iScience, Volume 24

# Supplemental information

## **Robust deep learning model**

#### for prognostic stratification of pancreatic

## ductal adenocarcinoma patients

Jie Ju, Leonoor V. Wismans, Dana A.M. Mustafa, Marcel J.T. Reinders, Casper H.J. van Eijck, Andrew P. Stubbs, and Yunlei Li

# Supplementary documents

Figure S1. The optimal number of clusters. This was the number of clusters that gave the largest scores by (A) silhouette width, and (B) Calinski-Harabasz methods. Related to Figure 2A and STAR Methods.



Figure S2. Survival differences between the prognosis-correlated subtypes identified using PCA and NMF for the multi-omics integration. Kaplan-Meier plots of the prognosis-correlated subtypes identified using A. PCA and B. NMF in the TCGA PAAD cohort. Related to Figure 2A.



| Variable             |                    | N   | Hazard ratio   |                     | р     |
|----------------------|--------------------|-----|--|---------------------|-------|
| tobacco_usage        | CURRENT NON-SMOKER | 101 | =  | Reference           |       |
|                      | CURRENT SMOKER     | 17  | <b>.</b>   | 1.39 (0.65, 2.98)   | 0.391 |
| age                  |                    | 146 | , in the second se | 1.02 (0.99, 1.04)   | 0.158 |
| gender               | FEMALE             | 68  |  | Reference           |       |
|                      | MALE               | 78  | i i  | 1.14 (0.67, 1.97)   | 0.625 |
| alcohol_usage        | NO                 | 50  |  | Reference           |       |
|                      | YES                | 84  | <b>*</b>   | 1.11 (0.61, 2.04)   | 0.730 |
| diabetes_diagnosis   | NO                 | 88  | i i i i i i i i i i i i i i i i i i i  | Reference           |       |
|                      | YES                | 33  | ÷  | 0.98 (0.50, 1.92)   | 0.958 |
| chronic_pancreatitis | NO                 | 105 |  | Reference           |       |
|                      | YES                | 11  | ı.   | 1.07 (0.45, 2.52)   | 0.878 |
| residual_tumor       | R0                 | 83  |  | Reference           |       |
|                      | R1                 | 47  | •  | 2.37 (1.31, 4.28)   | 0.004 |
|                      | R2                 | 5   |  | 0.54 (0.07, 3.95)   | 0.541 |
| т                    | T1                 | 4   |  | Reference           |       |
|                      | T2                 | 15  |  | 2.98 (0.35, 25.62)  | 0.319 |
|                      | Т3                 | 123 |  | 2.84 (0.39, 20.74)  | 0.304 |
|                      | Τ4                 | 3   | · <b>∔</b> _∎'   | 7.61 (0.46, 125.25) | 0.155 |
| N                    | T1                 | 4   |  | Reference           |       |
|                      | T2                 | 15  |  | 2.98 (0.35, 25.62)  | 0.319 |
|                      | Т3                 | 123 |  | 2.84 (0.39, 20.74)  | 0.304 |
|                      | Τ4                 | 3   |  | 7.61 (0.46, 125.25) | 0.155 |
| М                    | MO                 | 69  | i i i i i i i i i i i i i i i i i i i  | Reference           |       |
|                      | M1                 | 3   | $\longleftrightarrow$  | 0.00 (0.00, Inf)    | 0.997 |
| stage2               | Stage I            | 12  |  | Reference           |       |
|                      | Stage II           | 127 | H <b>a</b> ri  | 1.28 (0.46, 3.58)   | 0.638 |
|                      | Stage III          | 3   | <b>⊢</b>   | 3.39 (0.37, 31.27)  | 0.281 |
|                      | Stage IV           | 3   | $\longleftrightarrow$  | 0.00 (0.00, Inf)    | 0.997 |
| grade                | G1                 | 21  |  | Reference           |       |
|                      | G2                 | 83  | <b>⊢</b> ∎-1   | 1.61 (0.61, 4.22)   | 0.338 |
|                      | G3                 | 41  | ₽ <mark>₩</mark> -1  | 2.25 (0.83, 6.11)   | 0.110 |
|                      | G4                 | 1   |  | 1.94 (0.22, 16.94)  | 0.548 |

Figure S3. Impact of clinical risk factors on patient overall survival. The results of the univariate Cox-PH analysis (see Results) were given, which exhibited the impact of the clinical factors on patient actual OS individually. Related to Table 1.

0.01 0.1 1 10 100

Figure S4. Added value of the clinical factors to identified subtypes. The results of the multivariate Cox-PH analysis (see Results) were given, which showed how the clinical factors affect the OS when prognosiscorrelated survival subtypes are held constantly. Related to Table 1.

| Variable   |   | N   | Hazard ratio                        |   |   |
|--|---|---|-------------------------------------|---|---|
|  |   | 440   | Surv ~ age + class                  | 101/000 100   | 0.0   |
| age  | and sole  | 146   |                                     | 1.01 (0.99, 1.04)   | 0.3   |
| ciass  | moderate  | 103   |                                     | Reference; Variable logrank:  | -0.0  |
|  | aqqressive  | 43  | Surv ~ alcohol usage + class        | 4.02 (2.17, 7.40)   | <0.0  |
| alcohol usage  | NO  | 50  | Sul V alconor usage + class         | Reference: Variable loorank:  | 2 N   |
| alconol_usage  | VES   | 84  |                                     | 0.95 (0.51, 1.77)   | 0.8   |
| class  | moderate  | 03  |                                     | Reference: Variable lograph:  | 0.0   |
| ciass  | annreesive  | 41  |                                     | 3 78 (2 03 7 05)  | <0.0  |
|  | uqqroosire  |   | Surv ~ chronic pancreatitis + class | 0.10 (2.00, 1.00)   |   |
| chronic pancreatitis   | NO  | 105   |                                     | Reference; Variable logrank:  | ١   |
|  | YES   | 11  |                                     | 0.93 (0.39, 2.21)   | 0.8   |
| class  | moderate  | 77  |                                     | Reference: Variable logrank:  | 1   |
|  | aggressive  | 39  | · · · · ·                           | 3.70 (1.98, 6.92)   | <0.0  |
|  |   |   | Surv ~ diabetes diagnosis + class   |   |   |
| diabetes_diagnosis   | NO  | 88  |                                     | Reference; Variable logrank:  | 1   |
|  | YES   | 33  |                                     | 0.61 (0.30, 1.24)   | 0.1   |
| class  | moderate  | 81  |                                     | Reference; Variable logrank:  | 1   |
|  | aggressive  | 40  | !                                   | 4.83 (2.45, 9.51)   | <0.0  |
|  |   |   | Surv ~ gender + class               |   |   |
| gender   | FEMALE  | 68  |                                     | Reference; Variable logrank:  | 1   |
|  | MALE  | 78  |                                     | 0.78 (0.44, 1.40)   | 0.4   |
| class  | moderate  | 103   |                                     | Reference; Variable logrank:  | 1   |
|  | aggressive  | 43  | !                                   | 4.58 (2.38, 8.81)   | <0.0  |
|  |   |   | Surv ~ grade + class                |   |   |
| grade  | G1  | 21  |                                     | Reference; Variable logrank:  | . 8   |
|  | G2  | 83  |                                     | 1.22 (0.45, 3.32)   | 0.6   |
|  | G3  | 41  |                                     | 1.48 (0.52, 4.19)   | 0.4   |
|  | G4  | 1   |                                     | 2.17 (0.25, 19.22)  | 0.4   |
| class  | moderate  | 103   |                                     | Reference; Variable logrank:  |   |
|  | aggressive  | 43  |                                     | 3.99 (2.11, 7.54)   | < 0.0   |
|  |   |   | Surv ~ M + class                    |   |   |
| M  | MO  | 69  |                                     | Reference; Variable logrank:  | 3   |
|  | M1  | 3   | <                                   | → 0.00 (0.00, lnf)  | 0.9   |
| class  | moderate  | 49  |                                     | Reference; Variable logrank:  |   |
|  | aggressive  | 23  |                                     | 3.07 (1.36, 6.94)   | 0.0   |
|  | -   |   | Surv ~ N + class                    |   |   |
| N  | 11  | 4   |                                     | Reference; Variable logrank:  |   |
|  | 12  | 15  |                                     | 2.53 (0.29, 21.86)  | 0.3   |
|  | 13  | 123   |                                     | 2.46 (0.34, 18.06)  | 0.3   |
|  | 14  | 3   |                                     | 3.08 (0.18, 51.83)  | 0.4   |
| class  | moderate  | 102   |                                     | Reference; Variable logrank:  |   |
|  | aggressive  | 43  |                                     | 3.95 (2.14, 7.31)   | <0.(  |
| and deal designs   | 00  | 00  | Surv ~ residual tumor + class       | Defense Meridia Incomb  |   |
| residual_tumor   | RU  | 63  | <b>T H</b>                          | Reference; Variable logrank:  |   |
|  | R1  | 4/  |                                     | 2.63 (1.43, 4.85)   | 0.0   |
|  | RZ  | 5   |                                     | 0.74 (0.10, 5.56)   | 0.1   |
| class  | moderate  | 94  | <b>–</b>                            | Reference; Variable logrank:  |   |
|  | 0.000000000   |   |                                     | 4.27 (2.22, 8.23)   | <0.   |
|  | aqqressive  | 41  | Curry attanta lines                 |   |   |
| atagal   | Stage   | 41  | Surv ~ stage2 + class               | Deference: Veriable legraph:  | 1   |
| stage2   | Stage I   | 12  | Surv ~ stage2 + class               | Reference; Variable logrank:  |   |
| stage2   | Stage I<br>Stage I  | 12<br>127   | Surv ~ stage2 + class               | Reference; Variable logrank:<br>1.44 (0.51, 4.03)   | 0.4   |
| stage2   | Stage I<br>Stage II<br>Stage II   | 12<br>127<br>3  | Surv ~ stage2 + class               | Reference; Variable logrank:<br>1.44 (0.51, 4.03)<br>1.69 (0.18, 15.76)   | 0.4   |
| stage2   | Stage I<br>Stage I<br>Stage II<br>Stage II  | 12<br>127<br>3<br>3   | Surv ~ stage2 + class               | Reference; Variable logrank:<br>1.44 (0.51, 4.03)<br>1.69 (0.18, 15.76)<br>→ 0.00 (0.00, Inf)   | 0.4<br>0.6<br>0.9   |
| stage2<br>class  | Stage I<br>Stage II<br>Stage II<br>Stage II<br>Stage IV<br>moderate   | 41<br>12<br>127<br>3<br>3<br>102  | Surv ~ stage2 + class               | Reference;         Variable logranic           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           →         0.00 (0.00, Inf)           Reference;         Variable logranic   | 0.4<br>0.6<br>0.9   |
| stage2<br>class  | Stage I<br>Stage I<br>Stage II<br>Stage IV<br>moderate<br>aqgressive  | 41<br>12<br>127<br>3<br>3<br>102<br>43  | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, inf)         Reference;           Reference;         Variable logrank:           4.18 (2.26, 7.75)         1.10 (0. | 0.4<br>0.6<br>0.9<br><0.0   |
| stage2<br>class  | atgressive<br>Stage I<br>Stage II<br>Stage III<br>Stage IV<br>moderate<br>aqgressive<br>T   | 41<br>12<br>127<br>3<br>3<br>102<br>43  | Surv ~ stage2 + class               | Reference; Variable logrank:<br>1.44 (0.51, 4.03)<br>1.69 (0.18, 15.76)<br>0.00 (0.00, hf)<br>Reference; Variable logrank:<br>4.18 (2.26, 7.75)   | 0.4<br>0.6<br>0.9<br><0.0   |
| stage2<br>class<br>T   | adgressive<br>Stage I<br>Stage II<br>Stage III<br>Stage IV<br>moderate<br>aggressive  | 41<br>12<br>127<br>3<br>3<br>102<br>43<br>4   | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, hf)         Reference;           Variable logrank:         4.18 (2.26, 7.75)           Reference;         Variable logrank:           9.62 (0.26, 6.26, 9.275)         Reference;   | 0.4<br>0.6<br>0.9<br><0.0   |
| stage2<br>class<br>T   | Stage I<br>Stage I<br>Stage II<br>Stage III<br>Stage IV<br>moderate<br>aqgressive<br>T1<br>T2<br>T2<br>T2   | 41<br>12<br>127<br>3<br>102<br>43<br>4<br>15  | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, Inf)         Reference;           Reference;         Variable logrank:           4.18 (2.26, 7.75)         Reference;           Reference;         Variable logrank:           2.53 (0.29, 21.88)         2.64 (0.24.60)  | 0.4<br>0.6<br>0.9<br><0.0   |
| stage2<br>class<br>T   | adgressive<br>Stage I<br>Stage II<br>Stage III<br>Stage IV<br>moderate<br>aqgressive<br>T1<br>T2<br>T3<br>T4  | 41<br>12<br>127<br>3<br>3<br>102<br>43<br>4<br>15<br>123  | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, hf)         Reference;           Reference;         Variable logrank:           4.18 (2.26, 7.75)         Reference;           Reference;         Variable logrank:           2.53 (0.29, 21.86)         2.46 (0.34, 18.06)           2.96 (0.46, FL.00)         2.96 (0.46, FL.00)   | 0.4<br>0.6<br>0.9<br><0.0   |
| stage2<br>class<br>T   | Stage I<br>Stage I<br>Stage II<br>Stage III<br>Stage IV<br>moderate<br>aqqressive<br>T1<br>T2<br>T3<br>T4   | 41<br>12<br>127<br>3<br>3<br>102<br>43<br>4<br>15<br>123<br>3<br>102  | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, Inf)         Reference;           Variable logrank:         4.18 (2.26, 7.75)           Reference;         Variable logrank:           2.53 (0.29, 21.86)         2.46 (0.34, 18.06)           3.08 (0.18, 51.83)   | 0.4<br>0.6<br>0.9<br><0.0   |
| stage2<br>class<br>T<br>class  | Stage I<br>Stage I<br>Stage II<br>Stage III<br>Stage IV<br>moderate<br>aggressive<br>T1<br>T2<br>T3<br>T4<br>moderate   | 41<br>12<br>127<br>3<br>102<br>43<br>4<br>15<br>123<br>3<br>102<br>123<br>3<br>102                          | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, Inf)         Reference;           Reference;         Variable logrank:           4.18 (2.26, 7.75)         Reference;           Reference;         Variable logrank:           2.53 (0.29, 21.88)         2.46 (0.34, 18.06)           3.08 (0.18, 51.33)         Reference;           Reference;         Variable logrank:   | 0.4<br>0.6<br>0.5<br><0.0   |
| stage2<br>class<br>T<br>class  | adgressive Stage I Stage I Stage II Stage II Stage IV moderate aggressive T1 T2 T3 T4 moderate aggressive   | 41<br>12<br>127<br>3<br>3<br>102<br>43<br>4<br>15<br>123<br>3<br>102<br>43<br>102<br>43                     | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, 1nf)         Reference; Variable logrank:           4.18 (2.26, 7.75)         Reference; Variable logrank:           2.53 (0.29, 21.86)         2.46 (0.34, 18.06)           3.08 (0.18, 51.83)         Reference; Variable logrank:           3.95 (2.14, 7.31)         3.95 (2.14, 7.31)  | 0.4<br>0.6<br>0.9<br><0.0<br>0.3<br>0.3<br>0.3<br>0.4<br><0.0                                 |
| stage2<br>class<br>T<br>class  | Addressive Stage I Stage I Stage II Stage II Stage II Stage IV moderate aqqressive T1 T2 T3 T4 moderate aqqressive CUDPENT NON SHOREE   | 41<br>12<br>127<br>3<br>102<br>43<br>4<br>125<br>123<br>3<br>102<br>43<br>102<br>43                         | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, hf)         Reference;           Reference;         Variable logrank:           4.18 (2.26, 7.75)         Reference;           Reference;         Variable logrank:           2.46 (0.34, 18.06)         3.08 (0.18, 51.33)           Reference;         Variable logrank:           3.95 (2.14, 7.31)         Deference;   | 0.4<br>0.6<br>0.9<br><0.0<br>0.3<br>0.3<br>0.3<br>0.4<br><0.0                                 |
| stage2<br>class<br>T<br>class<br>class   | addressive Stage I Stage I Stage I Stage II Stage II Stage IV moderate aqcressive T1 T2 T3 T4 moderate aqcressive CURRENT NON-SMOKER CURRENT NON-SMOKER CURRENT NON-SMOKER  | 41<br>12<br>127<br>3<br>3<br>102<br>43<br>102<br>43<br>102<br>43<br>102<br>43<br>102<br>43                  | Surv ~ stage2 + class               | Reference; Variable logrank:           1.44 (0.51, 4.03)           1.69 (0.18, 15.76)           0.00 (0.00, lnf)           Reference; Variable logrank:           4.18 (2.26, 7.75)           Reference; Variable logrank:           2.53 (0.29, 21.86)           2.46 (0.34, 18.06)           3.08 (0.18, 51.33)           Reference; Variable logrank:           3.95 (2.14, 7.31)           Reference; Variable logrank:           4.04 (2.25)   | 0.4<br>0.6<br>0.9<br>1<br><0.0<br>1<br>0.3<br>0.3<br>0.3<br>0.4<br>1<br><0.0                  |
| stage2<br>class<br>T<br>class<br>tobacco_usage   | augressive Stage I Stage I Stage II Stage II Stage IV moderate aqgressive T1 T2 T3 T4 moderate aqgressive CURRENT NON-SMOKER CURRENT SMOKER moderate  | 41<br>12<br>127<br>3<br>102<br>43<br>102<br>43<br>102<br>43<br>102<br>43<br>102<br>101<br>101<br>101<br>101 | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, inf)         Reference; Variable logrank:           4.18 (2.26, 7.75)         Reference; Variable logrank:           2.53 (0.29, 21.86)         2.46 (0.34, 18.06)           3.08 (0.18, 51.83)         Reference; Variable logrank:           3.95 (2.14, 7.31)         Reference; Variable logrank:           3.95 (2.14, 7.31)         Reference; Variable logrank:           1.03 (0.47, 2.25)         Reference; Variable logrank:   | 1<br>0.4<br>0.6<br>0.9<br>1<br><0.0<br>1<br>0.3<br>0.3<br>0.3<br>0.4<br>1<br><0.0             |
| stage2<br>class<br>T<br>::<br>:lass<br>::<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>:<br>: | Addressive<br>Stage I<br>Stage II<br>Stage III<br>Stage III<br>Stage IV<br>moderate<br>aqcressive<br>T1<br>T2<br>T3<br>T4<br>moderate<br>aqcressive<br>CURRENT NON-SMOKER<br>CURRENT SMOKER<br>moderate<br>aqcressive | 41<br>12<br>127<br>3<br>102<br>43<br>102<br>43<br>102<br>43<br>102<br>43<br>101<br>17<br>80<br>20           | Surv ~ stage2 + class               | Reference;         Variable logrank:           1.44 (0.51, 4.03)         1.69 (0.18, 15.76)           0.00 (0.00, Inf)         Reference;           Variable logrank:         4.18 (2.26, 7.75)           Reference;         Variable logrank:           2.53 (0.29, 21.86)         2.46 (0.34, 18.06)           3.08 (0.18, 51.83)         Reference; Variable logrank:           3.95 (2.14, 7.31)         Reference; Variable logrank:           1.03 (0.47, 2.25)         Reference; Variable logrank:           2.35 (0.27, 2.14, 7.31)         Reference; Variable logrank:   | 1<br>0.4<br>0.6<br>0.9<br>1<br><0.0<br>1<br>0.3<br>0.3<br>0.3<br>0.4<br>1<br><0.0<br>1<br>0.9 |

Figure S5. The mutational profiles of relevant single-base substitution (SBS) signatures in **A**. the "moderate" subtype, and **B**. the "aggressive" subtype. Related to Figure 5.



Table S1. Comparison between supervised prognosis-correlated approach and the unsupervised approach. Log-rank p-values of the two subtypes identified on the training set (TCGA PAAD) by these two approaches are shown. Based on identified subtypes, the survival difference of the predicted groups in the test sets are also given by log-rank p-values. Related to Figure 2.

| Datasets  | Prognosis-correlated subtype<br>identification and prediction |                  | Unsupervised subtype<br>identification and prediction |                  |  |
|---|---|------------------|---|------------------|--|
|   | Number of<br>predictors                                       | Log-rank p-value | Number of<br>predictors                               | Log-rank p-value |  |
| TCGA PAAD<br>(n = 146)                            | ١   | 1e-6             | 1   | 0.005            |  |
| ICGC PACA-AU mRNA-<br>seq<br>(n = 59)             | 107   | 0.030            | 83  | 0.500            |  |
| ICGC PACA-AU mRNA<br>microarray<br>(n = 64)       | 99  | 0.031            | 66  | 0.050            |  |
| ICGC PACA-AU DNA<br>methylation array<br>(n = 57) | 81  | 0.036            | 17  | 0.400            |  |
| GEO GSE62452 mRNA<br>microarray<br>(n = 65)       | 113   | 0.007            | 85  | 0.180            |  |
| GEO GSE62498<br>microRNA<br>(n = 65)              | 14  | 0.029            | 14  | 0.200            |  |

Table S2. The proposed etiologies of the single-base substitution (SBS) signatures according to COSMIC database. Related to Figure 5.

|        | Proposed etiologies and comments   | The<br>percentage<br>of the SBS<br>in the<br>"moderate"<br>subtype | The<br>percentage<br>of the SBS in<br>the<br>"aggressive"<br>subtypes |
|--------|--|--|---|
| SBS1   | Clock-like mutational signature. This<br>signature correlates with the individual's<br>age, and might be a cell division/mitotic<br>clock.<br>The mutational process is initiated by the<br>G:T mismatches in double stranded DNA,<br>which is caused by the spontaneous or<br>enzymatic deamination of 5-<br>methylcytosine to thymine causes. Then<br>due to the failure to detect and remove<br>these mismatches prior to DNA replication,<br>the fixation of the T substitution for C<br>occurs. | 10.8%  | 24.2%   |
| SBS5   | Clock-like mutational signature. This signature correlates with the age and the tobacco smoking of the individual.   | 18.8%  | 32.8%   |
| SBS15  | Defective DNA mismatch repair.   | 39.9%  | 37.5%   |
| SBS10b | Polymerase epsilon exonuclease domain mutations.   | 7.8%   | 0.0%  |
| SBS14  | Concurrent polymerase epsilon mutation and defective DNA mismatch repair.  | 15.5%  | 0.0%  |
| SBS87  | Thiopurine chemotherapy treatment, experimentally validated.   | 1.8%   | 0.0%  |
| SBS49  | Possible sequencing artefact.  | 0.0%   | 0.9%  |
| SBS52  | Possible sequencing artefact.  | 0.5%   | 4.9%  |
| SBS59  | Possible sequencing artefact.  | 4.9%   | 0.0%  |