PREDICTING SURVIVAL OF TONGUE CANCER PATIENTS BY MACHINE LEARNING MODELS

Angelos Vasilopoulos[a] and Nan Miles Xi[a]

[a] Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL 60660, USA

INTRODUCTION

Tongue cancer is one of the most frequent head and neck malignancies, diagnosed in approximately 50 thousand patients and causing more than 10 thousand deaths annually in the United States, according to the American Cancer Society. The occurrence of new cases has risen in the last 20 years. Treatment primarily involves surgery, chemotherapy, and radiation therapy. The five-year relative survival rate after treatment was 68.8% between 2012 and 2018.

Machine learning is potentially effective in predicting survival from patient data and identifying important treatment factors. Literature has few machine learning models of tongue cancer survival. Some studies have identified survival and risk factors, but conclusions are usually based on descriptive statistics and linear models that ignore complex, nonlinear variable relationships.

Here we utilize a comprehensive machine learning framework to predict tongue cancer survival after treatment. We train five cutting-edge models and evaluate performance and uncertainty. Important prognostic factors are identified. Our models show high accuracy and consistency. Identified prognostic factors echo previous findings of clinical studies. Our method is accurate, interpretable, and thus useable as additional evidence in tongue cancer management.

Our dataset was collected at Chang Gung Memorial Hospitals, Taiwan, from 2004 to 2013 by Tsai et al (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5162395/). Among 1712 patients, 1280 survive (74.77%), and 432 do not (25.23%), with survival recorded at follow-up time (median for all and surviving patients 2.88 and 3.66 years, respectively). We treat survival as positive and non-survival as negative in our analysis. The original dataset contains 12 clinical and demographic variables. Area and occurrence of operation are the same for all subjects and therefore do not contribute to classification. We exclude these variables along with follow-up time, which is not known prior to patient survival. Eight variables are used for prediction: tumor stage, T stage, N stage, tumor grade, radiation therapy, chemotherapy, gender, and age.

Tumor stage and grade are assigned according to the American Joint Committee on Cancer's (AJCC) classification of malignant tumors. Radiation therapy is a binary variable, indicating the administration of ionizing radiation to control or kill malignant tumor cells. Chemotherapy is also a binary variable, indicating a regimen of one or more anti-cancer drugs.

METHODS

Patient survival prediction is a binary classification task. We utilize k-nearest neighbors (kNN), random forest, extreme gradient boosting (XGBoost), logistic regression with a L_2 penalty (logistic LASSO regression), and an ensemble of these four models. We evaluate model performance as accuracy, precision, recall, true negative rate (TNR), and area under the precision-recall curve (AUPRC).

We calculate the five measurements of each model by five-fold cross-validation and perform a grid search to finetune hyperparameters in the four individual models. All performance measurements are calculated using the hyperparameter combinations with highest AUPRC under five-fold cross-validation. Optimal hyperparameters are summarized in Table 1.

Table 1. Optimal hyperparameters for the four individual models. Each value is determined by five-fold cross validation with AUPRC as the optimization criterion.

Model	Hyperparameter	Optimal Values	Description
kNN	k	85	# of neighbors
Random forest	ntree	100	# of trees
	mtry	1	# of variables sampled
	nodesize	3	# of terminal observations
XGBoost	nrounds	7	Maximum # of iterations
	max.depth	2	Tree depth
	eta	0.5	Learning rate
LASSO	lambda	0.02	Shrinkage coefficient

RESULTS

Overall prediction performance. Table 2 summarizes the accuracy, AUPRC, precision, recall, and TNR of our five models. Among the four individual models, XGBoost achieves the highest accuracy (0.7664), AUPRC (0.8802), precision (0.7855), and TNR (0.2335). Random forest outperforms others in recall (0.9752). Among all models, the ensemble model achieves the highest AUPRC and close-to-top performance in terms of accuracy, precision, and recall. The leading performance of XGBoost and random forest demonstrates the strong nonlinear relationships between patient survival and other variables.

Table 2. Five measurements of model prediction performance. Each measurement is calculated by five-fold cross-validation. The highest values among the five models are underscored.

Model	Accuracy	AUPRC	Precision	Recall	TNR
kNN	0.7523	0.8726	0.7658	0.9652	0.1236
Random forest	0.7593	0.8791	0.7675	0.9752	0.1237
XGBoost	0.7664	0.8802	0.7855	0.9463	0.2335
LASSO	0.7553	0.8752	0.7820	0.9347	0.2257
Ensemble	0.7605	0.8855	0.7719	0.9658	0.1538

Flexibility between positive and negative predictions. The overall performance in Table 2 shows that all models have high recalls (above 0.9) but low TNRs (below 0.3), indicating an imbalance in predicting positive and negative patient. To examine the models' flexibility between positive and negative predictions, we adjust the probability cut-off from 0.5 to 0.9, with a step size of 0.1, and then calculate the corresponding precisions, recalls, and TNRs, respectively (Table 3). Under larger cut-offs, models predict fewer positive patients and more negative patients, resulting in lower recalls but higher TNRs. Model users can choose appropriate cut-offs based on their interest in positive or negative predictions.

Table 3. Three measurements of model performance under different cut-offs. The probability cut-off of positive patients is adjusted from 0.5 to 0.9. Then corresponding precisions, recalls, and TNRs are calculated for each model.

		Cut-off				
Model	Measurement	0.5	0.6	0.7	0.8	0.9
kNN	Precision	0.7658	0.7962	0.8355	0.8657	0.9013
	Recall	0.9652	0.8918	0.7853	0.6254	0.3603
	TNR	0.1236	0.3210	0.5406	0.7126	0.8836
Random forest	Precision	0.7675	0.7896	0.8094	0.8336	0.8562
	Recall	0.9752	0.9394	0.8948	0.8400	0.7548
	TNR	0.1237	0.2570	0.3725	0.5004	0.6208
XGBoost	Precision	0.7855	0.8277	0.8491	0.8779	0.9398
	Recall	0.9463	0.8570	0.7557	0.6650	0.0399
	TNR	0.2335	0.4677	0.6005	0.7229	0.9910
LASSO	Precision	0.7820	0.8212	0.8549	0.8720	0.0000
	Recall	0.9347	0.8711	0.7596	0.6790	0.0000
	TNR	0.2257	0.4359	0.6165	0.7023	1.0000
Ensemble	Precision	0.7719	0.8113	0.8339	0.8638	0.9062
	Recall	0.9658	0.8954	0.8227	0.7158	0.3380
	TNR	0.1538	0.3808	0.5122	0.6624	0.8963

Performance uncertainty measurement. We estimate the uncertainty of performance estimates by bootstrapping. Figure 1 presents visual comparisons of measurement distributions for each model. Table 4 includes 95% confidence intervals and means of performance measurements from bootstrapping. The five models show a slightly different asymptotical performance ranking compared with their point estimations. Logistic LASSO regression has the highest median accuracy, precision, and TNR. The ensemble model outperforms other models in median AUPRC and recall. We observe less performance variation in logistic LASSO regression and the ensemble model, an expected result given the stable model structure of logistic LASSO regression and the diverse model components of the ensemble model.

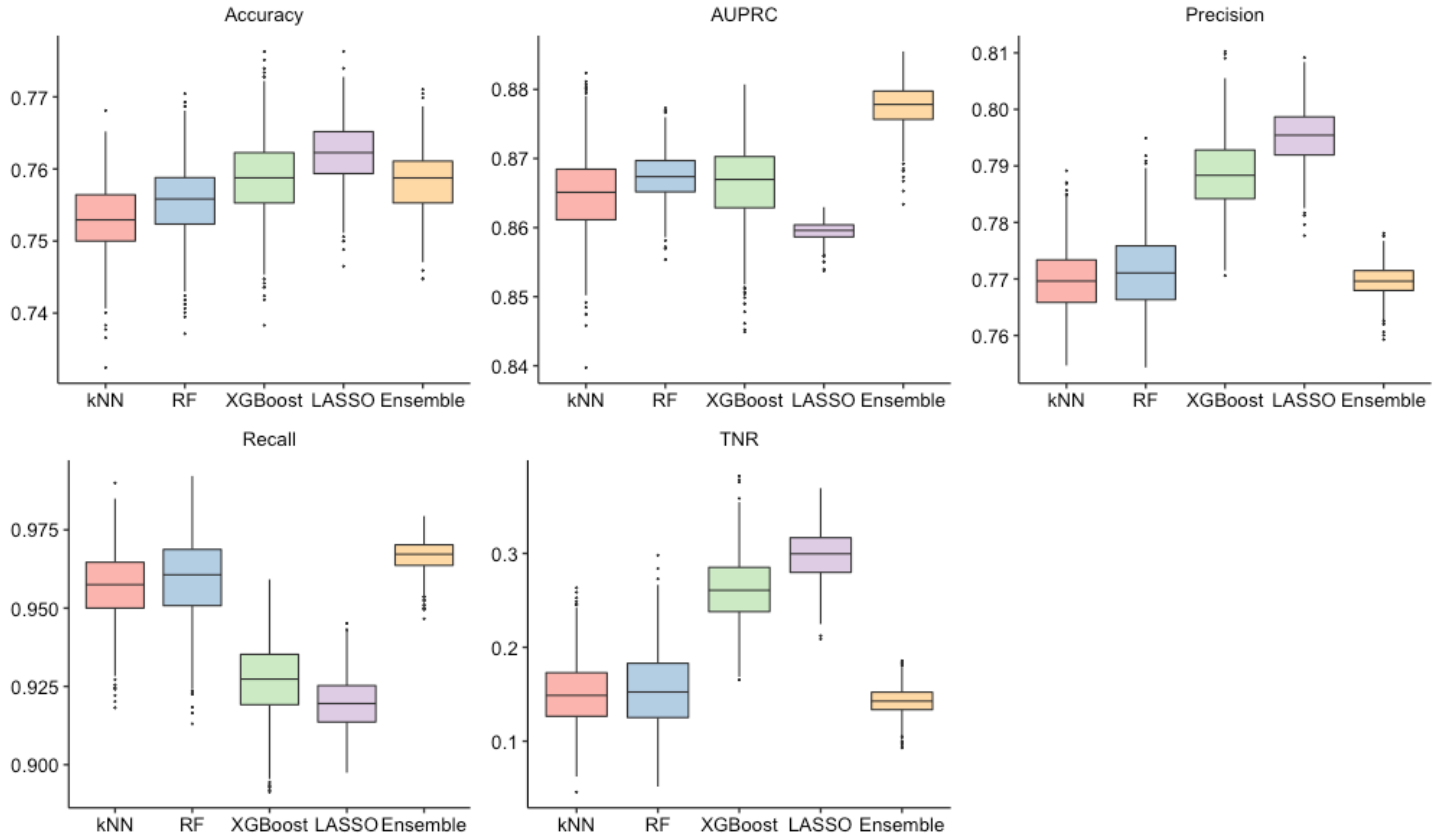

Figure 1. The empirical distributions of model performance computed by bootstrapping. Five performance measurements obtained from 1000 bootstrap iterations.

Table 4. Summary statistics of model performance calculated by bootstrapping. The empirical 95% confidence intervals and means of five measurements are calculated for each model.

Model	Measurement	Accuracy	AUPRC	Precision	Recall	TNR
kNN	95% CI	(0.7430, 0.7623)	(0.8533, 0.8752)	(0.7599, 0.7814)	(0.9338, 0.9776)	(0.0886, 0.2224)
	Mean	0.7531	0.8649	0.7698	0.9570	0.1504
Random forest	CI	(0.7465, 0.7658)	(0.8605, 0.8737)	(0.7593, 0.7855)	(0.9320, 0.9818)	(0.0815, 0.2403)
	Mean	0.7558	0.8674	0.7713	0.9593	0.1549
XGBoost	CI	(0.7477, 0.7693)	(0.8544, 0.8771)	(0.7773, 0.8011)	(0.9024, 0.9496)	(0.1974, 0.3312)
	Mean	0.7588	0.8665	0.7886	0.9270	0.2623
LASSO	CI	(0.7535, 0.7699)	(0.8568, 0.8620)	(0.7851, 0.8042)	(0.9045, 0.9360)	(0.2438, 0.3440)
	Mean	0.7622	0.8595	0.7951	0.9197	0.2973
Ensemble	CI	(0.7500, 0.7652)	(0.8711, 0.8834)	(0.7639, 0.7752)	(0.9569, 0.9757)	(0.1130, 0.1732)
	Mean	0.7583	0.8776	0.7697	0.9668	0.1431

Feature importance analysis. We utilize permutation feature importance to measure the contribution of each variable to predictive performance. Permutation feature importance is the decrease in AUPRC when a model predicts on a test set with one variable permuted. Because permutation breaks the relationship between variables and patient survival, a subsequent decrease in AUPRC indicates model dependency on that variable for prediction. For each variable, we average its permutation feature importance across the five models. We then divide those averages by the largest importance among all variables to obtain the normalized permutation feature importance.

Figure 2 shows variable ranking from most important to least important as normalized permutation feature importance. Tumor grade contributes the most to the prediction of patient survival, followed by N stage, T stage, chemotherapy, and radiation therapy. These variables are consistent with previous findings in clinical and modeling studies. For example, histological grading and the TNM staging system (i.e., tumor grade, N stage, and T stage) are well-established prognostic factors in oral cancer diagnosis. Numerous studies also suggest the importance of adjuvant therapy (i.e., chemotherapy and radiation therapy), especially for patients in advanced stages.

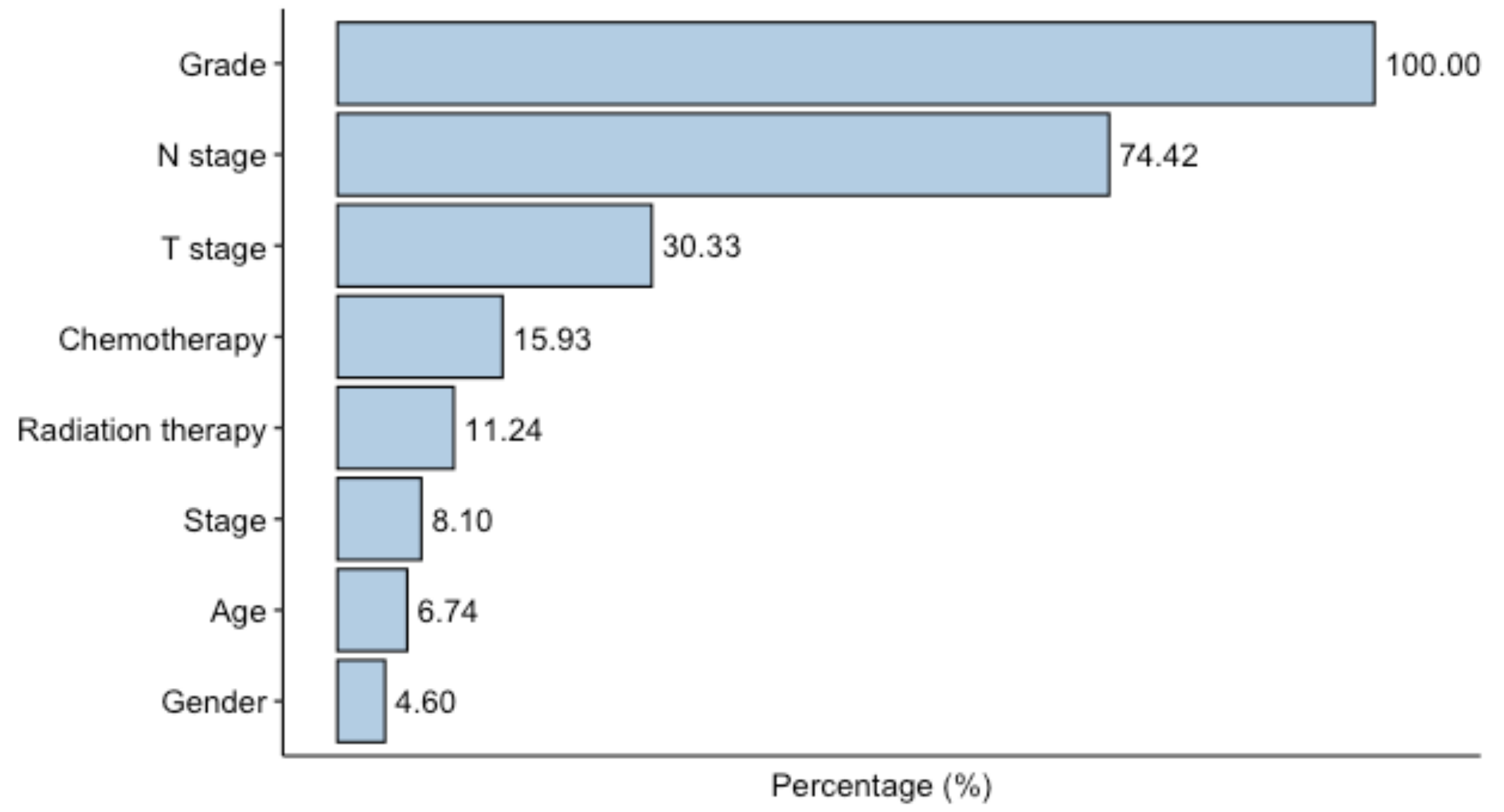

Figure 2. The normalized permutation feature importance. Variables are sorted from highest to lowest normalized importance.

CONCLUSION

The nonlinear models, XGBoost and random forest, exhibit greater overall accuracy. The linear model, logistic LASSO regression, provides more stable prediction in bootstrap analysis. The ensemble model improves accuracy and stability, incorporating the strengths of individual models. By adjusting the probability cut-off, our models offer flexibility in predicting positive and negative patients. Feature importance analysis identifies key predictors consistent with previous findings in clinical and modeling studies. The models show satisfactory performance, with average accuracy and AUPRC 0.7588 and 0.8785, respectively.