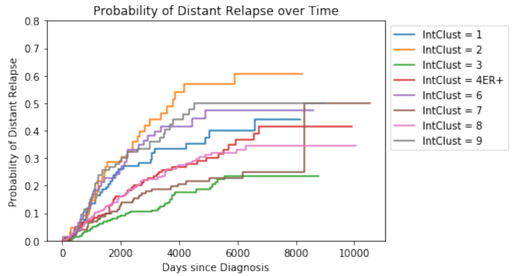
Predicting Risk of Breast Cancer Relapse from Copy Number

Soumya Kundu [1], Jose A. Seoane [2], and Christina Curtis [2]

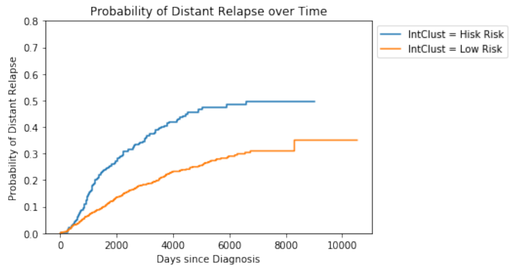
[1] Department of Computer Science, Stanford University, [2] Departments of Medicine and Genetics, Stanford University

Abstract

- The IntClust algorithm clusters breast cancer patients into 11 subtypes using gene expression and copy number from tumor [1]
- ER+/HER2- patients in 2 sets of 4 subtypes are enriched for high and low risk of distant relapse, respectively [2]
- Both gene expression and copy number data are not always available for a given cohort of patients [3]
- We use machine learning to classify patients as having high risk or low risk of distant relapse from just copy number and basic clinical data
- We evaluate the accuracy of our models using actual times for distant relapse in patients



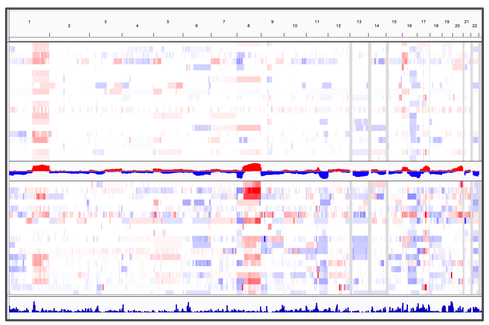
Cumulative Incidence Curves for IntClust subtypes



Cumulative Incidence Curves for IntClust risk labels

Features

- The primary set of features for our model is the mean copy number for different segments of the genome for each patient
- We start with a total of 1,244,072 different copy number values across the genomes of 1285 patients
- We use the iClusterPlus package to obtain the mean copy number of a set of 4794 consensus regions across all patients
- In addition, we use 4 clinical features that are known to influence breast cancer: age, tumor grade, tumor size, and number of tumor-positive lymph nodes

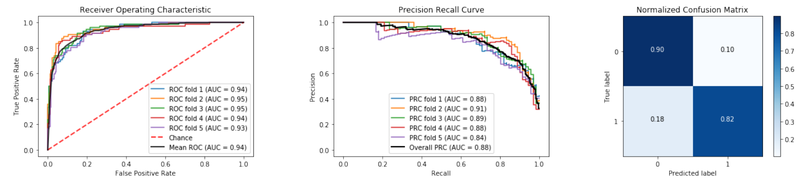


Copy number variation for low risk (top) and high risk (bottom) patients

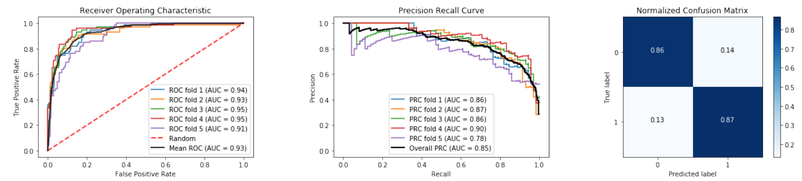
Total Number of Patients	1285
High Risk Patients (IntClust)	360
Patients with Observed Distant Relapse	354
Copy Number Segments	4794

Model Training

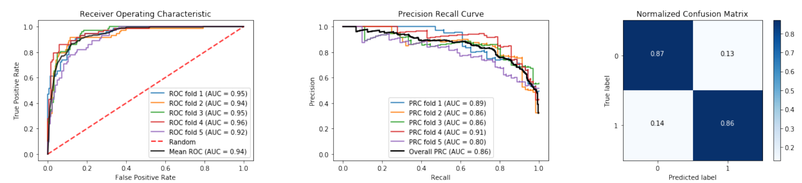
Logistic Regression with L1 Regularization



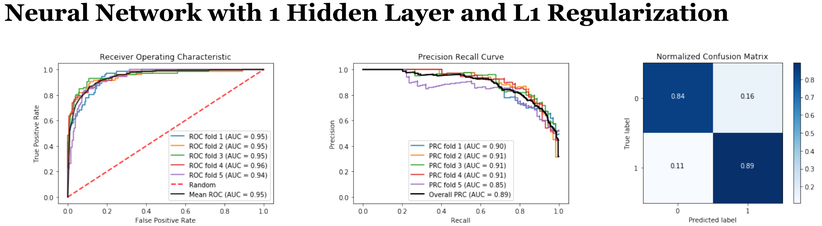
Support Vector Machine with Linear Kernel



Support Vector Machine with Gaussian Kernel



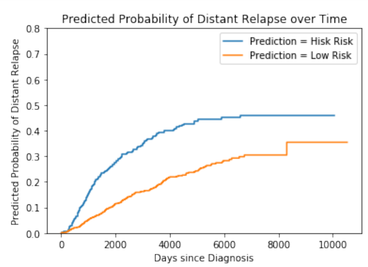
Neural Network with 1 Hidden Layer and L1 Regularization



Log-Rank Test

- Tests for significant difference between survival curves
- Predictions from all models resulted in significant p-values (< 0.005)

Model	$-\log_2$(p-value)
IntClust labels	36.12
Logistic Regression	33.73
SVM with Linear Kernel	36.43
SVM with Gaussian Kernel	37.17
Neural Network	39.88



Cumulative Incidence Curves for Neural Net predictions

Hazard Ratio & C-Index

- The Cox Proportional Hazards Model is a regression model that can predict event times in the context of survival analysis
- We fit this model to the distant relapse times for our patients using the predicted risk labels and clinical data as features
- The hazard ratio for the risk label indicates the relative risk of distant relapse for the high risk patients
- The C-Index for the Cox PH model indicates the model's accuracy in predicting relative event times for a group of patients

Model	Hazard Ratio
IntClust labels	1.64
Logistic Regression	1.68
SVM with Linear Kernel	1.72
SVM with Gaussian Kernel	1.73
Neural Network	1.79

Model	C-Index
IntClust labels	0.6756
Logistic Regression	0.6785
SVM with Linear Kernel	0.6811
SVM with Gaussian Kernel	0.6809
Neural Network	0.6874

Discussion

- We show that logistic regression, SVM, and neural network models can all learn to accurately predict risk of distant relapse in breast cancer patients from just copy number and basic clinical data
- While the neural network model performs best, all of the trained models perform better at predicting risk of distant relapse than the existing IntClust subtyping
- Future work will include evaluating these models on other datasets

References

1. Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, April 2012.
2. Oscar M. Rueda, Stephen-John Sammut, Jose A. Seoane, Suet-Feung Chin, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature, 567(7748):399–404, March 2019.
3. H Raza Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biology, 15(8), August 2014.