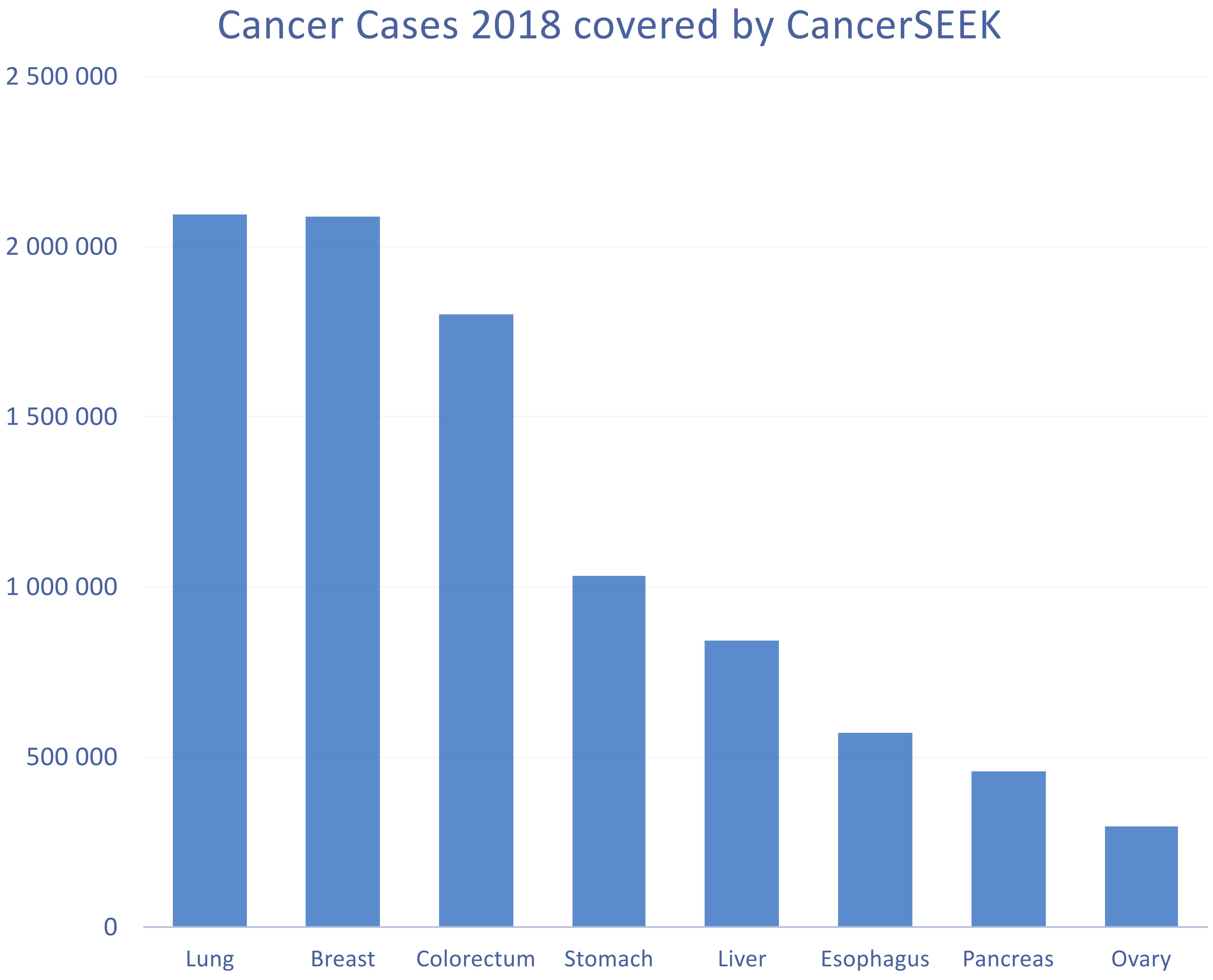
Cancer Detection with Machine Learning – CancerSEEK

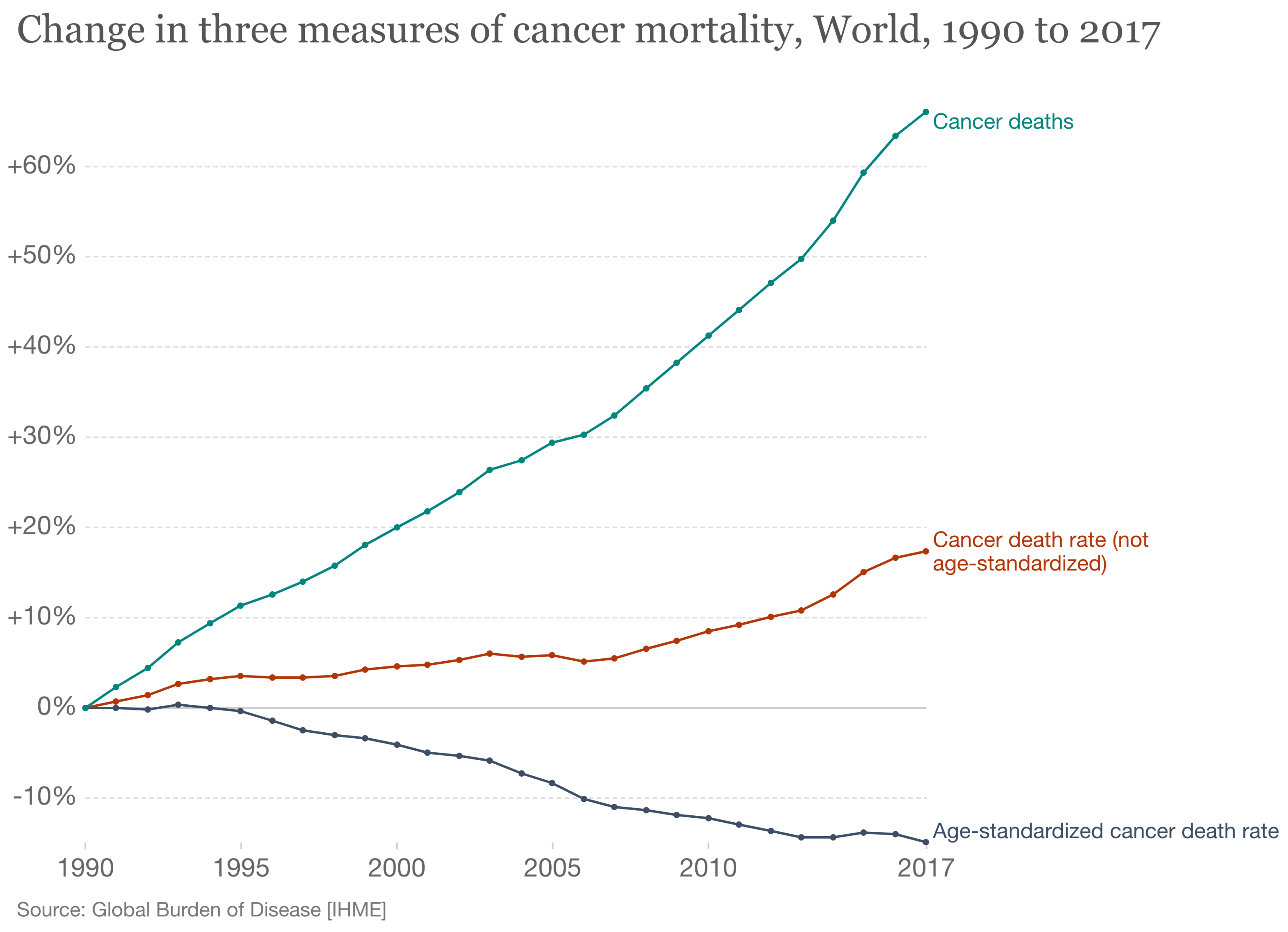
From a Technical Perspective

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





Source: Global Burden of Disease [IHME]

- 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)

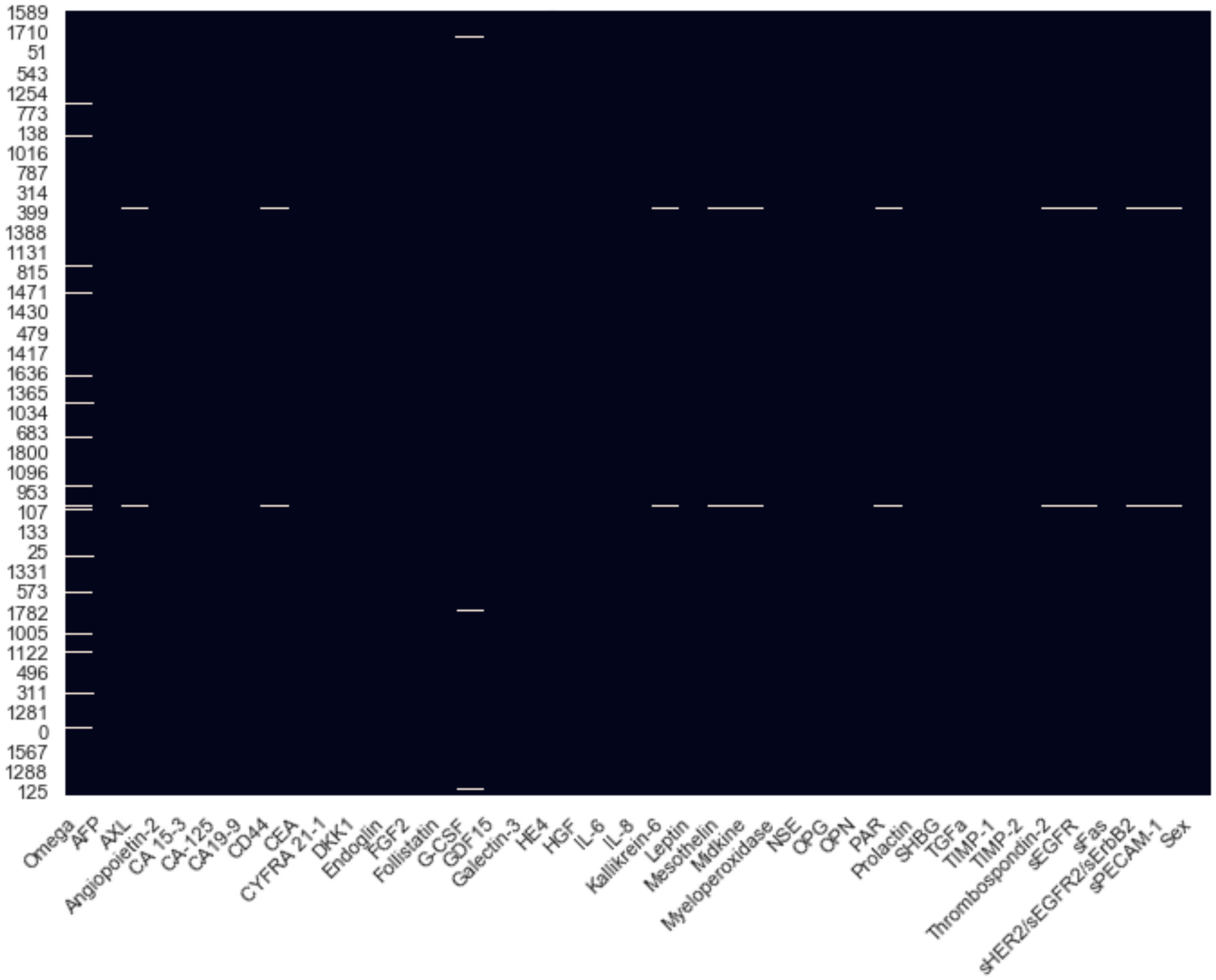
- Conclusions

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
```

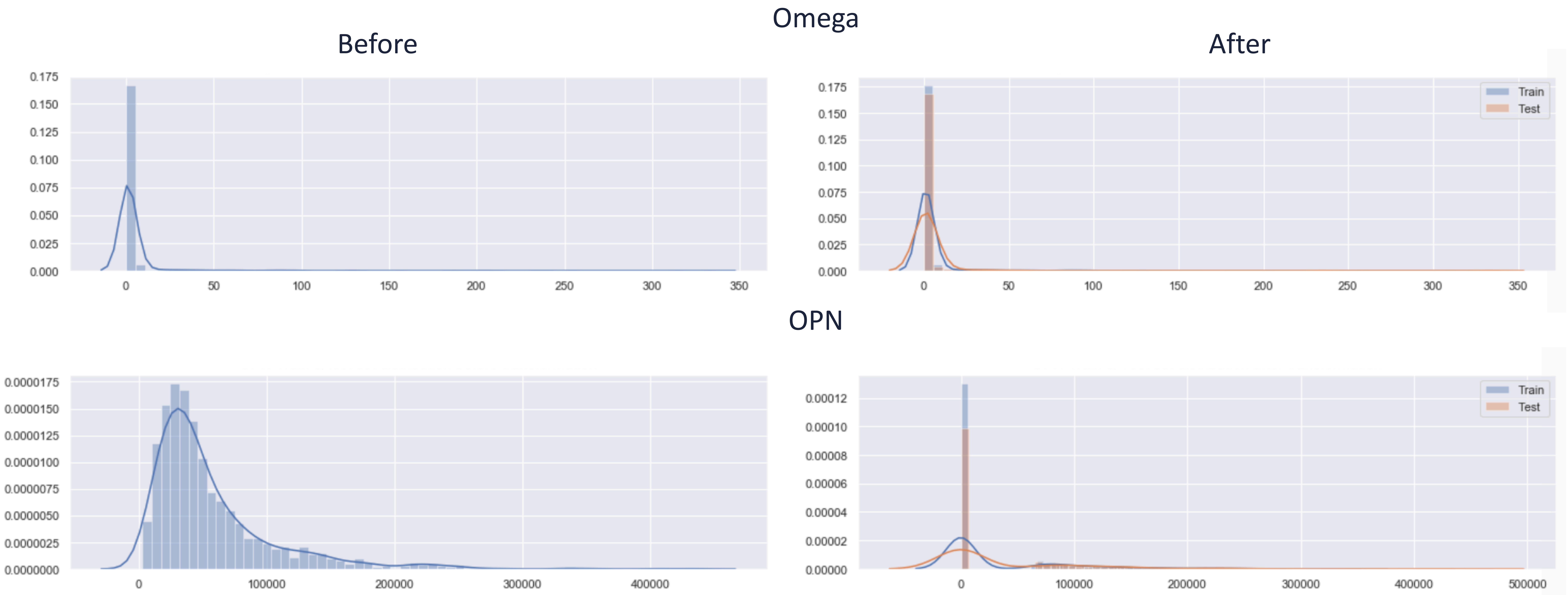
```
# Custom transform method to replace cancer values
# that are below the healthy 95th percentile
def transform(self, X, y=None):

    # 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

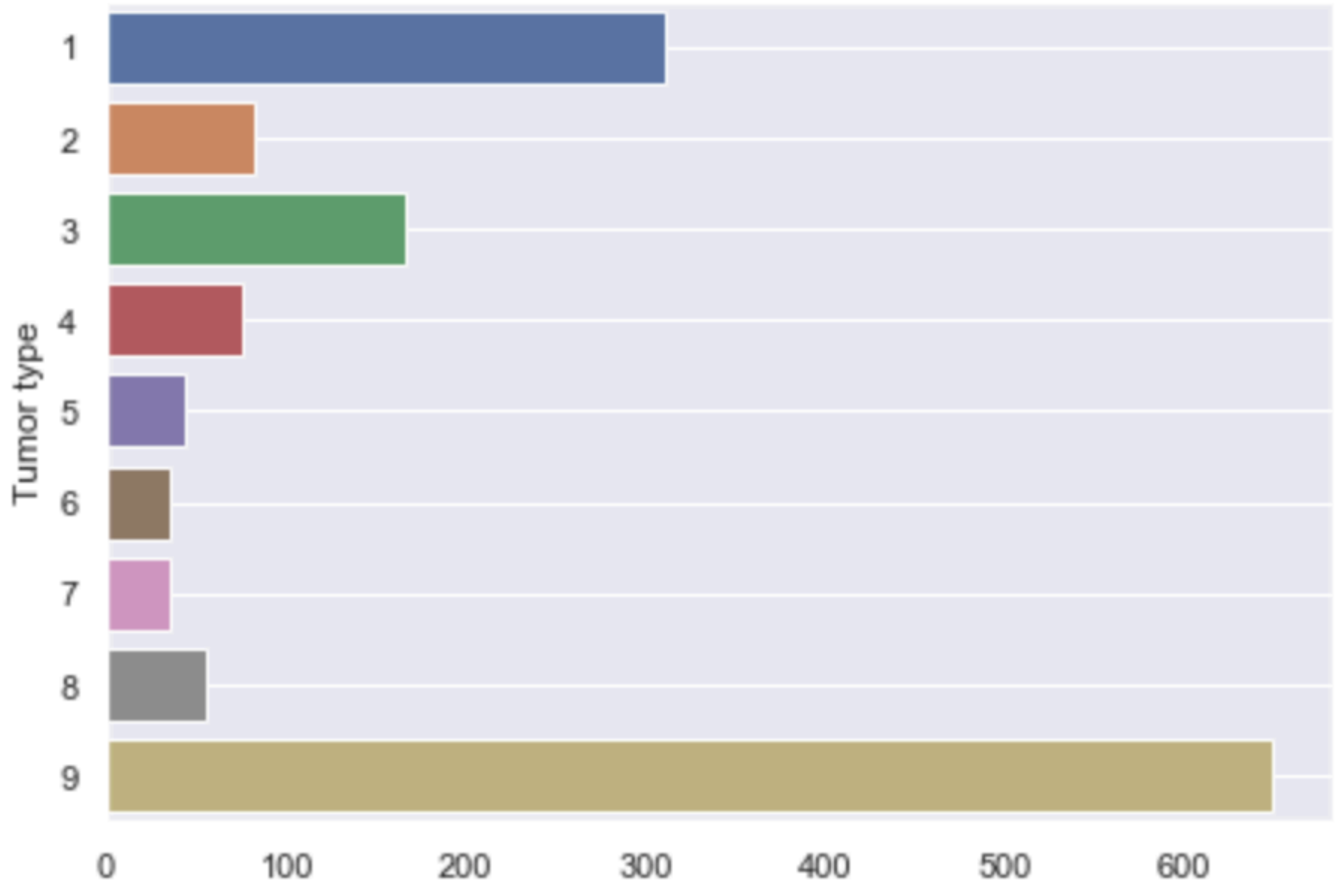


Common Steps (3)

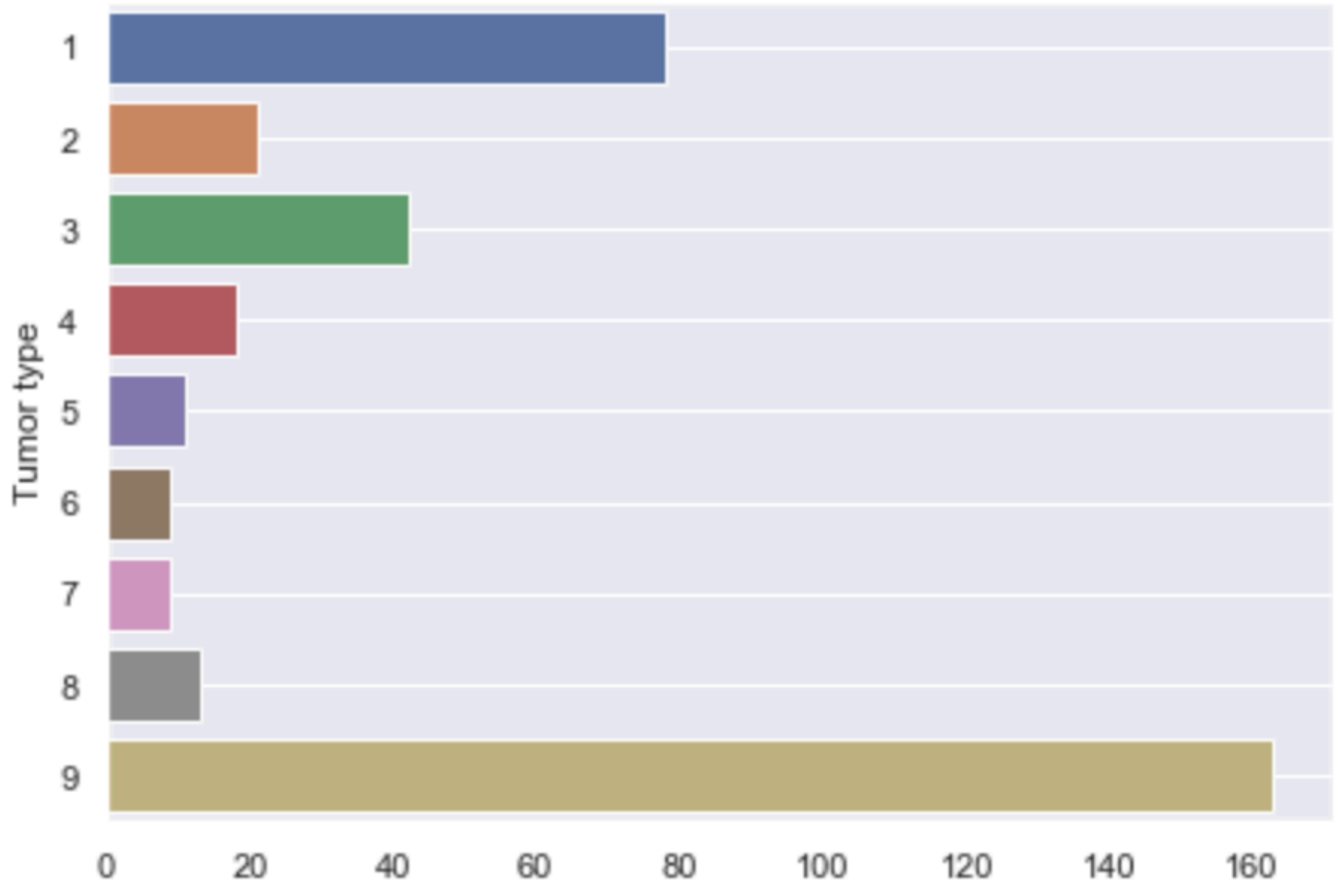
- Data Visualisation – Tumor Counts on Train & Test Sets

1005 cancer
812 normal samples, in total

Train

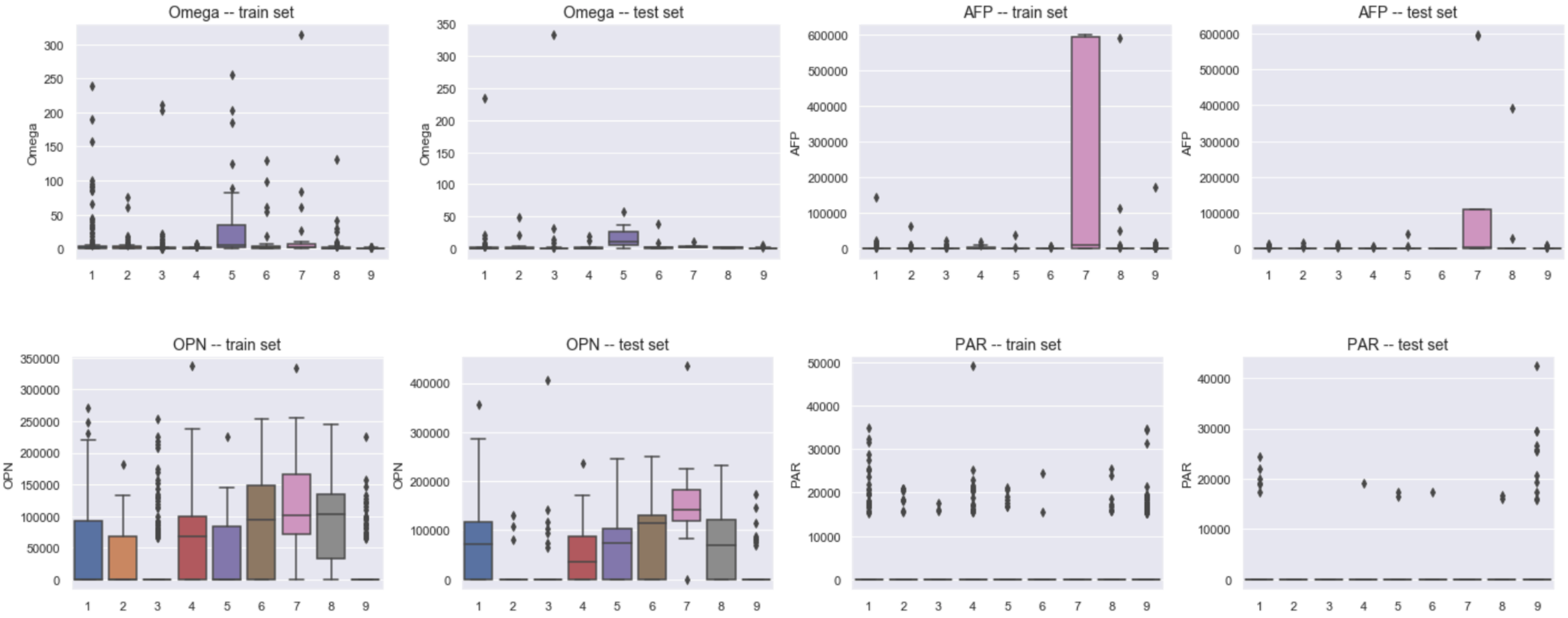


Test



Common Steps (3)

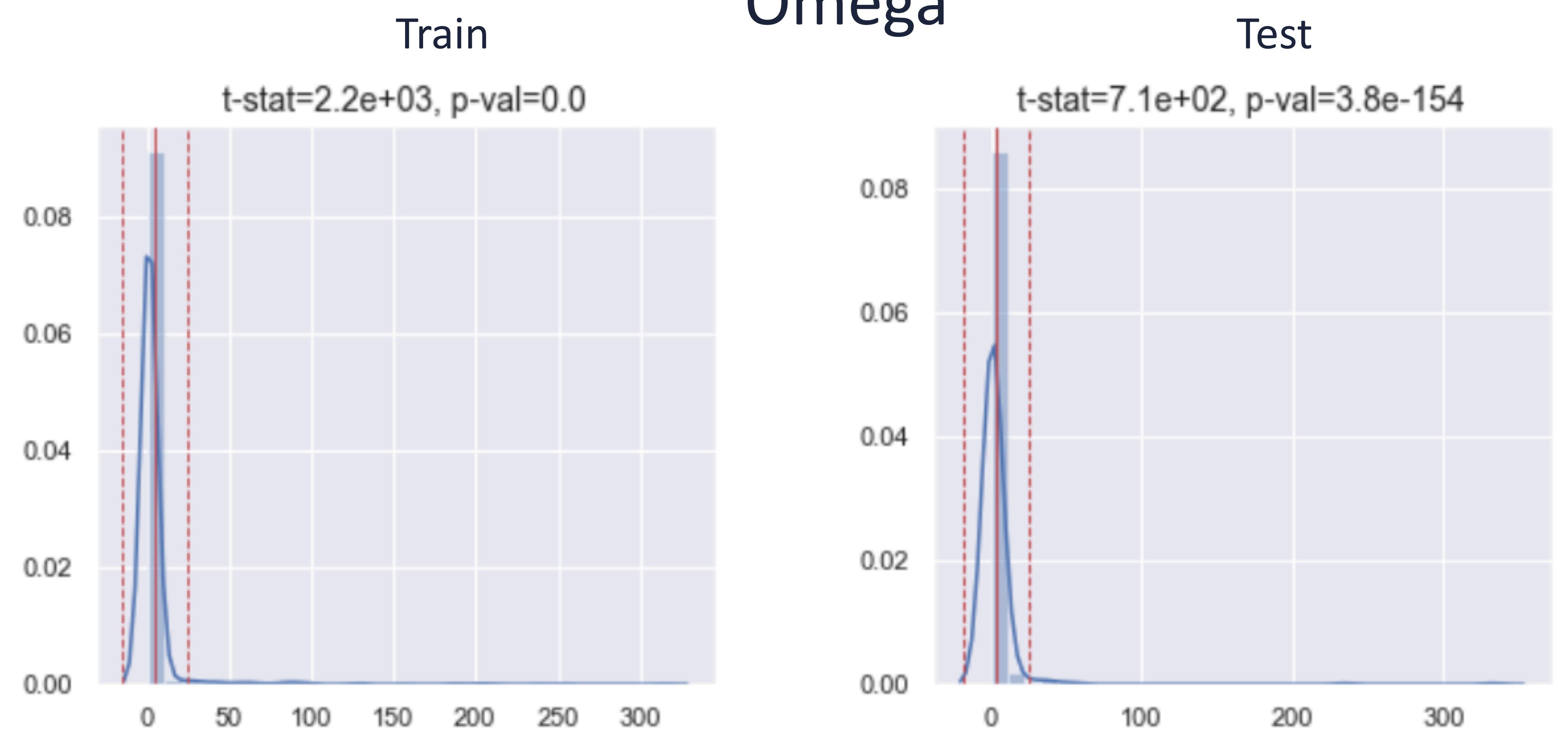
- Data Visualisation – Distribution per Tumor Type after Custom Transformation



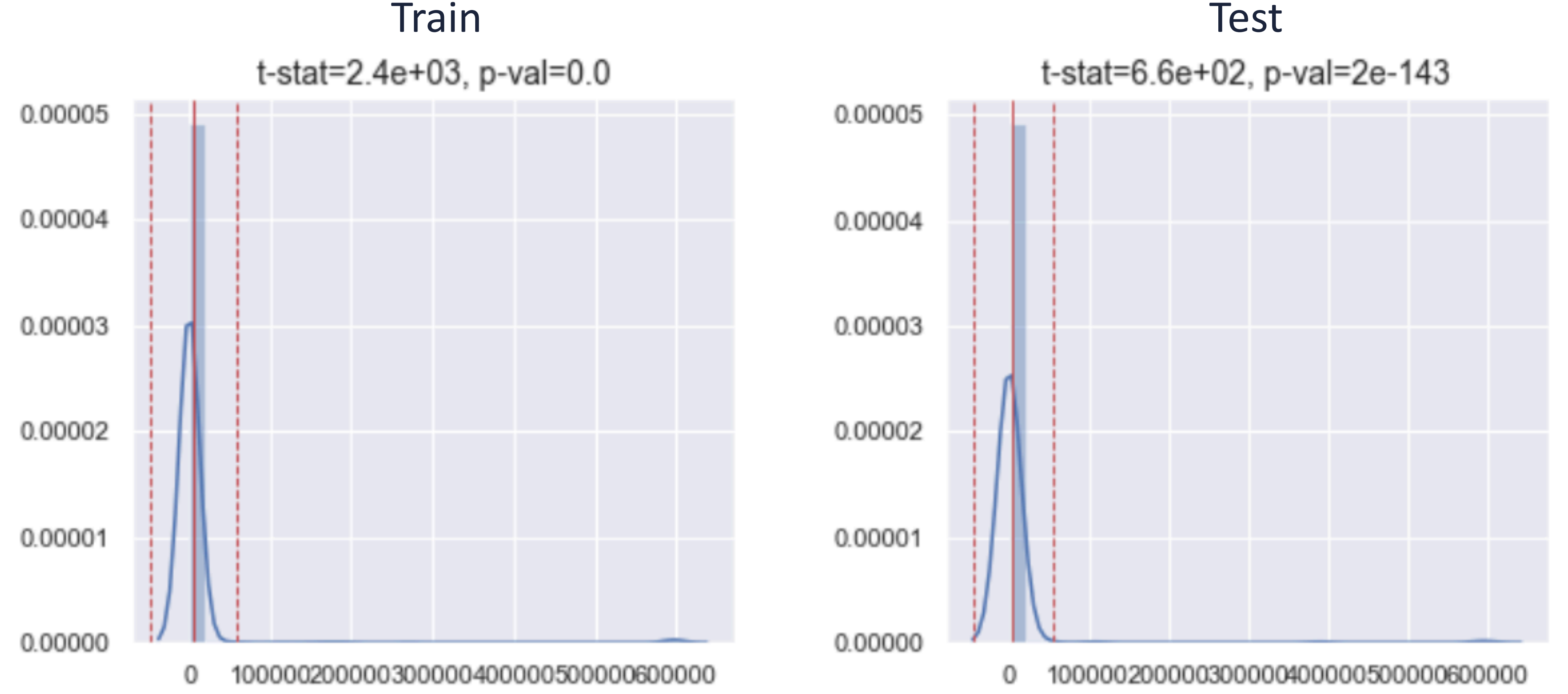
Common Steps (3)

- Data Visualisation – Test for Normality

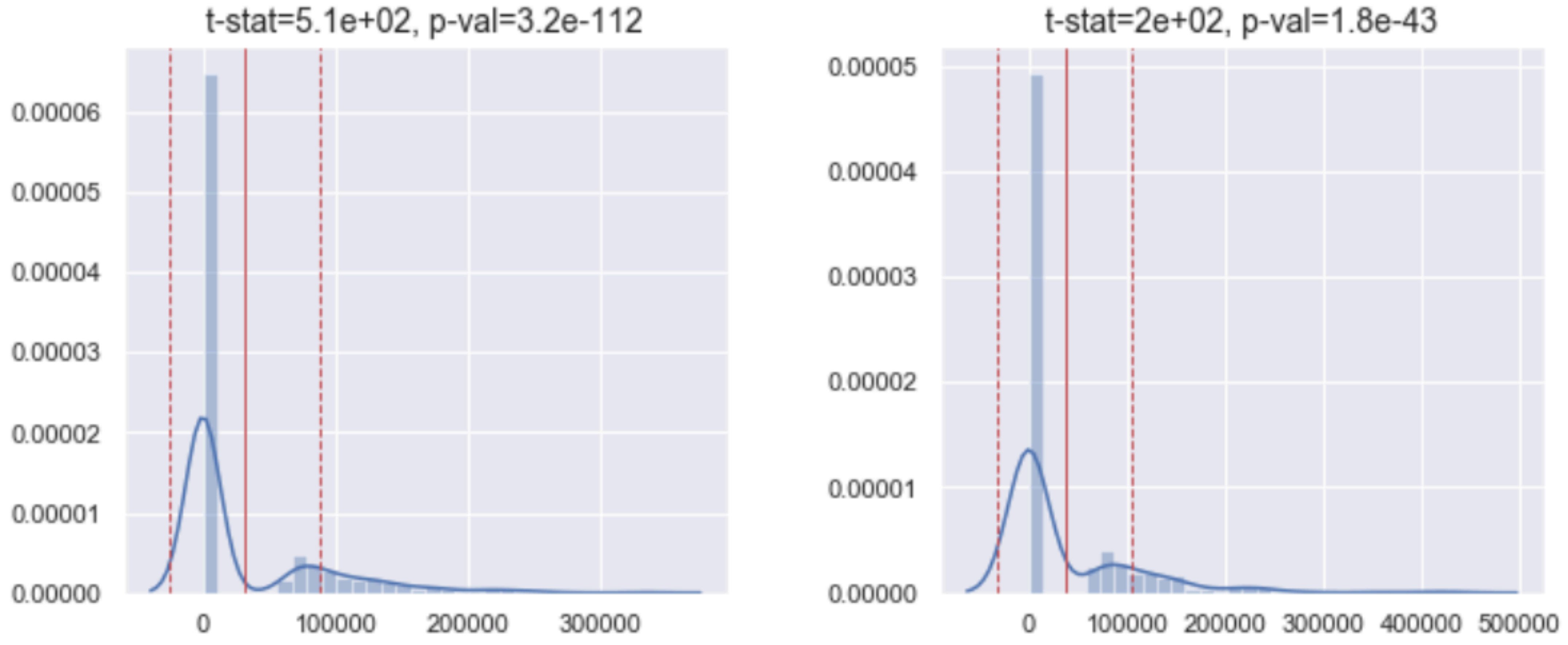
Omega



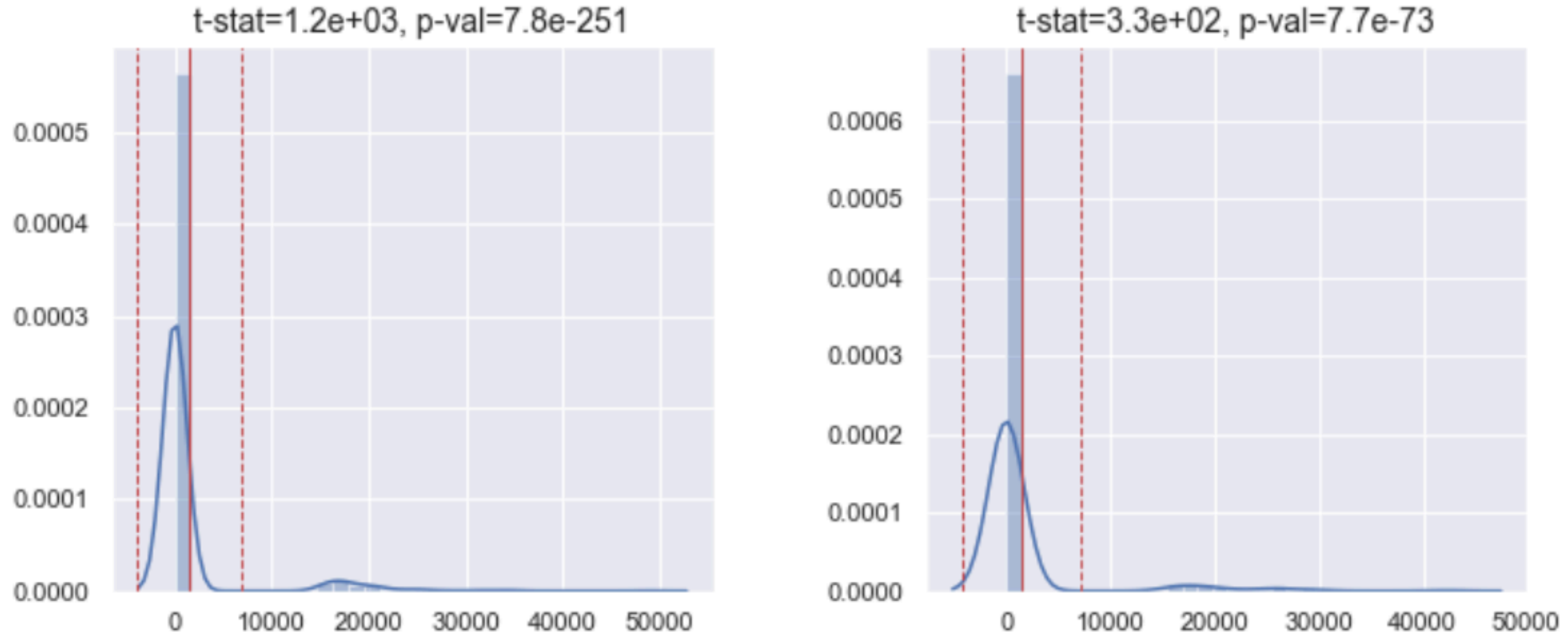
AFP



OPN



PAR



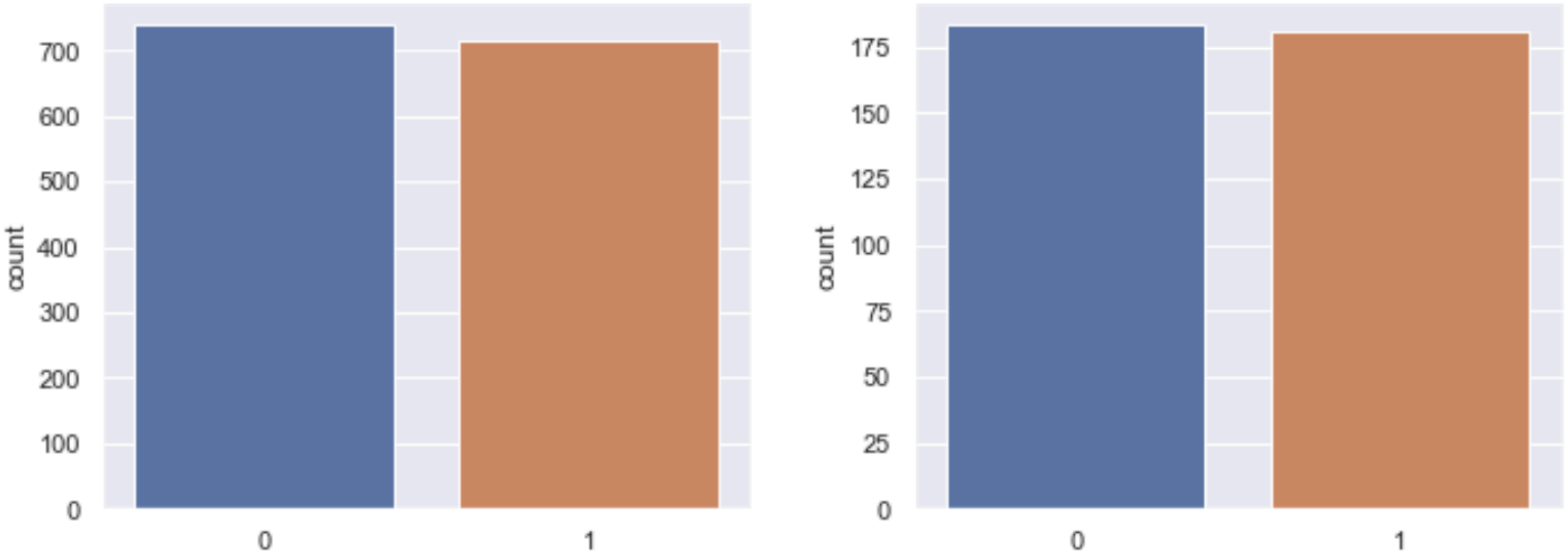
Common Steps (3)

- Data Visualisation – Categorical Variable’s Distribution

Sex

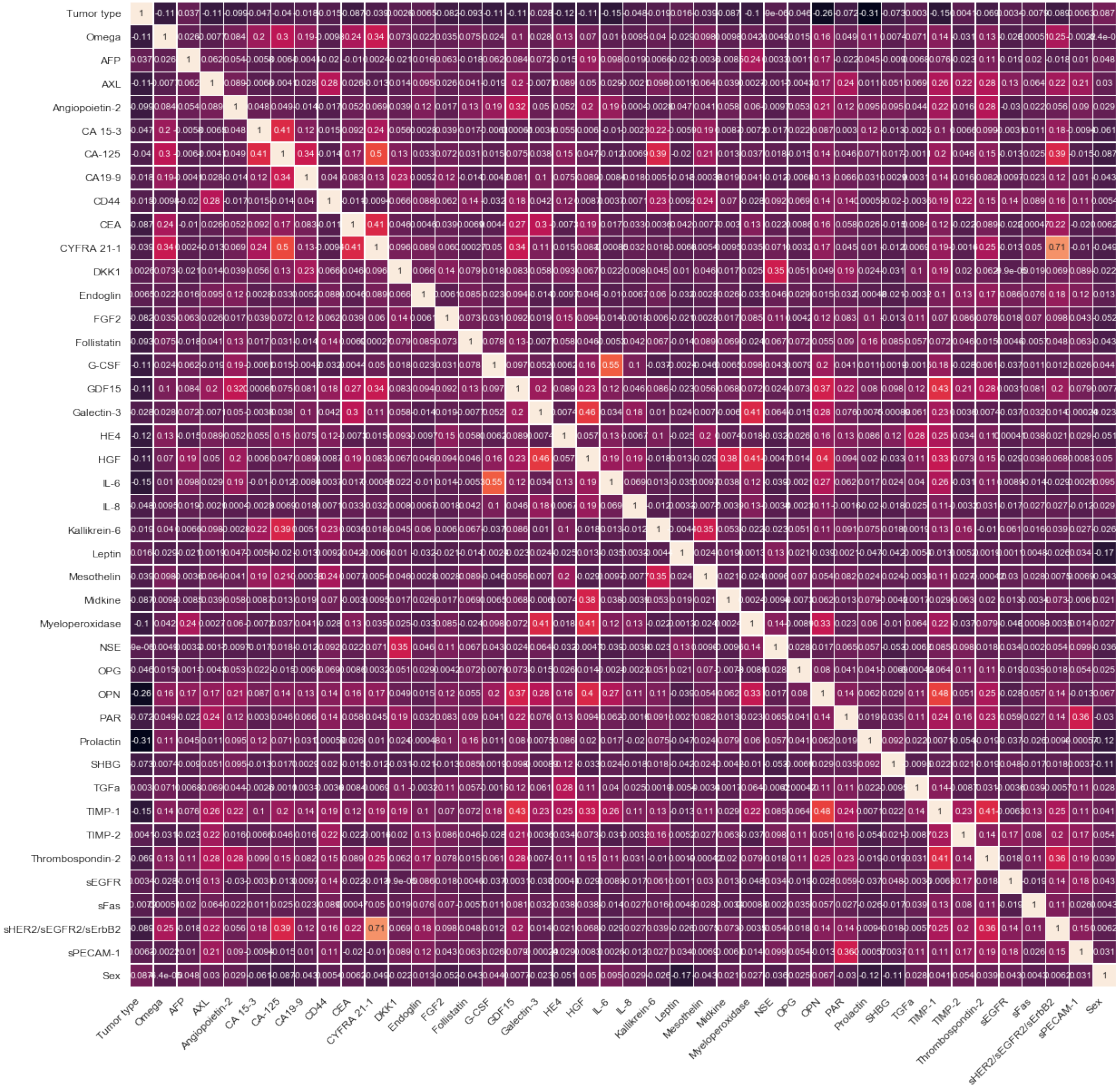
Train

Test



Common Steps (3)

- Data Visualisation – Correlation Matrix



Low multicollinearity

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

Common Steps (4)

Experimentation

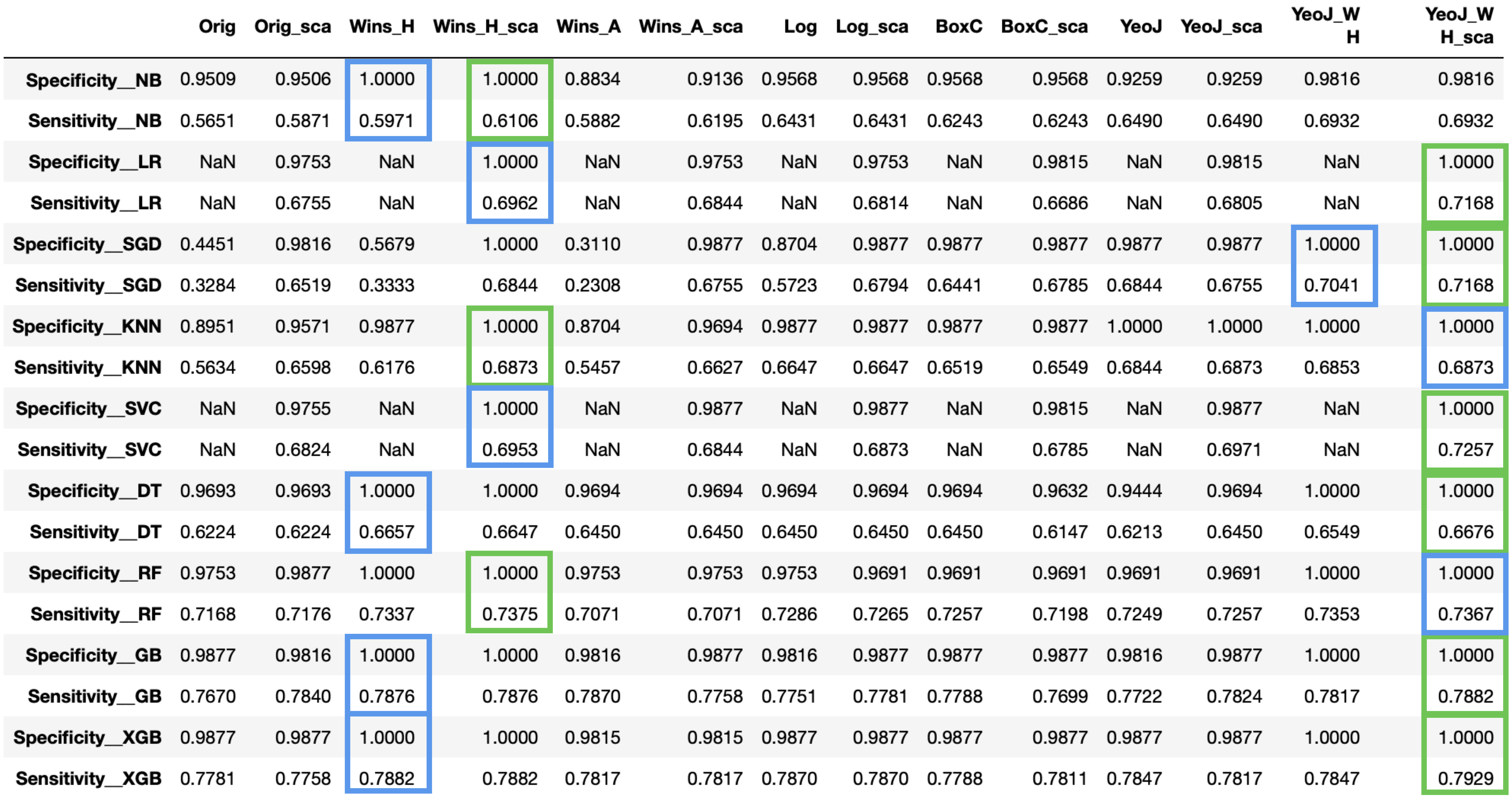
Data Transformation,
Feature Engineering,
Feature Selection
& Algorithms

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	0.9691	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

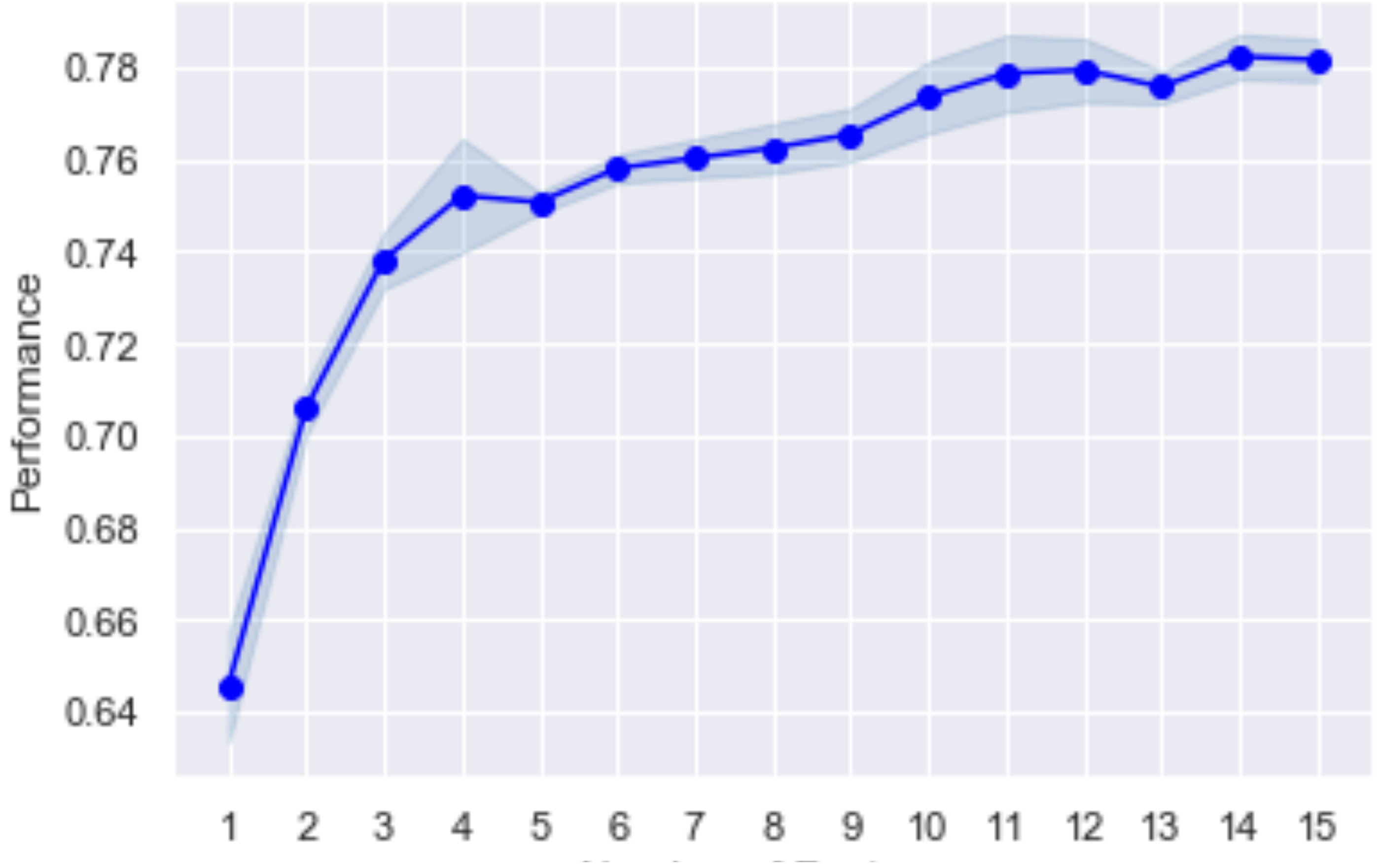
Feature Engineering

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

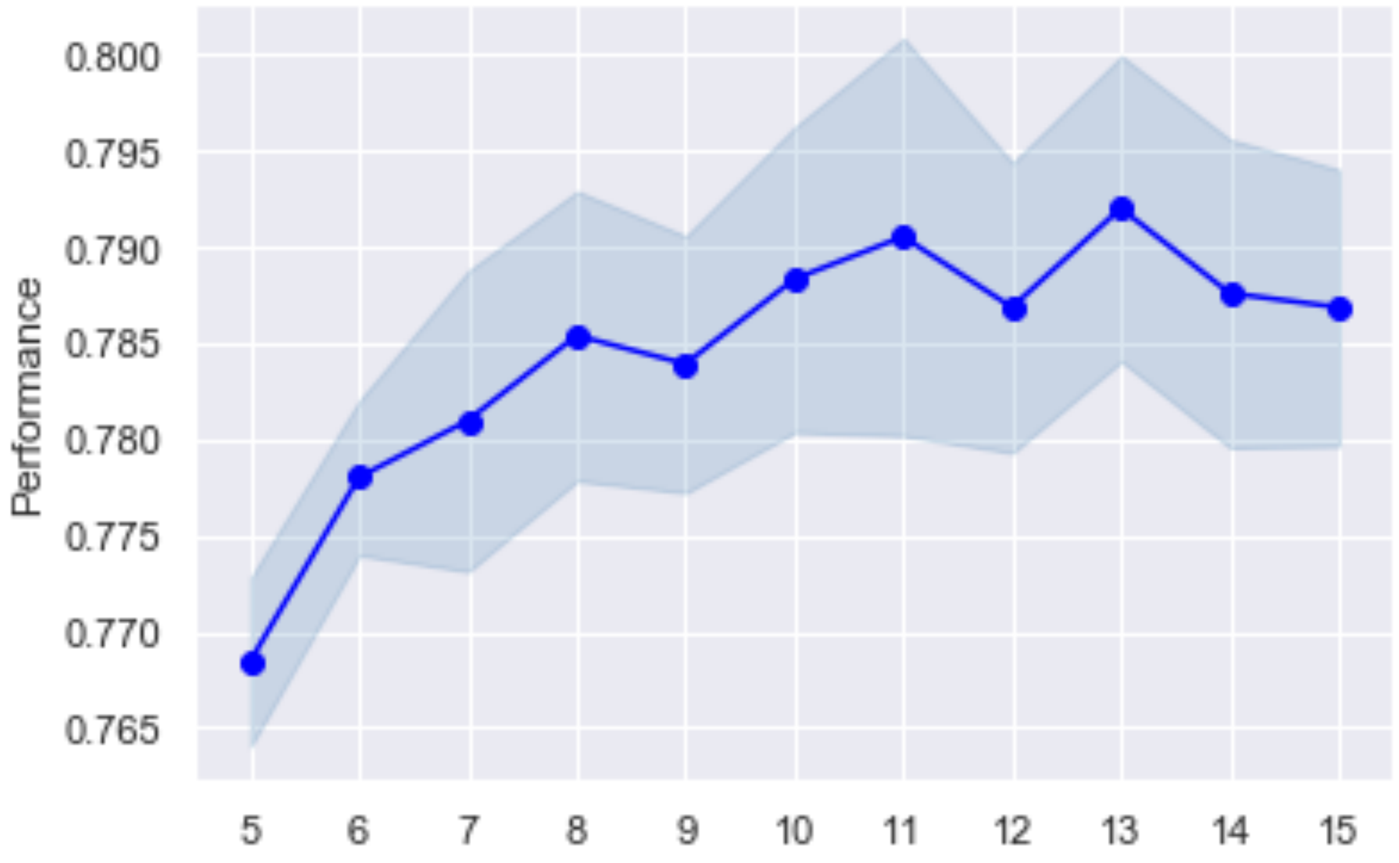
Feature Selection

- 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)

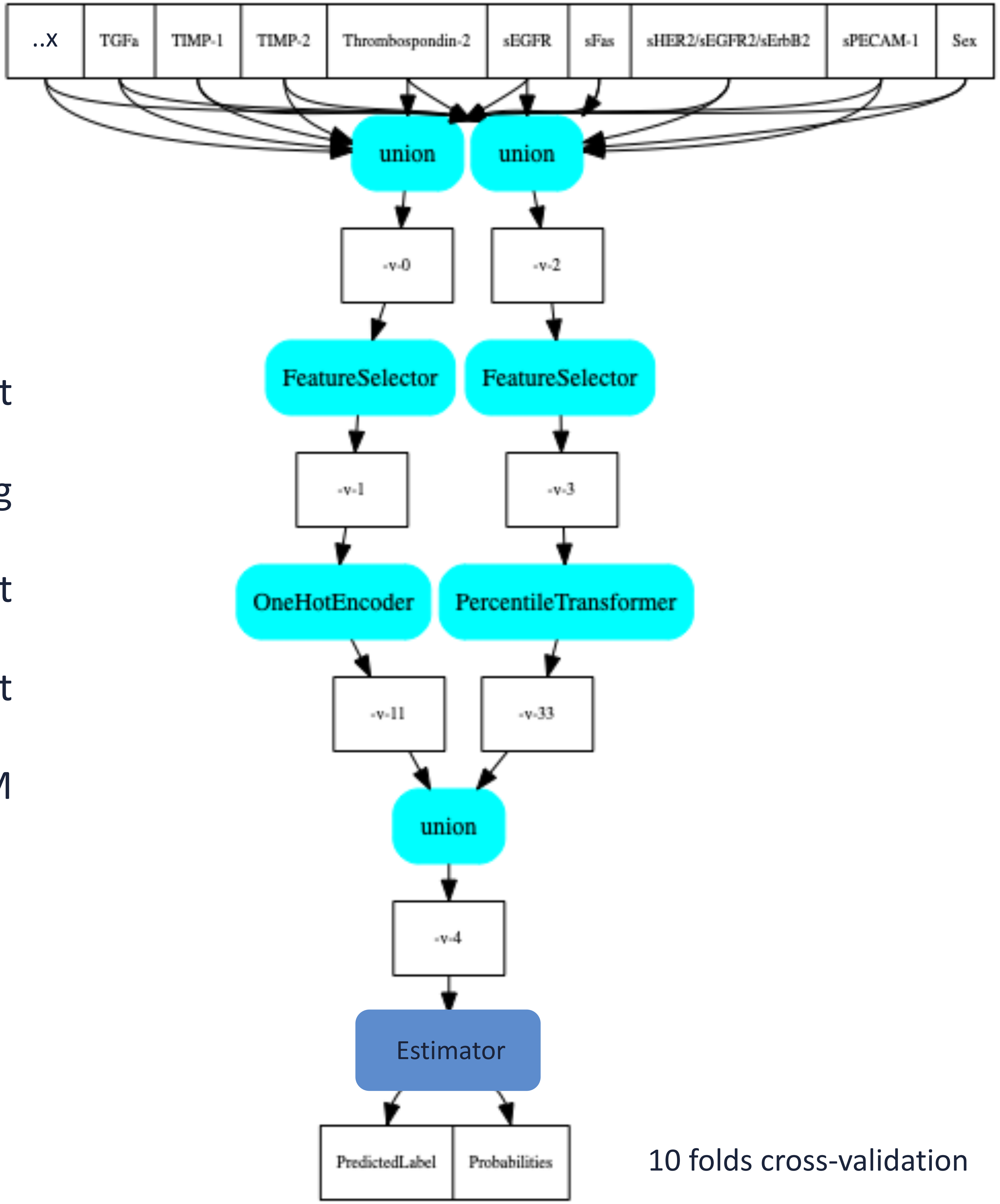
Pipeline

Common Steps (5)

- Pipeline

Naïve Bayes	Random Forest
Logistic Regression (+ SGD)	Gradient Boosting
K-Nearest Neighbors	XGBoost
Support Vector Machine	CatBoost
Decision Trees	LightGBM

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



10 folds cross-validation

Common Steps (5)

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

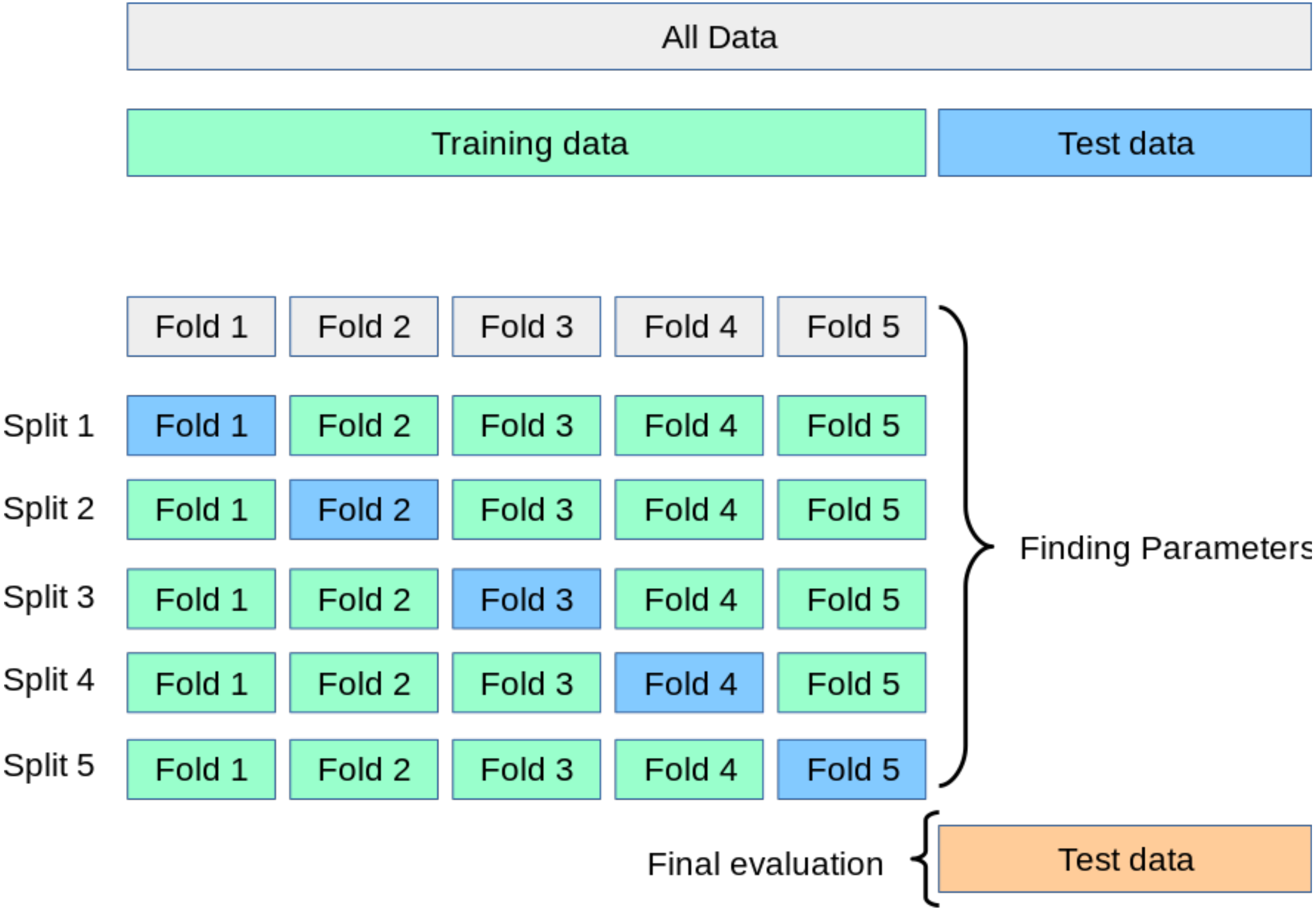
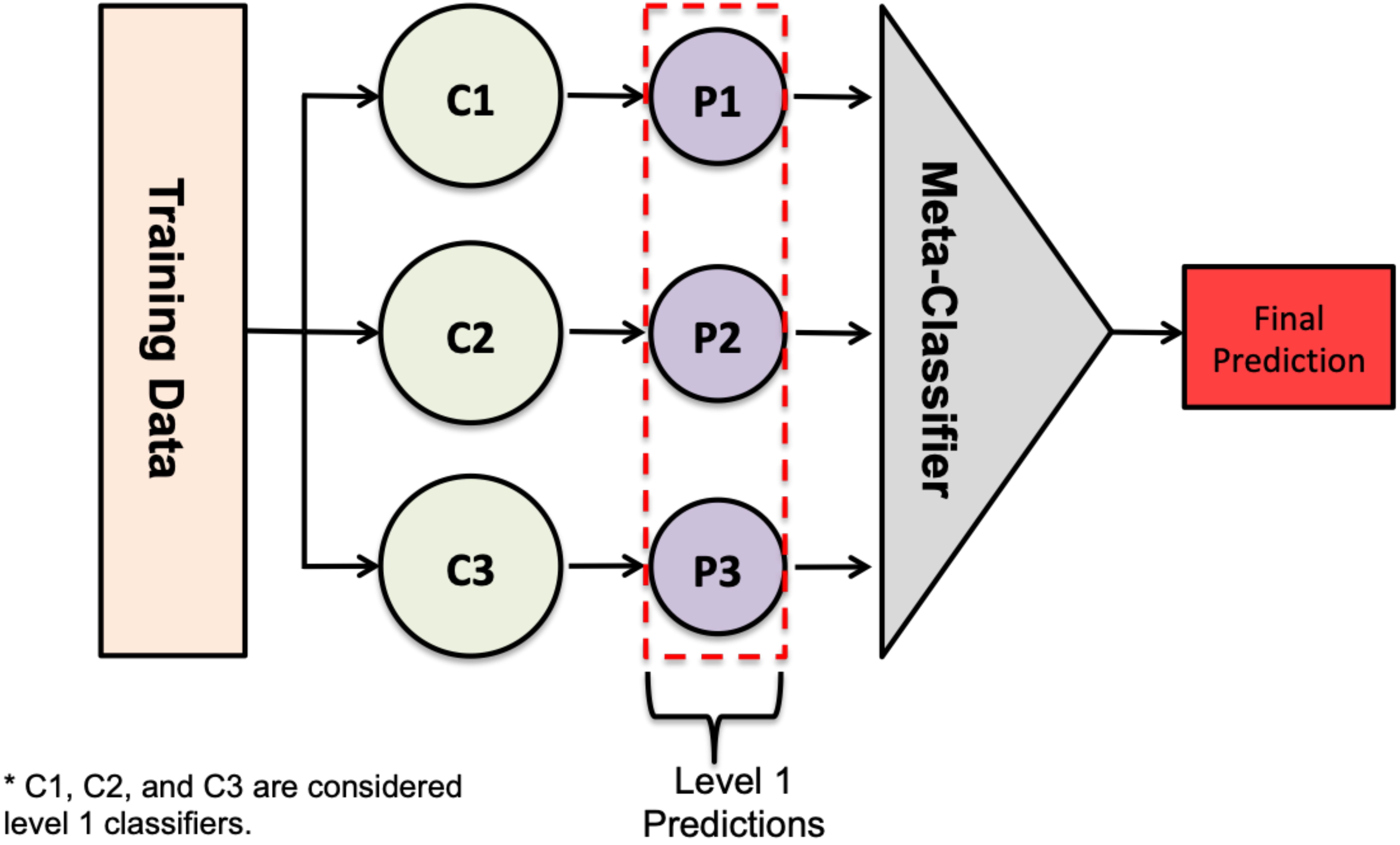


Image source: Scikit-Learn documentation

Common Steps (5)

- Stacking Classifier - Concept



* C1, C2, and C3 are considered level 1 classifiers.

Image source: Medium article*

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)

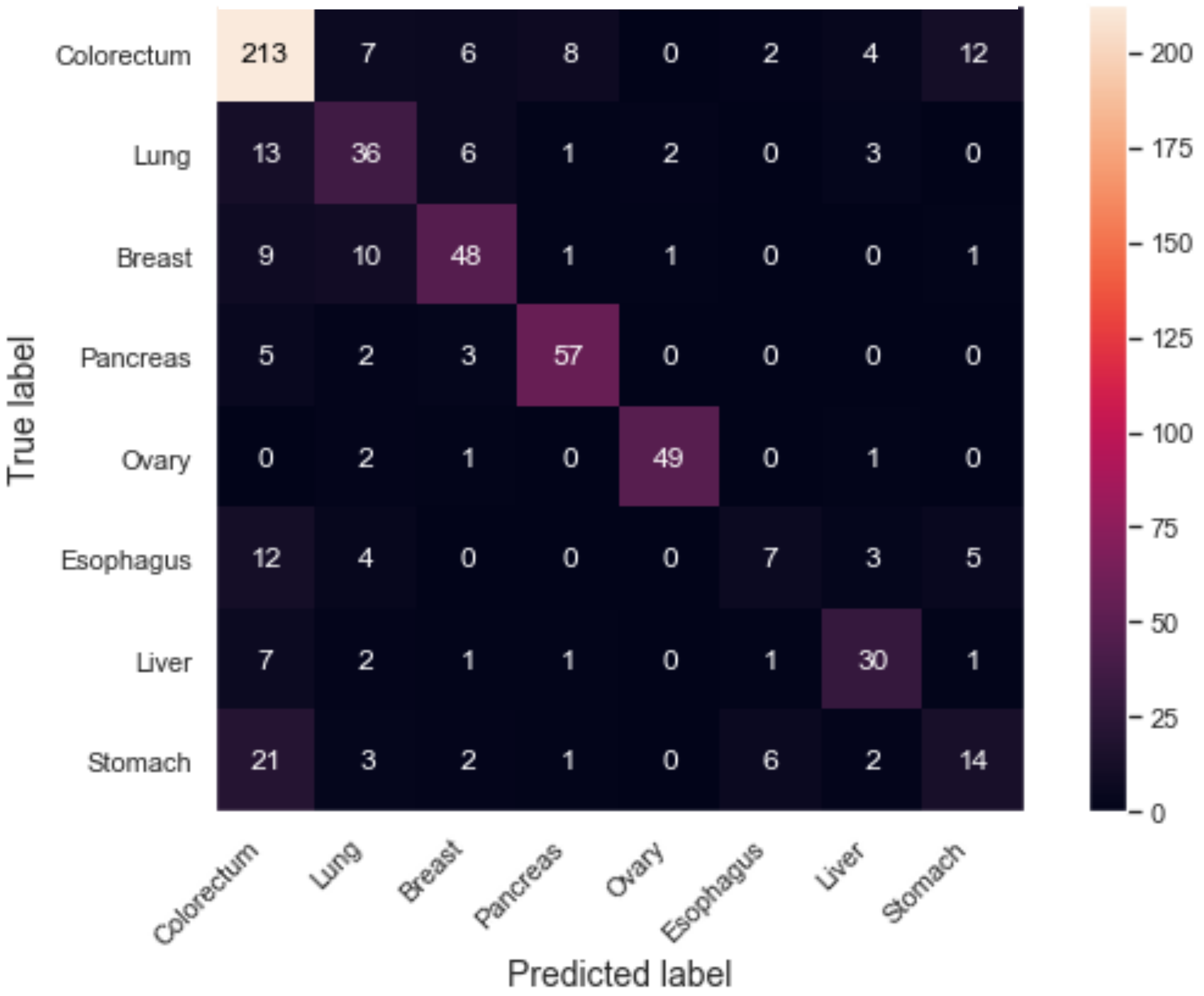
Cancer Type Classification

(as in publication)

Results (1)

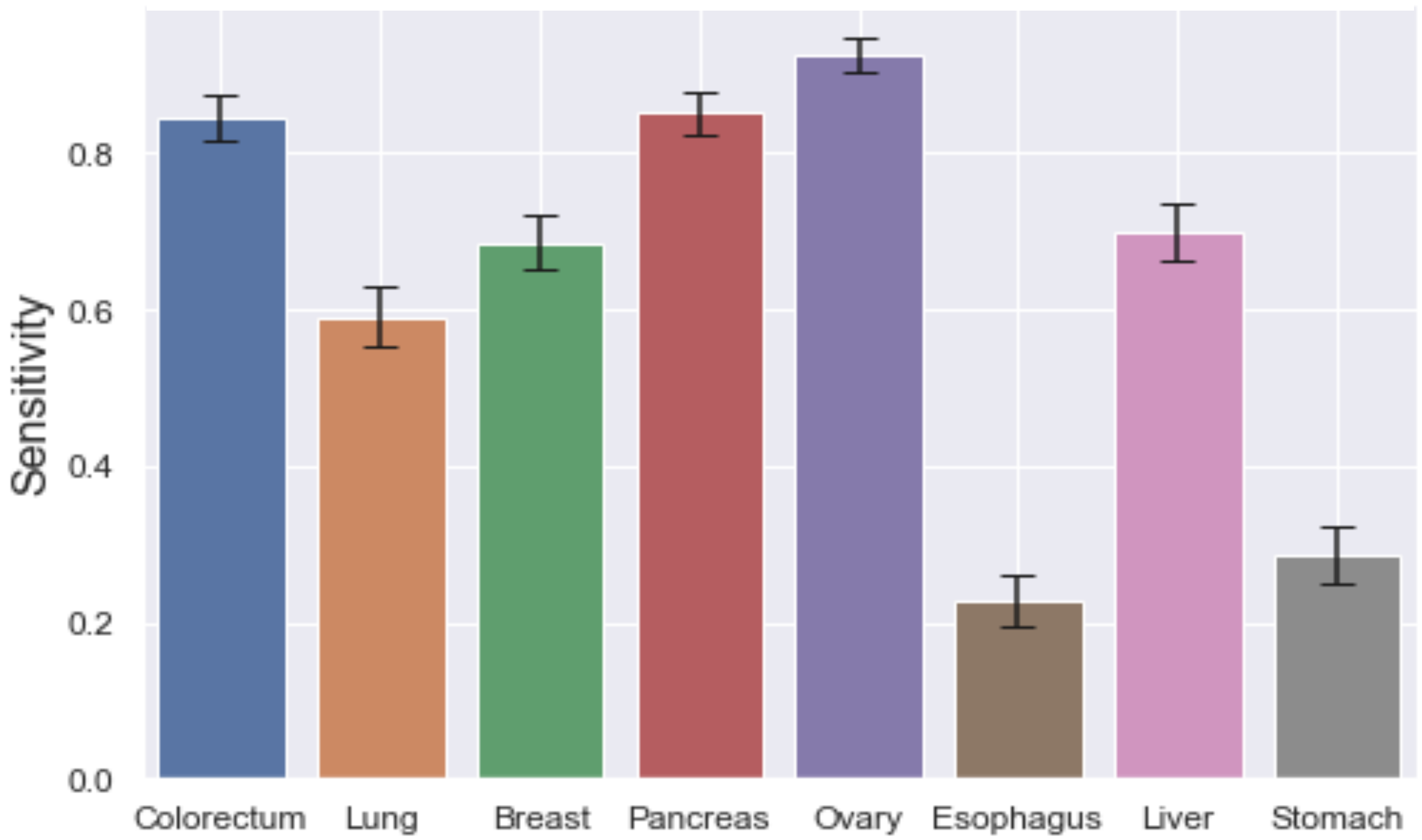
- Cancer Type Classification (as in publication)

Confusion Matrix



True label \ Predicted label	Colorectum	Lung	Breast	Pancreas	Ovary	Esophagus	Liver	Stomach
Colorectum	213	7	6	8	0	2	4	12
Lung	13	36	6	1	2	0	3	0
Breast	9	10	48	1	1	0	0	1
Pancreas	5	2	3	57	0	0	0	0
Ovary	0	2	1	0	49	0	1	0
Esophagus	12	4	0	0	0	7	3	5
Liver	7	2	1	1	0	1	30	1
Stomach	21	3	2	1	0	6	2	14

Sensitivity per Cancer type



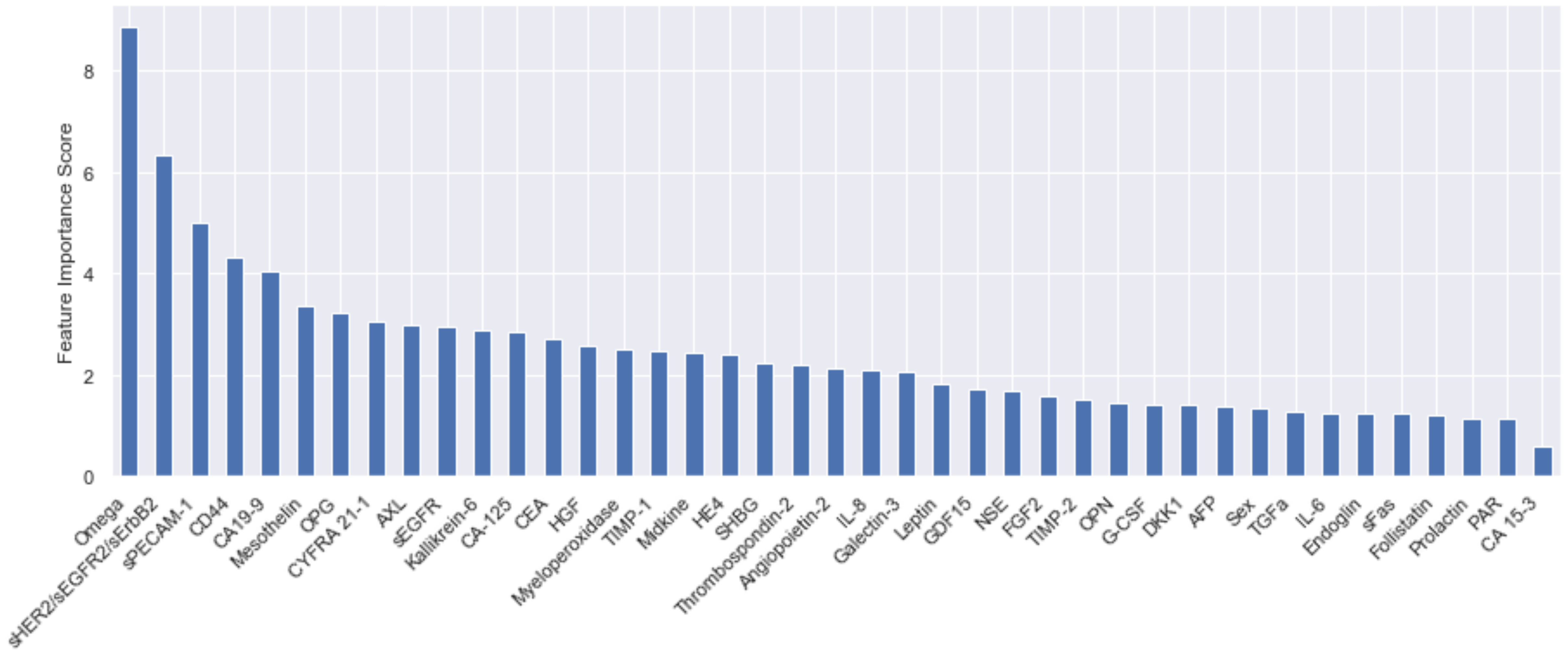
XGBoost

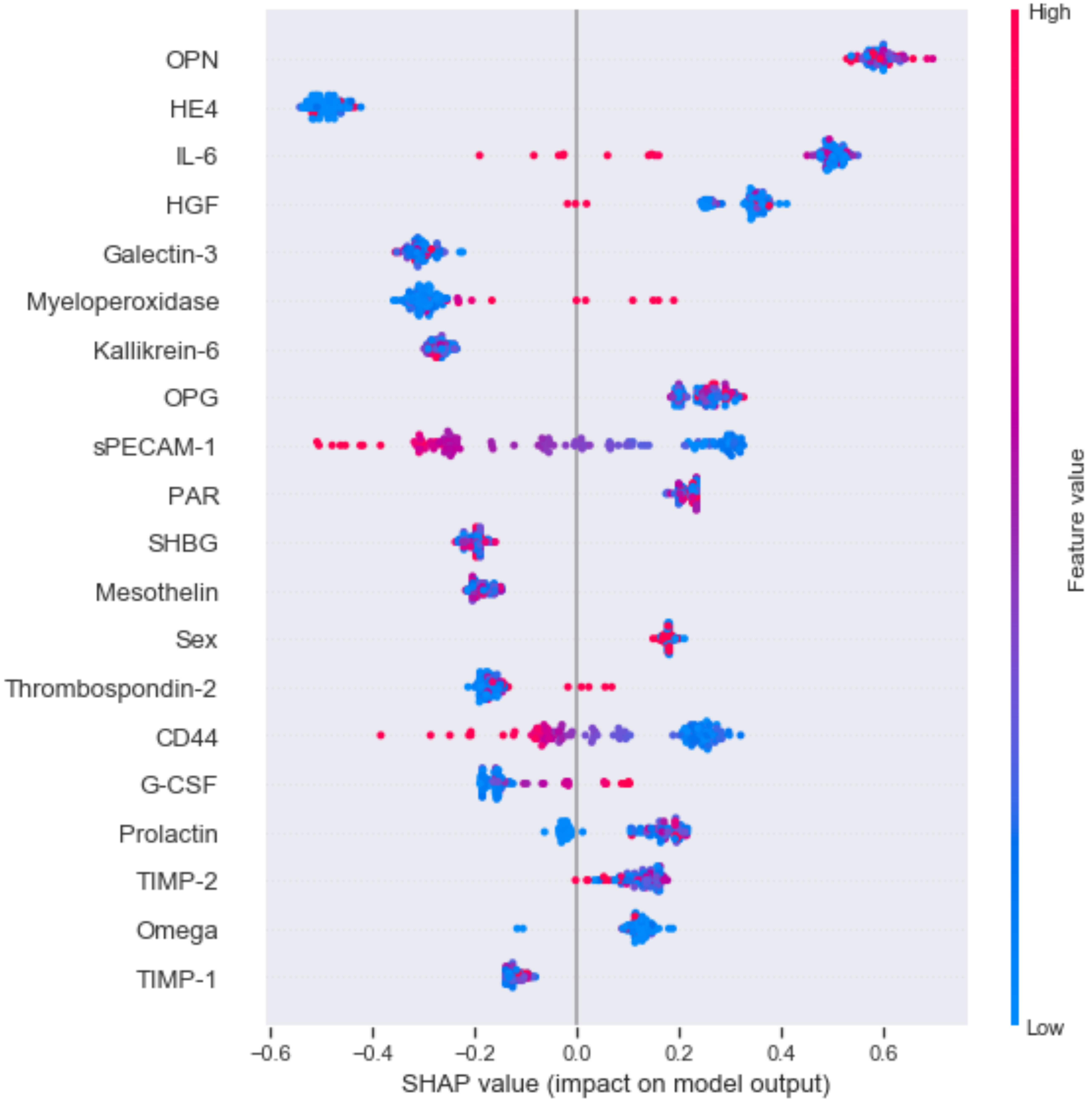
Gradient Boosting + CatBoost + LightGBM + XGBoost

Results (1)

- Cancer Type Classification
  (as in publication)
  Feature Importance

Feature Importance





Results (2)

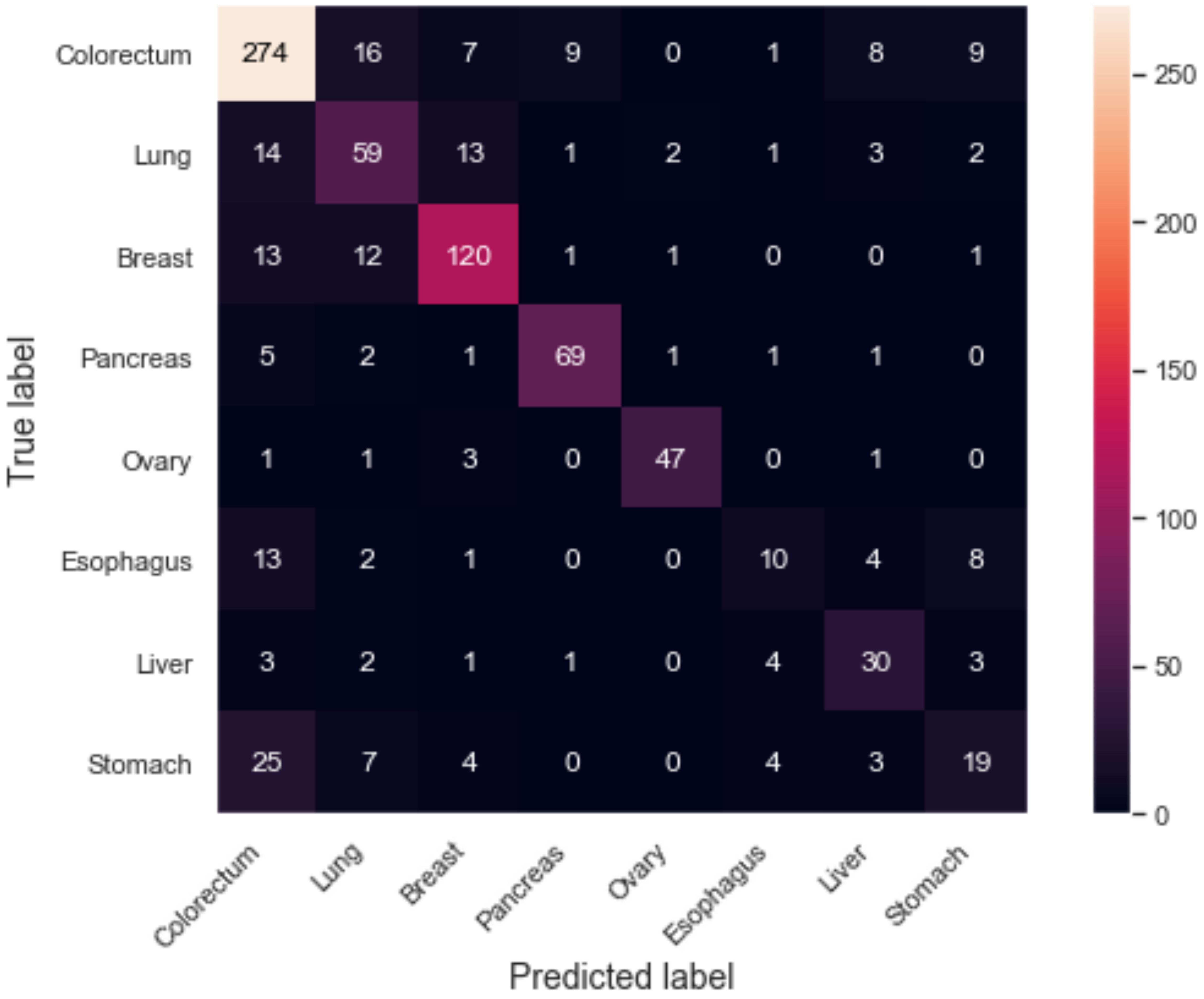
Cancer Type Classification

(On Full Dataset)

Results (2)

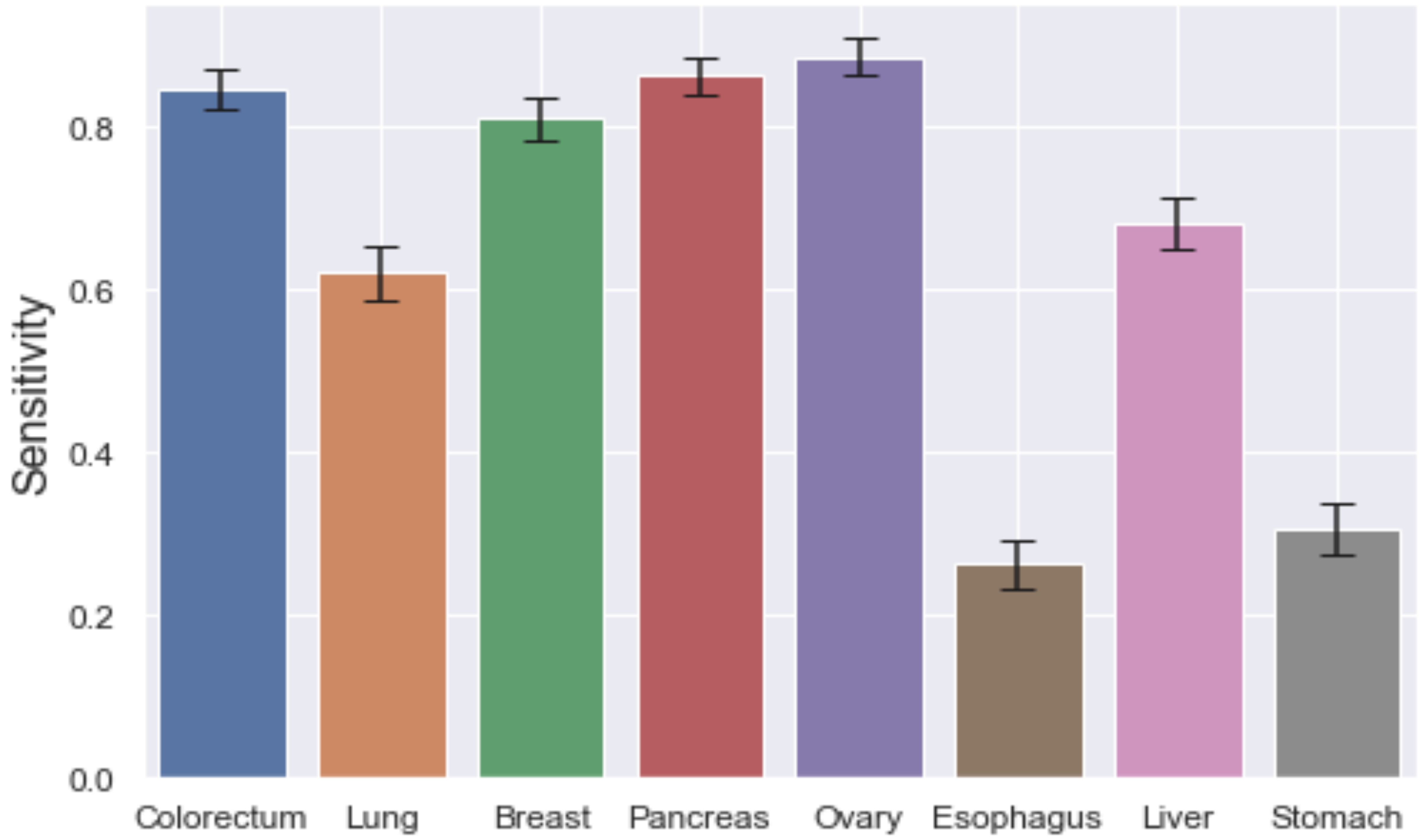
- Cancer Type Classification (on Full Dataset)

Confusion Matrix



True label \ Predicted label	Colorectum	Lung	Breast	Pancreas	Ovary	Esophagus	Liver	Stomach
Colorectum	274	16	7	9	0	1	8	9
Lung	14	59	13	1	2	1	3	2
Breast	13	12	120	1	1	0	0	1
Pancreas	5	2	1	69	1	1	1	0
Ovary	1	1	3	0	47	0	1	0
Esophagus	13	2	1	0	0	10	4	8
Liver	3	2	1	1	0	4	30	3
Stomach	25	7	4	0	0	4	3	19

Sensitivity per Cancer type

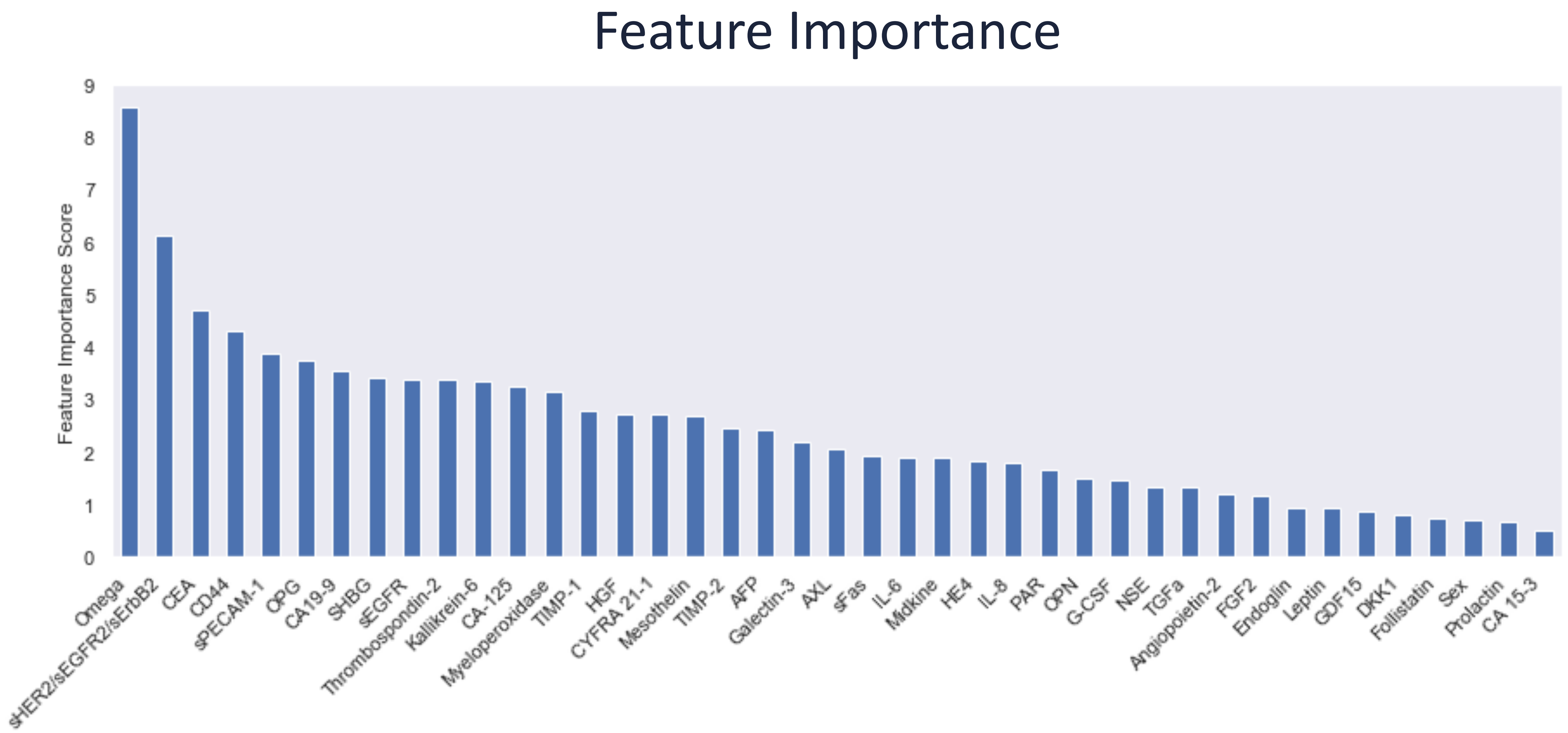


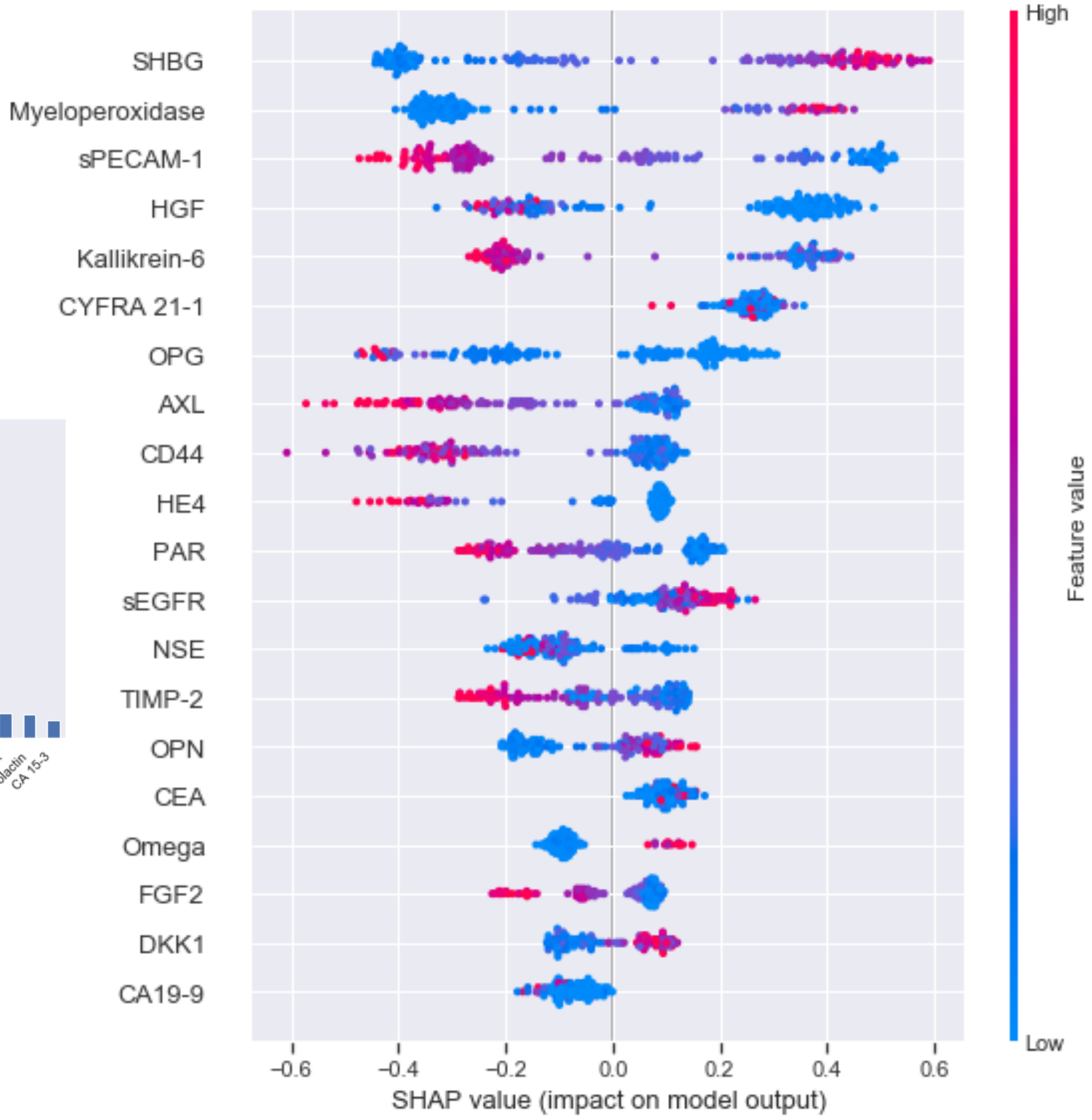
Cancer Samples Correctly Classified: 844 (84%) (626)

Specificity: 94% (99%)

Results (2)

- Cancer Type Classification
  (on Full Dataset)





Results (3)

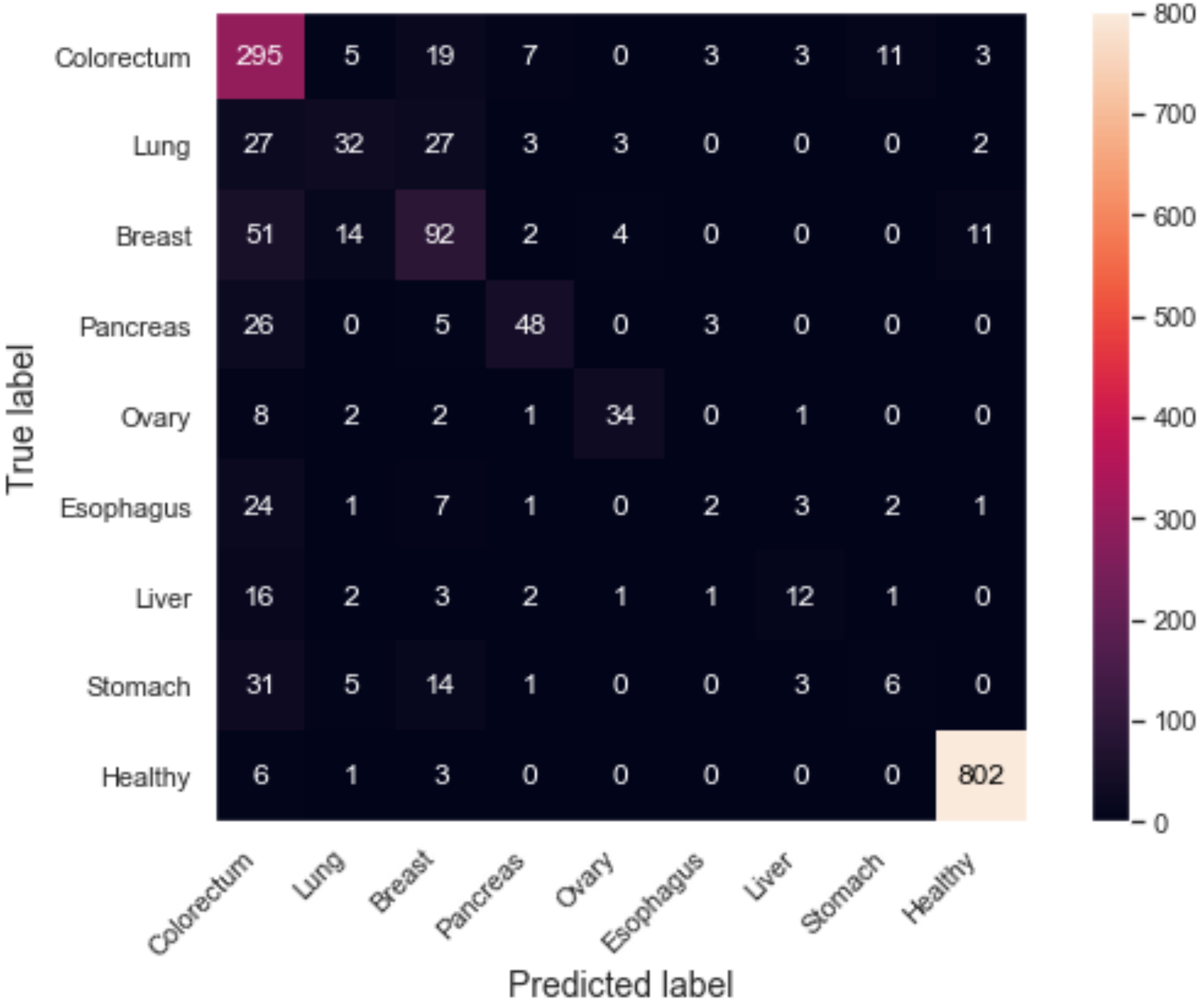
Cancer Type Classification

(On the Aneuploidy Dataset)

Results (3)

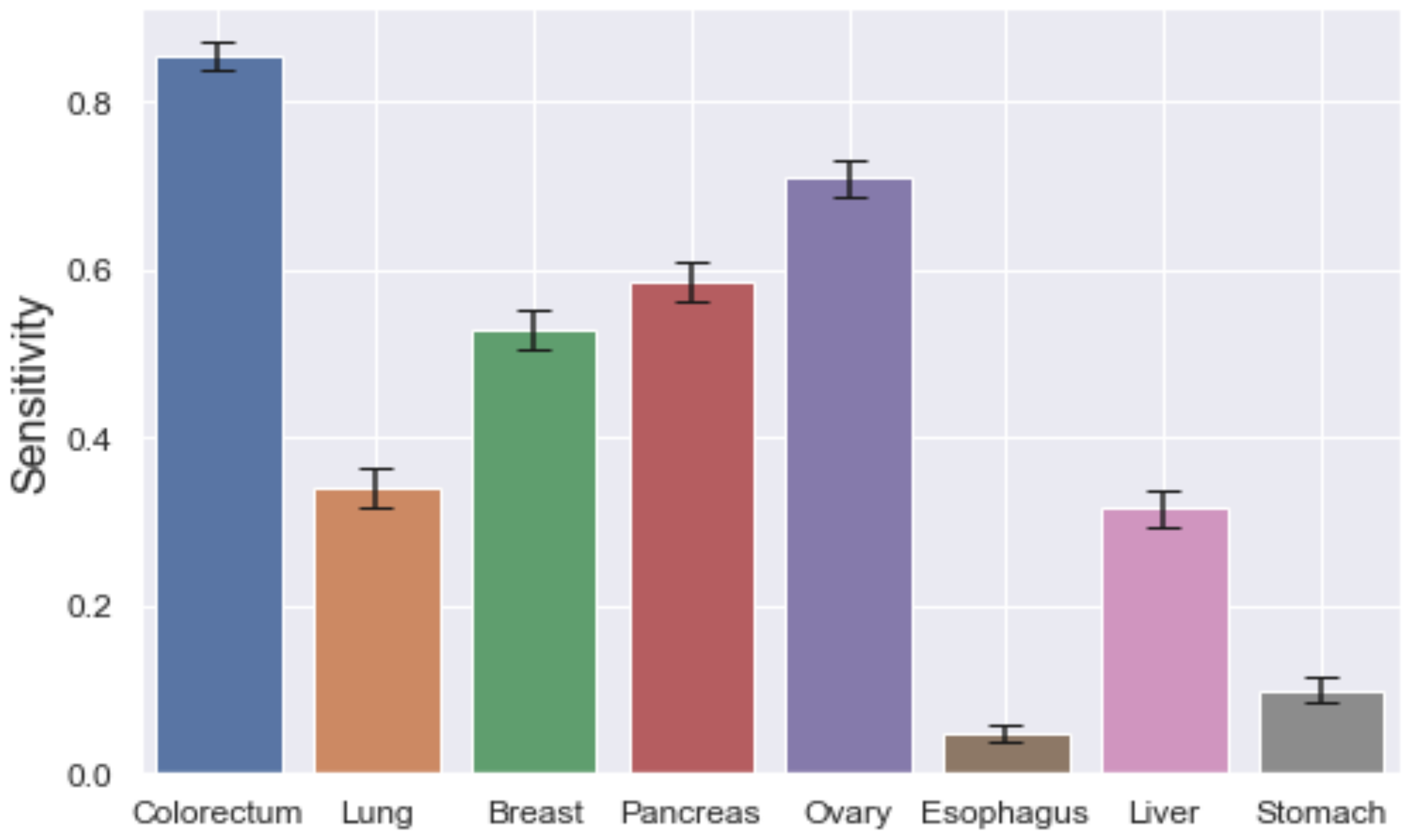
- Cancer Type Classification
  (on the Aneuploidy dataset)

Confusion Matrix



True label \ Predicted label	Colorectum	Lung	Breast	Pancreas	Ovary	Esophagus	Liver	Stomach	Healthy
Colorectum	295	5	19	7	0	3	3	11	3
Lung	27	32	27	3	3	0	0	0	2
Breast	51	14	92	2	4	0	0	0	11
Pancreas	26	0	5	48	0	3	0	0	0
Ovary	8	2	2	1	34	0	1	0	0
Esophagus	24	1	7	1	0	2	3	2	1
Liver	16	2	3	2	1	1	12	1	0
Stomach	31	5	14	1	0	0	3	6	0
Healthy	6	1	3	0	0	0	0	0	802

Sensitivity per Cancer type

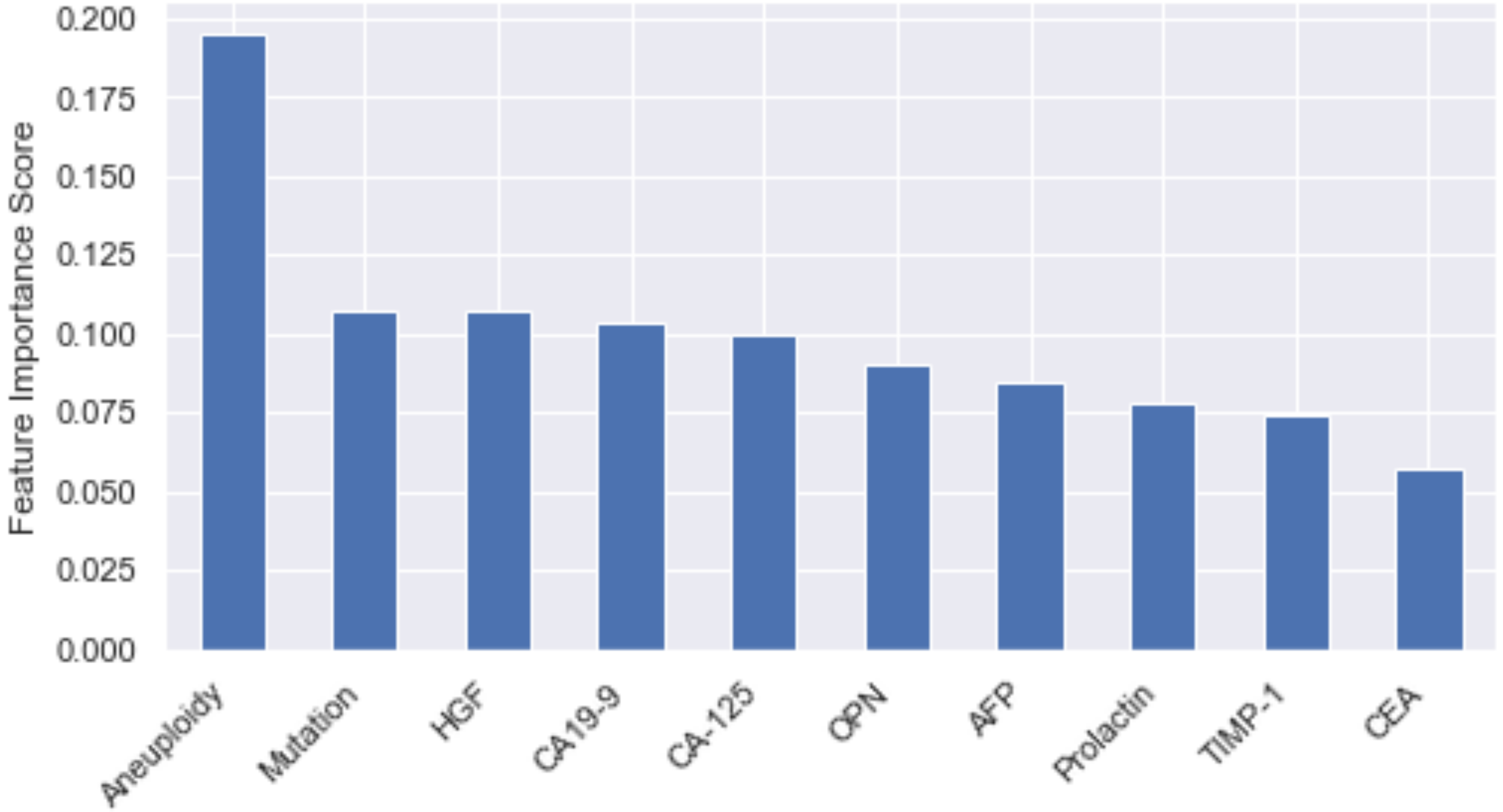


Sensitivity: 98%
Specificity: 99%
Precision (cancer): 99%

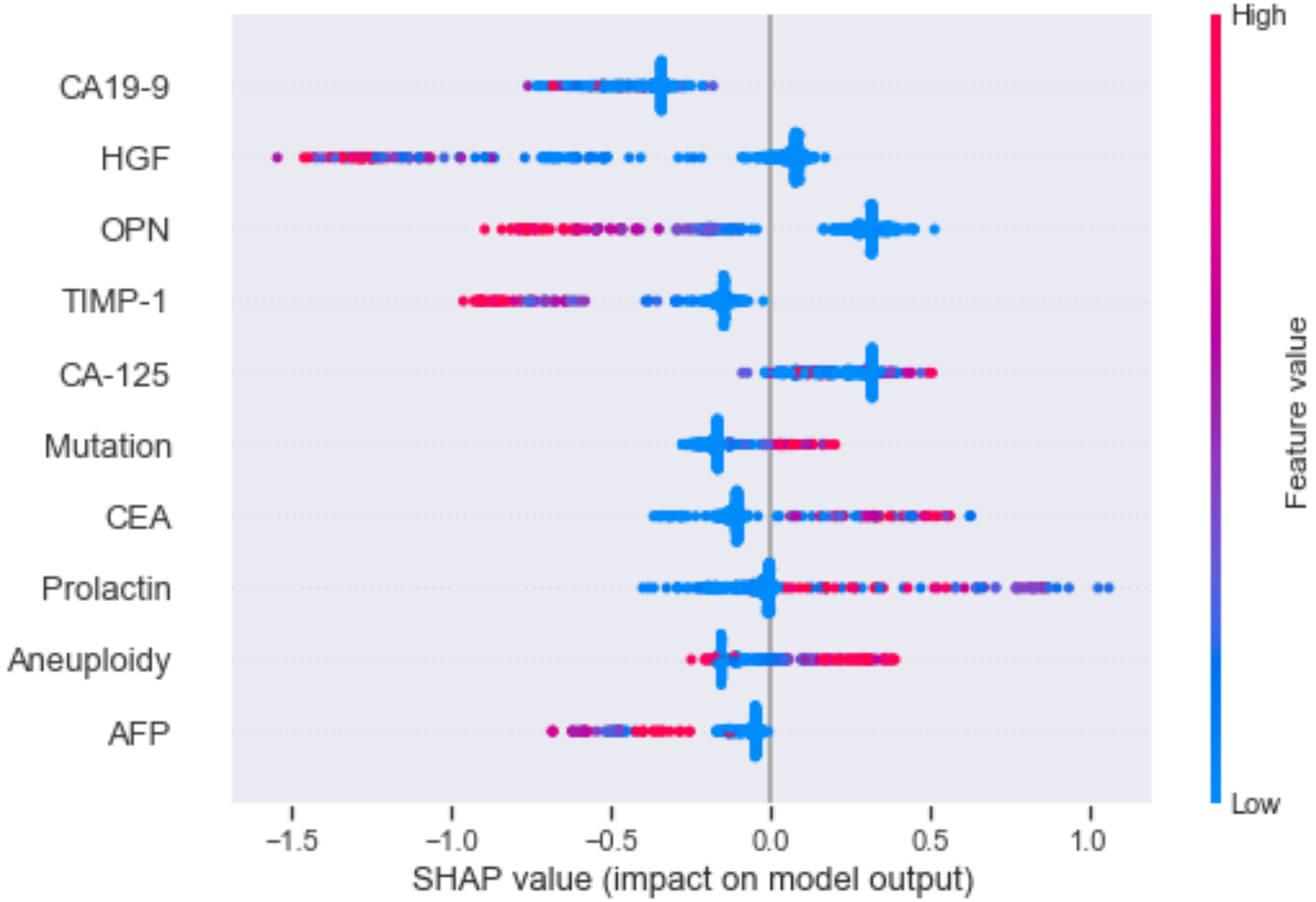
Results (3)

- Cancer Type Classification
  (on the Aneuploidy dataset)

Feature Importance



SHAP Values



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