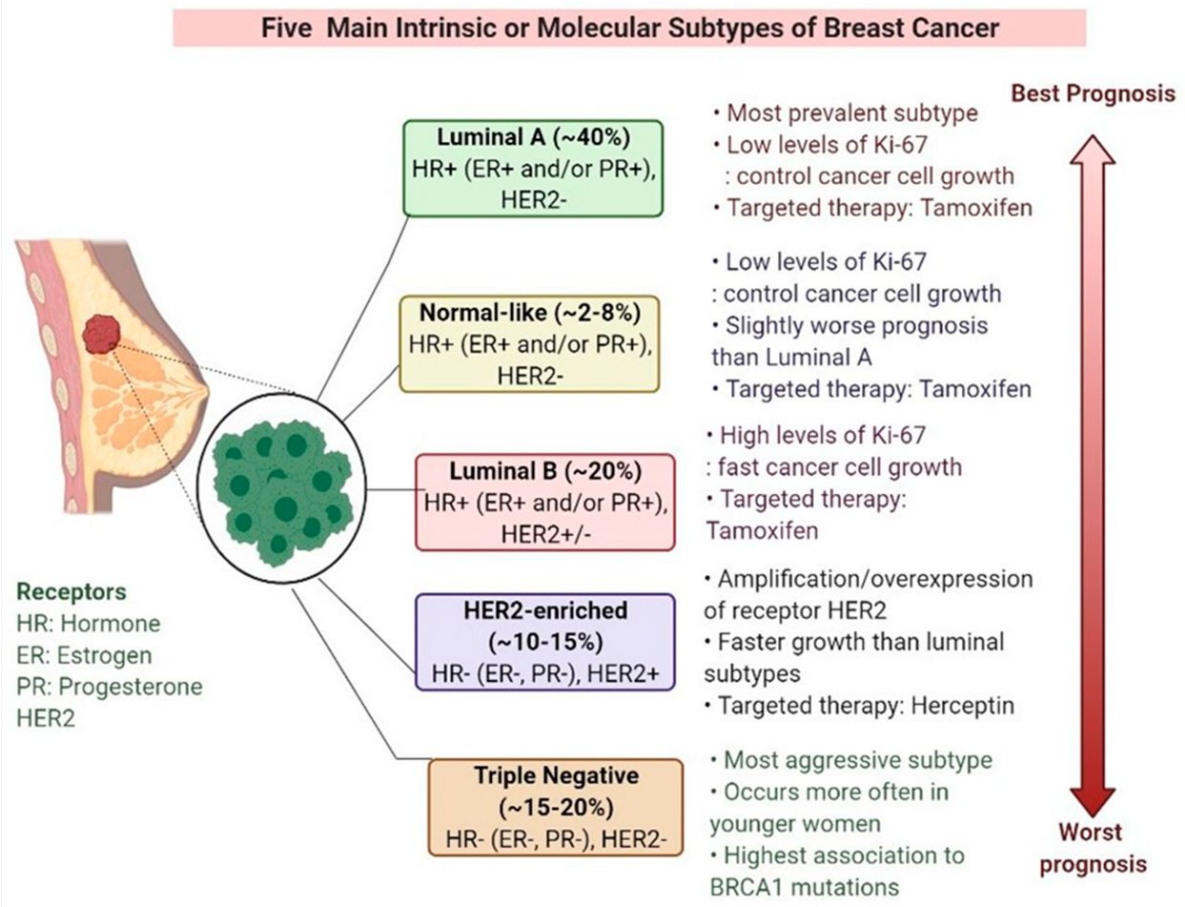
Breast Cancer 5 years Disease Specific Survival.

Group 5
賴世宗
羅偉嘉
龍品瑞
Berke Ugurlu
Tanoj Langore

Introduction of the dataset and research targets

Molecular subtypes of breast cancer



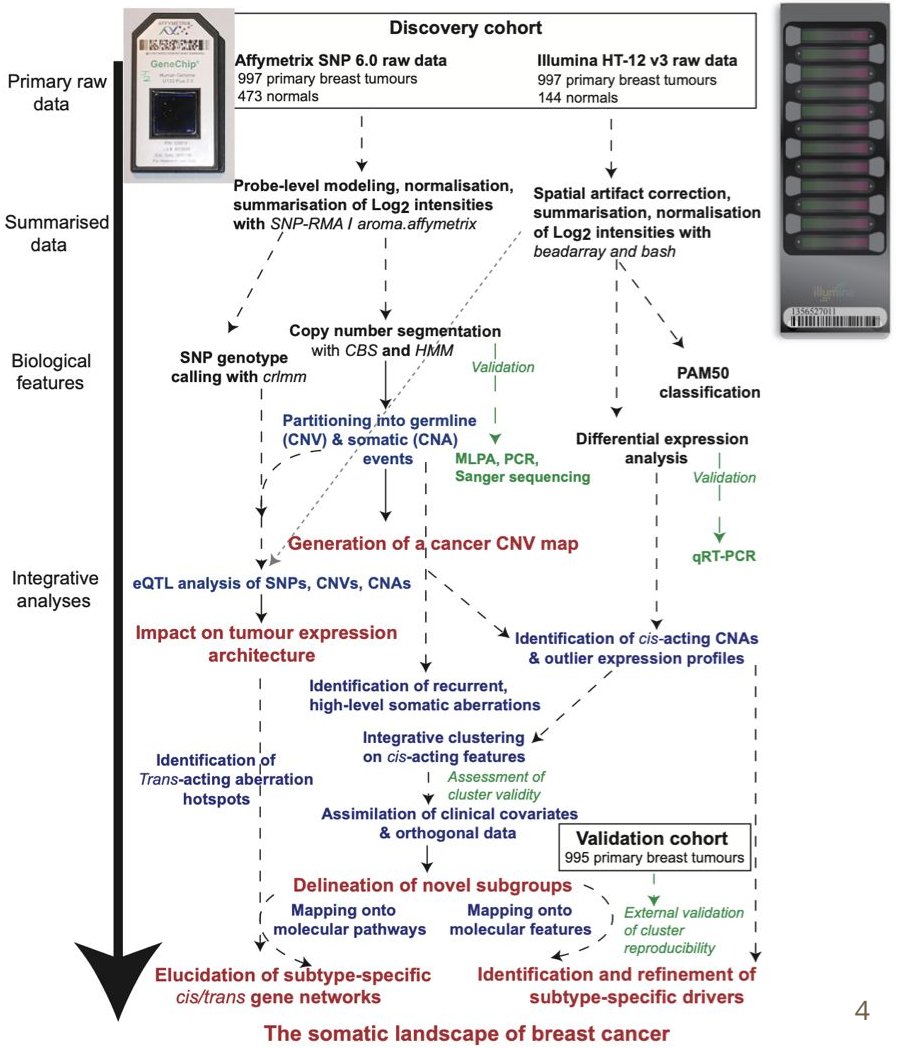
ARTICLE

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Nature **volume 486**, pages346–352 (2012)

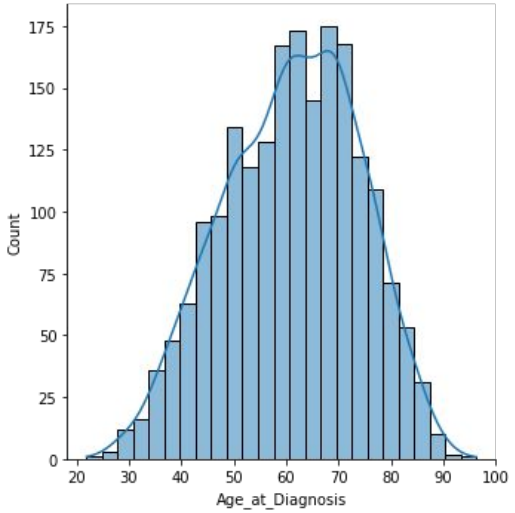
- METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)

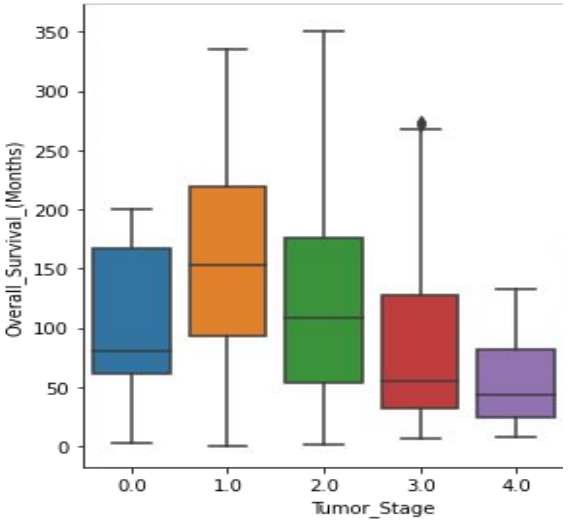
Factor	Discovery set	Validation set
Median (LQ, UQ)		
Age at diagnosis (years)	61.3 (51.15, 70.31)	62.61 (51.84, 70.90)
Follow-up all cases (years)	6.99 (3.88, 11.98)	7.32 (4.35, 11.91)
Follow-up still living (years)	9.98 (5.08, 13.13)	9.62 (5.72, 13.28)
Tumour size	23 (17, 30)	23 (17, 30)
NPI	4.05 (3.05, 5.05)	4.04 (3.04,5.04)
Median (95% CI)		
Survival (months)	149 (131, 161)	151 (142, 168)
Lymph nodes positive		
Number (0, 1, 2, >3)	514, 174, 96, 213	528, 164, 76, 220
Grade		
I	72	98
II	415	360
III	510	447
ER status		
Pos	801	707
Neg	196	244
TP53 status		
Mutated	99	-
Wildtype	721	-
NA	177	-
PAM50 subtype		
Basal	118	213
HER2	87	153
Luminal A	466	255
Luminal B	268	224
Normal	58	144
Not classified	0	6
Cellularity		
High	553	420
Moderate	444	294
Low	-	216

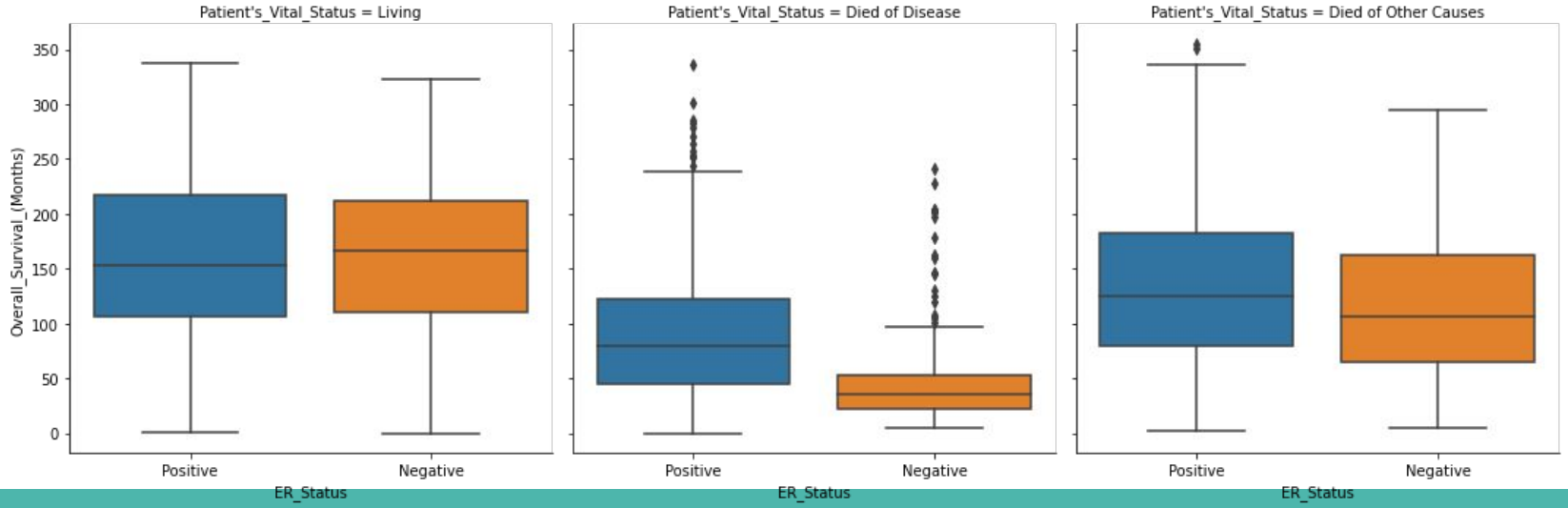


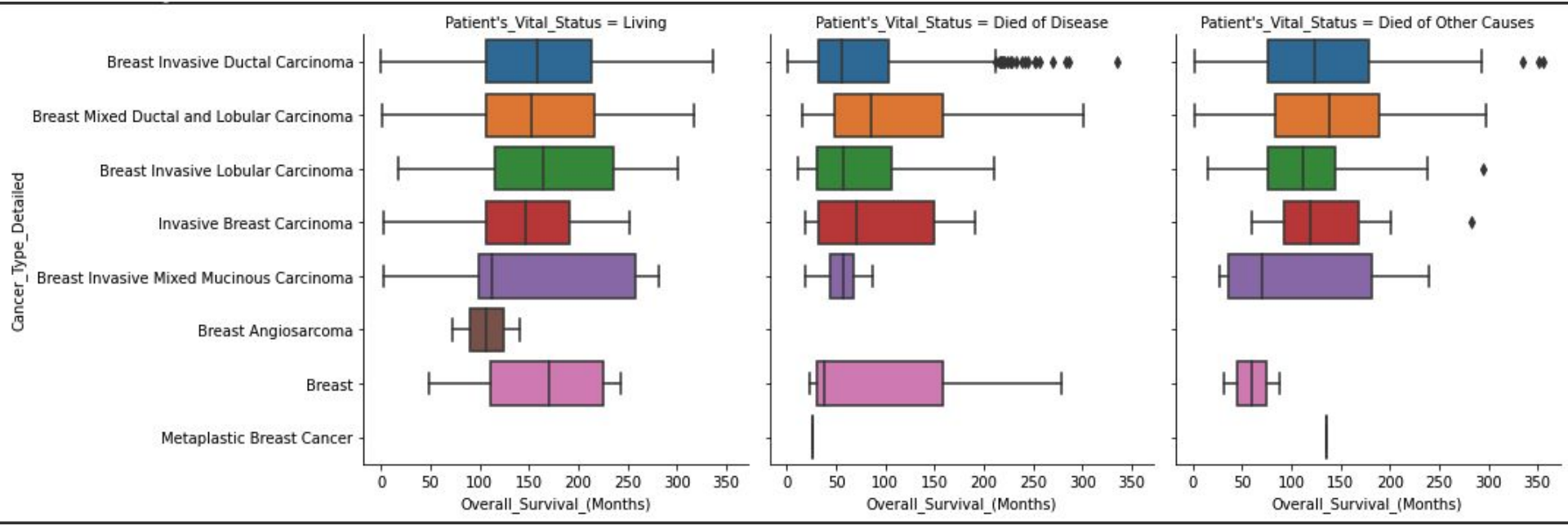
26 dimentional feature vectors were constructed from clinical variables

```
clinical_features = ['Age_at_Diagnosis',
                     'Cohort',
                     'Neoplasm_Histologic_Grade',
                     'Lymph_nodes_examined_positive',
                     'Mutation_Count',
                     'Nottingham_prognostic_index',
                     #'Relapse_Free_Status_(Months)','label_Relapse_Free_Status'
                     'Tumor_Size',
                     'Tumor_Stage',
                     'label_Type_of_Breast_Surgery',
                     'label_Cancer_Type',
                     'label_Cancer_Type_Detailed',
                     'label_Cellularity',
                     'label_Pam50_+_Claudin-low_subtype',
                     'label_Chemotherapy',
                     'label_ER_Status',
                     'label_HER2_status_measured_by_SNP6',
                     'label_HER2_Status',
                     'label_Hormone_Therapy',
                     'label_Inferred_Menopausal_State',
                     'label_Integrative_Cluster',
                     'label_Primary_Tumor_Laterality',
                     'label_Oncotree_Code',
                     'label_PR_Status',
                     'label_Radio_Therapy',
                     'label_Sex',
                     'label_Three_Gene_classifier_subtype']
```

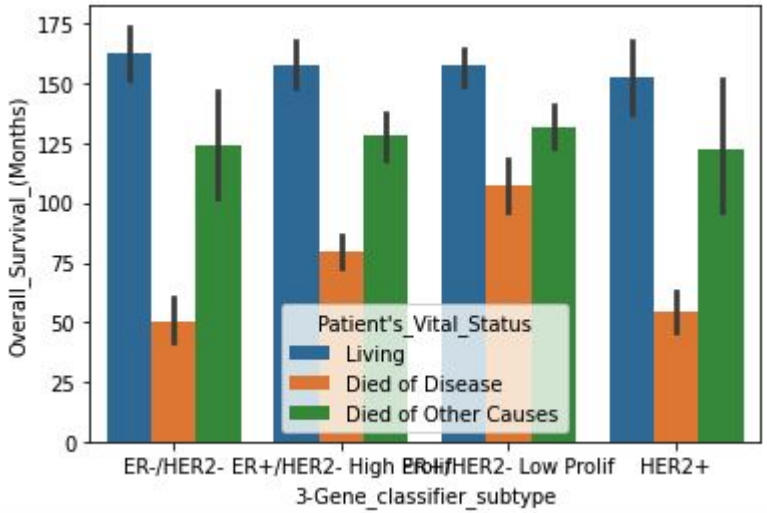

Count
Age_at_Diagnosis


Overall_Survival_(Months)
Tumor_Stage

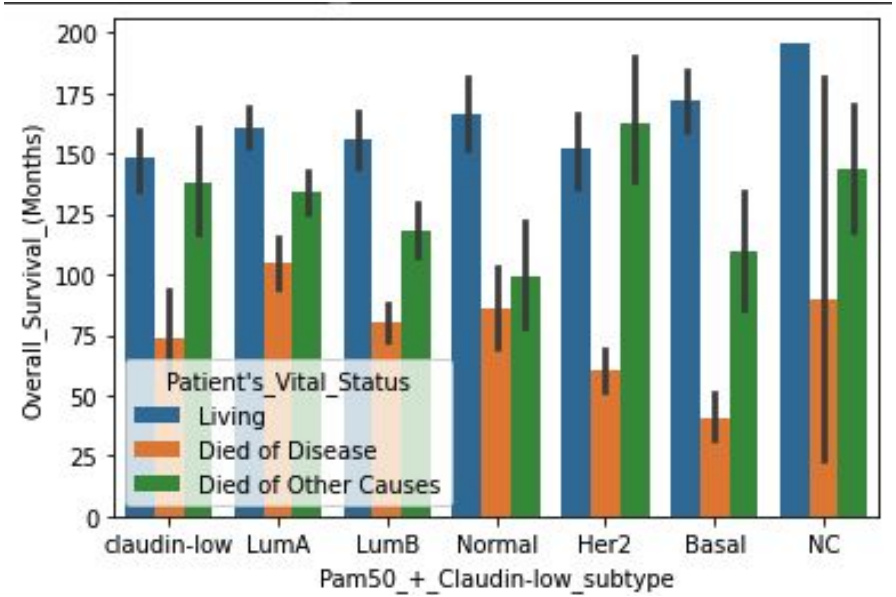

Patient's_Vital_Status = Living
Patient's_Vital_Status = Died of Disease
Patient's_Vital_Status = Died of Other Causes
Overall_Survival_(Months)
Positive
Negative
ER_Status


Patient's_Vital_Status = Living
Patient's_Vital_Status = Died of Disease
Patient's_Vital_Status = Died of Other Causes
Cancer_Type_Detailed
Breast Invasive Ductal Carcinoma
Breast Mixed Ductal and Lobular Carcinoma
Breast Invasive Lobular Carcinoma
Invasive Breast Carcinoma
Breast Invasive Mixed Mucinous Carcinoma
Breast Angiosarcoma
Breast
Metaplastic Breast Cancer
0
50
100
150
200
250
300
350
Overall_Survival_(Months)

3-gene classifier subtype

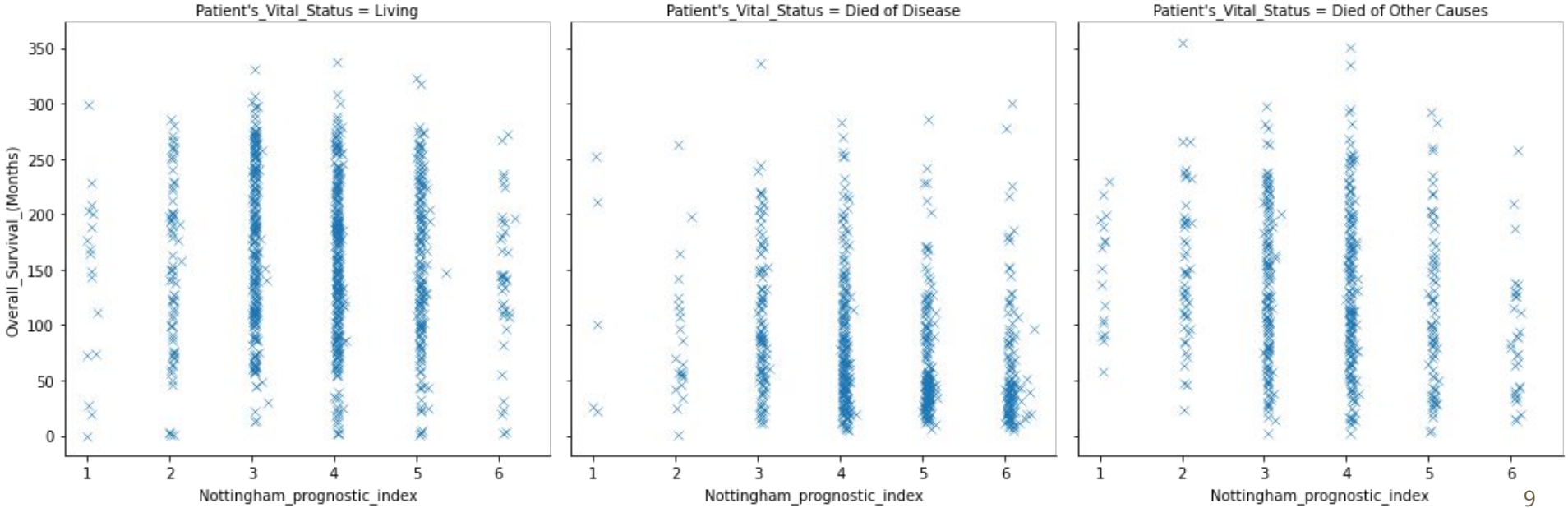


Pam50 + claudin-low



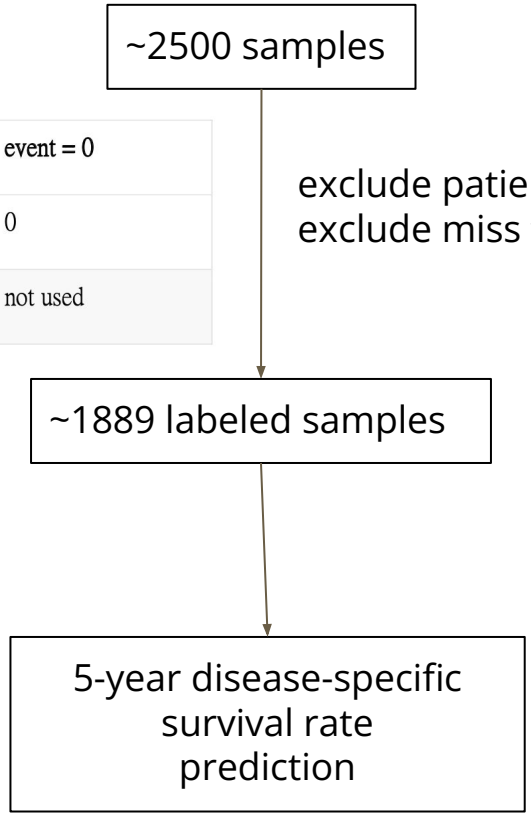
Nottingham prognostic index (NPI)

The Nottingham prognostic index (NPI) is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria: the size of the tumour; the number of involved lymph nodes; and the grade of the tumour.



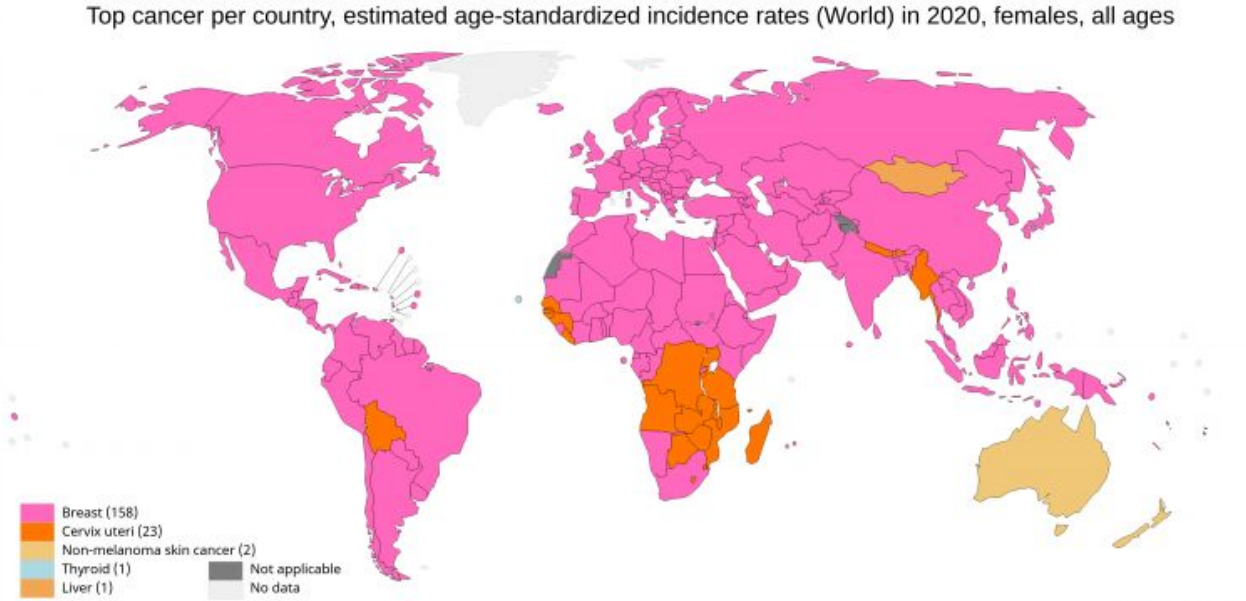
Research target

time \ event	event = 1	event = 0
time >= 60 months	0	0
time < 60 months	1	not used



Motivations

Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics released by the IARC in December 2020.



Motivation

- Breast cancer is a major public health problem in both developing and developed countries.
- The mortality rates of this disease are due to lack of awareness about screening methods and late detection of breast cancer.
- In 2020, there were 2.3 million women diagnosed with breast cancer and 685000 deaths globally.
- van't Veer LJ, Paik S, Hayes DF. Gene expression profiling of breast cancer: a new tumor marker. J Clin Oncol. 2005;23:1631–1635. [PubMed] [Google Scholar]

Motivation

- Breast Cancer: Small amount of research on overall Breast Cancer prognosis biomarkers.
- Immediate challenges in patient management are the determination of prognosis and identification of the most appropriate adjuvant systemic therapy.
- high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients.

Methodology

Methodology

1. Feature Selection by statistical method with medical knowledge
2. Bimodel learning model
    a. lightGBM
    b. SVM
    c. DNN

Cleaning Data

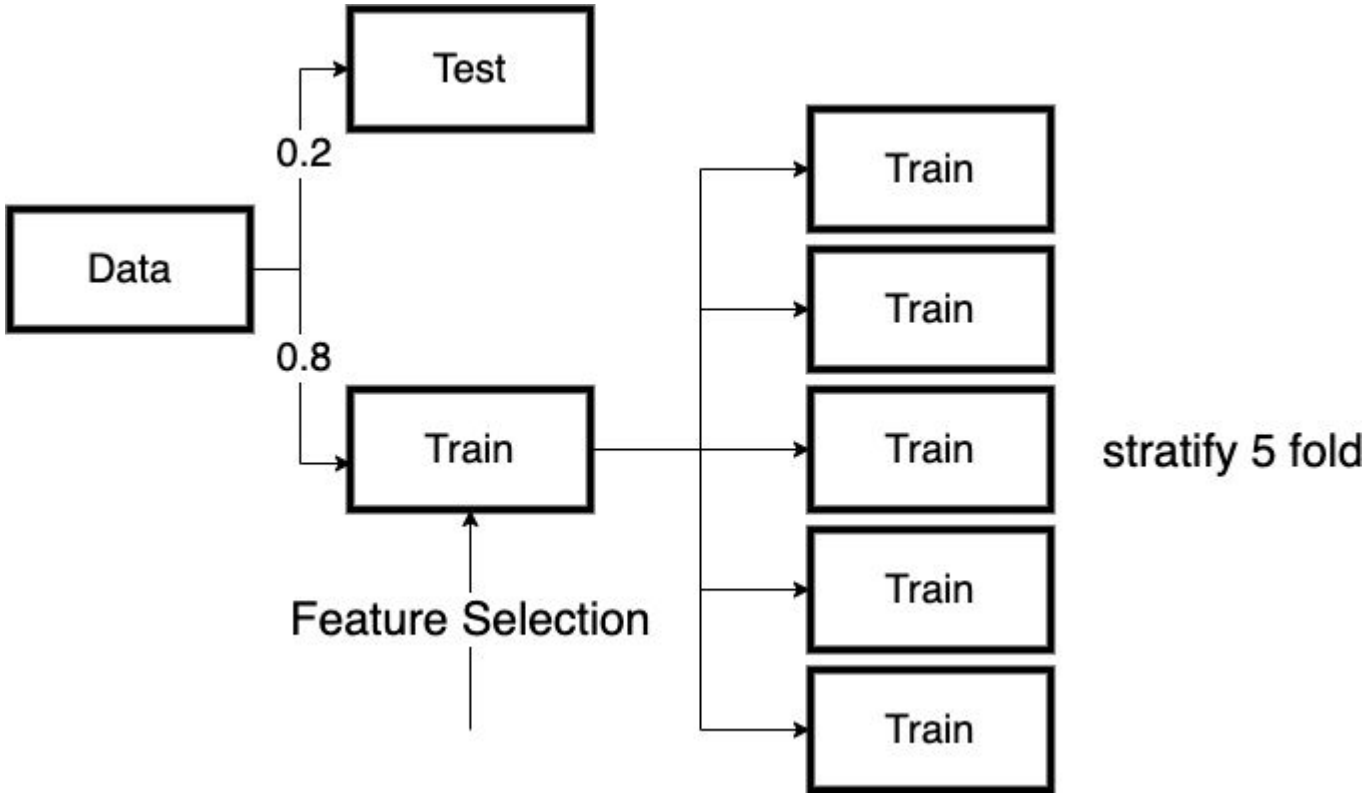
- After merge

Type_of_Breast_Surgery	25
Cellularity	63
ER_status_measured_by_IHC	43
Neoplasm_Histologic_Grade	88
Tumor_Other_Histologic_Subtype	44
Primary_Tumor_Laterality	110
Lymph_nodes_examined_positive	76
Mutation_Count	121
Relapse_Free_Status	1
3-Gene_classifier_subtype	216
Tumor_Size	26
Tumor_Stage	514
Patient's_Vital_Status	1

Cleaning Data

- Do label encoding for categorical features
- Drop poor nan value: Patient's vital status
- Drop duplicate/similar features
  - Tumor_Other_Histologic_Subtype == Cancer_Type_Detailed
  - ER_status_measured_by_IHC == ER_Status

Train Test Split



Bio-markers

- ER
- HER2
- PR
- BRCA1
- BRCA2
- MKI67
- MKI67IP
- PLAU
- PLAUR

- ESR1
- ERBB2
- PGB
- MUC1

Biomarker	Official gene name[a]	Clinical use	Cancer type	Source type
α-fetoprotein (AFP)	*AFP*	Staging	Nonseminomatous testicular	Serum
Human chorionic gonadotropin (hGC)	*CGB*	Staging	Testicular	Serum
Carbohydrate antigen 19-9 (CA19-9)		Monitoring	Pancreatic	Serum
Carbohydrate antigen 125 (CA125)	*MUC16*	Monitoring	Ovarian	Serum
Carcinoembryonic antigen (CEA)	*PSG2*	Monitoring	Colorectal	Tissue
Epidermal growth factor receptor (EGFR)	*EGFR*	Prediction	Colorectal	Tissue
v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (KIT)	*KIT*	Prediction	Gastrointestinal	Tissue
Thyroglobulin	*TG*	Monitoring	Thyroid	Serum
Prostate specific antigen (PSA)	*KLK3*	Screening and monitoring	Prostate	Serum
Carbohydrate antigen 15.3 (CA 15.3)	*MUC1*	Monitoring	Breast	Serum
Carbohydrate antigen 27.29 (CA27.29)	*MUC1*	Monitoring	Breast	Serum
Estrogen receptor (ER)	*ESR1*	Prognosis and prediction	Breast	Tissue
Progesterone receptor (PR)	*PGR*	Prognosis and prediction	Breast	Tissue
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (HER2-neu)	*ERBB2*	Prognosis and prediction	Breast	Tissue

https://www.frontiersin.org/articles/10.3389/fphar.2020.632079/full#h4

https://www.researchgate.net/figure/FDA-cleared-protein-cancer-biomarkers_tbl1_231225439

Feature Selection

Categorical features

- **Welch's t-test** which is popular statistical techniques used to test whether mean difference between two groups is statistically significant. useful when the there is a difference between the variations of two populations and also when their sample sizes are unequal.

- applied on ER, HER2, PR

Feature Selection

Numerical Features

- **Pearson's correlation coefficient** is the test statistics that measures the statistical relationship, or association, between two continuous variables.
- applied on MKI67, MKI67IP, BRCA1, BRCA2, PLAU, PLAUR, ESR1, PGR, ERBB2, MUC1,

Degree of correlation:

- **Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
- **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
- **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
- **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
- **No correlation:** When the value is zero.

```
BRCA2
[0.9999999999999999, 0.36551088320422453, 0.35775814641756265, 0.35085940672957316, 0.3
['BRCA2', 'PPAPDC1A', 'MFAP2', 'COL11A1', 'COPZ2', 'COL5A2', 'NID2', 'GJB2', 'PLAU', 'G
BRCA1
[0.9999999999999999, 0.5177260092419608, 0.5082786669984971, 0.4888362392627947, 0.4739
['BRCA1', 'C17orf53', 'TUBG1', 'TIMELESS', 'TOP2A', 'NUSAP1', 'PSMC3IP', 'C16orf59', 'A
MKI67
[1.0, 0.7684722242627228, 0.7439805410414201, 0.7427070736012766, 0.7378716734364885, 0
['MKI67', 'FOXM1', 'HJURP', 'AURKB', 'KIF20A', 'KIFC1', 'CEP55', 'TPX2', 'TROAP', 'KIF2
MKI67IP
[1.0, 0.6679072093800831, 0.6517132096647617, 0.6269926439712363, 0.6142224221948241, 0
['MKI67IP', 'SSB', 'ZC3H15', 'MMADHC', 'NUP35', 'WDR12', 'C2orf47', 'RPS27A', 'METTL5',
PLAU
[1.0, 0.7653317115953178, 0.7344702494487141, 0.7320108890508251, 0.7205804853897141, 0
['PLAU', 'ANTXR1', 'COL11A1', 'KIAA1199', 'SULF1', 'TGFBI', 'ITGA11', 'COL12A1', 'GJB2'
PLAUR
[0.9999999999999999, 0.6952061777327959, 0.6887836153452497, 0.6735154849518779, 0.6725
['PLAUR', 'CTSL1', 'SPP1', 'PLAU', 'DSE', 'FCGR1B', 'LOX', 'CD68', 'SERPINB8', 'GLIPR1'
ESR1
[0.9999999999999998, -0.8753213031020667, 0.8300076207640039, 0.8101626647966366, 0.789
['ESR1', 'label_ER_Status', 'GATA3', 'AGR3', 'CA12', 'TBC1D9', 'U79293', 'C6orf97', 'MY
MUC1
[1.0, 0.5988593982829881, 0.5984205298675195, 0.590649779875871, 0.5885040340015464, 0.
['MUC1', 'SELENBP1', 'REEP6', 'TJP3', 'SLC44A4', 'RORC', 'CREB3L4', 'MLPH', 'CFB', 'CAP
PGR
[1.0, 0.7741674387322155, 0.5731052597095384, 0.5601904223836532, 0.5521690051413889, 0
['PGR', 'label_PR_Status', 'SUSD3', 'MAPT', 'SERPINA11', 'CASC1', 'TMEM26', 'CHST8', 'A
ERBB2
[1.0, 0.8684552720961477, 0.8489522440664496, 0.7977138896327706, 0.7719228812119098, 0
['ERBB2', 'GRB7', 'PGAP3', 'label_HER2_Status', 'STARD3', 'C17orf37', 'ORMDL3', 'GSDMB'
```

Models

- LightGBM: A Highly Efficient Gradient Boosting Decision Tree

	CatBoost	LightGBM	XGBoost
Developer	Yandex	Microsoft	DMLC
Release Year	2017	2016	2014
Tree Symmetry	Symmetric	Asymmetric Leaf-wise tree growth	Asymmetric Level-wise tree growth
Splitting Method	Greedy method	Gradient-based One-Side Sampling (GOSS)	Pre-sorted and histogram-based algorithm
Type of Boosting	Ordered	-	-
Numerical Columns	Support	Support	Support
Categorical Columns	Support Perform one-hot encoding (default) Transforming categorical to numerical columns by border, bucket, binarized target mean value, counter methods available	Support, but must use numerical columns Can interpret ordinal category	Supports, but must use numerical columns Cannot interpret ordinal category, users must convert to one-hot encoding, label encoding or mean encoding
Text Columns	Support Support Bag-of-Words, Naïve-Bayes or BM-25 to calculate numerical features from text data	Do not support	Do not support
Missing values	Handle missing value Interpret as NaN (default) Possible to interpret as error, or processed as minimum or maximum values	Handle missing value Interpret as NaN (default) or zero Assign missing values to side that reduces loss the most in each split	Handle missing value Interpret as NaN (tree booster) or zero (linear booster) Assign missing values to side that reduces loss the most in each split

Advantages of LightGBM

- **Faster training speed and higher efficiency:** Light GBM use histogram based algorithm i.e it buckets continuous feature values into discrete bins which fasten the training procedure.
- **Lower memory usage:** Replaces continuous values to discrete bins which result in lower memory usage.
- **Better accuracy than any other boosting algorithm**: It produces much more complex trees by following leaf wise split approach rather than a level-wise approach which is the main factor in achieving higher accuracy. However, it can sometimes lead to overfitting which can be avoided by setting the max_depth parameter.
- **Compatibility with Large Datasets**: It is capable of performing equally good with large datasets with a significant reduction in training time as compared to XGBOOST.

LightGBM

- loss function: binary logloss
- Parameters
- 'learning_rate' : 0.1,
- 'n_estimators' : 1024,
- 'num_leaves' : 32,
- 'max_depth' : -1
- early_stopping_rounds=10
-

Experiment Result

lightGBM		valid_ACC	valid_AUC	test_ACC	test_AUC
Gene	top 5	0.816	0.720	0.818	0.674
	top 15	0.818	0.721	0.822	0.700
	top 50	0.822	0.710	0.824	0.695
	all	0.821	0.724	0.822	0.685
Clinical	all	0.837	0.795	0.827	0.705
Bi	top 5	0.832	0.784	0.819	0.715
	top 15	0.822	0.779	0.819	0.730
	top 50	0.839	0.776	0.824	0.719
	all	0.827	0.78	0.822	0.703
Ensemble	top 5			0.830	0.745
	top 15			0.828	0.766
	top 50			0.825	0.768
	all			0.827	0.740

SVM, DNN

Genetic Data (picked 15 genes per biomarker) results

- SVM :
  - Gene Expression
    - ACC 0.728
    - AUC 0.587
- DNN:
  - Gene Expression
    - ACC 0.790
    - AUC 0.564
- Outperform by LightGBM

Methodology

- Two-stage workflow model

  stage 1: survival classification

  stage 2: survival month regression

- Or **simply categorial classification**

- **Multi-model ensembling**

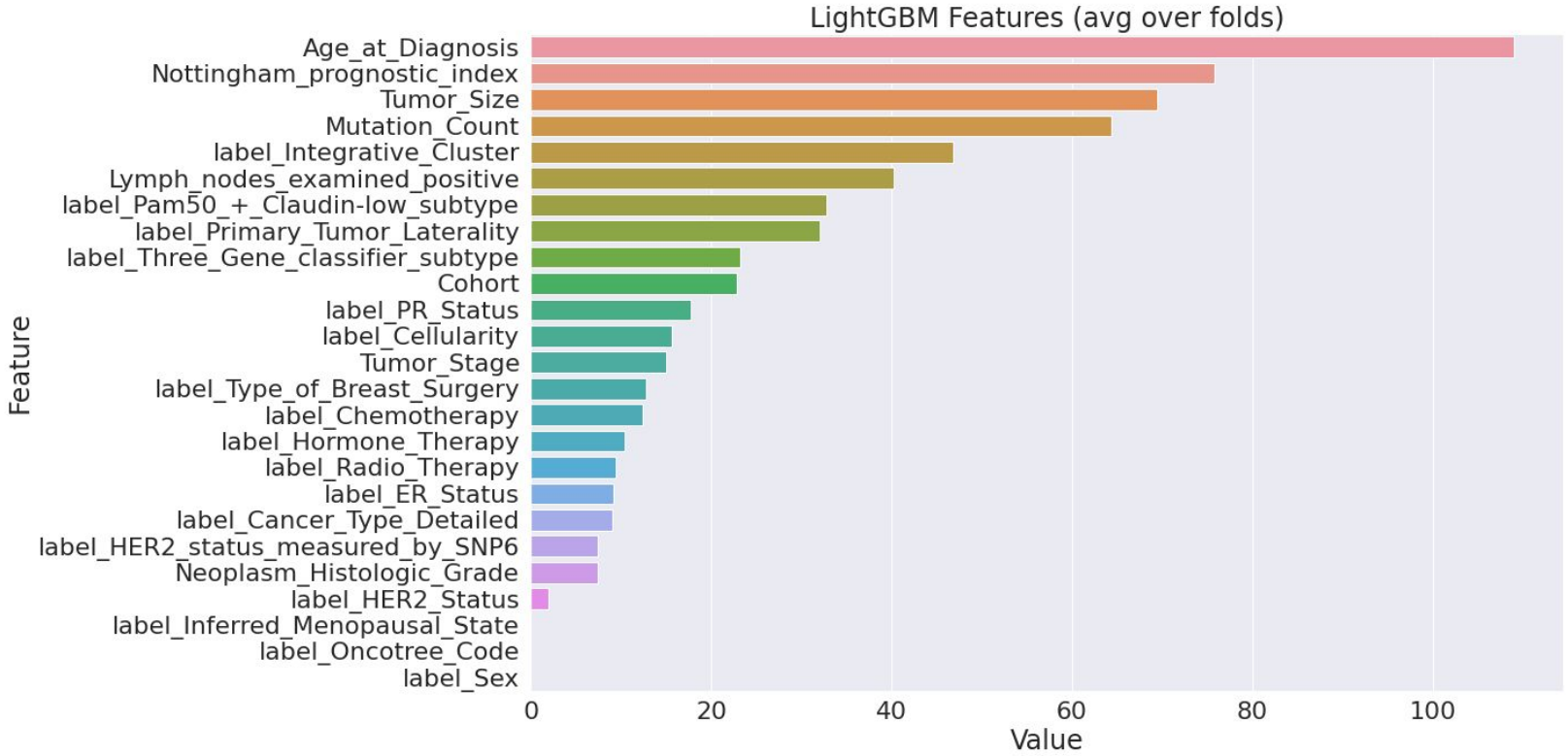
  model 1: Genetic Data

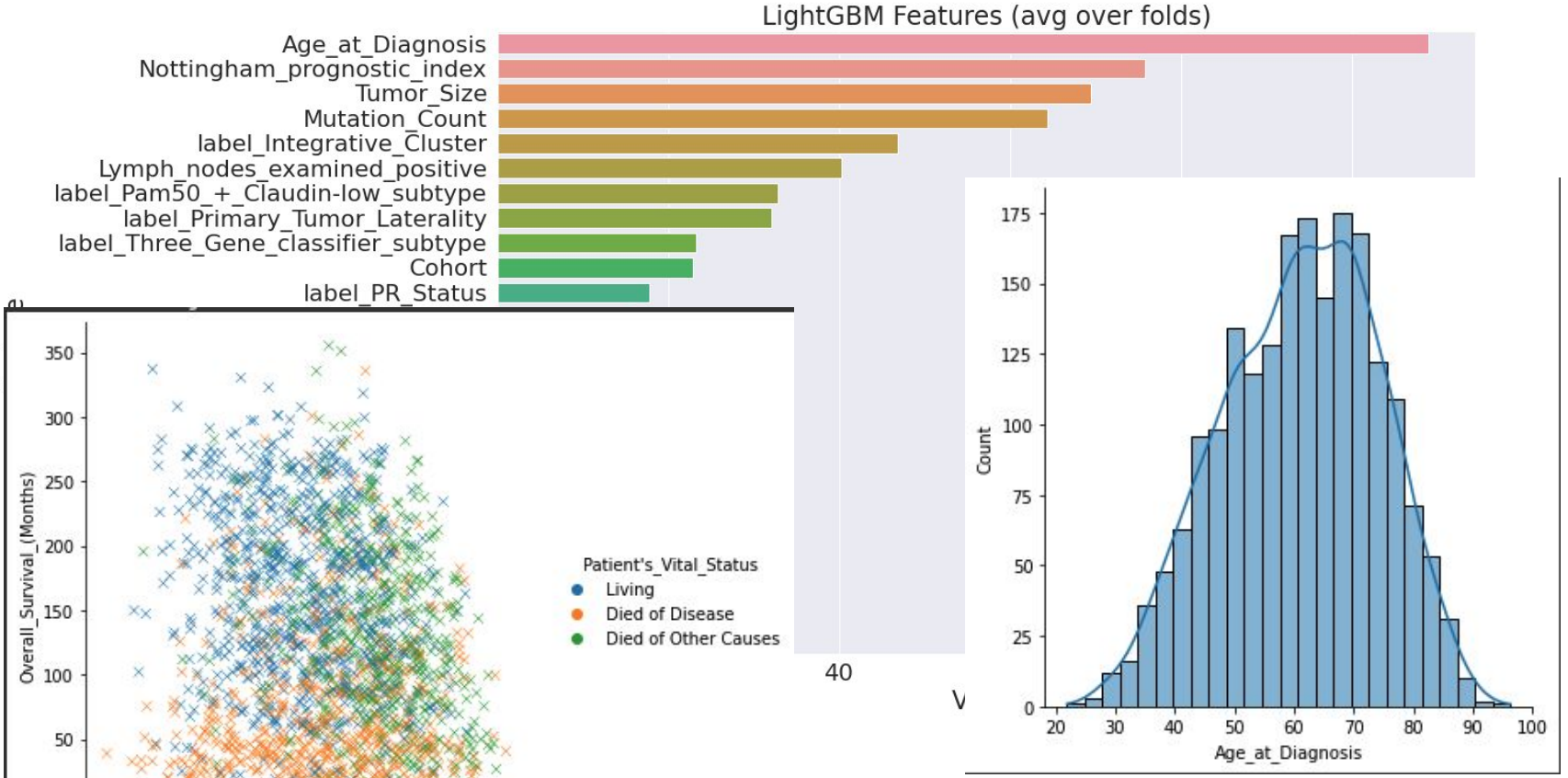
  model 2: Clinical Data

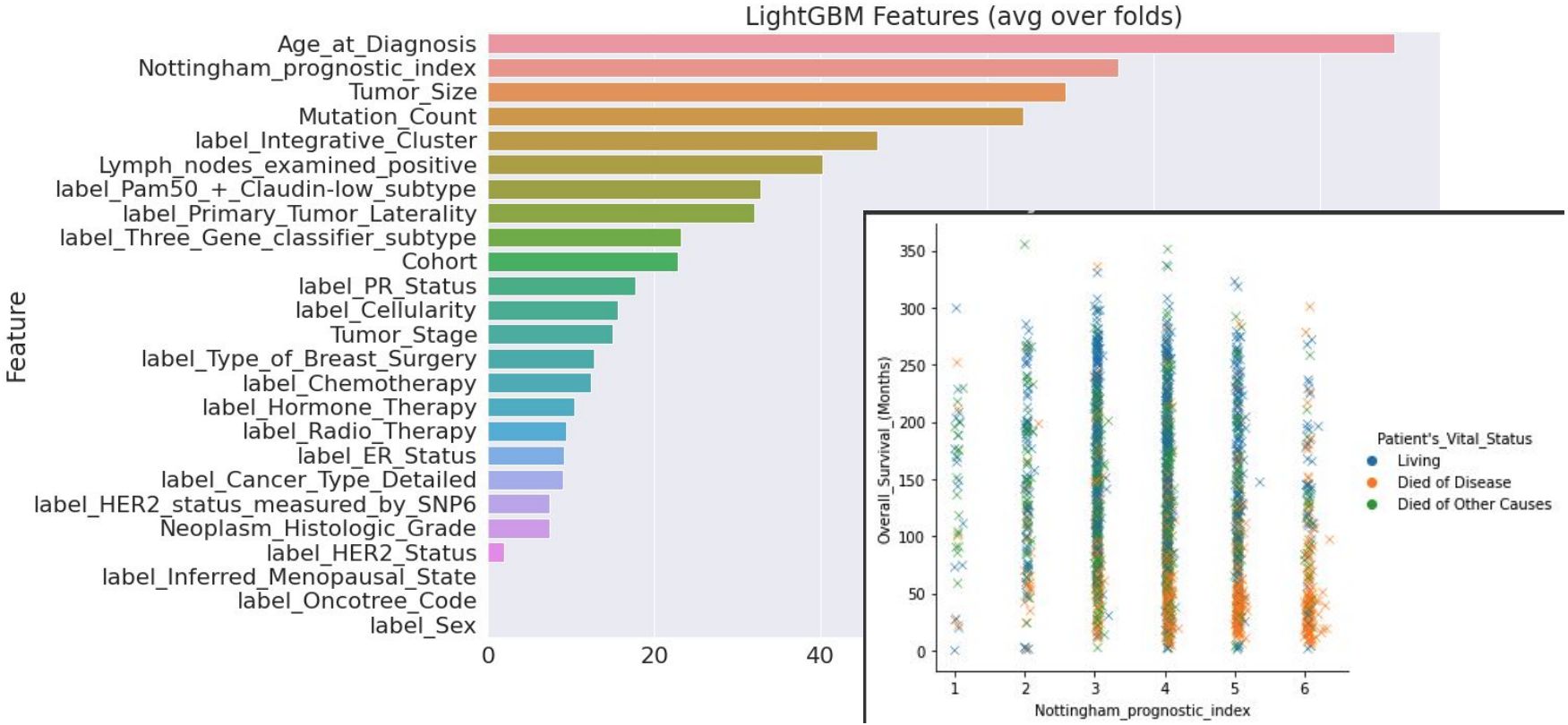
  model 3: Genetic Data + Clinical Data

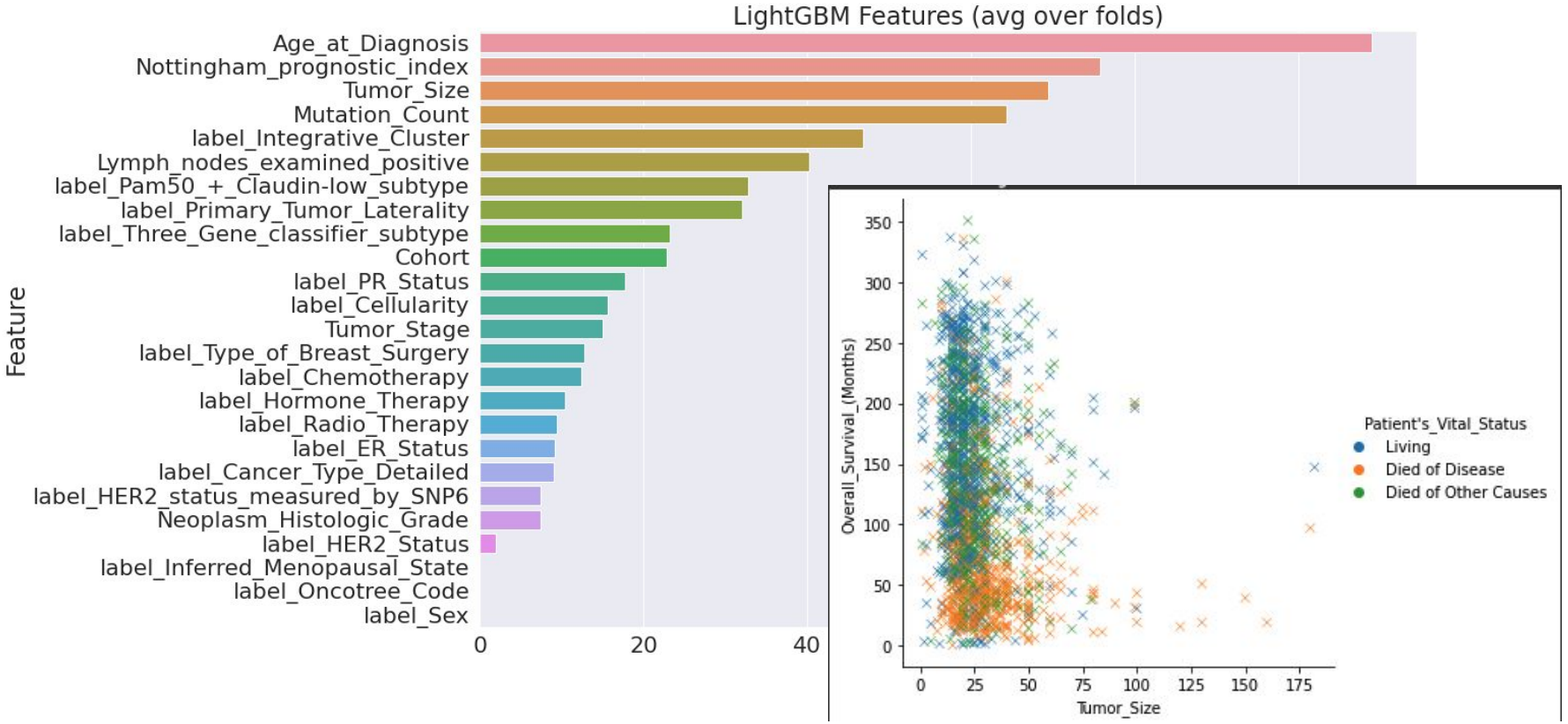
- Semi-Supervised learning

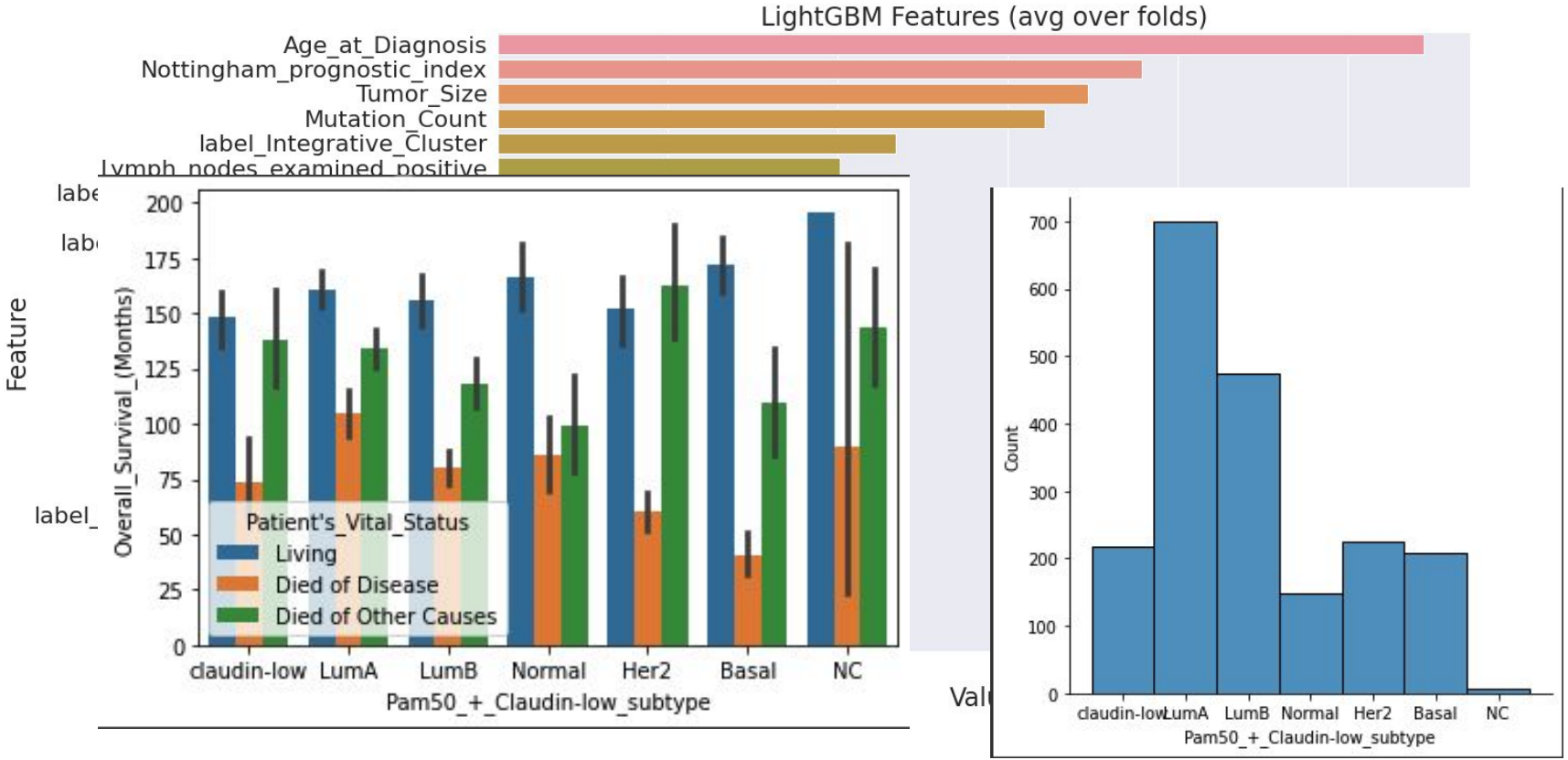
Clinical Feature Important

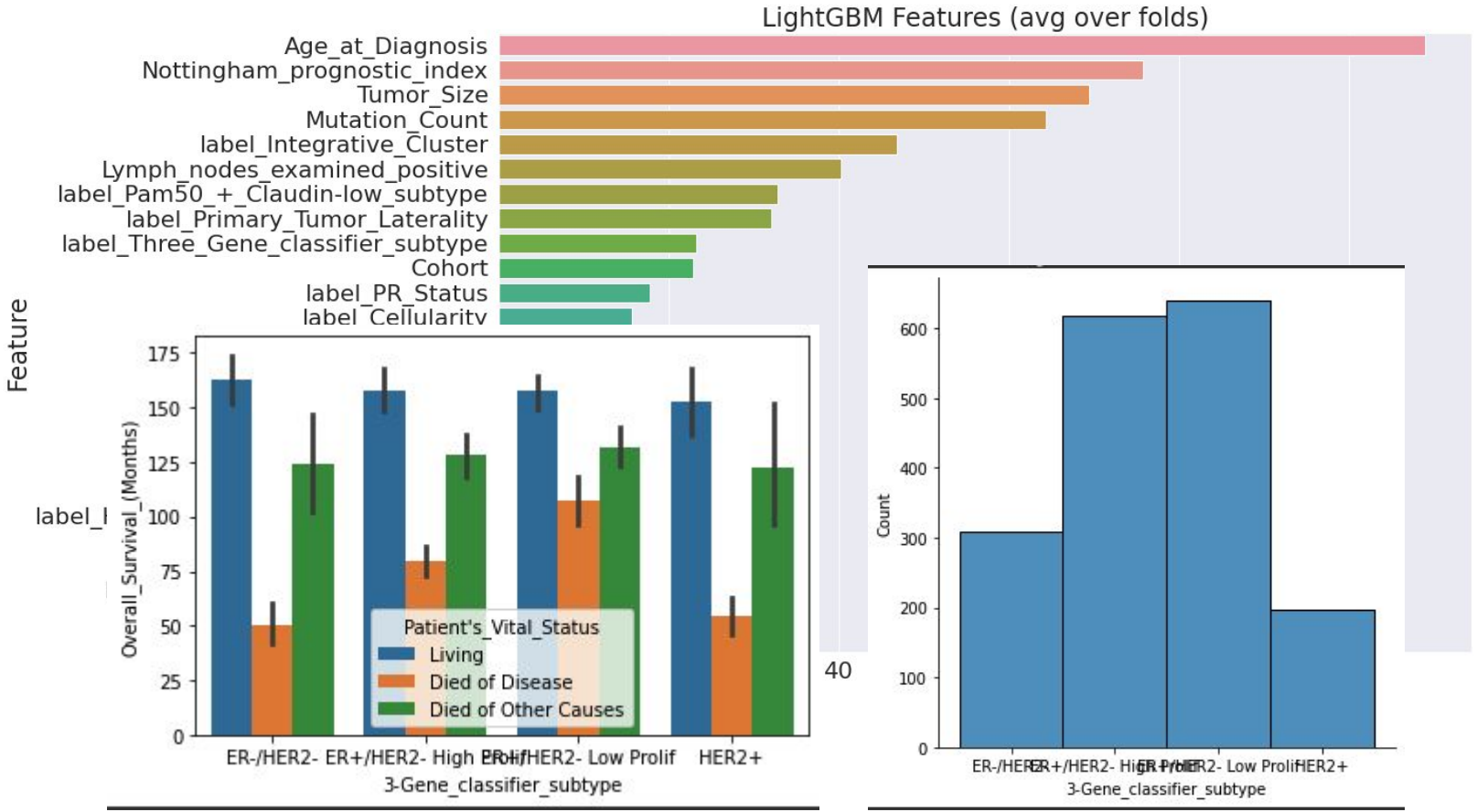



LightGBM Features (avg over folds)
Age_at_Diagnosis
Nottingham_prognostic_index
Tumor_Size
Mutation_Count
label_Integrative_Cluster
Lymph_nodes_examined_positive
label_Pam50_+_Claudin-low_subtype
label_Primary_Tumor_Laterality
label_Three_Gene_classifier_subtype
Cohort
label_PR_Status
Overall_Survival_(Months)
Patient's_Vital_Status
Living
Died of Disease
Died of Other Causes
Count
Age_at_Diagnosis

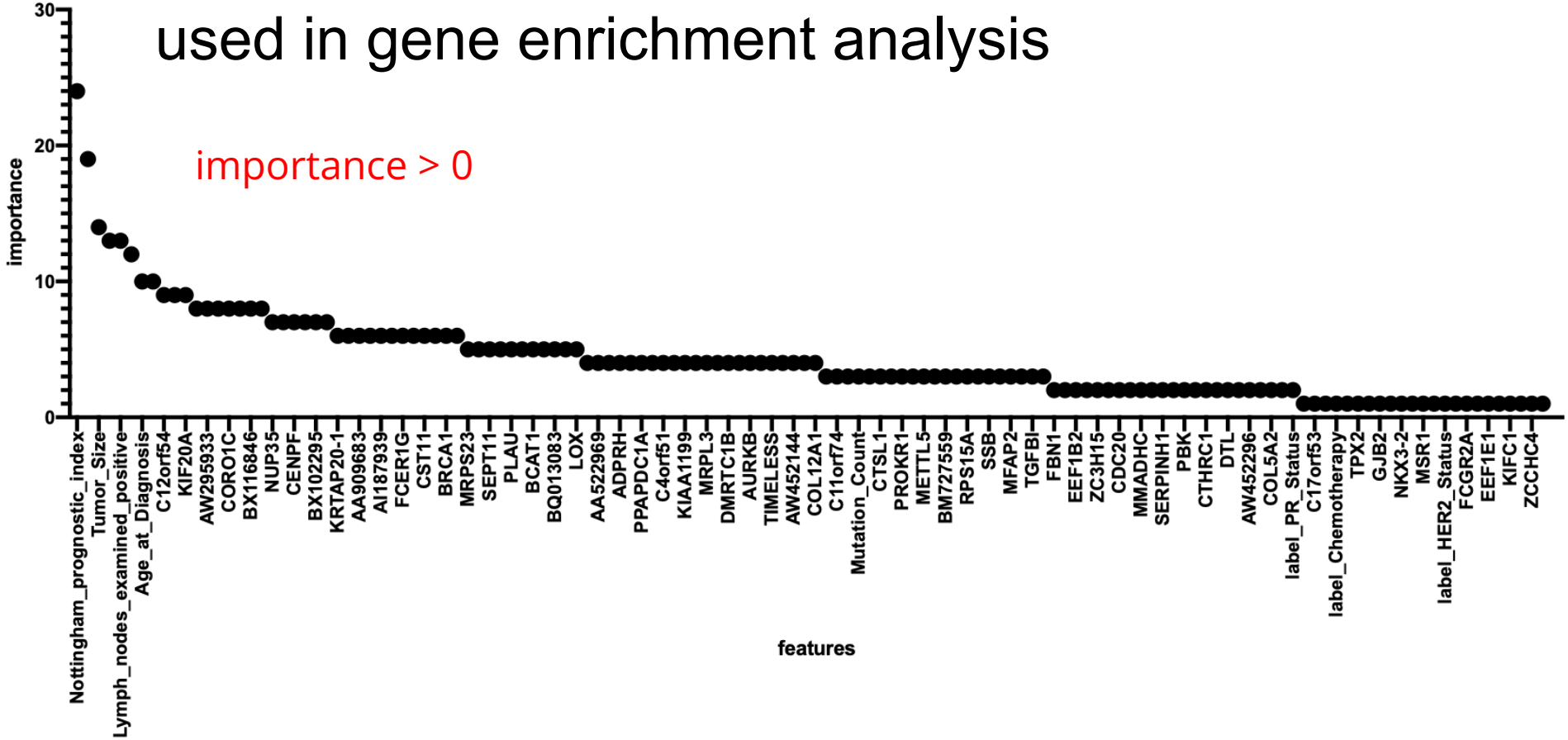

LightGBM Features (avg over folds)
Feature
Age_at_Diagnosis
Nottingham_prognostic_index
Tumor_Size
Mutation_Count
label_Integrative_Cluster
Lymph_nodes_examined_positive
label_Pam50_+_Claudin-low_subtype
label_Primary_Tumor_Laterality
label_Three_Gene_classifier_subtype
Cohort
label_PR_Status
label_Cellularity
Tumor_Stage
label_Type_of_Breast_Surgery
label_Chemotherapy
label_Hormone_Therapy
label_Radio_Therapy
label_ER_Status
label_Cancer_Type_Detailed
label_HER2_status_measured_by_SNP6
Neoplasm_Histologic_Grade
label_HER2_Status
label_Inferred_Menopausal_State
label_Oncotree_Code
label_Sex
0
20
40
Overall_Survival_(Months)
Nottingham_prognostic_index
Patient's_Vital_Status
Living
Died of Disease
Died of Other Causes


LightGBM Features (avg over folds)
Feature
Age_at_Diagnosis
Nottingham_prognostic_index
Tumor_Size
Mutation_Count
label_Integrative_Cluster
Lymph_nodes_examined_positive
label_Pam50_+_Claudin-low_subtype
label_Primary_Tumor_Laterality
label_Three_Gene_classifier_subtype
Cohort
label_PR_Status
label_Cellularity
Tumor_Stage
label_Type_of_Breast_Surgery
label_Chemotherapy
label_Hormone_Therapy
label_Radio_Therapy
label_ER_Status
label_Cancer_Type_Detailed
label_HER2_status_measured_by_SNP6
Neoplasm_Histologic_Grade
label_HER2_Status
label_Inferred_Menopausal_State
label_Oncotree_Code
label_Sex
0
20
40
Overall_Survival_(Months)
Tumor_Size
Patient's_Vital_Status
Living
Died of Disease
Died of Other Causes

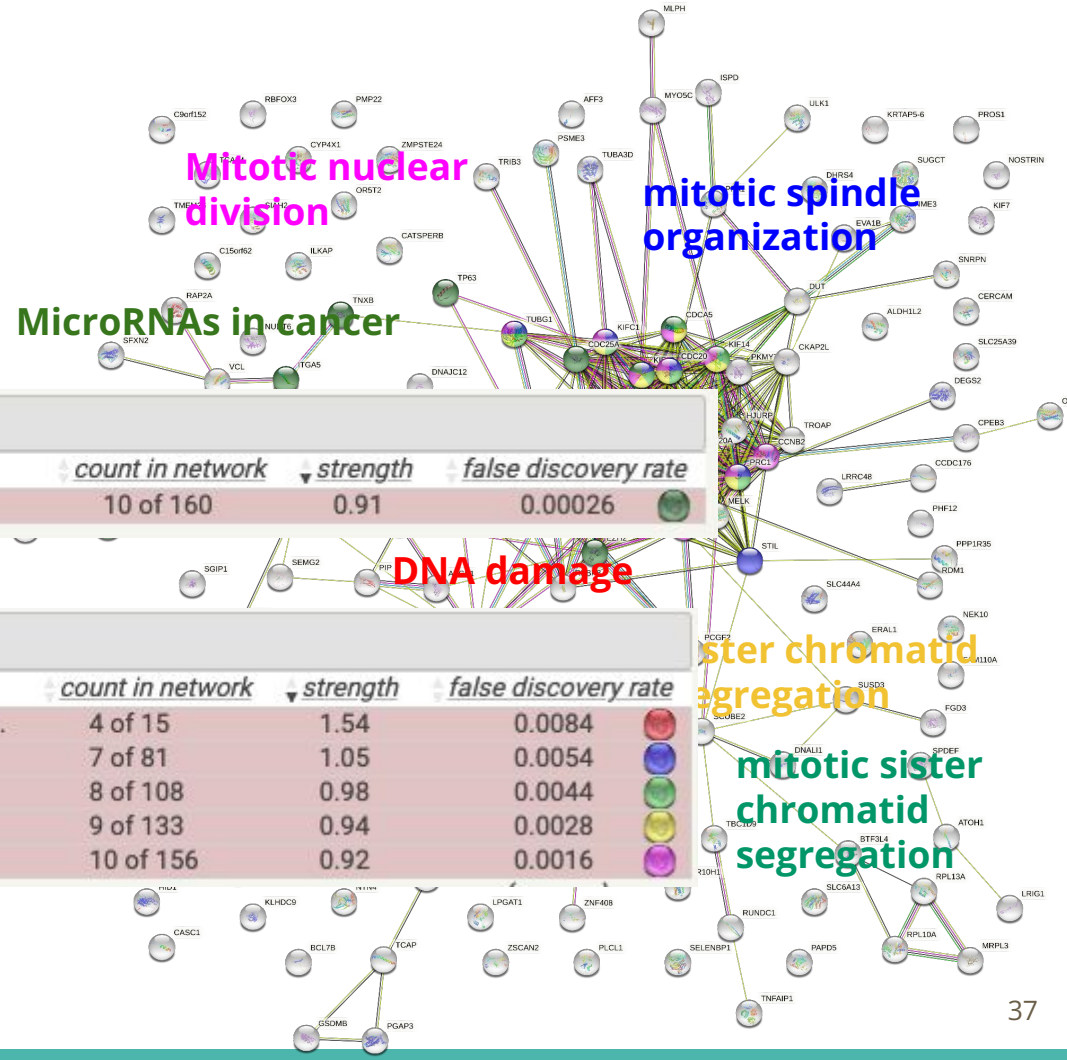

LightGBM Features (avg over folds)
Age_at_Diagnosis
Nottingham_prognostic_index
Tumor_Size
Mutation_Count
label_Integrative_Cluster
Lymph_nodes_examined_positive
Feature
Overall_Survival_(Months)
Patient's_Vital_Status
Living
Died of Disease
Died of Other Causes
claudin-low
LumA
LumB
Normal
Her2
Basal
NC
Pam50_+_Claudin-low_subtype
Count


LightGBM Features (avg over folds)
Age_at_Diagnosis
Nottingham_prognostic_index
Tumor_Size
Mutation_Count
label_Integrative_Cluster
Lymph_nodes_examined_positive
label_Pam50_+_Claudin-low_subtype
label_Primary_Tumor_Laterality
label_Three_Gene_classifier_subtype
Cohort
label_PR_Status
label_Cellularity
Feature
Overall_Survival_(Months)
Patient's_Vital_Status
Living
Died of Disease
Died of Other Causes
ER-/HER2-
ER+/HER2- High Prolif
HER2+
3-Gene_classifier_subtype
Count
ER+/HER2- Low Prolif

Features with importance greater than zero were used in gene enrichment analysis



50 genes per marker



KEGG Pathways

description	count in network	strength	false discovery rate
MicroRNAs in cancer	10 of 160	0.91	0.00026

Biological Process (Gene Ontology)

description	count in network	strength	false discovery rate
DNA damage response, signal transduction by p53 class m...	4 of 15	1.54	0.0084
Mitotic spindle organization	7 of 81	1.05	0.0054
Mitotic sister chromatid segregation	8 of 108	0.98	0.0044
Sister chromatid segregation	9 of 133	0.94	0.0028
Mitotic nuclear division	10 of 156	0.92	0.0016

15 genes per marker

sporadic breast cancer

SUMOylation

Nuclear division

Cell cycle

Biological Process (Gene Ontology)

description	count in network	strength	false discovery rate
Nuclear division	10 of 291	0.83	0.0366

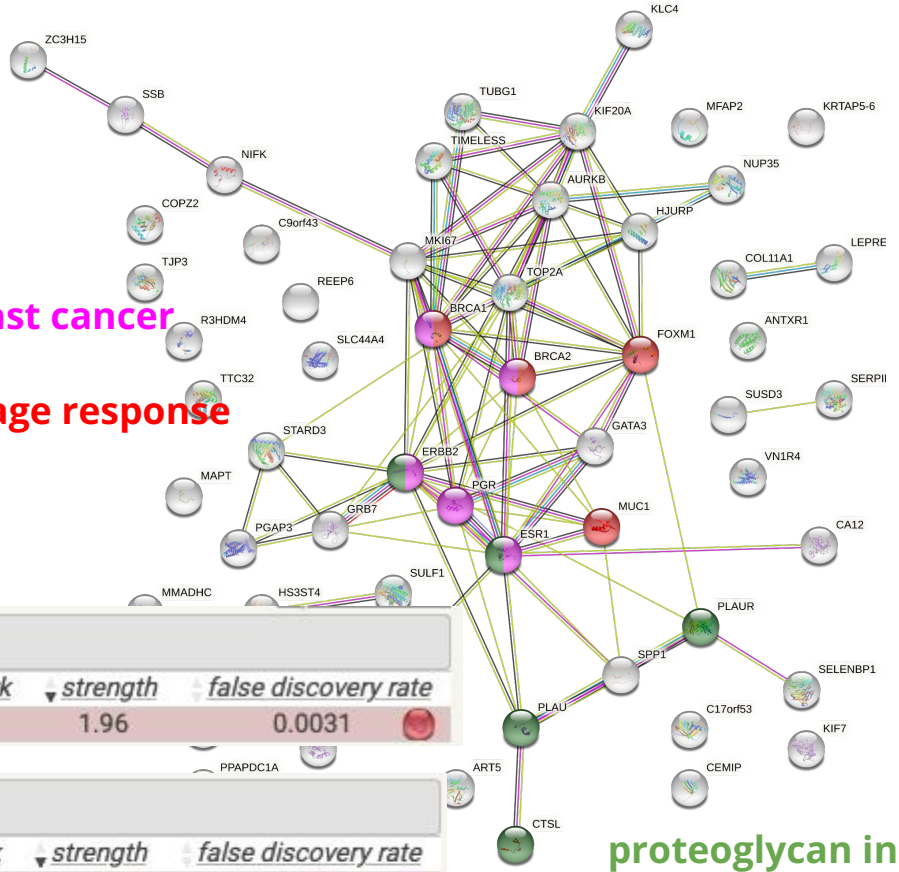
Reactome Pathways

description	count in network	strength	false discovery rate
SUMO E3 ligases SUMOylate target proteins	7 of 166	0.92	0.0296
Cell Cycle	14 of 647	0.63	0.0121

Disease-gene associations (DISEASES)

description	count in network	strength	false discovery rate
Sporadic breast cancer	3 of 3	2.3	0.0107
Male breast cancer	3 of 4	2.17	0.0107
Breast carcinoma	4 of 28	1.45	0.0284

5 genes per marker



Breast cancer

DNA damage response

proteoglycan in cancer

Biological Process (Gene Ontology)

description	count in network	strength	false discovery rate
DNA damage response, signal transduction by p53 class m...	4 of 15	1.96	0.0031

KEGG Pathways

description	count in network	strength	false discovery rate
Breast cancer	5 of 145	1.07	0.0249
Proteoglycans in cancer	5 of 196	0.94	0.0490

Surveyed Literature

1. **Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction**

   Li-Hsin Cheng , Te-Cheng Hsu & Che Lin

   (https://doi.org/10.1038/s41598-021-92864-y)

2. **Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data**

   Barbara Pes, Nicoletta Dessì, Marta Angioni

   (https://doi.org/10.1016/j.inffus.2016.10.001)

Surveyed Literature (cont'd)

3. **High-Dimensional Feature Selection for Genomic Datasets**

   Majid Afshar & Hamid Usefi

   (https://arxiv.org/pdf/2002.12104.pdf)

4. **Selecting Genes by Test Statistics**

   Dechang Chen, Zhenqiu Liu, Xiaobin Ma, and Dong Hua

   (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1184045/)

Surveyed Literature (cont'd)

5. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts

M.Dashtban, MohammadaliBalafar

(https://www.sciencedirect.com/science/article/pii/S0888754317300046?via%3Dihub)

Thanks