Appendix A. - Study Flow Diagram

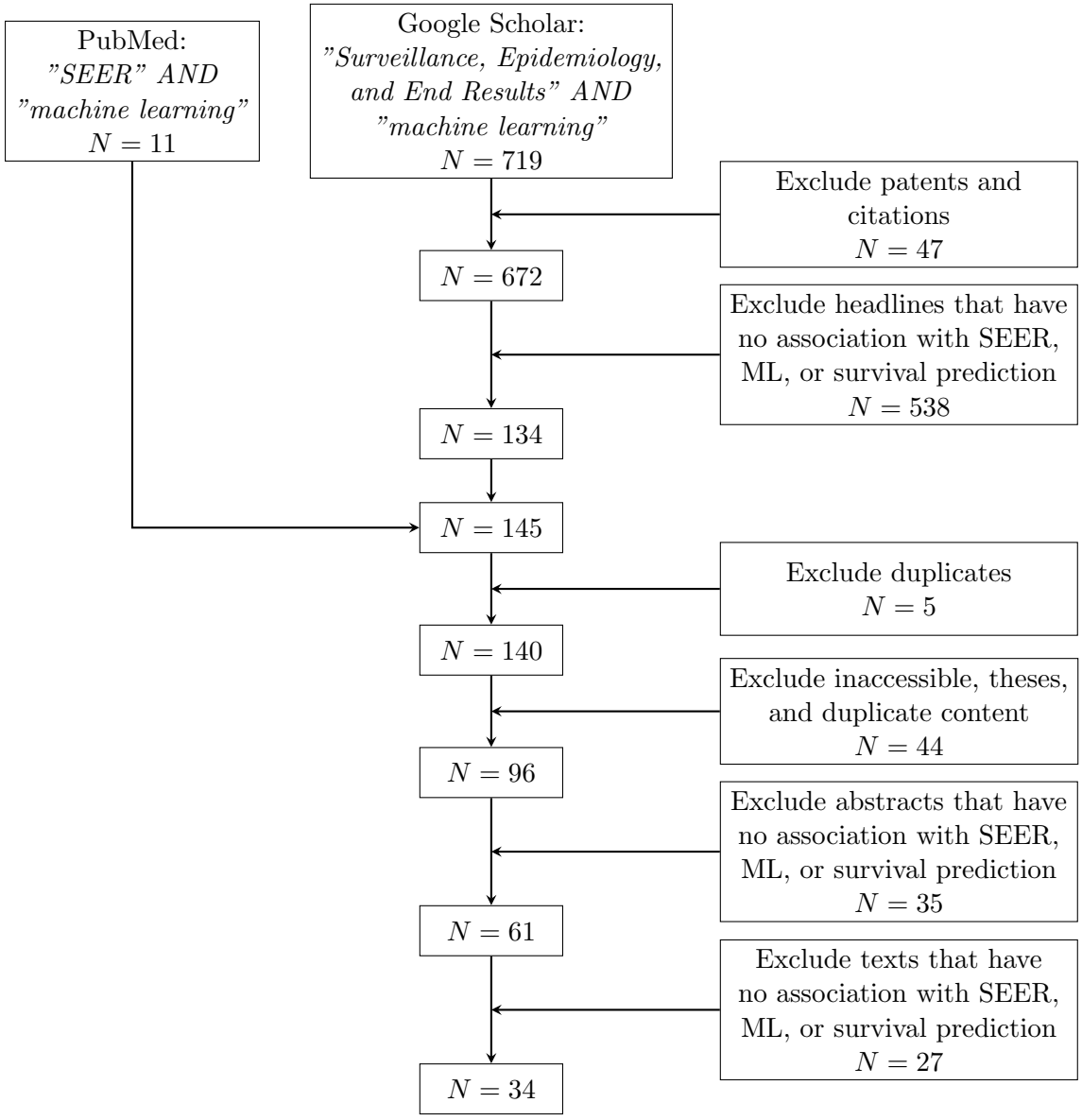

Figure 3: Inclusion and exclusion criteria for literature review to identify studies applying machine learning for survival prediction with SEER cancer data.

Appendix B. - Model Categories for Literature Review

1-NN - 1-nearest neighbor
ANFIS - Adaptive neuro fuzzy inference system
BAG - Bagging
BN - Bayesian (belief) network
BOO - Boosting
CFS - Correlation-based feature selection
Co-train - Co-training
DSO - Density-based synthetic oversampling
DT - Decision tree (including variants)
EV - (Ensemble) voting
GBM - Gradient boosting machine
HC - Hierarchical clustering
InfoGain - Information gain
k-means - k-means clustering
k-NN - k-nearest neighbor
MBC - Model-based clustering
MLP - Multilayer perceptron
NB - Naives Bayes
NNMF - Non-negative matrix factorization
LinR - Linear regression
LogR - Logistic regression
PCA - Principal component analysis
PSO - Particle swarm optimization
RF - Random forest SMOTE - Synthetic minority oversampling technique
SSL - Semi-supervised learning
SOM - Self-ordering maps
ST - Stacking
SVM - Support vector machine

Appendix C. - Input Attributes Overview

Table 4: Input attributes for breast cancer 1-year survival prediction (270,107 cases).

Column	Minimum	Maximum	Mean	Std	Values	Empty	Status
Age at diagnosis	10	99	60.07	13.97	88	0	continuous
Age recode <1 year olds	3	18	12.6	2.77	16	0	continuous
CS Tumor size	0	999	96.84	258.39	241	0	continuous
EOD 10 - number of lymph nodes examined	0	99	9.29	16.35	79	0	continuous
EOD 10 - positive lymph nodes examined	0	99	13.97	32.68	66	0	continuous
Year of birth	1904	1998	1945.98	14.06	92	0	continuous
Year of diagnosis	2004	2009	2006.56	1.71	6	0	continuous
Adjusted AJCC 6th M (1988+)	0	99	3.36	16.67	4	0	categorical
Adjusted AJCC 6th N (1988+)	0	99	11.15	24.34	6	0	categorical
Adjusted AJCC 6th Stage (1988+)	0	99	30.37	24	11	0	categorical
Adjusted AJCC 6th T (1988+)	0	99	25.37	20.66	15	0	categorical
AYA site recode/WHO 2008	13	56	36.19	1.99	11	0	categorical
Behavior Code ICD-O-2	0	3	3	0.02	3	0	categorical
Broad Histology recode	0	39	8.73	1.61	26	0	categorical
CS Extension	50	999	153.04	179.35	40	0	categorical
CS Lymph Nodes	0	999	153.29	255.75	39	0	categorical
CS Mets at DX	0	99	4.9	18.53	9	0	categorical
CS Site-Specific Factor 1	10	999	84.01	256.29	7	0	categorical
CS Site-Specific Factor 2	10	999	94.06	270.1	7	0	categorical
CS Site-Specific Factor 3	0	99	15.78	34.57	68	0	categorical
CS Site-Specific Factor 4	0	987	372.2	478.1	5	0	categorical
CS Site-Specific Factor 5	0	987	371.99	478.26	4	0	categorical
CS Site-Specific Factor 6	0	987	64.66	209.21	8	0	categorical
CS Version Input Current	20510	20550	20514.53	9.83	5	0	categorical
CS Version Input Original	937	20550	10455.66	1307.25	17	0	categorical
Derived AJCC M	0	99	3.39	16.67	4	0	categorical
Derived AJCC N	0	99	9.53	19.67	20	0	categorical
Derived AJCC Stage Group	0	99	30.42	24.06	11	0	categorical
Derived AJCC T	0	99	23.44	18.56	16	0	categorical
Derived SS1977	1	9	2.13	1.79	6	0	categorical
Derived SS2000	1	9	2.12	1.78	6	0	categorical
ER Status Recode Breast Cancer (1990+)	1	4	1.42	0.83	4	0	categorical
Grade	1	9	2.75	2.03	5	0	categorical
Histologic Type ICD-O-2	8000	9580	8483.67	98.94	116	0	categorical
Histologic Type ICD-O-3	8000	9580	8486.61	94.57	132	0	categorical
Historic SSG 2000 Stage	1	9	1.77	1.64	4	0	categorical
ICCC site rec extended ICD-O-3/ WHO 2008	21	999	104.75	49.06	20	0	categorical
ICCC site recode ICD-O-3/WHO 2008	33	999	118.68	48.29	12	0	categorical
IHS link	-1	1	0	0.06	3	287	categorical
Insurance Recode (2007+)	-1	5	1.1	2.09	6	130628	categorical
Laterality	1	9	1.55	0.76	5	0	categorical
Marital status at diagnosis	1	9	2.83	1.81	7	0	categorical
Month of diagnosis	1	12	6.47	3.43	12	0	categorical
NHIA Derived Hisp Origin	0	8	0.45	1.55	9	0	categorical
Origin Recode NHIA	0	1	0.1	0.3	2	0	categorical
PR Status Recode Breast Cancer (1990+)	1	4	1.56	0.87	4	0	categorical
Race recode A	1	9	1.31	0.82	4	0	categorical
Race recode Y	1	9	1.38	0.99	5	0	categorical
Race/ethnicity	1	99	2.92	11.58	30	0	categorical
Recode ICD-O-2 to 10	-1	9999	-0.67	57.72	2	270098	categorical
Recode ICD-O-2 to 9	175	9999	1738.41	116.37	19	0	categorical
SEER historic stage A	1	9	1.64	1.25	4	0	categorical
SEER registry	1501	1547	1532.56	13.45	18	0	categorical
Sex	1	2	1.99	0.08	2	0	categorical
State-county recode	2900	53073	17874.25	14081.77	614	0	categorical
Type of reporting source	1	8	1.12	0.78	6	0	categorical
Survived cancer for 12 months	0	1	0.97	0.17	2	0	target

Table 5: Input attributes for breast cancer 5-year survival prediction (248,751 cases).

Column	Minimum	Maximum	Mean	Std	Values	Empty	Status
Age at diagnosis	10	99	59.35	13.54	88	0	continuous
Age recode <1 year olds	3	18	12.46	2.7	16	0	continuous
CS Tumor size	0	999	96.49	257.97	237	0	continuous
EOD 10 - number of lymph nodes examined	0	99	9.34	16.32	79	0	continuous
EOD 10 - positive lymph nodes examined	0	99	13.15	31.76	66	0	continuous
Year of birth	1904	1998	1946.68	13.61	92	0	continuous
Year of diagnosis	2004	2009	2006.53	1.7	6	0	continuous
Adjusted AJCC 6th M (1988+)	0	99	3.24	16.26	4	0	categorical
Adjusted AJCC 6th N (1988+)	0	99	11	24	6	0	categorical
Adjusted AJCC 6th Stage (1988+)	0	99	30.24	23.82	11	0	categorical
Adjusted AJCC 6th T (1988+)	0	99	25.21	20.39	15	0	categorical
AYA site recode/WHO 2008	13	56	36.19	1.99	11	0	categorical
Behavior Code ICD-O-2	0	3	3	0.02	3	0	categorical
Broad Histology recode	0	39	8.73	1.61	26	0	categorical
CS Extension	50	999	152.45	178.48	40	0	categorical
CS Lymph Nodes	0	999	153.29	254.51	39	0	categorical
CS Mets at DX	0	99	4.85	18.27	9	0	categorical
CS Site-Specific Factor 1	10	999	82.56	253.86	7	0	categorical
CS Site-Specific Factor 2	10	999	92.82	268.17	7	0	categorical
CS Site-Specific Factor 3	0	99	14.96	33.74	68	0	categorical
CS Site-Specific Factor 4	0	987	374.43	478.66	5	0	categorical
CS Site-Specific Factor 5	0	987	374.22	478.82	4	0	categorical
CS Site-Specific Factor 6	0	987	64.73	209.07	8	0	categorical
CS Version Input Current	20510	20550	20514.42	9.69	5	0	categorical
CS Version Input Original	937	20550	10445.79	1276.85	17	0	categorical
Derived AJCC M	0	99	3.27	16.27	4	0	categorical
Derived AJCC N	0	99	9.46	19.41	20	0	categorical
Derived AJCC Stage Group	0	99	30.28	23.88	11	0	categorical
Derived AJCC T	0	99	23.35	18.42	16	0	categorical
Derived SS1977	1	9	2.13	1.79	6	0	categorical
Derived SS2000	1	9	2.13	1.78	6	0	categorical
ER Status Recode Breast Cancer (1990+)	1	4	1.42	0.82	4	0	categorical
Grade	1	9	2.74	2.03	5	0	categorical
Histologic Type ICD-O-2	8000	9580	8483.79	98.71	114	0	categorical
Histologic Type ICD-O-3	8000	9580	8486.66	94.45	130	0	categorical
Historic SSG 2000 Stage	1	9	1.78	1.64	4	0	categorical
ICCC site rec extended ICD-O-3/ WHO 2008	21	999	104.72	48.77	20	0	categorical
ICCC site recode ICD-O-3/WHO 2008	33	999	118.65	48	12	0	categorical
IHS link	-1	1	0	0.06	3	266	categorical
Insurance Recode (2007+)	-1	5	1.07	2.09	6	122081	categorical
Laterality	1	9	1.55	0.77	5	0	categorical
Marital status at diagnosis	1	9	2.79	1.78	7	0	categorical
Month of diagnosis	1	12	6.45	3.43	12	0	categorical
NHIA Derived Hisp Origin	0	8	0.45	1.55	9	0	categorical
Origin Recode NHIA	0	1	0.1	0.3	2	0	categorical
PR Status Recode Breast Cancer (1990+)	1	4	1.56	0.86	4	0	categorical
Race recode A	1	9	1.31	0.81	4	0	categorical
Race recode Y	1	9	1.38	0.99	5	0	categorical
Race/ethnicity	1	99	2.92	11.57	30	0	categorical
Recode ICD-O-2 to 10	-1	9999	-0.64	60.15	2	248742	categorical
Recode ICD-O-2 to 9	175	9999	1738.82	114.57	19	0	categorical
SEER historic stage A	1	9	1.64	1.23	4	0	categorical
SEER registry	1501	1547	1532.49	13.46	18	0	categorical
Sex	1	2	1.99	0.08	2	0	categorical
State-county recode	2900	53073	17840.35	14106.57	614	0	categorical
Type of reporting source	1	8	1.12	0.77	6	0	categorical
Survived cancer for 60 months	0	1	0.87	0.33	2	0	target

Table 6: Input attributes for lung cancer 1-year survival prediction (215,630 cases).

Column	Minimum	Maximum	Mean	Std	Values	Empty	Status
Age at diagnosis	0	99	68.24	11.57	98	0	continuous
Age recode <1 year olds	0	18	14.23	2.29	19	0	continuous
CS Tumor size	0	999	327.69	437.19	283	0	continuous
EOD 10 - number of lymph nodes examined	0	99	11.15	28.92	86	0	continuous
EOD 10 - positive lymph nodes examined	0	99	76.83	40.09	36	0	continuous
Year of birth	1904	2007	1937.75	11.68	101	0	continuous
Year of diagnosis	2004	2009	2006.5	1.71	6	0	continuous
AYA site recode/WHO 2008	13	56	36.15	4.85	17	0	categorical
Behavior Code ICD-O-2	0	3	3	0.02	3	0	categorical
Broad Histology recode	0	39	2.76	2.31	31	0	categorical
CS Extension	100	999	518.96	313.5	48	0	categorical
CS Lymph Nodes	0	999	276.64	331.83	7	0	categorical
CS Mets at DX	0	99	25.62	29.12	28	0	categorical
CS Site-Specific Factor 1	0	999	964.6	154.43	7	0	categorical
CS Version Input Current	20510	20550	20512.62	6.05	5	0	categorical
CS Version Input Original	937	20550	10551.98	1619.7	17	0	categorical
Derived AJCC M	0	99	13.74	27.55	4	0	categorical
Derived AJCC N	0	99	25	31.55	6	0	categorical
Derived AJCC Stage Group	12	99	58.43	24.34	10	0	categorical
Derived AJCC T	0	99	39.93	29.05	7	0	categorical
Derived SS1977	1	9	5.28	2.59	6	0	categorical
Derived SS2000	1	9	5.31	2.56	6	0	categorical
First malignant primary indicator	0	1	1	0	2	0	categorical
Grade	1	9	6.11	3.21	5	0	categorical
Histologic Type ICD-O-2	8000	9581	8090.01	103.02	127	0	categorical
Histologic Type ICD-O-3	8000	9581	8096.95	100.61	148	0	categorical
Historic SSG 2000 Stage	1	9	5.09	2.76	4	0	categorical
ICCC site rec extended ICD-O-3/ WHO 2008	21	999	100.73	4.88	34	0	categorical
ICCC site recode ICD-O-3/WHO 2008	33	999	116.31	3.73	17	0	categorical
IHS link	-1	1	0	0.06	3	299	categorical
Insurance Recode (2007+)	-1	5	1.05	2.12	6	107475	categorical
Laterality	0	9	1.88	1.82	6	0	categorical
Marital status at diagnosis	1	9	2.97	1.77	6	0	categorical
Month of diagnosis	1	12	6.38	3.43	12	0	categorical
NHIA Derived Hisp Origin	0	8	0.24	1.15	9	0	categorical
Origin Recode NHIA	0	1	0.05	0.23	2	0	categorical
Primary by International Rules	0	1	1	0	2	0	categorical
Race recode A	1	9	1.24	0.64	4	0	categorical
Race recode Y	1	9	1.3	0.81	5	0	categorical
Race/ethnicity	1	99	1.94	7.15	29	0	categorical
Recode ICD-O-2 to 10	-1	9999	-0.4	77.64	2	215617	categorical
Recode ICD-O-2 to 9	193	9999	1625.12	67.67	15	0	categorical
SEER historic stage A	1	9	3.32	1.83	4	0	categorical
SEER registry	1501	1547	1533.49	13.31	18	0	categorical
Sex	1	2	1.47	0.5	2	0	categorical
State-county recode	2900	53073	18313.01	13210.1	614	0	categorical
Type of reporting source	1	8	1.15	0.76	6	0	categorical
Survived cancer for 12 months	0	1	0.44	0.5	2	0	target

Table 7: Input attributes for lung cancer 5-year survival prediction (205,554 cases).

Column	Minimum	Maximum	Mean	Std	Values	Empty	Status
Age at diagnosis	0	99	68.11	11.58	98	0	continuous
Age recode <1 year olds	0	18	14.2	2.29	19	0	continuous
CS Tumor size	0	999	334.17	439.59	281	0	continuous
EOD 10 - number of lymph nodes examined	0	99	11.12	28.98	85	0	continuous
EOD 10 - positive lymph nodes examined	0	99	77.57	39.56	35	0	continuous
Year of birth	1904	2007	1937.87	11.69	101	0	continuous
Year of diagnosis	2004	2009	2006.5	1.71	6	0	continuous
AYA site recode/WHO 2008	13	56	36.15	4.84	17	0	categorical
Behavior Code ICD-O-2	0	3	3	0.02	3	0	categorical
Broad Histology recode	0	39	2.75	2.31	31	0	categorical
CS Extension	100	999	526.12	312.05	48	0	categorical
CS Lymph Nodes	0	999	282.23	332.69	7	0	categorical
CS Mets at DX	0	99	26.28	29.07	28	0	categorical
CS Site-Specific Factor 1	0	999	965.08	152.95	7	0	categorical
CS Version Input Current	20510	20550	20512.61	6.04	5	0	categorical
CS Version Input Original	937	20550	10549.43	1613.83	17	0	categorical
Derived AJCC M	0	99	13.92	27.55	4	0	categorical
Derived AJCC N	0	99	25.42	31.6	6	0	categorical
Derived AJCC Stage Group	12	99	59.21	23.81	10	0	categorical
Derived AJCC T	0	99	40.35	29.02	7	0	categorical
Derived SS1977	1	9	5.37	2.55	6	0	categorical
Derived SS2000	1	9	5.4	2.51	6	0	categorical
First malignant primary indicator	0	1	1	0	2	0	categorical
Grade	1	9	6.14	3.2	5	0	categorical
Histologic Type ICD-O-2	8000	9581	8089.43	102.64	127	0	categorical
Histologic Type ICD-O-3	8000	9581	8096.4	100.2	148	0	categorical
Historic SSG 2000 Stage	1	9	5.18	2.72	4	0	categorical
ICCC site rec extended ICD-O-3/ WHO 2008	21	999	100.73	4.94	34	0	categorical
ICCC site recode ICD-O-3/WHO 2008	33	999	116.31	3.8	17	0	categorical
IHS link	-1	1	0	0.06	3	288	categorical
Insurance Recode (2007+)	-1	5	1.04	2.12	6	102909	categorical
Laterality	0	9	1.89	1.84	6	0	categorical
Marital status at diagnosis	1	9	2.96	1.77	6	0	categorical
Month of diagnosis	1	12	6.38	3.43	12	0	categorical
NHIA Derived Hisp Origin	0	8	0.24	1.15	9	0	categorical
Origin Recode NHIA	0	1	0.05	0.23	2	0	categorical
Primary by International Rules	0	1	1	0	2	0	categorical
Race recode A	1	9	1.25	0.64	4	0	categorical
Race recode Y	1	9	1.3	0.81	5	0	categorical
Race/ethnicity	1	99	1.94	7.15	29	0	categorical
Recode ICD-O-2 to 10	-1	9999	-0.37	79.52	2	205541	categorical
Recode ICD-O-2 to 9	193	9999	1625.17	69.3	15	0	categorical
SEER historic stage A	1	9	3.37	1.81	4	0	categorical
SEER registry	1501	1547	1533.48	13.31	18	0	categorical
Sex	1	2	1.47	0.5	2	0	categorical
State-county recode	2900	53073	18293.32	13211.73	614	0	categorical
Type of reporting source	1	8	1.16	0.76	6	0	categorical
Survived cancer for 60 months	0	1	0.16	0.37	2	0	target

Appendix D. - Attribute Importance

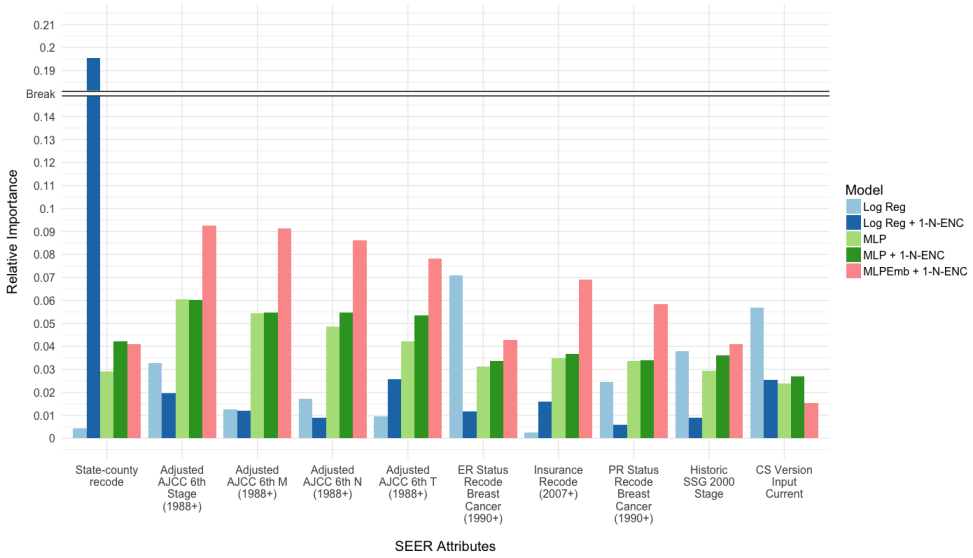

Figure 4: Relative attribute importance for 1-year survival prediction of breast cancer.

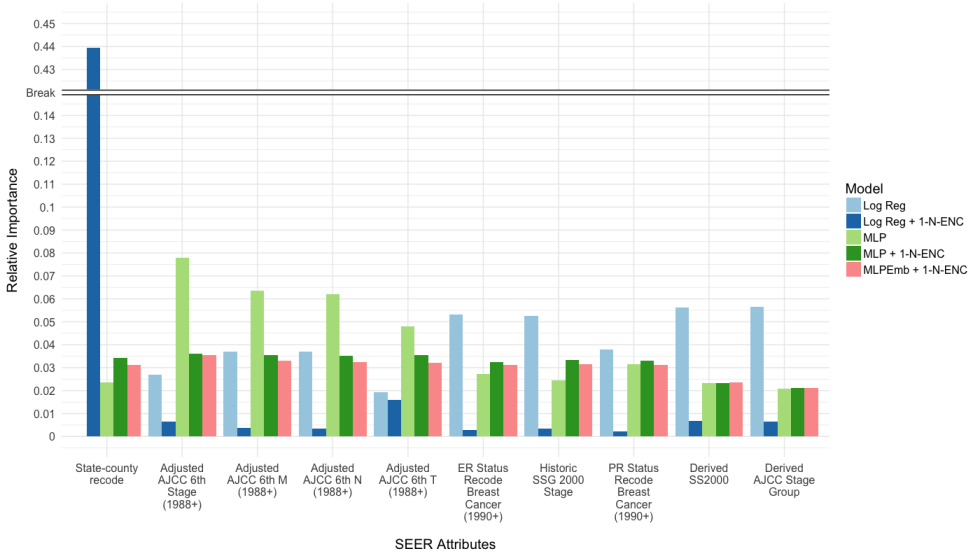

Figure 5: Relative attribute importance for 5-year survival prediction of breast cancer.

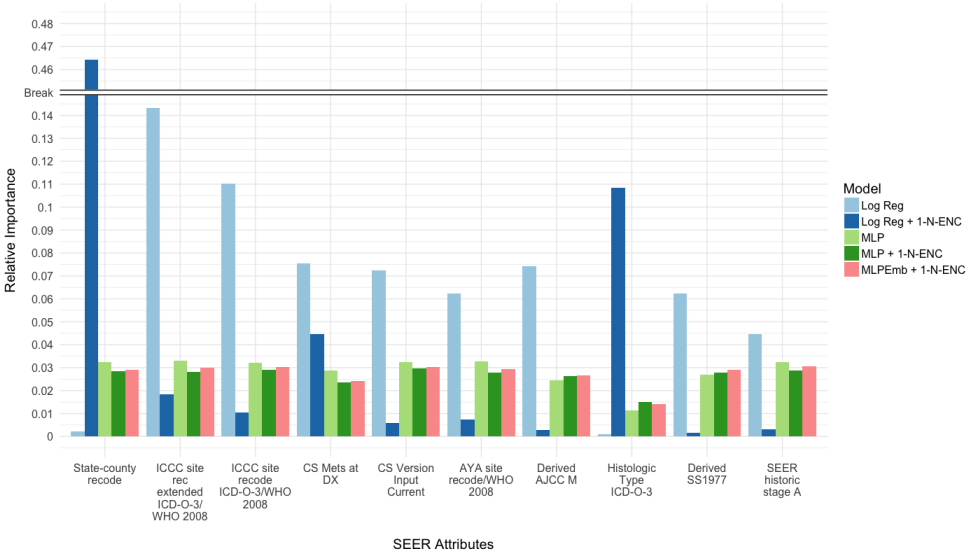

Figure 6: Relative attribute importance for 1-year survival prediction of lung cancer.

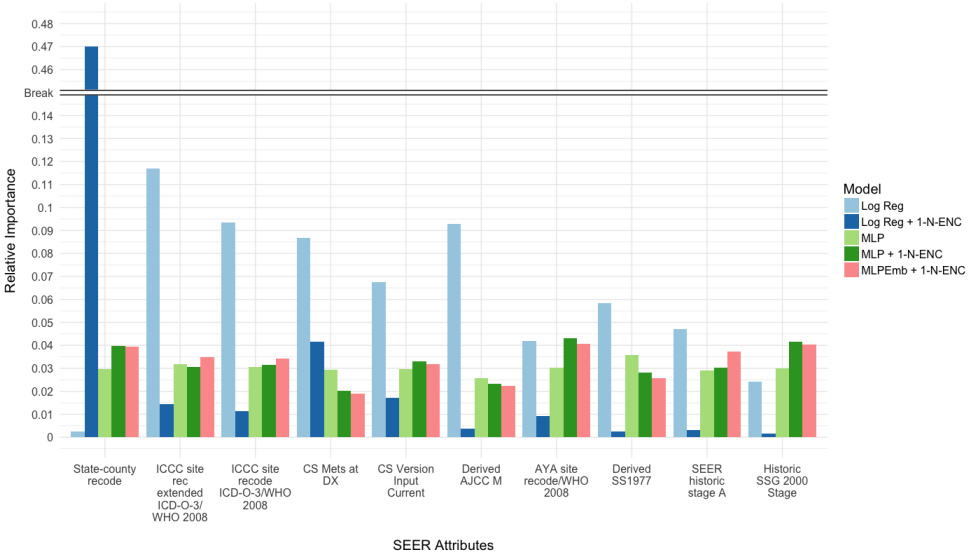

Figure 7: Relative attribute importance for 5-year survival prediction of lung cancer.

Appendix E. - Github Readme File

MLHC 2018 - Reproducible Survival Prediction with SEER Cancer Data

(https://github.com/stefanhgm/MLHC2018-reproducible-survival-seer)

This repository contains the code that was used for experiments reported in *Reproducible Survival Prediction with SEER Cancer Data* submitted to the Machine Learning for Healthcare 2018 conference.

Repository overview:

- **/bin/cluster**: Slurm submission scripts for all parameter tuning experiments on the HPC cluster.
- **/cohort**: SEER*Stat session files to reproduce cohort selections.
- **/example**: Randomly generated SEER example to test the software without sensitive data.
- **/example/CASES.csv**: Example case export. To reproduce experiments, this should be generated for each cohort by loading the provided session files into SEER*Stat, executing the case listing, and exporting it via Matrix→Export→Results as Text File... with "CSV Defaults".
- **/example/INCIDENCES.txt**: Example SEER incidences. To reproduce experiments, this should contain all incidences provided by SEER 1973-2014 data (November 2016 submission) in ASCII format (e.g. by merging them into a single file). The according ASCII data files are available from SEER on request.
- **/lib**: Python classes and functions used for the experiments.
- **main.py**: Main routine to perform the experiments.
- **requirements.txt**: Python dependencies (can be installed with pip, e.g. in a virtual environment).

To execute main.py and reproduce our experiments Python3 (we used version 3.5.2) is necessary and all dependencies in requirements.txt must be satisfied. The easiest way would be to setup an according virtual environment and to install requirements with pip (https://docs.python.org/3/tutorial/venv.html).

The option -h gives an overview of all command line arguments. Note that this code provides some additional functionality such as SVM models and survival regression that were not used for the paper's experiments.

```
$ python main.py -h
```

An experiment with the randomly generated examples and an MLP model can be performed as shown below. This will produce a folder in the current directory containing results and a plot for the AUC score.

```
$ python main.py --incidences example/INCIDENCES.txt --specifications example/read.seer.
    research.nov2016.sas --cases example/CASES.csv --task survival60 --oneHotEncoding --
    model MLP --mlpLayers 2 --mlpWidth 20 --mlpEpochs 1 --mlpDropout 0.1 --importance --
    plotData --plotResults
[...]
Raw data: (10000; 133) cases and attributes
Filtered SEER*Stat cases from ASCII: (5000; 133) cases and attributes
Remove irrelevant, combined, post-diagnosis, and treatment attributes: (5000; 960) cases
    and attributes
Create target label indicating cancer survival for survival60: (2831; 959) cases and
    attributes
Remove inputs with constant values: (2831; 925) cases and attributes
Data:  (2831, 925) -> x:(2831, 924), y:(2831,)
Train: x:(2264, 924), y:(2264,)
Valid: x:(283, 924), y:(283,)
Test:  x:(284, 924), y:(284,)

Train on 2264 samples, validate on 283 samples
Epoch 1/1
 - 1s - loss: 0.4241 - acc: 0.8913 - val_loss: 0.2623 - val_acc: 0.9293
Validation results: auc = 0.48878326996197724, f1 = 0.9633699633699635, acc =
    0.9293286219081273
```