Feature extraction

Histogram and textural features presented in our paper are identical to the ones detailed in (*1*). Here, we use "gray level" without "-" consistently. Furthermore, emphasis names are called without using the "/" sign in between the "zone" and "grey" phrases. In addition, this study uses the expression "gray" instead of "grey". Compactness (Shape) was calculated as defined in (*2*). The Spherical dice coefficient (Shape) was defined in this study as the Dice coefficient (*3*) of the volume of interest (VOI) voxels compared to the hypothetical sphere voxels fit in the bounding box of the given VOI.

Model error estimation

SUPPLEMENTAL TABLE 1. Interpretation of True Positive, True Negative, False Positive and False Negative confusion matrix entries based on 36 months dichotomized reference survival as well as machine learning-predicted survival values.

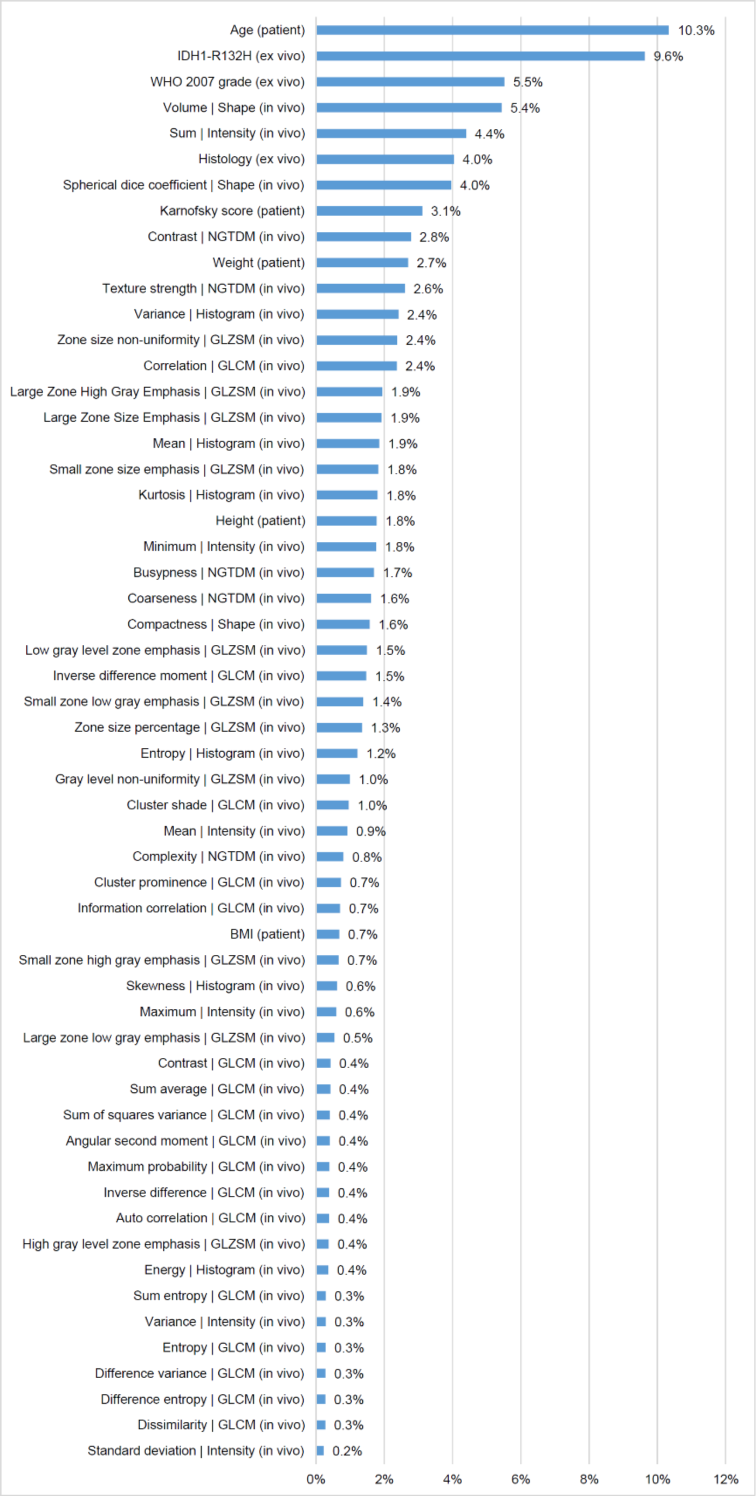
Confusion matrix values	Reference survival	Predicted survival
True Positive (TP)	Survived	Survived
True Negative (TN)	Did not survive	Did not survive
False Positive (FP)	Did not survive	Survived
False Negative (FN)	Survived	Did not survive

Feature selection and weight estimation

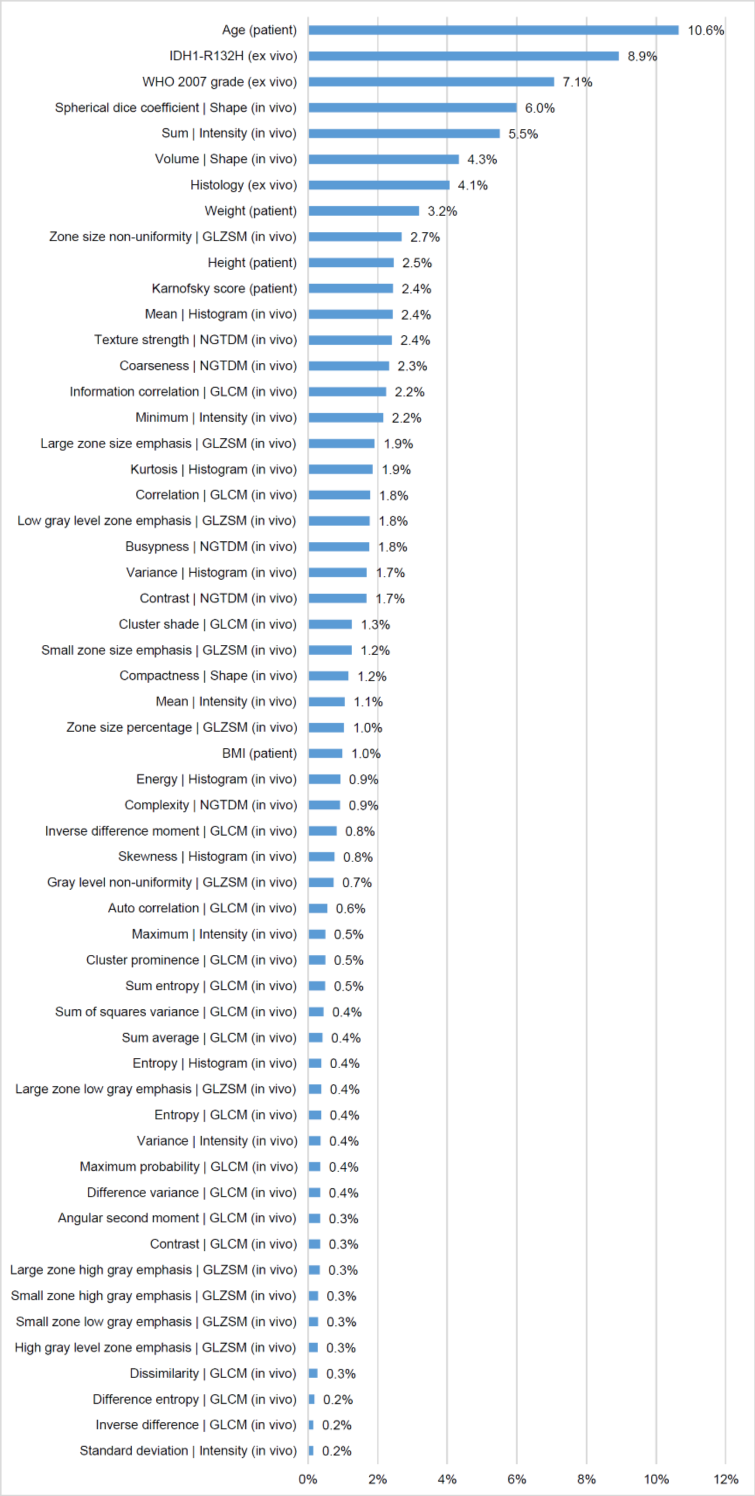
The Genetic Algorithm (GA) population size of our feature selection step was 10 with 5 GA iterations. The initial population was established by selecting only those 10 features that had the lowest overlap of survival groups based on the kernel density estimation (KDE) approach (*4*).

The functional tolerance our second ML layer was 0.0005 with maximum 56000 iterations (1000 iterations per feature). Initial parameters for each weight were 0.0 with initial scale values of 10.0 for each weight.

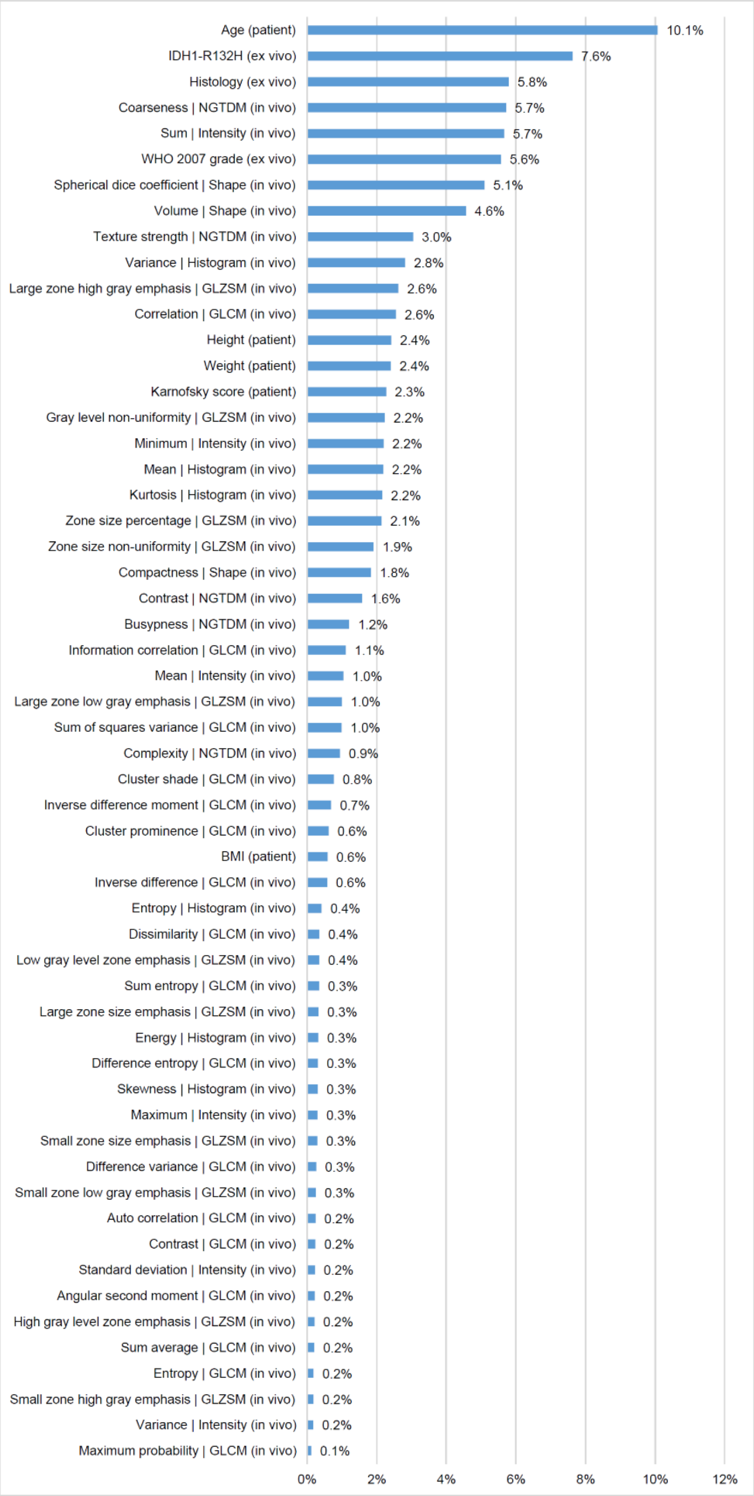
The weight estimation step was regulated so that negative weights were not allowed to be created during the search of the $\overline{w}$ weights. The fitness value of each $\overline{m}$ candidate in the GA populations was the reciprocal of the error measurement provided by the NM algorithm over the given model. The resulted $\overline{w}$ weight values were normalized of each model to the sum of 1.0.



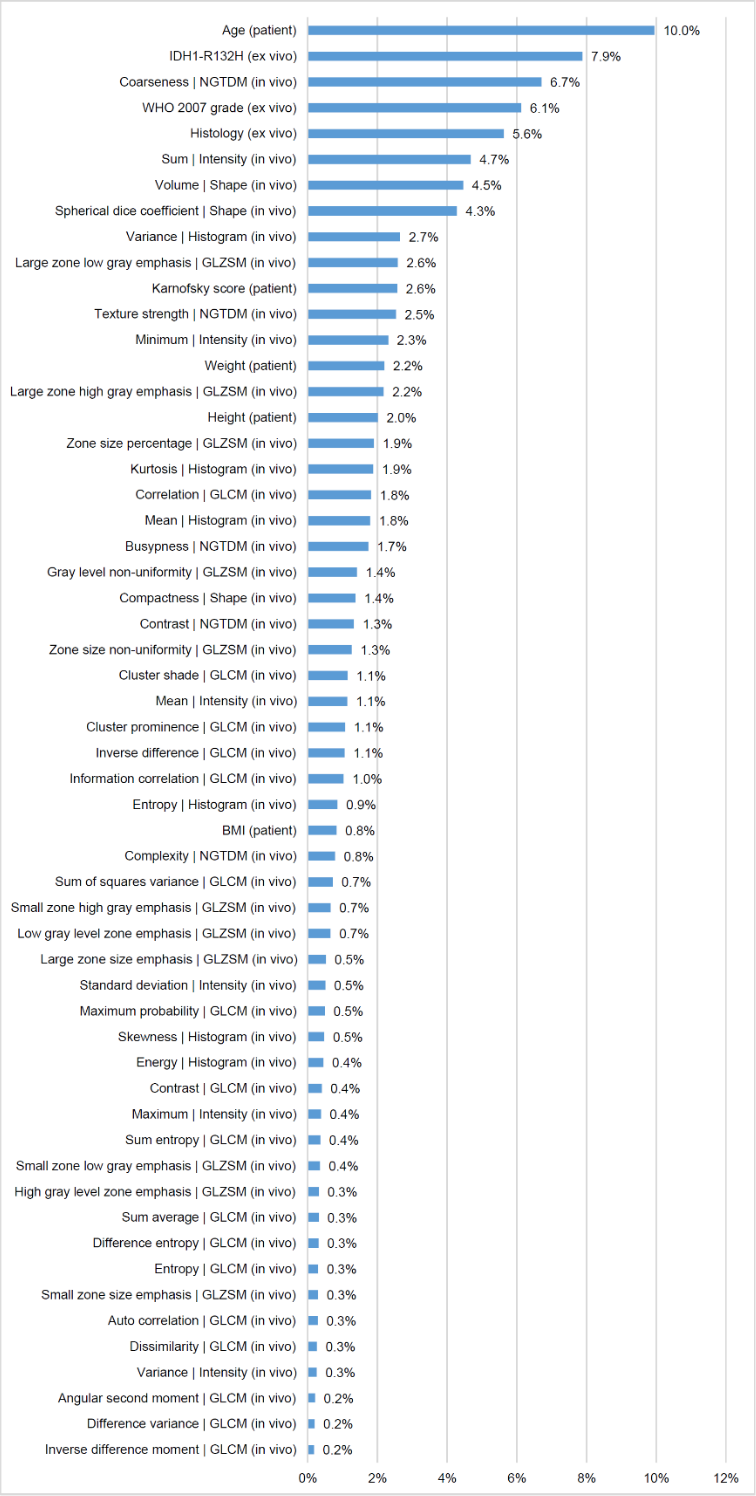
SUPPLEMENTAL FIGURE 1. Average weights of 112 model variants (14 folds x 8 machine learning optimization) with bin size 64, bin width 0.12 TBR. The weights of each 112 models were normalized to the sum of 1.0 before averaging them.



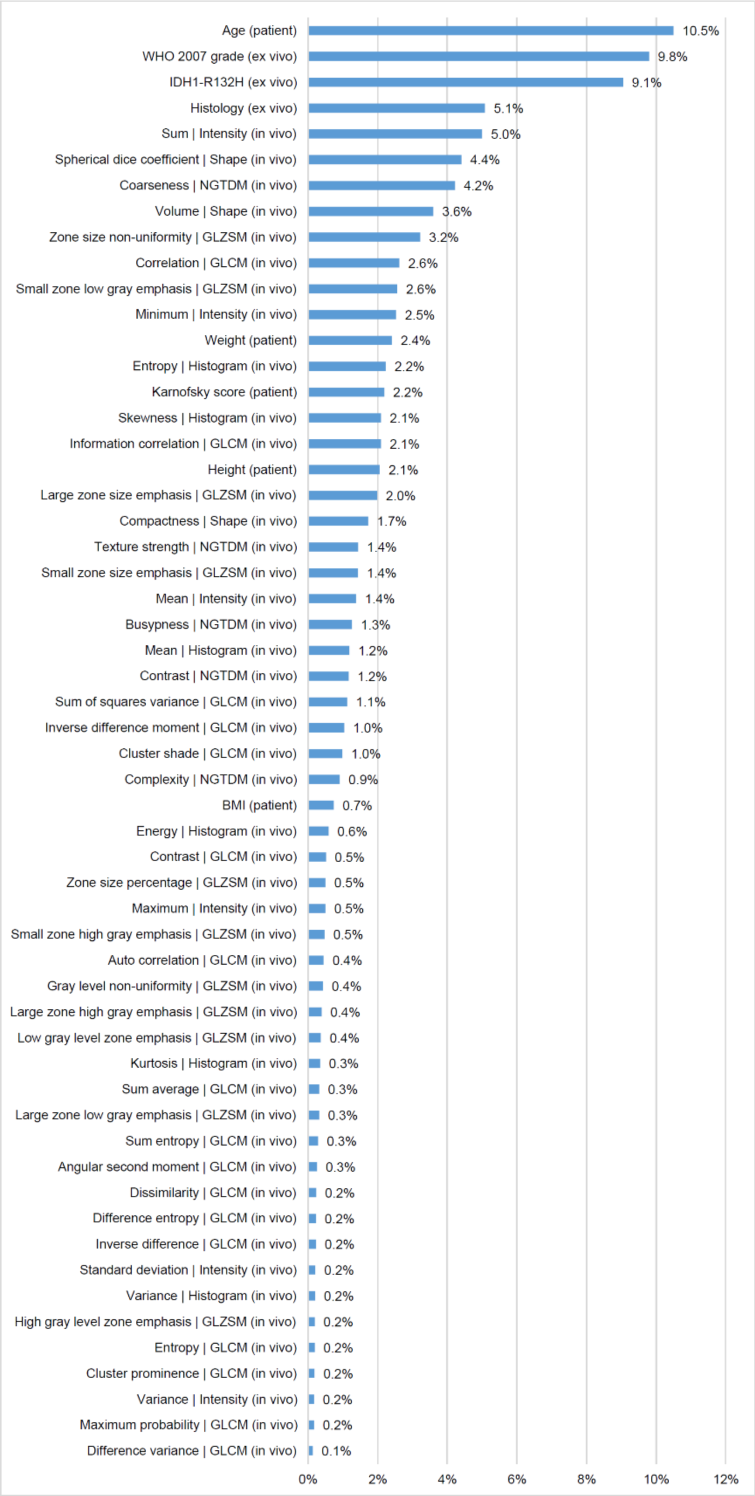
SUPPLEMENTAL FIGURE 2. Average weights of 112 model variants (14 folds x 8 machine learning optimization) with bin size 150, bin width 0.05 TBR. The weights of each 112 models were normalized to the sum of 1.0 before averaging them.



SUPPLEMENTAL FIGURE 3. Average weights of 112 model variants (14 folds x 8 machine learning optimization) with bin size 375, bin width 0.02 TBR. The weights of each 112 models were normalized to the sum of 1.0 before averaging them.



SUPPLEMENTAL FIGURE 4. Average weights of 112 model variants (14 folds x 8 machine learning optimization) with bin size 512, bin width 0.014 TBR. The weights of each 112 models were normalized to the sum of 1.0 before averaging them.



SUPPLEMENTAL FIGURE 5. Average weights of 112 model variants (14 folds x 8 machine learning optimization) with bin width 0.05 TBR and tumor-specific bin size. The weights of each 112 models were normalized to the sum of 1.0 before averaging them.

References

1. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? *Eur J Nucl Med Mol Imaging*. 2017;44:151-165.

2. Bribiesca E. An easy measure of compactness for 2D and 3D shapes. *Pattern Recognit*. 2008;41:543-554.

3. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26:297-302.

4. Geng X, Hu G. Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting. *Biomed Signal Process Control*. 2012;7:112-117.