Advances in Artificial Intelligence and Machine Learning; Research 3 (1) 853-867 Received 19-01-2023; Accepted 15-03-2023; Published 22-03-2023

Predicting Survival of Tongue Cancer Patients by Machine Learning Models

Angelos Vasilopoulos avasilopoulos1@luc.edu
Department of Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660, USA

Nan Miles Xi mxi1@luc.edu
Department of Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660, USA

Corresponding Author: Nan Miles Xi

Abstract

Tongue cancer is a common oral cavity malignancy that originates in the mouth and throat. Much effort has been invested in improving its diagnosis, treatment, and management. Surgical removal, chemotherapy, and radiation therapy remain the major treatment for tongue cancer. The treatment effect is determined by patients' survival status. Previous studies have identified certain survival and risk factors based on descriptive statistics, ignoring the complex, nonlinear relationship among clinical and demographic variables. In this study, we utilize five cutting-edge machine learning models and clinical data to predict the survival of tongue cancer patients after treatment. Five-fold cross-validation, bootstrap analysis, and permutation feature importance are applied to estimate and interpret model performance. The prognostic factors identified by our method are consistent with previous clinical studies. Our method is accurate, interpretable, and thus useable as additional evidence in tongue cancer treatment and management.

Keywords: Tongue cancer, Machine learning, Survival prediction, Prognostic Factors, Cancer treatment

1. INTRODUCTION

Tongue cancer is one of the most frequent head and neck malignancies. According to the American Cancer Society, tongue cancer is diagnosed in approximately 20,000 patients and causes more than 2,700 deaths annually in the United States [1]. The average age of diagnosis is around 63, and 20% of cases occur in patients younger than 55 [2]. The overall rate of new cases has risen in the last 20 years due to smoking, drinking alcohol, and human papillomavirus infection, the three major risk factors [3]. The clinical community groups tongue cancer into two types based on its location: oral

Citation: Angelos Vasilopoulos and Nan Miles Xi. Predicting Survival of Tongue Cancer Patients by Machine Learning Models Advances in Artificial Intelligence and Machine Learning. 2023;3(1):53.

cancer, beginning in the front two-thirds of the tongue, and oropharyngeal cancer, beginning at the back third of the tongue [3]. Tongue cancer typically originates in the squamous cells that line the tongue's surface, and the types of cells affected may determine prognosis and treatment [4]. Tongue cancer treatment primarily involves surgical removal, chemotherapy, and radiation therapy [5]. The 5-year relative survival rate after treatment was 68.8% between 2012 and 2018 [6].

Accurate survival prediction is crucial for the treatment design and management of tongue cancer given its wide occurrence. In recent years, machine learning has been successfully applied in related medical fields, including cancer therapy [7], drug development [8, 9], and precision medicine [10]. Machine learning is potentially effective in predicting survival from patient data and identifying important factors in treatment. However, the literature has few machine learning models dedicated for tongue cancer survival. Although some studies have identified certain survival and risk factors, the conclusions are usually based on descriptive statistics and linear models, ignoring the complex, nonlinear relationship among clinical and demographic variables [2, 11, 12].

In this paper, we utilize a comprehensive machine learning framework to predict tongue cancer survival after treatment. The analysis is performed on a real clinical dataset containing information on 1712 patients receiving curative tongue cancer surgery. We train five cutting-edge machine learning models on this dataset and provide unbiased prediction performance by five-fold cross-validation. We further quantify the uncertainty of model performance by bootstrap analysis. The important prognostic factors in model prediction are identified by permutation feature importance. We also utilize three sampling schemes to generate data with balanced patients' survival status and examine its effects on model performance. Overall, the proposed machine learning models show high accuracy in predicting patient survival. The prediction is consistent in terms of point estimation and uncertainty measurement. The identified prognostic factors echo previous findings in clinical studies. Our method is accurate, interpretable, and thus useable as additional evidence in tongue cancer treatment and management.

The paper is outlined as follows. Section two describes the dataset used in this study and introduces data cleaning and preprocessing. Section three summarizes the machine learning models and prediction measurements. Section four shows the main results from five perspectives. Section five concludes the study and discusses potential future work.

2. DATASET AND PREPROCESSING

In this study, we analyze a dataset collected from 1712 tongue cancer curative surgery recipients at Chang Gung Memorial Hospitals, Taiwan, from 2004 to 2013 [2]. Among all patients, 1280 survive after surgery (74.77%), and 432 fail to do so (25.23%). We treat survival as positive and non-survival as negative in our analysis. Survival information is recorded at follow-up time for each patient. The follow-up time is the period from the cancer diagnosis until death or the last follow-up visit. The median follow-up time of all patients is 2.88 years. The exact death causes of each patient are not provided in the datasets due to privacy concerns. In the original data collection, patients with poor performance status (ECOG score ≥ 3 [13]), end-stage renal disease, Child-Pugh C liver cirrhosis, and heart or lung malfunction are removed to reduce the impact of factors besides tongue cancer. The original dataset contains 12 variables of patients' medical records and demographic information. The area of operation and occurrence of operation are the same for all subjects and therefore do not

Table 1: Summary statistics of all variables used in this study.

Variable	**Value**	**Percentage**
Age (years)		
Mean ± SD	51.83 ± 11.29	
Range	21 – 92	
Gender		
Male	1504	87.85%
Female	208	12.15%
Stage		
1	611	35.69%
2	406	23.71%
3	233	13.61%
4	1	0.06%
4A	453	26.46%
4B	6	0.35%
4C	2	0.12%
T stage		
1	657	38.38%
2	629	36.74%
3	156	9.11%
4	270	15.77%
N stage		
0	1202	70.21%
1	180	10.51%
2	327	19.10%
3	3	0.18%
Grade		
1	516	30.14%
2	1033	60.34%
3	163	9.52%
Radiation therapy		
Yes	612	35.75%
No	1100	64.25%
Chemotherapy		
Yes	469	27.39%
No	1243	72.61%
Survival		
Yes	1280	74.77%
No	423	25.23%

contribute to classification. We exclude these variables from analysis along with follow-up time, which is not known prior to patient survival. Eight variables are used to predict patient survival, including tumor stage, T stage, N stage, tumor grade, radiation therapy, chemotherapy, gender, and age. The summary statistics of all variables are shown in TABLE 1.

Tumor stage is assigned according to the American Joint Committee on Cancer's (AJCC) TNM classification of malignant tumors [14]: in stage 0, tumors have not grown or spread; in stage 1, tumors are small and have not spread; in stage 2, tumors have grown but not spread; in stage 3, tumors are larger and have spread to surrounding tissue or lymph nodes of the immune system; in stages 4A, 4B, and 4C, tumors are larger and have spread to at least one other body organ (secondary or metastatic cancer). T stage is the size of a tumor, labeled by numbers one (small) to four (large). N stage indicates whether cancer has metastasized to lymph nodes, labeled by numbers zero (no metastasis to lymph nodes) to three (metastasis to multiple lymph nodes). Tumor grade is associated with the rate of cancer metastasis to other organs, represented by grades one (cancer cells resemble normal cells and are not proliferating) to three (cancer cells look abnormal and spread aggressively). 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 multiple anti-cancer drugs.

3. METHODS

In machine learning terminology, the prediction of patient survival is a binary classification task. We utilize k-nearest neighbors (kNN) [15], random forest [16], extreme gradient boosting (XGBoost) [17], logistic regression with a L_2 penalty (logistic LASSO regression) [15], and an ensemble of these four models to tackle this task using the dataset described in the last section.

k-nearest neighbors (kNN). kNN classifies patient survival according to their distance from neighbors. For one patient with a variable vector X, its binary survival $Y(X)$ is predicted as

$$Y(X) = \begin{cases} 1\ (survial) & if\ \frac{1}{k} \sum\limits_{X_i \in N_k(X_i)} Y_i\ (X_i) \geq 0.5 \\ 0\ (non-survial) & otherwise \end{cases}$$

where $N_k\ (X_i)$ is the neighborhood of that patient defined by the k closest patients. The distance between patients is calculated by the Euclidean distance in terms of patients' variables. We implement kNN using the function `knn` in R package `class`.

Random forest. Random forest classifies patient survival according to the most frequent outcome of a set of decision trees. A decision tree assigns each patient to one class based on split rules defined on the variable space. Suppose that there are P variables $X_1, X_2, \ldots, X_P$ in the dataset, and we split the variable space into two regions, R_1 and R_2, according to variable X_t and threshold s:

$$R_1(x, t, s) = \{x | X_t \leq s\}$$
$$R_2(x, t, s) = \{x | X_t > s\}$$

where x denotes patients. Then for any region R_m with N_m patients, let $\hat{p}_{mr}$ be the proportion of class r in region R_m:

$$\hat{p}_{mr} = \frac{1}{N_m} \sum_{x_i \in R_m} I\ (y_i = r)$$

where x_i and y_i are the variable vector and survial of patient i, respectively. $I(x)$ is the indicator function. The survival of patient x in region R_m is predicted as:

$$Y(x) = argmax_r\ \hat{p}_{mr},$$

where $r \in \{1, 0\}$. In each split generating regions R_1 and R_2, we seek the variable X_t and threshold s by solving the following optimization problem:

$$min_{t,s}\left[\sum_{x_i \in R_1(t,s)} L\left(Y\left(x_i\right),\, y_i\right) + \sum_{x_l \in R_2(t,s)} L\left(Y\left(x_l\right),\, y_l\right)\right]$$

where $L(x,\, y)$ is the misclassification error. The splitting process continues until it satisfies certain stopping rules, usually the maximum number of splits (tree depth) or the minimum number of observations per region (leaf size). To build a random forest from multiple decision trees, we generate bootstrap samples and apply the previous splitting rule to build one decision tree for each bootstrap sample. Instead of searching all P variables, we randomly selected $\sqrt{P}$ variables in each split to reduce the correlation among different decision trees. The prediction of random forest is the majority vote of all decision trees. We implement random forest using the function `randomForest` in R package `randomForest`.

Extreme gradient boosting (XGBoost). Similar to random forest, XGBoost assigns patient survival according to the outcomes of multiple decision trees. Different from random forest, decision trees are grown sequentially with each tree gradually reducing the misclassification errors to avoid overfitting. We implement XGBoost using the function `xgboost` in R package `xgboost`.

Logistic LASSO regression. Logistic LASSO regression is based on regular logistic regression, in which one patient's survival probability is modeled by:

$$P(Y = 1) = \frac{1}{1 + e^{-X\beta}}$$

where X is the patient's variable vector and β is the vector of model parameters. The model parameter vector β is estimated by maximum likelihood estimation. With the estimated model, the patient survival is calculated by:

$$Y = \begin{cases} 0 & if\ P(Y = 1) < 0.5 \\ 1 & if\ P(Y = 1) \geq 0.5 \end{cases}$$

Logistic LASSO regression regularizes logistic regression by introducing the L_2 norm of model parameter β into the likelihood function, which aims to reduce overfitting and accomplish feature selection. We implement logistic LASSO regression using the function `glmnet` in R package `glmnet`.

Ensemble model. An ensemble model is a combination of the aforementioned four models, i.e., the survival probabilities output by four models are averaged before conversion to patient survival status. Suppose that one patient's survival probabilities output by kNN, random forest, XGBoost, and logistic LASSO regression are $P_{kNN}(Y = 1)$, $P_{RF}(Y = 1)$, $P_{XGBoost}(Y = 1)$, $P_{LASSO}(Y = 1)$, respectively. Then the patient's survival probability of ensemble model is

$$P_{ensemble}(Y = 1) = \frac{P_{kNN}(Y = 1) + P_{RF}(Y = 1) + P_{XGBoost}(Y = 1) + P_{LASSO}(Y = 1)}{4}$$

Same as other models, the patient survival is predicted by:

$$Y = \begin{cases} 0 & if\ P(Y = 1) < 0.5 \\ 1 & if\ P(Y = 1) \geq 0.5 \end{cases}$$

Table 2: **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	Number of neighbors
Random forest	ntree	100	Number of trees
	mtry	1	Number of variables sampled
	nodesize	3	Minimum observations in a terminal node
XGBoost	nrounds	7	Maximum number of iterations
	max.depth	2	Tree depth
	η	0.5	Learning rate
Logistic LASSO regression	λ	0.02	Shrinkage coefficient

We refer to the four non-ensemble models as individual models moving forward.

Prediction performance measurement. We evaluate the model performance by six measurements: accuracy, precision, recall, true negative rate (TNR), balanced accuracy, and area under the precision-recall curve (AUPRC). Accuracy is the ratio of true predictions to the total number of patients; precision is the ratio of true positive predictions to all positive predictions; recall is the ratio of true positive predictions to all positive patients; TNR is the ratio of true negative predictions to all negative patients; balanced accuracy is the average between recall and TNR; AUPRC measures the overall capacity of a binary predictive model and adjusts for class imbalance in the data.

Cross-validation and hyperparameter tuning. We calculate the six measurements of each model by five-fold cross-validation. First, patients are split into five groups, or folds. Then, five different combinations of four folds are used as training sets in each of five successive cross-validation iterations, with the remaining fold in each iteration as a test set for performance assessment. Finally, we average performance measurements over five iterations. We also conduct a grid search to finetune hyperparameters in the four individual models. In each iteration of five-fold cross-validation, we train models with different hyperparameter combinations on the four folds of training set and calculate the AUPRC on the one fold of validation set. After cross-validation, we average the AUPRC across five validation sets to obtain the final performance for each hyperparameter combination. All model prediction measurements in the following analysis are calculated using the hyperparameter combinations with the highest AUPRC under five-fold cross-validation. The hyperparameters we finetune for each model and their optimal values are summarized in TABLE 2. It should be noted that the hyperparameter tuning could slightly cause the overestimation of model predictive power due to data leakage from training set to validation set. On the other hand, models with finetuned hyperparameters provide an upper bound of predictive power, showing how well machine learning models can achieve in predicting tongue cancer survival after surgery.

Table 3: **Six measurements of model prediction performance.** Each measurement is calculated by five-fold cross-validation using optimal hyperparameters. The highest values among the five models are underscored.

Model	Accuracy	AUPRC	Precision	Recall	TNR	Balanced accuracy
kNN	0.7523	0.8726	0.7658	0.9652	0.1236	0.5444
Random forest	0.7593	0.8791	0.7675	<u>0.9752</u>	0.1237	0.5495
XGBoost	<u>0.7664</u>	0.8802	<u>0.7855</u>	0.9463	<u>0.2335</u>	<u>0.5899</u>
Logistic LASSO regression	0.7553	0.8752	0.7820	0.9347	0.2257	0.5802
Ensemble	0.7605	<u>0.8855</u>	0.7719	0.9658	0.1538	0.5598

4. RESULTS

Overall prediction performance. TABLE 3 summarizes the accuracy, AUPRC, precision, recall, TNR, and balanced accuracy of the five models proposed in the last section. All measurements are calculated by five-fold cross-validation using the optimal hyperparameters in TABLE 2. Among four individual models, XGBoost achieves the highest accuracy (0.7664), AUPRC (0.8802), precision (0.7855), TNR (0.2335), and balanced accuracy (0.5899), indicating a solid overall capacity of differentiating between positive and negative patients. Random forest outperforms others in recall (0.9752), showing its strength in identifying surviving patients. Combining the four individual models as an ensemble model improves AUPRC over the best-performing dividual model. It also provides close-to-top performance in terms of accuracy, precision, and recall. Overall, the machine learning models show mixed performance in predicting patient survival. There is no single model dominating others in all six measurements. The leading performance of XGBoost and random forest demonstrate the strong nonlinear relationship between patient survival and other variables, which are largely ignored in previous studies. The ensemble model balances different measurements by utilizing the strengths of individual models.

FIGURE 1 shows the calibration curves of five models. The calibration curve visualizes the quality of the model's predicted probability by plotting the true frequency of positives against its predicted probability [18]. Specifically, the data are split into ten groups based on the positive probability output by each model. The x-axis represents the average predicted positive probability in each group and the y-axis is the fraction of true positives in each group, both ordered from 0.1 to 1 with a step size of 0.1. A well-calibrated model has a calibration curve close to the 45-degree diagonal line, representing consistency between predicted probability and true frequency of positives. Among the five models, the ensemble model and XGBoost exhibit the best calibration. Random forest produces many false negatives in the probability interval from 0.2 to 0.3, resulting in an early peaking in its calibration curve. All models' calibration curves start from 0.2 or larger values at the x-axis due to the lack of predicted positive probability below 0.2. In other words, models tend to predict more positives than negatives, which is also reflected by high precision and recall but low TNR in TABLE 3.

Flexibility between positive and negative predictions. The overall performance in TABLE 3, shows that all models have high recalls (above 0.9) but low TNRs (below 0.3), indicating an imbalance in predicting positive and negative patients. The reason is that all models, by default, use 0.5 as the probability cutoff for assigning binary outcome labels. A patient with a positive probability

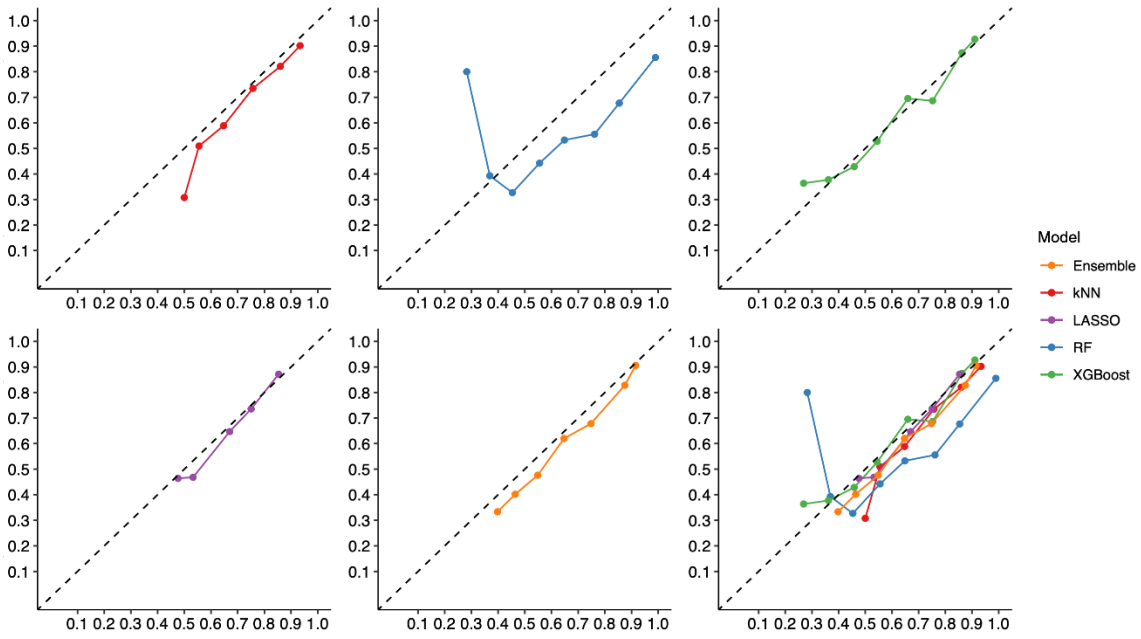

Figure 1: **The calibration curves of five models.** The data are split into ten groups based on predicted probabilities. The x-axis represents the average predicted positive probability in each group. The y-axis is the fraction of true positives in each group. The dashed line is the 45-degree diagonal line.

greater than 0.5 is predicted as positive (survival); otherwise, negative (not survival). The 0.5 cutoff forces the models to predict more positive patients than negative patients, resulting in low TNRs. To examine the models' flexibility between positive and negative predictions, we adjust the probability cutoff from 0.5 to 0.9, with a step size of 0.1, and then calculate the corresponding precisions, recalls, TNRs, and balanced accuracies, respectively (TABLE 4). Under larger cutoffs, models predict fewer positive patients and more negative patients, resulting in lower recalls but higher TNRs. For example, TABLE 4, shows that when the cutoff is changed from 0.5 to 0.9, the TNR of XGBoost increases from 0.2335 to 0.9910, and its recall decreases from 0.9463 to 0.0399. Model users can choose appropriate cutoffs based on their interest in positive or negative predictions. Another benefit of larger cutoffs is that they improve the accuracy of predicted positives, i.e., higher precision, through more cautious positive prediction. As the cutoff increases from 0.5 to 0.9, TABLE 4, shows that the precision of XGBoost increases from 0.7855 to 0.9398. We also observe that the 0.8 cutoff achieves the highest balanced accuracy for kNN, XGBoost, logistic LASSO regression, and ensemble model.

Prediction uncertainty measurement. The six measurements in TABLE 3, are point estimations of model performance. We further estimate the uncertainty of model prediction by bootstrapping. Specifically, we repeat five-fold cross-validation 1000 times, with training data resampled with replacement in every iteration. Each resampling is used to train one of the previous five models before assessing accuracy, AUPRC, precision, recall, TNR, and balanced accuracy on the test set in cross-validation. All models use their optimal hyperparameters from TABLE 2. We determine the model uncertainty from the resulting empirical distribution of five-fold measurements. FIGURE 2

Table 4: **Four measurements of model performance under different cutoffs.** The probability cutoff of positive patients is adjusted from 0.5 to 0.9. Then corresponding precisions, recalls, TNRs, and balanced accuracies are calculated for each model.

Model	**Measurement**	**Cutoff**				
		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
	Balanced accuracy	0.5444	0.6064	0.6630	0.6690	0.6220
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
	Balanced accuracy	0.5495	0.5982	0.6337	0.6702	0.6878
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
	Balanced accuracy	0.5899	0.6624	0.6781	0.6940	0.5155
Logistic LASSO regression	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
	Balanced accuracy	0.5802	0.6535	0.6881	0.6907	0.5000
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
	Balanced accuracy	0.5598	0.6381	0.6675	0.6891	0.6172

presents visual comparisons of measurement distributions for each model. TABLE 5 includes empirical 95% confidence intervals and means of performance measurements from bootstrapping. The five models show a slightly different asymptotical performance ranking compared with their point estimation in TABLE 2. Logistic LASSO regression has the highest mean accuracy, precision, TNR, and balanced accuracy. The ensemble model outperforms other models on mean AUPRC and recall. The performance differences among the five models are moderate in accuracy and AUPRC, the two overall measurements. However, the gaps are more significant in precision, recall, TNR, and balanced accuracy, indicating diverse model behavior in distinguishing positive and negative patients. We also observe less performance variation in logistic LASSO regression and ensemble model, an expected result given the stable model structure of logistic LASSO regression and the diverse model components of the ensemble model.

Feature importance analysis. We utilize the permutation feature importance [16] to measure the contribution of each variable to survival prediction. Permutation feature importance is the decrease of AUPRC when the model predicts 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 five models. We then divide those averages by

Table 5: **Summary statistics of model performance calculated by bootstrapping.** The empirical 95% confidence intervals and means of six measurements are calculated for each model. The highest values among the five models are underscored.

Model	**Measurement**	**Accuracy**	**AUPRC**	**Precision**	**Recall**	**TNR**	**Balanced accuracy**
kNN	95% CI	(0.7430, 0.7623)	(0.8533, 0.8752)	(0.7599, 0.7814)	(0.9338, 0.9776)	(0.0886, 0.2224)	(0.5307, 0.5798)
	Mean	0.7531	0.8649	0.7698	0.9570	0.1504	0.5537
Random forest	CI	(0.7465, 0.7658)	(0.8605, 0.8737)	(0.7593, 0.7855)	(0.9320, 0.9818)	(0.0815, 0.2403)	(0.5296, 0.5883)
	Mean	0.7558	0.8674	0.7713	0.9593	0.1549	0.5571
XGBoost	CI	(0.7477, 0.7693)	(0.8544, 0.8771)	(0.7773, 0.8011)	(0.9024, 0.9496)	(0.1974, 0.3312)	(0.5703, 0.6203)
	Mean	0.7588	0.8665	0.7886	0.9270	0.2623	0.5946
Logistic LASSO regression	CI	(0.7535, 0.7699)	(0.8568, 0.8620)	(0.7851, 0.8042)	(0.9045, 0.9360)	(0.2438, 0.3440)	(0.5876, 0.6270)
	Mean	0.7622	0.8595	0.7951	0.9197	0.2973	(0.6085)
Ensemble	CI	(0.7500, 0.7652)	(0.8711, 0.8834)	(0.7639, 0.7752)	(0.9569, 0.9757)	(0.1130, 0.1732)	(0.5408, 0.5681)
	Mean	0.7583	0.8776	0.7697	0.9668	0.1431	0.5549

the largest importance among all variables to obtain the normalized permutation feature importance. FIGURE 2 represents variable ranking from most important to least important in terms of 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 and treatment [19, 20]. Numerous studies also suggest the importance of adjuvant therapy (i.e., chemotherapy and radiation therapy), especially for patients in advanced stages [19, 21]. On the other hand, age and gender are among the least important variables in the model prediction, making less than 10% of the contribution of the top variable, tumor grade. These two variables are also found to have less impact on patient survival in clinical studies [22, 23].

The variable importance obtained by permutation feature importance needs to be cautiously interpreted due to its limitations. First, the permutation feature importance does not measure the variable necessity. Highly correlated variables produce positive feature importance because they are utilized jointly by machine learning models in prediction. However, the model performance likely remains the same after removing one of those variables from the training set. The models still obtain similar information from other highly correlated variables. Second, permutation feature importance tends to underestimate the importance of variables with moderate correlation. When one variable is permuted, the model still has access to that variable through its correlated counterparts. This will result in a lower decrease in prediction performance and thus importance values for both variables, even if they are actually important. Third, permutation feature importance is specific to the data used to calculate it. A different dataset may produce different feature importance for the same variable.

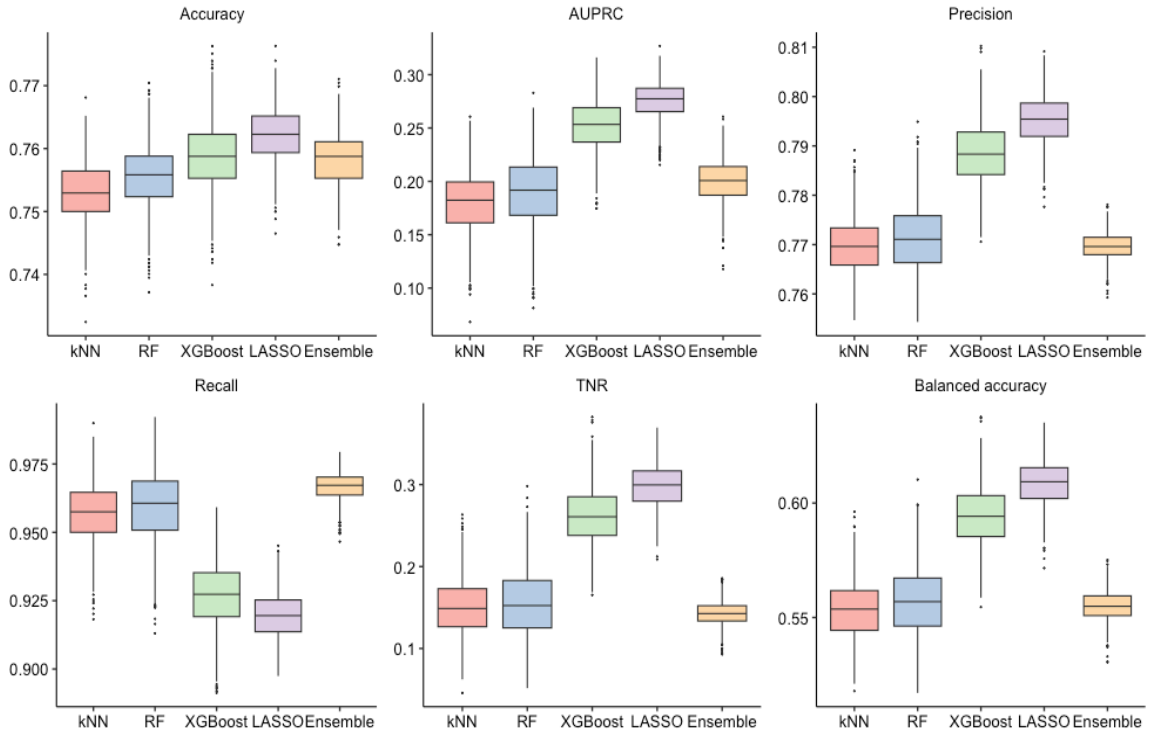

Figure 2: **The empirical distributions of model performance computed by bootstrapping.** Six performance measurements of the five models are obtained from 1000 bootstrap iterations.

Sampling schemes for predicting imbalanced data. One characteristic of the dataset used in this study is the imbalanced positive and negative classes. As shown in TABLE 1, the ratio between survival patients (positive) and non-survival patients (negative) is three to one. Such imbalance will bias the model prediction toward the majority positive class, which is reflected by the high precision and recall but low TNR in TABLES 3–5, and FIGURE 2. We adopt three sampling schemes, over-sampling, under-sampling, and hybrid sampling to mitigate the issue of imbalanced prediction. Over-sampling randomly duplicates data points of the minority class (negative) and adds them to the training set. Under-sampling randomly removes data points of the majority class (positive) from the training set. Hybrid sampling combines these two schemes by duplicating data points of the minority class and removing data points of the majority class simultaneously. All three schemes achieve equally sized positive and negative classes in the training set without changing the class distribution in the test set.

TABLE 6 compares the six prediction measurements of three sampling schemes and no sampling using random forest. All measurements are calculated by five-fold cross-validation with sampling schemes conducted in the training set of each iteration. We set the hyperparameters of random forest to their default setting in the R package randomForest. The default hyperparameters provide a fair comparison between sampling schemes and avoid overfitting, as suggested by previous studies [24]. We find that, compared with no sampling, all sampling schemes significantly improve the prediction of minority class (negative), reflected by higher TNRs. Although recalls are lower after

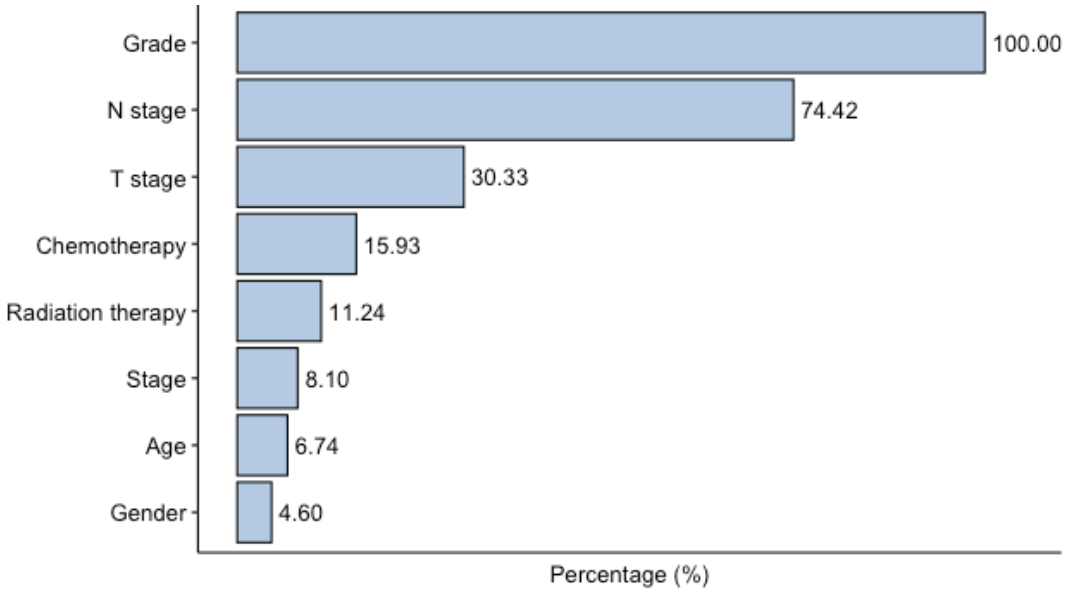

Figure 3: **The normalized permutation feature importance**. Variables are sorted from highest to lowest in terms of normalized importance.

Table 6: **Six measurements of model prediction performance under different sampling schemes.** Each measurement is calculated by five-fold cross-validation using random forest with default hyperparameters. The highest values among the five models are underscored.

Sampling scheme	Accuracy	AUPRC	Precision	Recall	TNR	Balanced accuracy
No sampling	<u>0.7623</u>	0.8719	0.7950	<u>0.9198</u>	0.2963	0.6081
Over-sampling	0.7115	0.8702	0.8489	0.7477	0.6020	<u>0.6748</u>
Under-sampling	0.6957	0.8648	<u>0.8522</u>	0.7178	<u>0.6272</u>	0.6725
Hybrid sampling	0.7030	<u>0.8770</u>	0.8470	0.7359	0.6019	0.6689

applying sampling schemes, the balanced accuracies are higher than no sampling, indicating the overall benefits of balanced training data. The three sampling schemes also improve the precision thanks to the more cautious prediction of the positive class. We observe no significant performance difference among the three sampling schemes.

5. CONCLUSIONS

In this study, we utilize machine learning models to predict the survival of tongue cancer patients after receiving curative surgery. Our models are built on a clinical dataset with 1712 patients. We use six measurements to provide a comprehensive evaluation of model performance. Although no individual model outperforms others in all measurements, the nonlinear models, i.e., XGBoost

and random forest, exhibit better overall accuracy. The linear predictive model, logistic LASSO regression, provides more stable prediction in bootstrap analysis. We also find that the ensemble model improves accuracy and stability by incorporating the strength of individual models. By adjusting the probability cutoff, our models offer flexibility in predicting positive and negative patients. Our feature importance analysis identifies key variables in predicting patient survival, consistent with previous findings in clinical and modeling studies. Overall, the machine learning models show satisfactory prediction performance. The average accuracy and AUPRC of the five models are 0.7588 and 0.8785, respectively. In practice, AUPRC greater than 0.8 indicates excellent discrimination between binary outcomes, especially given the imbalanced dataset in our study [25].

Several topics are worth exploring in future studies. First, all 1712 patients in the current dataset are from Taiwan. Collecting larger datasets from a more diverse population will increase the generality of machine learning models for new patients. Second, additional variables about patients' lifestyles and physical features, including smoking habits, body mass index, and occupational history, could contribute to creating superior models. Third, emerging genomic technology, e.g., single-cell RNA-sequencing (scRNA-seq), can be applied to reveal the transcriptomes of tongue cancer patients and identify molecular biomarkers [26–30]. Finally, our machine learning framework can evaluate the quality of clinical data in the survival diagnosis of tongue cancer. A high-quality dataset contains clinical information for models to accurately predict patient survival. The model prediction accuracy can serve as a proxy for the data quality of different datasets.

Data and Code Availability

The data used in this study is available at Zenodo repository:

`https://zenodo.org/record/7450476#.Y532FHaZMuJ`

The source code that implemented the result in this study is available at GitHub repository:

`https://github.com/angvasilop/tongue_cancer`

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72:7-33.

[2] Tsai MS, Lai CH, Lee CP, Yang YH, Chen PC, et al. Mortality in tongue cancer patients treated by curative surgery: a retrospective cohort study from CGRD. PeerJ. 2016;4:e2794.

[3] Gonzalez M, Riera March A. Tongue Cancer. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023

[4] Johnson DE, Burtness B, Leemans CR, Lui VW, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat. Rev. Dis. Primers. 2020;6:92

[5] Sultana J, Bashar A, Molla MR. New management strategies of oral tongue cancer in Bangladesh. J Maxillofac Oral Surg. 2014;13:394-400.

[6] Geum DH, Roh YC, Yoon SY, Kim HG, Lee JH, et al. The impact factors on 5-year survival rate in patients operated with oral cancer. J Korean Assoc Oral Maxillofac Surg. 2013;39:207-216.

[7] Rafique R, Islam SMR, Kazi JU. Machine learning in the prediction of cancer therapy. Comput Struct Biotechnol J. 2021;19:4003-4017.

[8] Xi NM, Hsu YY, Dang Q, Huang DP. Statistical learning in preclinical drug proarrhythmic assessment. J Biopharm Stat. 2022;32:450-473.

[9] Foster-Burns J, Xi NM. Prediction of drug-induced TdP risks using machine learning and rabbit ventricular wedge assay. Int J Multidiscip Res Anal. 2022;5:2862-2867.

[10] Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci. 2021;14:86-93.

[11] Geleijnse G, Chiang RC, Sieswerda M, Schuurman M, Lee KC, et al. Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure. Sci Rep. 2020;10:20526.

[12] Bai XX, Zhang J, Wei L. Analysis of primary oral and oropharyngeal squamous cell carcinoma in inhabitants of Beijing, China-a 10-year continuous single-center study. BMC Oral Health. 2020;20:208.

[13] Azam F, Latif MF, Farooq A, Tirmazy SH, AlShahrani S, et al. Performance status assessment by using ECOG (Eastern Cooperative Oncology Group) score for cancer patients by oncology healthcare professionals. Case Rep Oncol. 2019;12:728-736.

[14] Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, et al. AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more ”personalized” approach to cancer staging. CA Cancer J Clin. 8th ed. 2017;67:93-99.

[15] Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Springer. 2009.

[16] Breiman L. Random forests. Mach Learn. 2001;45:5-32.

[17] Chen T, Guestrin C. XG Boost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 785-794. Association for Computing Machinery. 2016:785-794.

[18] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on machine learning – ICML ’05. ACM Press. 2005:625-632.

[19] Boonpoapichart S, Punyavong P, Jenwitheesuk K, Surakunprapha P, Winaikosol K. Significant prognostic factors influencing the survival difference of oral tongue squamous cell carcinoma. Plast Reconstr Surg Glob Open. 2021;9:e3889.

[20] AlTuwaijri AA, Alessa MA, Abuhaimed AA, Bedaiwi RH, Almayouf MA, et al. Lymph node yield as a prognostic factor in clinically node negative oral cavity squamous cell carcinoma. Saudi Med J. 2021;42:1357-1361.

[21] Zanoni DK, Montero PH, Migliacci JC, Shah JP, Wong RJ, et al. Survival outcomes after treatment of cancer of the oral cavity (1985-2015). Oral Oncol. 2019;90:115-121.

[22] Suresh GM, Koppad R, Prakash BV, Sabitha KS, Dhara PS. Prognostic indicators of oral squamous cell carcinoma. Ann Maxillofac Surg. 2019;9:364-370.

[23] Garavello W, Spreafico R, Somigliana E, Gaini L, Pignataro L, et al. Prognostic influence of gender in patients with oral tongue cancer. Otolaryngol Head Neck Surg. 2008;138:768-771.

[24] Bodenhofer U, Haslinger-Eisterer B, Minichmayer A, Hermanutz G, Meier J. Machine learning-based risk profile classification of patients undergoing elective heart valve surgery. Eur J Cardiothorac Surg. 2021;60:1378-1385.

[25] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861-874.

[26] Xi NM, Li JJ. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. Star Protoc. 2021;2:100699.

[27] Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst. 2021;12:176-194.e6.

[28] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50:1-14.

[29] Baldini E, Tuccilli C, Prinzi N, Sorrenti S, Falvo L, et al. Deregulated expression of Aurora kinases is not a prognostic biomarker in papillary thyroid cancer patients. PLOS ONE. 2015;10:e0121514.

[30] Baldini E, Tuccilli C, Pironi D, Catania A, Tartaglia F, et al. Expression and clinical utility of transcription factors involved in epithelial-mesenchymal transition during thyroid cancer progression. J Clin Med. 2021;10.