Supplementary

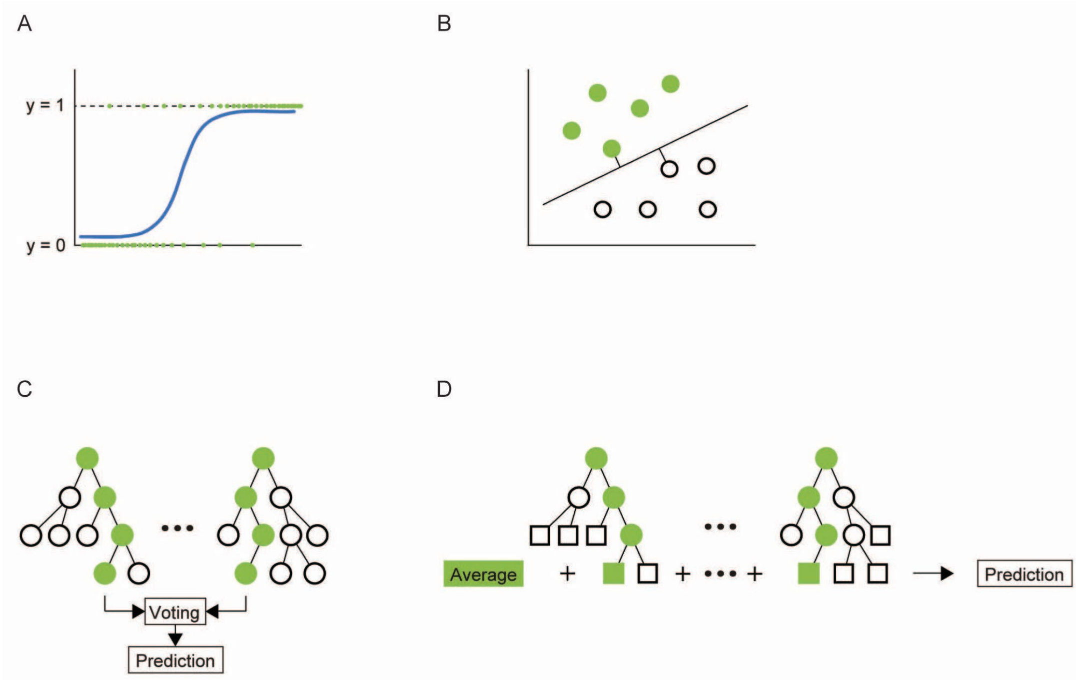

Figure S1 Schematic of each model. (A) LR is a regression model where the objective variable is between 0 and 1. The logistic function ensures that the output falls between 0 and 1 could be interpreted as a probability. (B) SVM maps data into a high-dimensional space and divides it by finding optimal boundaries between different classes. (C) RF develops multiple decision trees and classifies based on their respective predictions. (D) Both GB and XGB are ensemble learning and combine boosting and decision trees. Decision trees are created to modify one previous decision tree and are updated with a smaller error between the prediction and the correct value. The main differences are that GB uses gradient descent, while XGB uses newton boosting, and XGB has effective regularization term, which controls the complexity of the model. GB, gradient boosting; LR, logistic regression; RF, random forest; SVM, support vector machine; XGB, eXtreme Gradient Boosting.

Table S1 Parameters in the feature selection methods

Feature selection	Parameters
LASSO	• alpha: {0.05}
RFE	• model: {RF} • n_estimators of RF: {100} • n_feature_to_select: {10}
RF	• n_estimators: {100} • threshold: {0.05}
GB	• n_estimators: {100} • threshold: {0.05}
Boruta	• model: {RF} • n_estimators: {auto} • alpha: {0.05} • perc: {90}

LASSO, least absolute shrinkage and selection operator; RFE, recursive feature elimination; RF, random forest; GB, gradient boosting.

 https://dx.doi.org/10.21037/jtd-23-1840

Table S2 Tuned hyperparameters in the machine learning models

Model	Hyperparameters
LR	• C: {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000} • max_iter: {1000}
SVM	• C: {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000} • kernel: {rbf} • gamma: {0.001, 0.01, 0.1, 1, 5, 10}
KNN	• n_neighbors: {1, 3, 5, 7, 10, 15, 20} • weights: {uniform, distance}
NB	• var_smoothing: {1e-09, 1e-07, 1e-05, 1e-03, 0.1}
RF	• max_depth: {3, 4, 5, 6, 7, 8} • n_estimators: {1000} • max_features: {3, 4, 5, 6, 7, 8, 9, 10}
GB	• loss: {deviance} • learning_rate: {0.01} • n_estimators: {1000} • min_samples_split: {2, 3, 4, 5, 6, 7} • min_samples_leaf: {2, 5, 10} • max_depth: {3, 4, 5, 6, 7} • subsample: {0.6, 0.7, 0.8, 0.9, 1.0}
XGB	• objective: {binary} • eval_metric: {logloss} • max_depth: {4, 5, 6, 7} • n_estimators: {1000} • colsample_bytree: {0.6, 0.7, 0.8, 0.9} • subsample: {0.6, 0.7, 0.8, 0.9} • min_child_weight: {0.01} • reg_alpha: {0.01, 0.1, 1} • lambda: {1} • gamma: {0.1, 0.2, 0.3}

LR, logistic regression; SVM, support vector machine; KNN, k-nearest neighbor; NB, Naïve Bayes; RF, random forest; GB, gradient boosting; XGB, eXtreme Gradient Boosting.

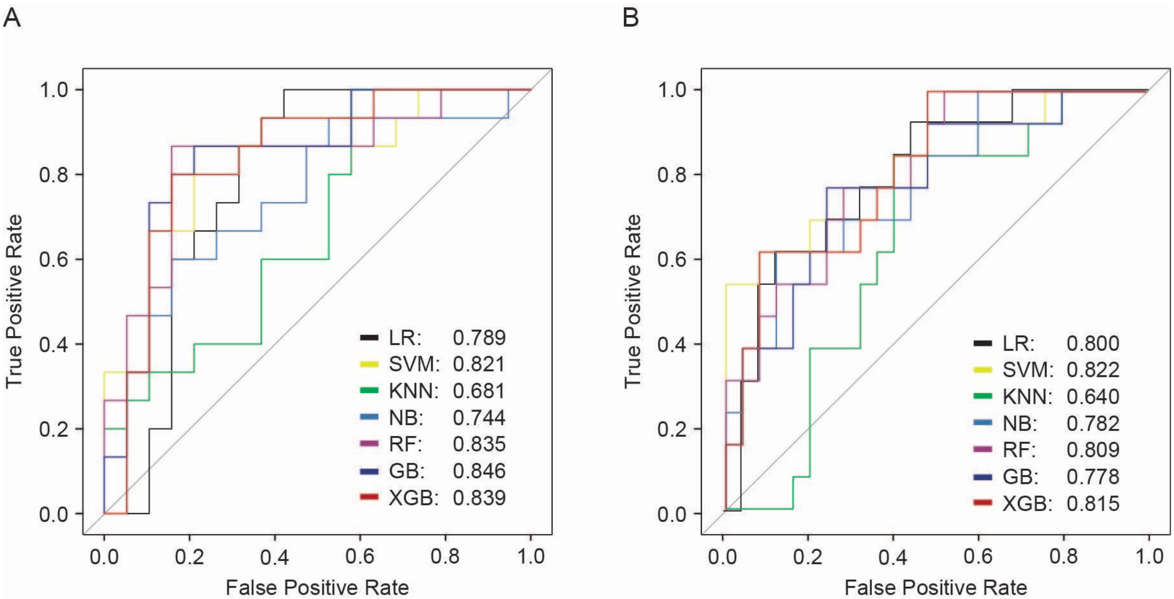

Figure S2 The prediction results for TCR for each hospital in the validation cohort. ROC curves show the results of each machine learning model with the best performing feature selection for Chiba Cancer Center (A) and Kimitsu Chuo Hospital (B). GB, gradient boosting; KNN, k-nearest neighbor; LR, logistic regression; NB, Naïve Bayes; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; TCR, tumors requiring combined resection of adjacent structure; XGB, eXtreme Gradient Boosting.

 https://dx.doi.org/10.21037/jtd-23-1840