iScience, Volume *26*

Supplemental information

Gene expression of non-homologous end-joining pathways in the prognosis of ovarian cancer

Ethan S. Lavi, Z. Ping Lin, and Elena S. Ratner

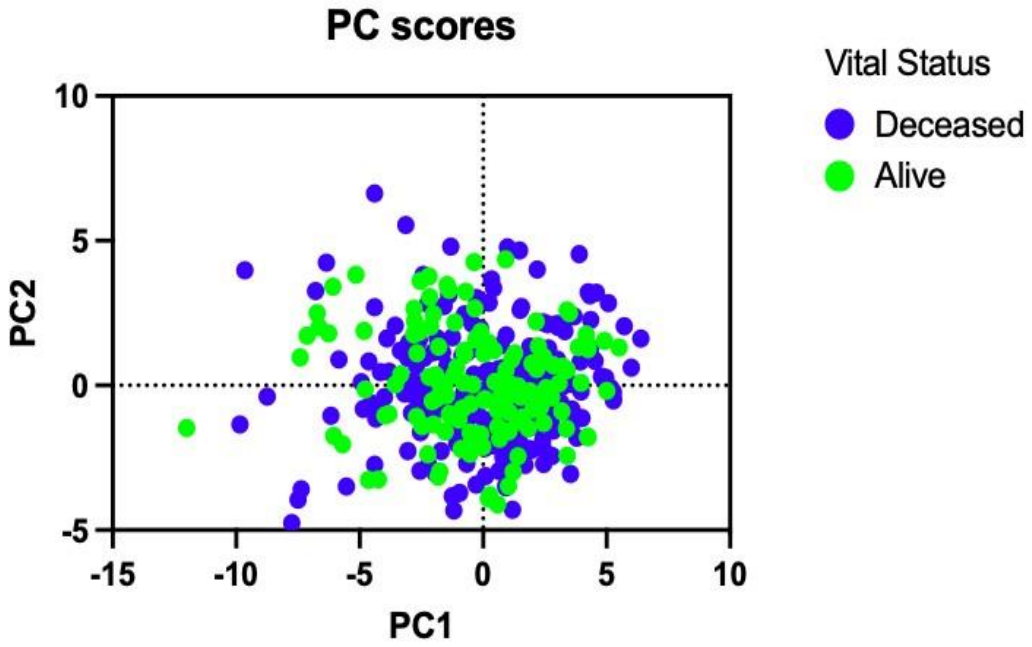

Figure S1: PC Scores from the PCA analysis, related to Figure 3 and S2. PC scores show the ability of the principal components to separate the deceased patients from living ones.

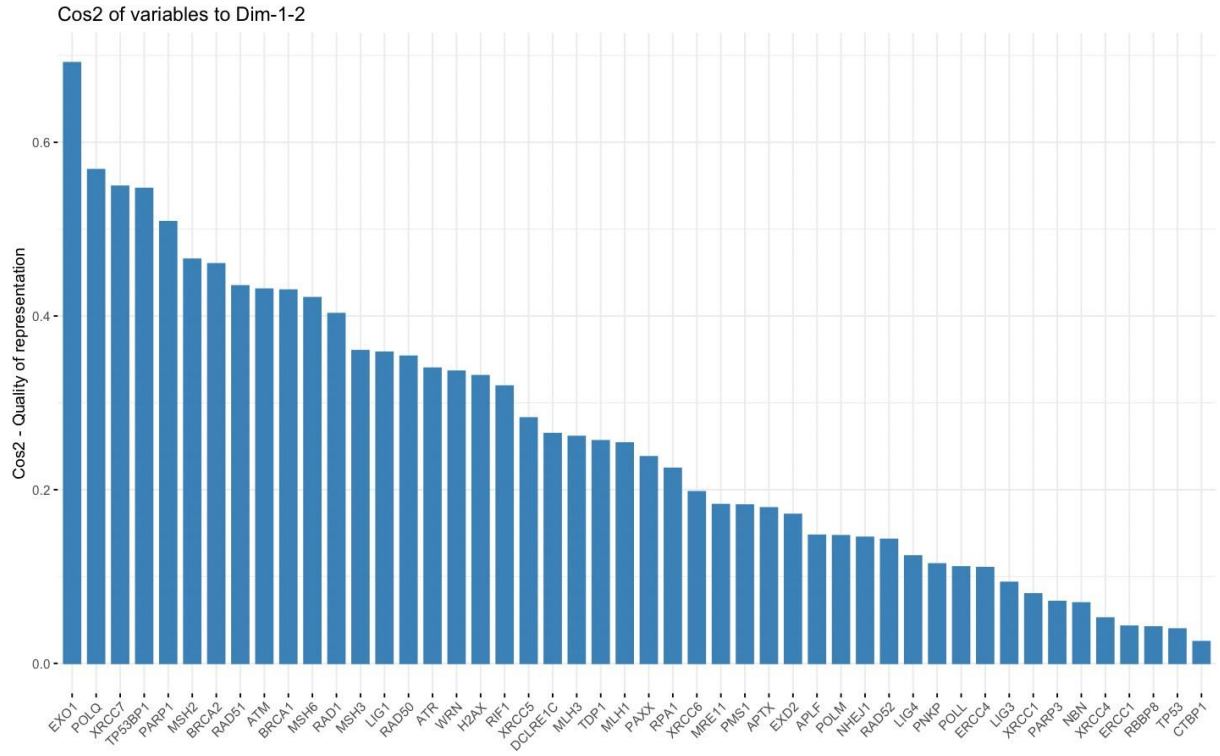

Figure S2: Quality of representation in PCA, related to Figure 3 and S1. The figure demonstrated the amount of importance each variable is to the principal components.

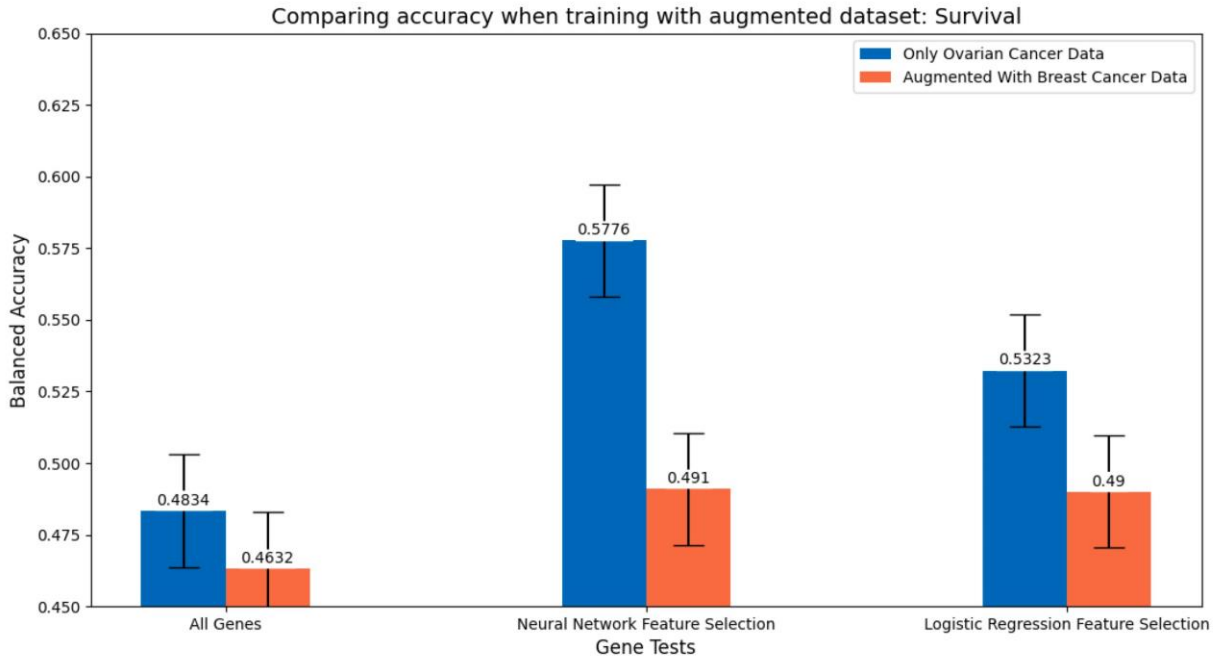

Figure S3: Bar Graph of Neural Network Performance on Survival, related to Figure 4 and Table 1. A bar graph was generated to compare the balanced accuracy of the neural network with and without the training augmentation of breast cancer data. Three gene sets were tested: all the genes, the best from the forward feature selection, and the logistic regression feature selection. The error bars represent a 95% confidence interval of the accuracy. The y-range was set from 0.45 to 0.65 for readability.

Table S1: Linear regression of 48 genes into survival outcome conducted on the ovarian cancer dataset, related to Table 1. Linear regression for the gene dataset was conducted and ordered based on their estimate. Positive estimates of gene expression are associated with improved survival whereas the negative estimates contribute to poor survival probability. Standard error, a t-value, and a p-value for each gene in the linear regression model was generated. p-values under 0.1 were bolded.

Gene	Estimate	Std. Error	t value	Pr(>|t|)
TP53BP1	**-559.54**	**258.8**	**-2.162**	**0.0314**
RAD52	**-417.58**	**202.6**	**-2.061**	**0.0401**
MRE11	-348.71	243.7	-1.431	0.1534
POLQ	-335.92	311.1	-1.080	0.2811
XRCC5	-319.71	232.5	-1.375	0.1702
XRCC1	-255.42	226.2	-1.129	0.2598
XRCC7	-193.28	260.2	-0.743	0.4581
APLF	-169.09	209.8	-0.806	0.4210
EXD2	-149.32	215.0	-0.694	0.4879
RIF1	-103.60	245.9	-0.421	0.6739
PAXX	-87.85	211.8	-0.415	0.6786
LIG3	-81.23	179.7	-0.452	0.6517
ERCC1	-80.89	229.1	-0.353	0.7243
MSH6	-76.40	392.4	-0.195	0.8458
LIG4	-73.72	198.5	-0.371	0.7107
RBBP8	-70.87	183.1	-0.387	0.6991
ATR	-70.15	257.9	-0.272	0.7858
XRCC6	-38.29	207.3	-0.185	0.8536
EXO1	-37.27	318.7	-0.117	0.9070
H2AX	-31.00	219.7	-0.141	0.8879
NBN	-28.98	197.7	-0.147	0.8836
XRCC4	11.11	225.4	0.049	0.9607
RPA1	27.99	203.4	0.138	0.8907
BRCA1	34.42	239.0	0.144	0.8856
RAD50	41.05	211.2	0.194	0.8460
PNKP	48.57	239.1	0.203	0.8392
RAD1	80.71	215.7	0.374	0.7086
NHEJ1	97.53	228.2	0.427	0.6694
MLH1	111.45	220.1	0.506	0.6130
MSH3	115.70	259.4	0.446	0.6559
TP53	117.34	180.3	0.651	0.5157
ERCC4	119.53	195.0	0.613	0.5404
LIG1	132.53	268.3	0.494	0.6217
MSH2	132.93	393.9	0.337	0.7360
RAD51	137.34	248.2	0.553	0.5805
DCLRE1C	148.59	218.8	0.679	0.4976
WRN	201.48	214.6	0.938	0.3487
PARP1	224.75	265.2	0.847	0.3974
POLM	225.56	202.9	1.111	0.2672
PARP3	245.27	194.0	1.264	0.2070
TDP1	265.13	229.4	1.156	0.2487
POLL	267.05	196.6	1.358	0.1753
CTBP1	281.03	191.7	1.466	0.1437
MLH3	281.51	218.2	1.290	0.1981
BRCA2	311.74	244.0	1.278	0.2023
PMS1	**345.51**	**191.3**	**1.806**	**0.0719**
APTX	**347.87**	**198.9**	**1.748**	**0.0814**
ATM	366.40	278.3	1.316	0.1891

Table S2: Aggregated balanced accuracies, related to Star Methods. The balanced accuracy metric from the logistic regression, decision tree, naïve-bayes, support vector machine, and neural network classifiers over survival, progression and recurrence were aggregated into a table. The best performance in each of the three prognosis outcomes categories are bolded.

Model	Survival	Progression	Recurrence
Logistic Regression	0.5000	0.5185	**0.6397**
Decision Tree	0.5386	0.5605	0.5101
Naive Bayes	0.5244	0.5040	0.5536
SVM	**0.5887**	0.5300	0.6301
Neural Network	0.5776	**0.6310**	0.5953

Table S3: Model Baseline Accuracies, related to Table S2 and Star Methods. The assorted machine learning models for this paper were recreated on 100 datasets with randomly generated data. Then, the calculated balanced accuracies along with the number of genes with a significant P-Value were recorded. Using the resulting values, the average and 95% percentile was calculated.

Model	Average	95% Percentile
Logistic Regression	0.4985	0.5223
Decision Tree	0.4966	0.5546
Naive Bayes	0.4978	0.5480
SVM (Linear Kernel)	0.5000	0.5000
SVM (Radial Kernel)	0.4994	0.5322
SVM (Sigmoid Kernel)	0.4990	0.5425
SVM (Polynomial Kernel)	0.5008	0.5271
Significant P-Value Count	4.64	9