Survival analysis of tcga

What I’ve done- summary

In order to have a wide infrastructure and arsenal for the various tasks at hand, I’ve implemented multiple methods and a feature selection process below on each omic and on the merged omics for each cancer type.

All applied methods mentioned below are pending hyperparameter tuning.

Feature selection and cleaning

- Log transformation to mirna and exp files
- Handle samples with negative ‘last_contact_days_to’
- Removing duplicates
- Removing features with std == 0
- Normalization (scaling) of features
- Calculate mutual information and remove insignificant features relatively to the event the duration
- Correlation between all remaining features and remove highly correlated features (>0.9)

Methods

COX

- No Regularization
- Ridge
- Lasso
- Elastic net

Gradient Boosting

- No Regularization
- Dropout
- Subsample
- Learning rate

Random Survival Forest

Vanilla NN

Details about the methods

The Cox model

The Cox model with Ridge regularization coefficients optimize the problem below:

$$\arg\max_{\beta} \quad \log \mathrm{PL}(\beta) - \frac{\alpha}{2} \sum_{j=1}^{p} \beta_j^2$$

α≥0 is a hyper-parameter that controls the amount of shrinkage.
I checked the alphas [0.01,0.1,0.3,0.5,0.7].

The Cox model with Lasso regularization coefficients optimize the problem below:

$$\arg\max_{\beta} \quad \log \mathrm{PL}(\beta) - \alpha \sum_{j=1}^{p} |\beta_j|$$

α≥0 is a hyper-parameter that controls the amount of shrinkage.
I checked 50 α values that give up to 5% of the estimated maximum.

The Cox model with Elastic net regularization coefficients optimize the problem below:

$$\arg\max_{\beta} \quad \log \mathrm{PL}(\beta) - \alpha \left(r \sum_{j=1}^{p} |\beta_j| + \frac{1-r}{2} \sum_{j=1}^{p} \beta_j^2 \right)$$

r∈[0,1] is the relative weight of the L1 and L2 penalty.
I chose r = 0.5 and checked 50 α values that give up to 5% of the estimated maximum.

Gradient Boosting

The loss function is the partial likelihood loss of Cox’s proportional hazards model . Therefore, the objective is to maximize the log partial likelihood function, but replacing the traditional linear model with the additive model.
$f(\mathbf{x})$:

$$\arg\min_{f} \quad \sum_{i=1}^{n} \delta_i \left[f(\mathbf{x}_i) - \log \left(\sum_{j \in \mathcal{R}_i} \exp(f(\mathbf{x}_j)) \right) \right]$$

I used a test portion of 0.15 of the data.
I checked the range of the number of estimators (weak learners).
The range is between 25 to 160 with jumps of 5 estimators in each iteration.

I tried each of the number of estimators on each of the regularizations:

The Gradient boosting model with dropout forces the base learners to account for some of the previously fitted base learners to be missing.
I chose a dropout rate of 0.1.

The Gradient boosting model using subsample uses a subsample of less than 1 such that each iteration only a portion of the training data is used.
I chose a subsample of 0.5 in each iteration.

The Gradient boosting model using a learning rate less than 1 to restrict the influence of individual base learners.
I chose a learning rate of 0.1.

Random Survival Forest
I used a test portion of 0.20 of the data.
I checked 1000 estimators as the number of estimators.

Vanilla NN
I used a general NN from this link.
I chose to split the duration to 50 intervals.
I used a validation portion of 0.20 and a test portion of 0.15 of the data.
I'm planning to extend and explore this and other networks.

The Tasks

TASK 1
For this task I will choose the best model based on data of the merged omics.

The Results
COX

- No Regularization
- Ridge
- Lasso
- Elastic net

Cancer type	**Regularization**	**alpha**	**Concordance index**
BLCA	no regularization		**0.551**
	ridge	0.3	**0.658**
	elastic net	0.04	**0.626**
BRCA	no regularization		**0.652**
	ridge	0.1	**0.658**
	lasso	0.003	**0.645**
	elastic net	0.005	**0.692**
HNSC	no regularization		**0.565**
	ridge	0.05	**0.627**
	lasso	0.004	**0.575**
	elastic net	0.006	**0.599**
LAML	no regularization		**0.53**
	ridge	0.7	**0.608**
	lasso	0.027	**0.596**
	elastic net	0.011	**0.58**

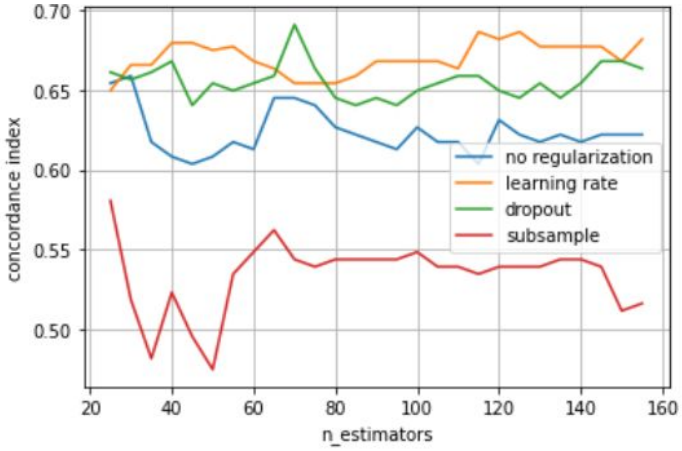
LGG	no regularization		**0.757**
	ridge	0.3	**0.886**
	lasso	0.01	**0.873**
	elastic net	0.012	**0.88**
LUAD	no regularization		**0.448**
	ridge	0.3	**0.55**
	lasso	0.001	**0.61**
	elastic net	0.0318	**0.588**

Boosting

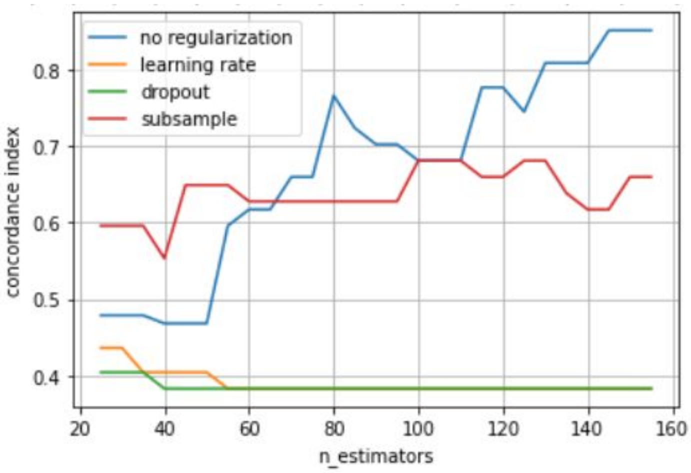
- No Regularization
- Dropout
- Subsample
- Learning rate

Cancer type	**Regularization**	**Number of estimators**	**Concordance index**
BLCA	dropout	70	**0.691**
BRCA	no regularization	145,150,150	**0.851**
LAML	no regularization	65	**0.862**
LGG	no regularization	90,95,120	**0.796**
	dropout	25,30	
LUAD	subsample	40	**0.75**
HNSC	no regularization	150	**0.783**

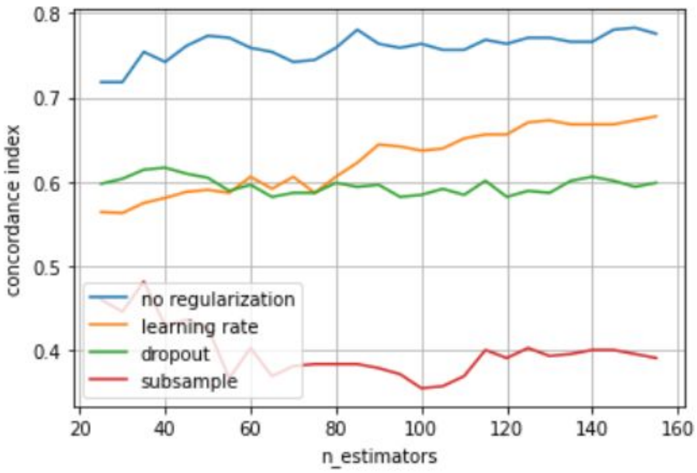
BLCA:



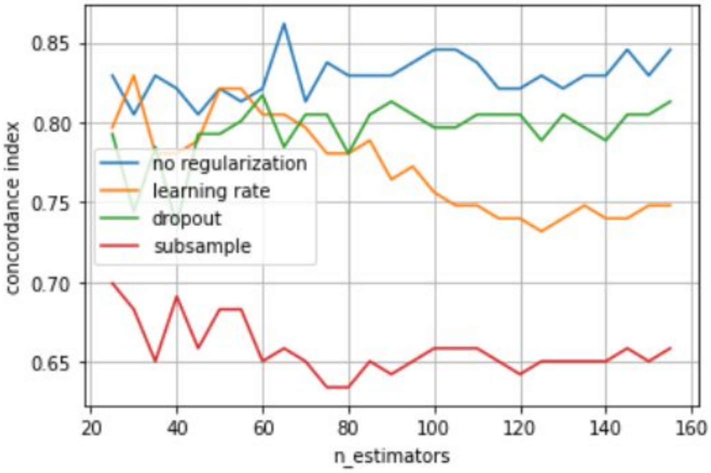
BRCA:



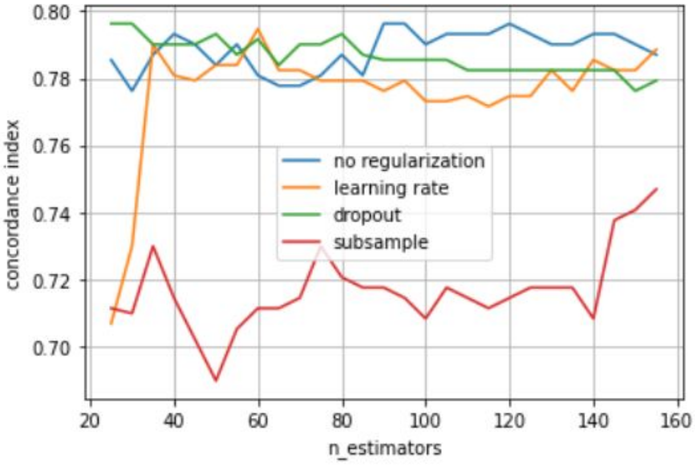
HNSC:



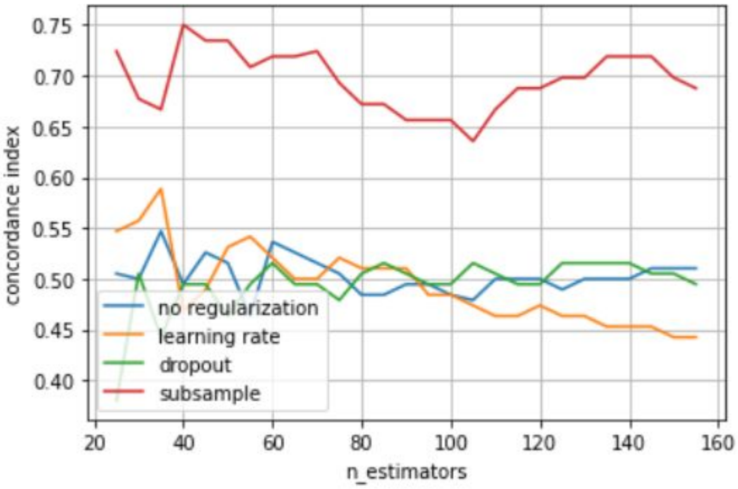
LAML:


concordance index
0.85
0.80
0.75
0.70
0.65
n_estimators
20
40
60
80
100
120
140
160
no regularization
learning rate
dropout
subsample

LGG:


concordance index
0.80
0.78
0.76
0.74
0.72
0.70
n_estimators
20
40
60
80
100
120
140
160
no regularization
learning rate
dropout
subsample

LUAD:


concordance index
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
n_estimators
20
40
60
80
100
120
140
160
no regularization
learning rate
dropout
subsample

Random Survival Forest

Cancer type	Random Survival Forest
BLCA	**0.596**
BRCA	**0.658**
LAML	**0.741**
LGG	**0.839**
LUAD	**0.72**
HNSC	**0.444**

Vanilla NN

Cancer type	Concordance index
BLCA	**0.92**
BRCA	**0.964**
LAML	**0.874**
LGG	**0.623**
LUAD	**0.932**
HNSC	**0.972**

Task 1 best results using 5-cross validation concordance index

Cancer type	Regularization	Hyper parameter	Concordance index
LGG	ridge	Alpha 0.3	0.886
LAML	Boosting no regularization	65	0.862
BRCA	Boosting no regularization	145 150 150	0.851
HNSC	Boosting no regularization	150	0.783
LUAD	Boosting subsample	40	0.75
BLCA	Boosting dropout	70	0.691

Plans

- Check tuning of hyper parameters
- Add features from the clinical data as gender and age in RSF - if that's improves the model - make a predictor to the clinical data based on the omics and then integrate it as a feature
- Check methods to consider the multi view omics- probably use the baseline predictors for each omic

Task 2
I build a baseline survival predictor based on each omic for each cancer type. For now the end results of this task are the best model based on data of gene expression.

The Results
COX

- No Regularization
- Ridge
- Lasso
- Elastic net

Cancer type	**Regularization**	**alpha**	**Concordance index**
BLCA	no regularization		**0.579**
	ridge	0.05	**0.679**
	lasso	0.004	**0.613**
	elastic net	0.005	**0.681**
BRCA	no regularization		**0.529**
	ridge	0.05	**0.679**
	lasso	0.004	**0.613**
	elastic net	0.006	**0.66**
LAML	no regularization		**0.574**
	ridge	0.05	**0.641**
	lasso	0.004	**0.608**
	elastic net	0.01	**0.606**
LGG	no regularization		**0.696**
	ridge	0.1	**0.877**
	lasso	0.006	**0.857**

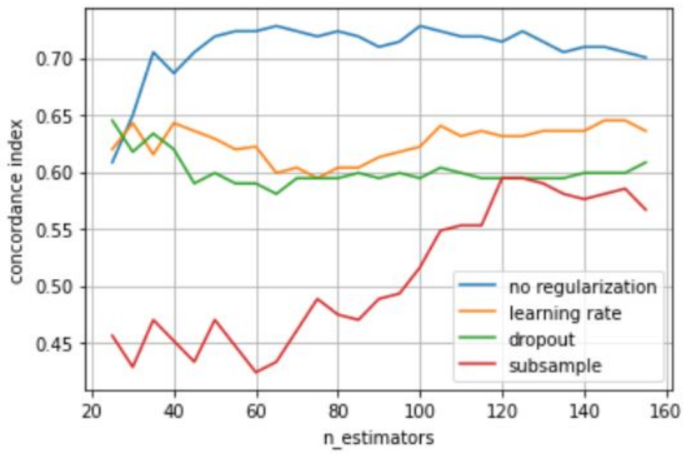
	elastic net	0.006	**0.868**
LUAD	no regularization		**0.464**
	ridge	0.7	**0.551**
	lasso	0.016	**0.584**
	elastic net	0.033	**0.588**

Boosting

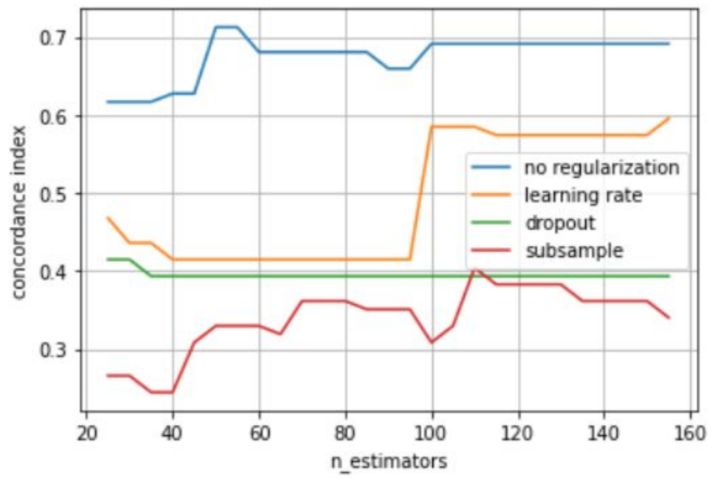
- No Regularization
- Dropout
- Subsample
- Learning rate

Cancer type	**Regularization**	**Number of estimators**	**Concordance index**
BLCA	no regularization	65, 100	**0.728**
BRCA	no regularization	50, 55	**0.712**
LAML	subsample	30	**0.837**
LGG	no regularization	45, 55	**0.855**
LUAD	learning rate	25, 30, 35	**0.599**
HNSC	no regularization	95	**0.795**

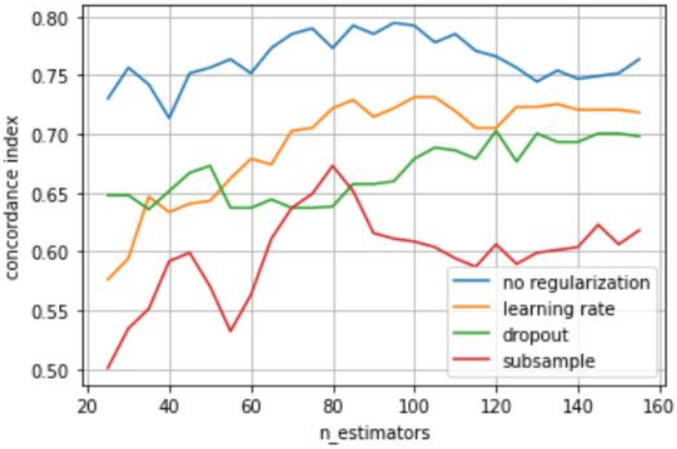
BLCA:


concordance index
0.70
0.65
0.60
0.55
0.50
0.45
20
40
60
80
100
120
140
160
n_estimators
no regularization
learning rate
dropout
subsample

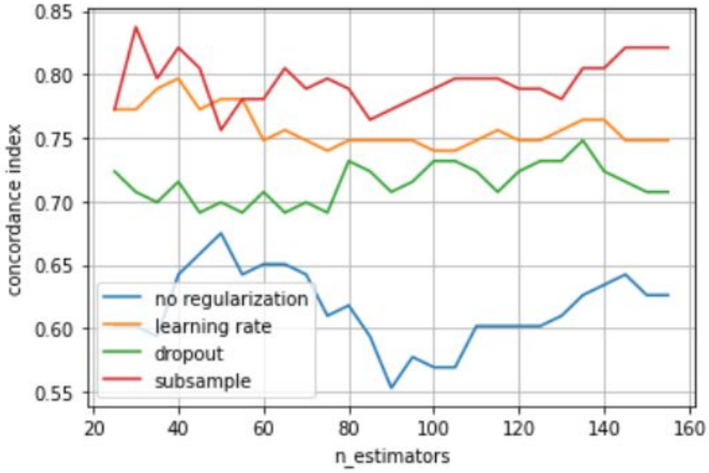
BRCA:


concordance index
0.7
0.6
0.5
0.4
0.3
20
40
60
80
100
120
140
160
n_estimators
no regularization
learning rate
dropout
subsample

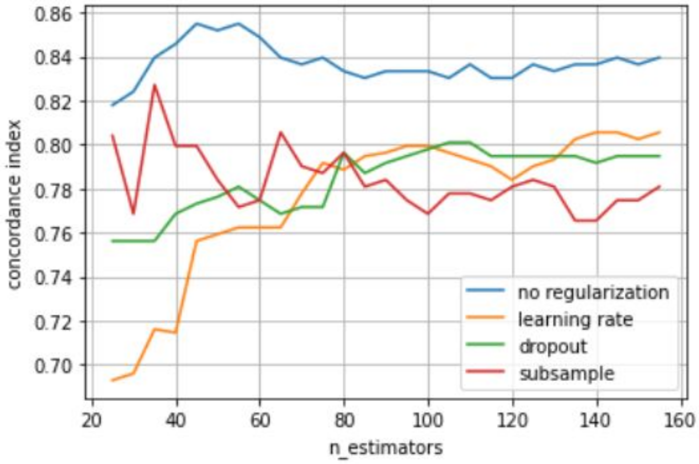
HNSC:



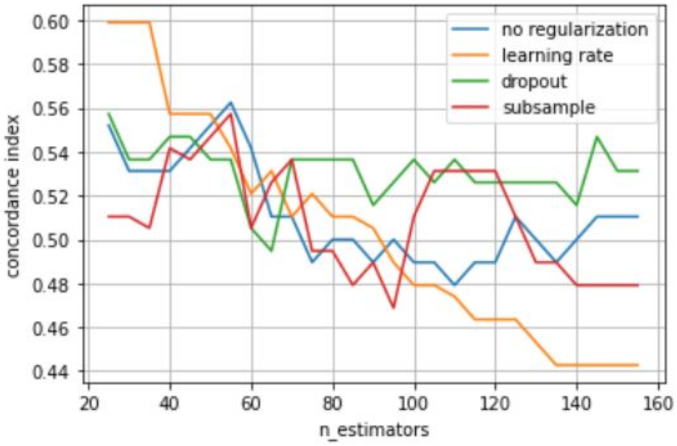
LAML:



LGG:



LUAD:


no regularization
learning rate
dropout
subsample
concordance index
0.60
0.58
0.56
0.54
0.52
0.50
0.48
0.46
0.44
20
40
60
80
100
120
140
160
n_estimators

Random Survival Forest

Cancer type	**RF**
BLCA	**0.606**
BRCA	**0.726**
LAML	**0.666**
LGG	**0.833**
LUAD	**0.672**
HNSC	**0.399**

Vanilla NN

Cancer type	**Concordance Index**
BLCA	**0.773**
BRCA	**0.99**
LAML	**0.951**
LGG	**0.965**
LUAD	**0.954**
HNSC	**0.978**

Task 2 best results using 5-cross validation concordance index

Cancer type	Regularization	alpha	Concordance index
BLCA	no regularization	65,100	0.728
BRCA	no regularization	50,55	0.712
HNSC	no regularization	95	0.795
LAML	subsample	30	0.837
LGG	ridge	0.1	0.877
LUAD	learning rate	25,30,35	0.599

Plans

- Check tuning of hyper parameters
- Learn the other omics representation based on the tested omic
  - NN
  - Regression
  - RF Regressor
  - Gradient Boosting Regressor
- Add the representation of the other omics the survival predictor of the tested omic
- Compare the results

Task 3

I build a baseline survival predictor based on all the omics for each cancer type.
For now the end results of this task are the best model based on data of tested cancer type - the same results as in task 1.

Plans

- Check tuning of hyper parameters
- Learn the representation of the other cancer types survival based on the tested on
    - NN
    - Regression
    - RF regressor
    - Gradient Boosting Regressor
- Add the representation of the other cancer types to the survival predictor of the tested cancer type
- Compare results to baseline and to each other