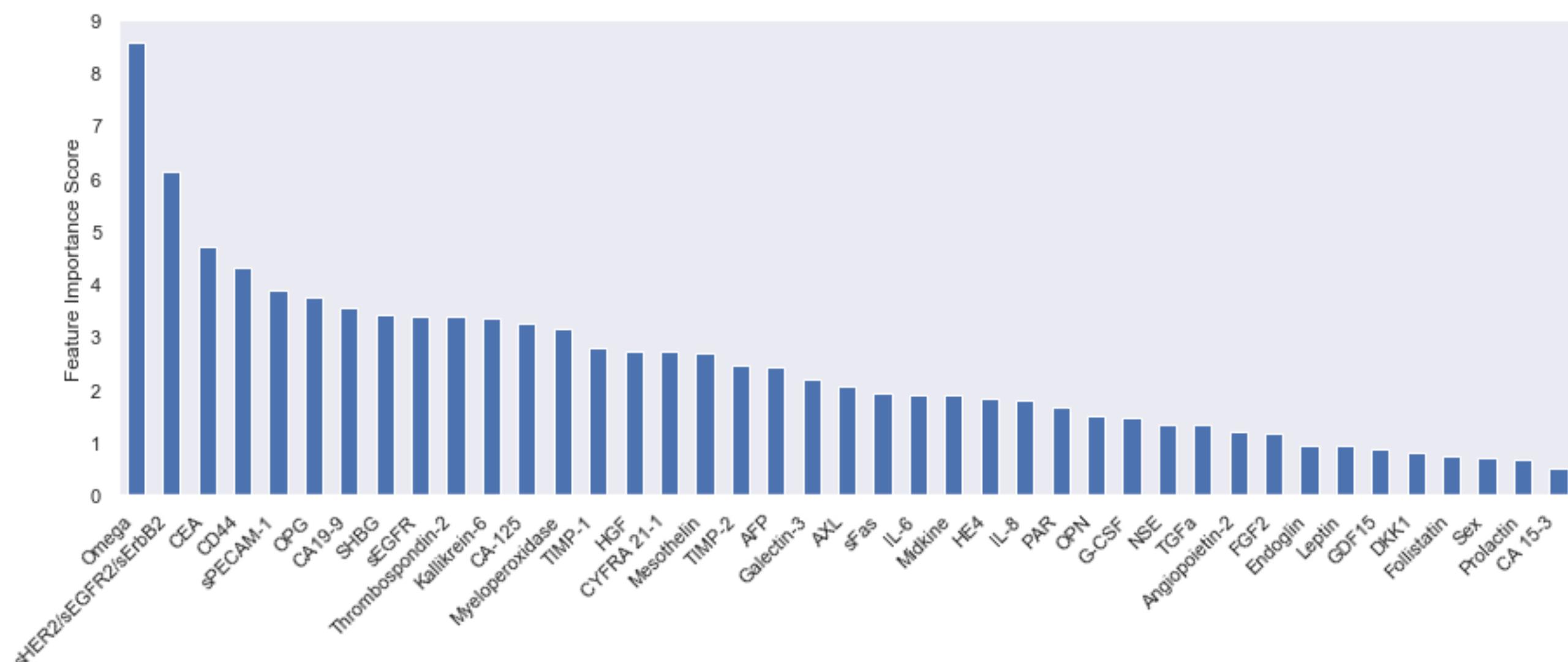


Cancer Samples Correctly Classified: 844 (84%) (626)
Specificity: 94% (99%)

Results (2)

- Cancer Type Classification
(on Full Dataset)

Feature Importance



Results (3)

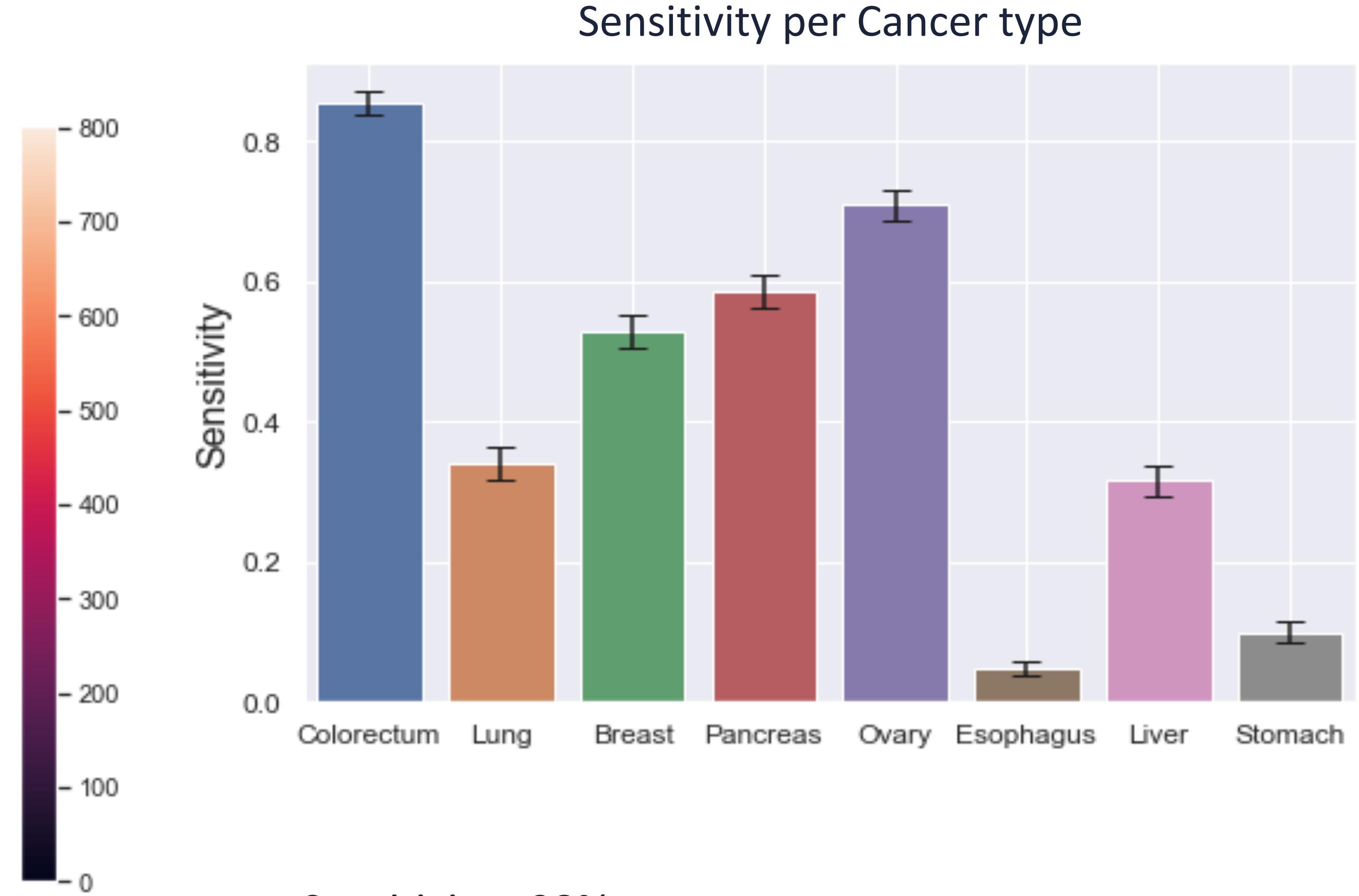
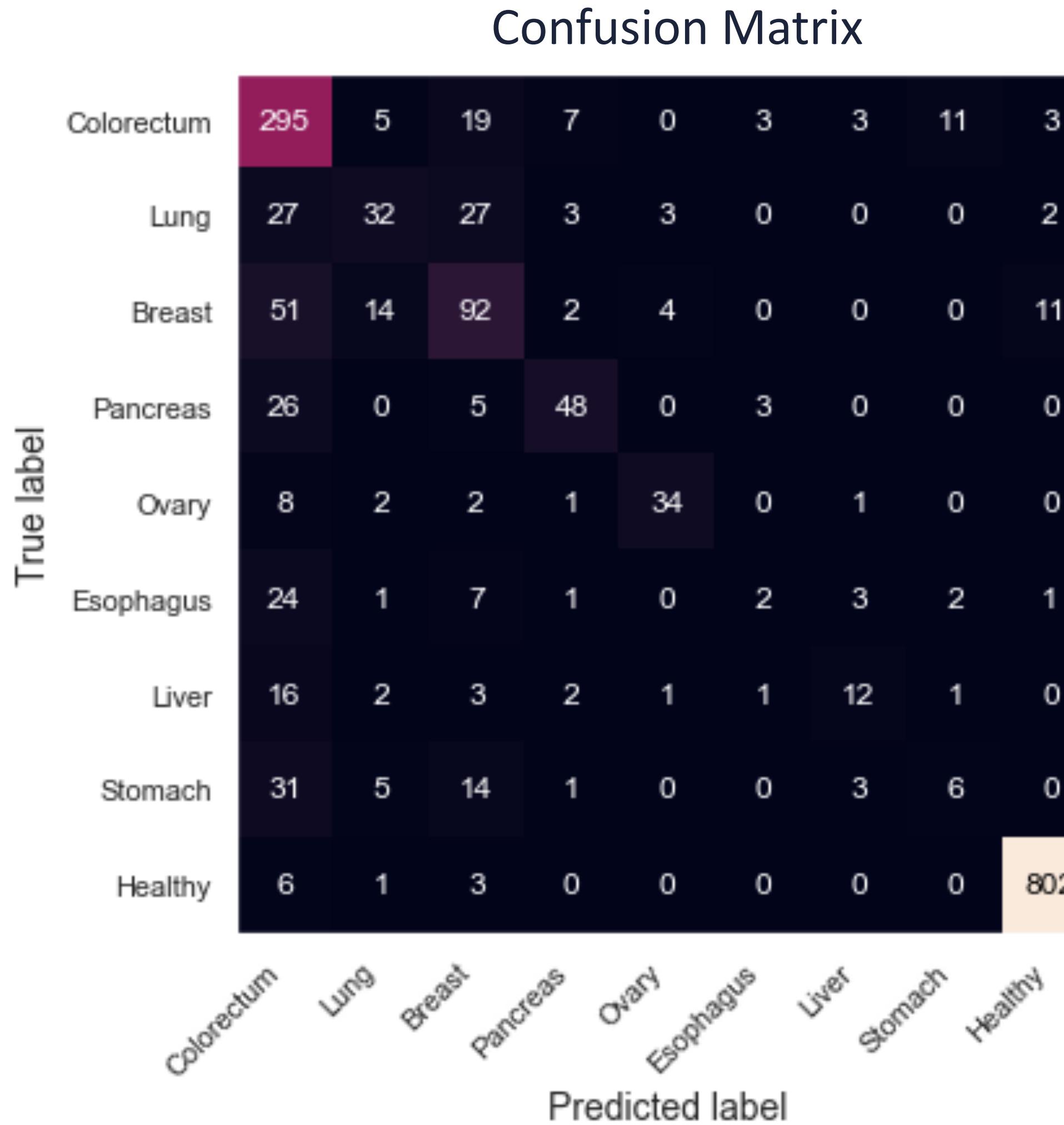


Cancer Type Classification

(On the Aneuploidy Dataset)

Results (3)

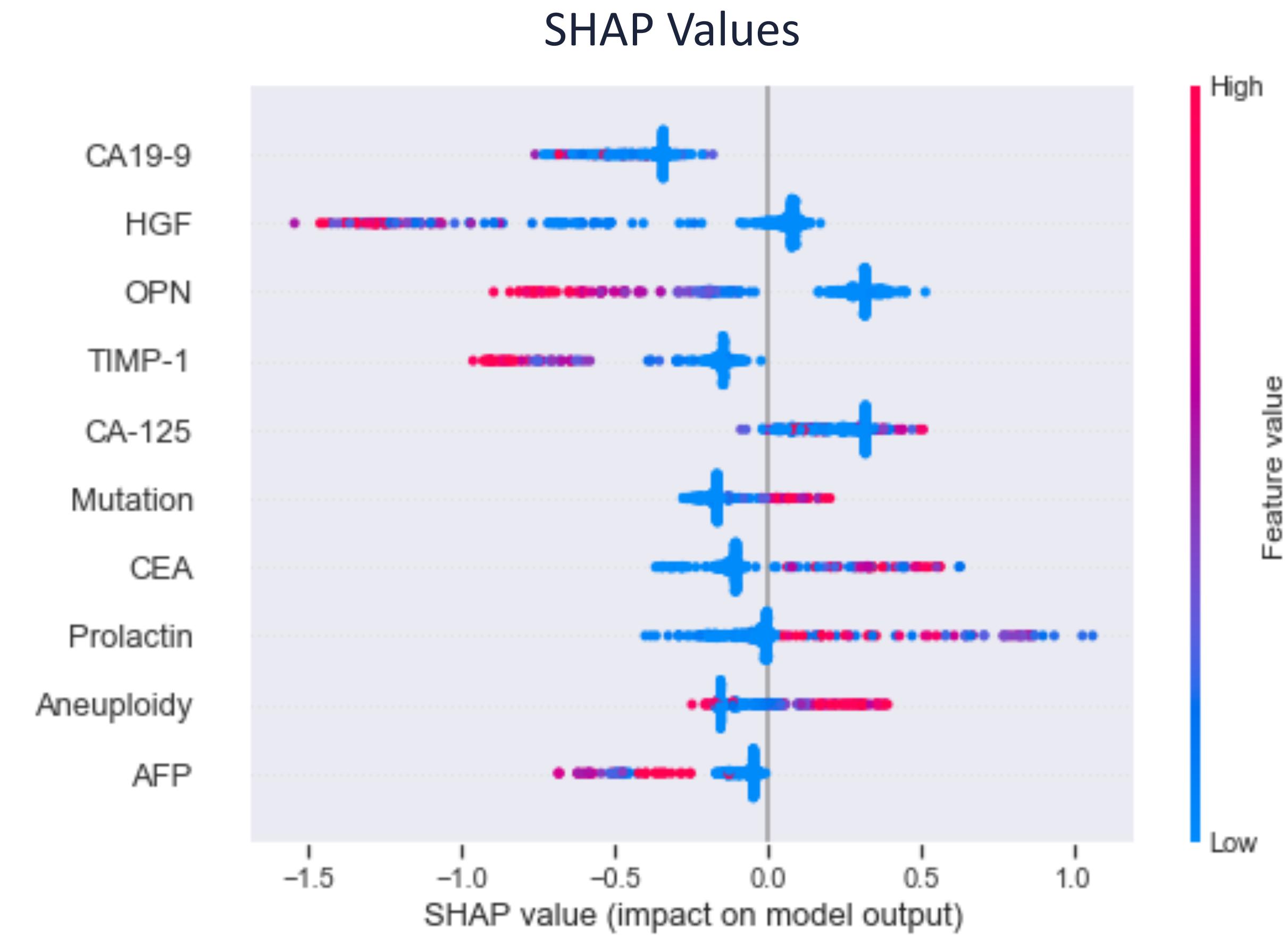
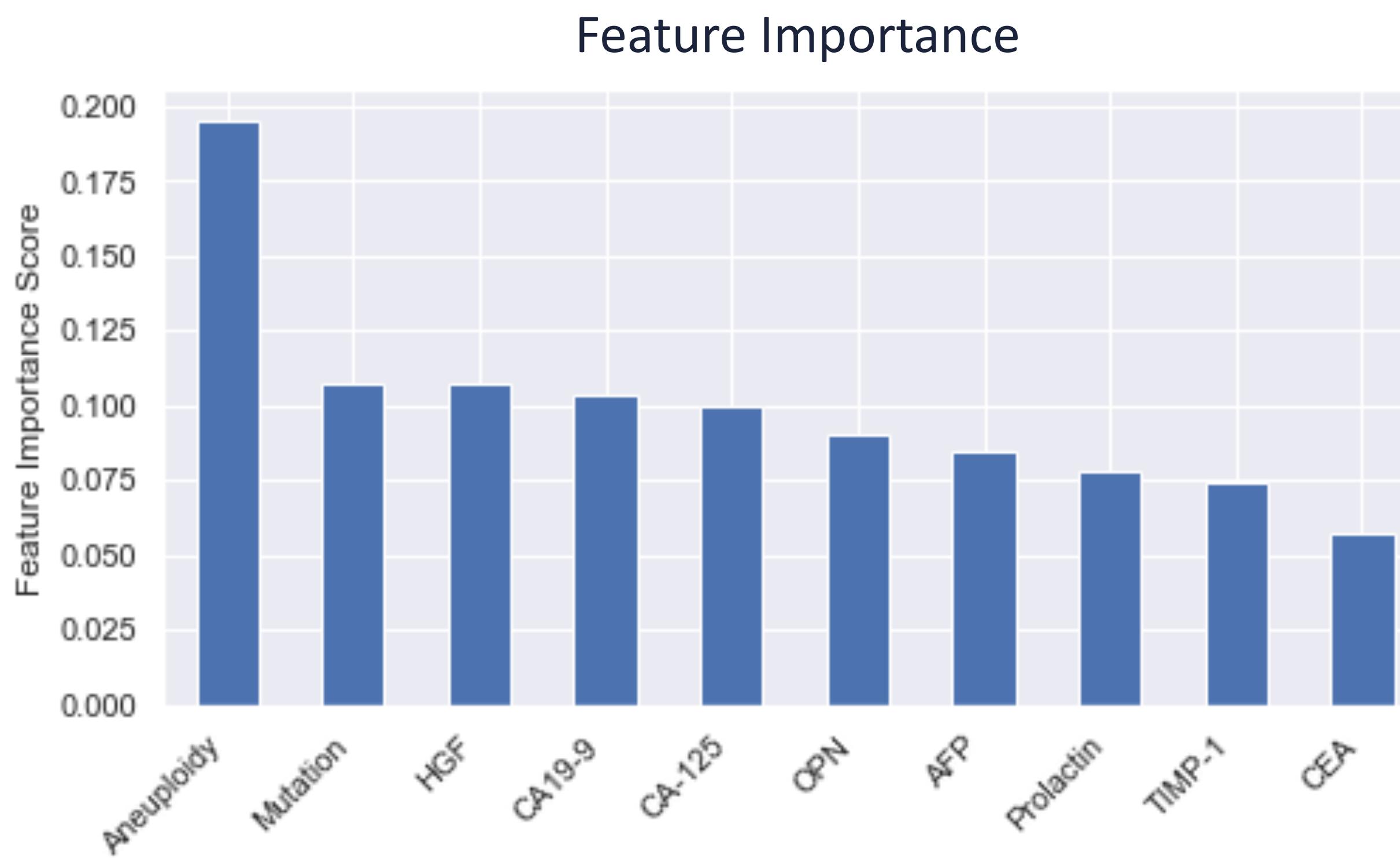
- Cancer Type Classification
(on the Aneuploidy dataset)



Sensitivity: 98%
Specificity: 99%
Precision (cancer): 99%

Results (3)

- Cancer Type Classification
(on the Aneuploidy dataset)



Conclusions

- Background
- Different Approaches
- Common Steps
 - Missing Values
 - Feature Transformation
 - Data Visualisation
 - Experimentation
 - Pipeline
- Results
 - Cancer Type Classification (as in publication)
 - Cancer Type Classification (full dataset)
 - Cancer Type Classification (Aneuploidy dataset)
- **Conclusions**

Conclusions

- Improvements on **Colorectum** (31%), **Breast** (100%) and **Pancreas** (21%)
- They correspond to more than 4 million new cases 2018
- CatBoost and XGBoost are in general most performant, along with Stacking Classifiers
- Feature Engineering seemingly doesn't improve the results.
- Some feature selection techniques are extremely time consuming
- Very poor distribution on the continuous variables; somehow amazing that the models can make sense out of it
- Pipelines makes life easier
- Doubts around SHAP values' consistency