의학석사 학위논문

단일 기관 레지스트리를 활용한 머신 러닝 기반 두경부암 환자의 생존 분석

Machine Learning-based Survival Analysis of
Head and Neck Cancer Patients
from a Single-Institution Cancer Registry

울산대학교 대학원
의 학 과
이 영 주

단일 기관의 레지스트리를 활용한 머신 러닝 기반 두경부암 환자의 생존 분석

지도교수 최 승 호

이 논문을 의학석사 학위 논문으로 제출함.

2024 년 2 월

울산대학교 대학원

의 학 과

이 영 주

이영주의 의학석사 학위 논문을 인준함.

심사위원 최 승 호 (인)

심사위원 정 영 호 (인)

심사위원 이 윤 세 (인)

울 산 대 학 교 대 학 원

2024 년 2 월

영문 요약

Background

Accurate prognosis estimation for patients with head and neck (H&N) cancer is crucial for clinical decision-making. The aim of this study was to identify an effective machine learning (ML) model for predicting the 5-year survival of H&N cancer patients.

Methods

We reviewed the records of 3,019 patients from the Asan Medical Center H&N Cancer registry collected between 2007-17. The feature set used to compare the performance of various ML algorithms comprised demographic characteristics, past and social history, primary site clinical and histopathologic attributes, and treatment modalities. We applied a total of five ML models to the dataset to classify H&N cancers based on their 5-year survival status. These models included two recently developed gradient-boosting models (XGBoost and LightGBM) and three commonly used tree-based models, Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB). We implemented 10-fold cross validation to measure model performance.

Results

After exclusion, the final study population comprised 1,287 patients. Of the five models, LightGBM showed the best performance. We evaluated model performance using four metrics: sensitivity, accuracy, F1 score, and AUC-ROC. The average model performance scores were as follows: SVM, 0.745; RF, 0.823; NB, 0.686; XGBoost, 0.827; and LightGBM, 0.839. The SHapley Additive exPlanations (SHAP) values were then calculated for the LightGBM model, and these results indicated that the top six (of 16) most important features were, in descending order: age, body mass index (BMI), Primary site, Clinical N stage, Overall stage, and Clinical T stage.

Conclusion

In this study, we found that features associated with staging were the most important for predicting H&N cancer survivability. The LightGBM model, combined with an appropriate dataset, can be used to construct an accurate prognostic model for patients diagnosed with H&N cancer. This can be applied in clinical practices in the future.

차 례

Table Contents

Figure Contents

INTRODUCTION

Head and neck (H&N) cancer comprises a diverse group of diseases, including cancers of the oral cavity, pharynx, larynx, and other tissues. It ranks as the seventh most prevalent cancer globally, with over 660,000 new diagnoses and approximately 300,000 fatalities every year[1]. Moreover, the global incidence of H&N cancer is increasing. Due to its location, H&N cancer greatly affects patient quality of life, and its treatment is often complex. For localized cancer cases, the 5-year survival rate is 85.1%, whereas for distant cases, the rate drops to 40.1%[2].

Over the past few decades, machine learning has played an increasingly prominent role in many scientific fields, including medicine[3]. ML methods have been proven effective in predicting the susceptibility, recurrence, and survival rates of various types of malignancies[4]. Moreover, several H&N research studies have performed survival prediction using ML algorithms. For example, Alabi *et al.* analyzed the performance of ML models in predicting overall survival for tongue cancer patients using the National Cancer Institute Database[5], and Adeoye *et al.* conducted a study on survival prediction using various ML algorithms to accurately predict survival of patients with oral cavity cancers[6]. Recently, the utility of ML in predicting 2-year survival has been reported in patients with cancers of the oral cavity, oropharynx, nasopharynx, hypopharynx, and larynx who underwent radiotherapy[7].

A better understanding of 5-year survival predictions can help clinicians and patients to

make informed treatment choices, ultimately leading to improved patient outcomes. In this study, we analyzed ML algorithm performance in predicting 5-year survival rates of H&N cancer patients using clinical features obtainable at the treatment modality decision point. We used five representative ML algorithms, ranging from established algorithms such as support vector machine and random forest algorithms to more recent gradient-boosting models[8]. Furthermore, we applied the best algorithm to each individual primary site and compared the results.

MATERIALS AND METHODS

Data and feature sets

We reviewed the records of 3,019 patients from the Asan Medical Center Head and Neck Cancer registry between 2007 and 2017. We included only representative types of H&N cancer (i.e., cancers of the oral cavity, oropharynx, hypopharynx, larynx, nasopharynx, salivary gland, and nasal cavity). Survival time was calculated on the basis of the end of treatment. Cases with a survival time of five years or more were classified as patients showing "Survival," while cases with a survival time of less than five years were classified as "Dead." Since our primary outcome was 5-year survival prediction, patients who were lost to follow-up before the 5-year period were excluded.

Feature sets included demographic variables (i.e., age, sex, and body mass index (BMI)), patient medical and social history (i.e., hypertension (HTN), Diabetes mellitus (DM), hepatitis, tuberculosis (TB), smoking status, and alcohol use), primary cancer site (i.e., oral cavity, oropharynx, hypopharynx, larynx, nasopharynx, salivary gland, or nasal cavity including the paranasal sinus), clinical and histopathologic attributes (i.e., T stage, N stage, M stage, differentiation, and overall stage), and treatment modalities (i.e., surgery, radiation therapy (RT), concurrent chemoradiation therapy (CCRT), and palliative chemotherapy). Feature sets were then used to compare the performance of different ML algorithms. The staging of the cancer was assessed in accordance with the 8th edition of the American Joint

Commission on Cancer (AJCC) TNM staging guidelines[9]. This study was conducted with the approval of the Institutional Review Board of Asan Medical Center (Approval No. 2023-0963)

Data pre-processing

Our datasets contained missing data, which we handled using data imputation (Supplemental Table 1). For the nearest neighbor algorithm predictions of unknown values used the values of the most similar neighbors[10]. For K-NN imputation, missing data are imputed using the K most similar neighbor values (where "K" represents the number of nearest neighbors). The missing values in each dataset were imputed using the most similar neighbor values from that dataset. Thus, we replaced the missing data with the value of the closest single neighbor (i.e., K = 1).

Since classification algorithms in ML are prone to overfitting when using unbalanced datasets, we used the synthetic minority oversampling technique-nominal continuous (SMOTE-NC) technique to balance our datasets[11]. Using this technique, the number of data points for the minority group ("Dead") in the training set was increased by creating new synthetic data.

Machine learning models

1.Support vector machine

The support vector machine (SVM) algorithm is one of the most commonly used in ML for classification tasks[12,13]. The SVM algorithm primarily aims to find the widest margin between two classes. This margin is defined as the distance between the closest data points from each class. These specific points are referred to as "support vectors," and serve to define the decision boundary. The implementation of the SVM in this study was performed using the scikit-learn package as implemented in Python.

2.Random forest

The random forest (RF) algorithm is a type of ensemble method commonly used in machine learning[14]. It operates by constructing multiple decision trees and controlling for overfitting by taking the average of their results. The three key steps in the Random Forest algorithm are bootstrap sampling, decision tree construction, and prediction. Bootstrap sampling involves randomly selecting samples from a given dataset. For each bootstrap sample, a decision tree is constructed. Instead of considering all features at each node, only a random subset of features is considered. Each tree's prediction is then treated as a "vote." The class that receives the most votes is selected as the final prediction. In this study, the RF algorithm was implemented using the scikit-learn package in Python.

3. Naive Bayes

The naïve Bayes (NB) algorithm is a probabilistic classification method based on Bayes'

theorem, and involves an assumption of independence between features. It is termed “naïve” due to the often unrealistic assumption that all features are conditionally independent given the class label[15]. Despite its simplicity, the model is particularly effective in text categorization tasks. Therefore, NB has traditionally been widely used for classification via ML. Here, NB was also implemented using the scikit-learn package in Python.

4.XGboost

XGBoost, which stands for “Extreme Gradient Boosting,” is a more advanced version of the Gradient Boosting algorithm[16]. This algorithm sequentially adds new models to correct errors from previous ones and includes regularization to prevent overfitting and enhance generalization. XGBoost is efficient, with fast computation and automatic handling of missing data, and uses tree pruning and built-in cross-validation routines to improve accuracy. Here, XGBoost models were built using the XGBoost package in Python.

5.LightGBM

LightGBM is a modern gradient-boosting machine that uses two innovative techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). These strategies accelerate the training speed of the boosting process. GOSS excludes a considerable fraction of data instances with small gradients to increase speed, whereas EFB groups together mutually exclusive features to minimize the total number of features, further

enhancing efficiency[17]. To use this algorithm we used the LightGBM package in Python.

Hyperparameter tuning

Hyperparameters are the configuration settings or variables that control the model training process and dramatically influence model performance. Since the selection of hyperparameter values can significantly impact a ML model's ability to learn and make accurate predictions, hyperparameter tuning is necessary to optimize model performance[18]. There are two conventional methods for conducting hyperparameter optimization: grid search and random search. Grid search optimizes hyperparameters via systematic exploration of all possible combinations of predefined values to find the best configuration. However, this exhaustive search can become computationally demanding, particularly when there are numerous hyperparameters or various potential values for each hyperparameter[19]. Random search involves the random selection of hyperparameters; however, this method may overlook important regions in the hyperparameter space, leading to suboptimal performance[20]. Unlike random and grid search, Bayesian optimization uses a probabilistic model to predict the performance of a specific set of hyperparameters before conducting an actual test. Bayesian optimization iteratively evaluates promising hyperparameter configurations based on prior information, including previous configurations and their corresponding loss of objective function. It then updates this information for subsequent iterations. This method allows Bayesian optimization to use the outcomes of previous

evaluations to guide the selection of the next set of hyperparameters. By using accumulated knowledge, this approach provides more direction to the search process and increases efficiency relative to the indiscriminate methods used by the random and grid search approaches. In addition, Bayesian optimization is capable of effectively handling continuous parameters, unlike grid search, which is limited to discrete parameters. Moreover, while a random search can handle continuous parameters, it does not use prior information, resulting in a less efficient process. In this study, we implemented Bayesian hyperparameter optimization using the bayes_opt package in Python. The set of hyperparameters that yielded the best performance across the 100 trials was then selected as the optimal set of hyperparameters for the model.

Model evaluation

In our study, we employed 10-fold cross validation to evaluate the performance of the ML models under consideration. We opted to use 10-fold cross validation since it can provide a comprehensive assessment of model performance and reduce the risk of model overfitting[17]. Therefore, the original dataset was first randomly partitioned into 10 equal-sized subsets. Of these, nine subsets were used for training the model, and the remaining subset was used for model testing the model. This process was repeated ten times (i.e., "10-fold"), with each of the ten subsets used exactly once as the test set. These results were then averaged to produce a single estimate. Figure 1 shows an overview of the testing process used for this

study.

Performance evaluation

The classification efficacy of the models in this study was assessed by metrics including sensitivity, accuracy, F1 score, and the area under the receiver operating characteristic curve (AUC).

1. Sensitivity (also known as Recall): the proportion of actual positive cases that are correctly identified.

Sensitivity = True Positives / (True Positives + False Negatives)

2. Accuracy: the ratio of correctly predicted instances to total instances within the dataset.

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

3. F1 Score: the harmonic mean of precision (i.e., the positive predictive value) and sensitivity.

*F1 Score = 2 * (Precision * Recall) / (Precision + Recall)*

Here,Precision is defined as: True Positives / (True Positives + False Positives)

4. AUC is a single metric that quantifies how well a model distinguishes between positive and negative classes across all threshold settings. Higher AUC values indicate better model performance, with 1.0 representing a perfect classifier and 0.5 indicating a model that performs no better than random chance.

Feature importance

Feature importance quantifies the influence of individual input features on the predictions of a ML model, and can be used to identify the most significant contributors. However, feature importance must be interpreted carefully; for this reason we used the SAP (SHapley Additive exPlanations) approach, which offers a sophisticated method, rooted in game theory, of measuring feature importance. Briefly, SHAP assigns importance scores to each feature for specific predictions, ensuring a consistent and fair distribution of contributions across potential predictions[21]. This approach improves local interpretability and elucidates how particular features impact individual predictions. Moreover, SHAP effectively captures complex feature interactions and can be used to generate intuitive visualizations. In this study, SHAP was implemented using the SHAP package in Python.

Our primary outcome was to identify the most suitable ML algorithm for predicting 5-year (overall) survival and to determine the importance of associated features. Subsequently, we applied the optimal algorithm to each site. The secondary outcome was to apply the optimal algorithm to each primary site and analyze the results.

RESULTS

Of all 3,019 patients included in the raw dataset, 1,287 were selected for inclusion in the study dataset (Figure 2). Based on survival at the 5-year point, of 1,287 patients 951 (73.9%) had “Survived” while 339 (26.1%) were “Dead.” Our data incorporated 16 features, categorized under five main headings: demographics, medical and social history, primary site, clinical and histopathologic attributes, and Initial Treatment Modality (Table 2). With respect to primary site, the Larynx was the most common, with 353 cases (27.4%), followed by the oral cavity with 291 cases (22.6%). Regarding treatment modality, surgery was the most prevalent, with 814 instances (63.2%).

Optimal hyperparameter values

Table 2 displays the optimal hyperparameters identified during the tuning process for the SVM, NB, RF, XGBoost, and LightGBM ML algorithms. We used various hyperparameters for each algorithm:

Support Vector Machine (SVM): Four parameters - C, gamma, degree, and kernel

Naïve Bayes (NB): One parameter - var_smoothing

Random Forest (RF): Four parameters: n_estimators, max_depth, min_samples_split, and min_samples_leaf.

XGBoost: Seven parameters: n_estimators, max_depth, gamma, min_child_weight,

subsample, colsample_bytree, and learning_rate.

LightGBM: Seven parameters: n_estimators, num_leaves, max_depth, min_child_weight, subsample, colsample_bytree, and learning_rate.

Model performances

Table 3 and Figure 3 present model performance measures derived from 10-fold cross validation, including the receiver operating characteristic (ROC) curve, AUC, sensitivity, accuracy, and F1 score. Based on average scores, LightGBM was the top-performing algorithm during the validation phase with a score of 0.839. This was followed by XGBoost (0.827), RF (0.823), SVM (0.745), and NB (0.686).

Feature importance

Figure 4 shows that the six features played a pivotal role in determining the performance of our ML models in predicting the survival outcomes of H&N cancer patients. These features included age, BMI, primary site, clinical N stage, overall stage, and clinical T stage.

Analysis at each primary site

Next, we used the top-performing Light GBM model to analyze each distinct primary site.

As shown in Table 4, this model's predictions were most accurate for 5-year survival prediction of salivary gland cancer patients. However, its predictions for 5-year survival of the hypopharynx and nasal cavity were less accurate than for other sites.

Analyzing by Primary Site

We analyzed the 5-year survival rate for each primary site. The Salivary gland show the highest survival rate at 0.84, followed closely by the Larynx at 0.83 and the Oropharynx at 0.80. Conversely, the Hypopharynx presented the lowest 5-year survival rate (0.57) (Supplement Table 2). The survival curves specific to each primary site can be viewed in Supplement Figure 1.

Upon evaluating the treatment modalities by primary site, it was observed that surgical treatments were predominant for both Oral cavity and Salivary gland, with over 90% of cases undergoing surgical intervention. For Hypopharynx, there was a near-even distribution between surgical treatments and CCRT (42.4% vs. 46.7%). In the case of Nasopharynx, a significant 76.8% of cases opted for CCRT as the primary treatment modality (detailed in Supplement Figure 2).

DISCUSSION

In this study, we found that ML algorithms exhibited high performance in predicting 5-year survival of H&N cancer patients. Notably, gradient-boosting models yielded outstanding results, with the Light GBM algorithm demonstrating the best performance. Among various primary sites, this model predicted 5-year survival most accurately for cancers of the salivary gland, whereas its performance was lower for cancers with the hypopharynx or nasal cavity as primary sites. The most critical feature for prediction was age, followed by BMI, primary site, clinical N stage, overall stage, and clinical T stage, all of which were significant predictive variables.

In this study, the gradient boosting and random forest models, two representative ensemble learning techniques, outperformed the other algorithms. The intrinsic strength of ensemble techniques is rooted in their ability to combine insights from multiple weak learners, culminating in a more robust and accurate predictive model[22]. Using this aggregation-based approach, the learning process inherently diversifies, concurrently reducing model bias and variance while increasing generalizability. Furthermore, Gradient Boosting uses a sequential learning strategy to correct prior errors, methodically enhancing its accuracy with each iteration. In contrast to the Random Forest approach, which constructs independent trees and aggregates their outputs afterward, Gradient-Boosting iteratively refines predictions by crafting trees that address the inaccuracies of predecessors.

This adaptability positions Gradient Boosting to concentrate on more challenging predictions, affording it a distinctive advantage in various situations. Among the evaluated models, the Light GBM model stood out as the top performer. Designed for expedited processing and efficiency, Light GBM constructs trees in a leaf-wise fashion, and can therefore potentially generate deeper and more precise trees than conventional level-wise tree development[17]. Furthermore, Light GBM's proficiency in managing sizable datasets and high cardinality categorical variables might have played a pivotal role in its performance in this study.

It is well known that managing missing and inconsistent data is crucial when dealing with medical data, particularly for clinical datasets. If the proportion of missing values exceeds a given threshold value (commonly set at 50%), it is generally advisable to delete those instances. However, for datasets with a relatively low percentage of missing data, the use of appropriate imputation methods is generally preferred[8]. Here, our dataset contained a mix of continuous and categorical variables. Given this composition, we opted for the Nearest Neighbor algorithm for imputation, using values from the most proximate neighbors.

In our dataset, the categories "Survived" and "Dead" were unbalanced, with respective proportions of 73.9% and 26.1% of the dataset. Given that ML classification is susceptible to overfitting in the presence of imbalanced datasets, it is imperative to seek equilibrium in the dataset using appropriate sampling techniques. The Synthetic Minority

Oversampling Technique (SMOTE) is a widely recognized resampling method that has been employed in numerous cancer prediction studies, including studies by Wang et al.[23], Santos et al.[24] and Doja et al.[25] By applying SMOTE, it is possible to enhance the representation of the minority class by creating novel synthetic instances. However, the SMOTE approach cannot handle nominal features; therefore, we used the Synthetic Minority Oversampling TEchnique-Nominal Continuous [SMOTE-NC][11]. This allowed us to avoid the danger of overfitting. In addition, we also used cross validation, resulting in significant improvements in both classification accuracy and the model's potential for generalization.

Moreover, it is also essential to adjust hyperparameters, train various models using diverse value combinations, and subsequently evaluate their performance to identify optimal ML model for highly specialized approaches. In the past, two representative methods have been widely used: grid search and random search[20]. Grid search systematically trains a ML model using every permutation of hyperparameter values on the training data, and its performance is then assessed based on preset metrics via cross validation. This means that the grid search approach can identify the specific hyperparameter set that delivers the best results. However, as the number and range of hyperparameters expands, the efficiency of this algorithm rapidly diminishes. In contrast, a random search explores random hyperparameter combinations within specified bounds. Although it operates more efficiently in high-dimensional spaces than the grid search

approach, the random search algorithm can sometimes generate inconsistent results when training intricate models. Bayesian optimization integrates a priori knowledge of a function with sampled data (evidence) to derive the function's posterior probability. This posterior is then used to determine where the function attains a maximum value based on a specific criterion[26]. Using Bayesian optimization for hyperparameter tuning is therefore often a more practical and highly effective approach. Before finalizing the model, this step is necessary since it significantly improves model performance.

In our study, age was identified as the most crucial factor for model prediction. Numerous studies have demonstrated a significant positive association between age and poorer patient survival outcomes[27,28]. Similarly, it has already been established that pretreatment BMI is closely related to survival rate[29] Moreover, clinical N stage was also found to be one of the most important features for predicting cancer survival. In numerous solid tumors, including H&N cancer, lymph node metastasis is widely recognized as a prognostic indicator[30]. For example, it has been established that the number of lymph nodes after neck dissection is directly correlated with prognosis[31]. Similarly, the T stage is also an important factor for assessing the prognosis of H&N cancer. Interestingly, within the tumor staging system, the M stage did not play as significant a role in predicting survival outcome as the T and N stages. This may be attributed to the fact that only 2% of patients were classified as having distant metastasis. Based on these findings, tumor staging was a crucial determinant in forecasting 5-year survival.

Among various primary sites, the salivary gland exhibited outstanding performance in predicting 5-year survival. This may be attributed to the fact that for the salivary gland, surgical intervention is the predominant treatment modality. Moreover, compared to other sites, the salivary gland presents a more uniform distribution in overall stages, potentially making predictions more straightforward. Conversely, model performance for cancers of the hypopharynx and nasal cavity was inferior relative to other sites. These sites exhibited a tendency toward lower survival rates than other types of cancer. Additionally, in the case of the nasal cavity, this diminished efficacy may be attributed to the inclusion of various paranasal sinuses within the nasal cavity, leading to increased heterogeneity.

In the 5-year survival rates based on primary sites, the salivary gland (0.84) and larynx (0.83) demonstrated the highest survival probabilities. According to the 5-year survival rates for different types of cancer reported in South Korea in 2020, the 5-year survival rate for larynx cancer from 2016 to 2020 was noted as 80.0%[32]. While our observed survival rate for the larynx was slightly higher, the findings were generally similar. Additionally, the reported 5-year survival rates for the oral cavity and pharynx were 65.5% for 2011-2015 and 69.4% for 2016-2020. The survival rate for the oral cavity, oropharynx, nasopharynx, and hypopharynx in our data was found to be 68.5%, which aligns closely with these reported rates. There have been multiple studies reported on salivary gland tumors. In the case of Western Europe, a 5-year survival rate of up to 81% has been reported [33]. While our results showed a slightly higher rate of 84%, they are nonetheless in line with

these findings.

O Our study has several limitations. First, although we implemented imputation procedures to manage missing data, there may still be constraints in adequately addressing all data gaps. In addition, our analyses were based on retrospective cohorts. Patients with follow-up durations of fewer than five years were also excluded, and this may introduce bias that should be considered. In addition, our analysis was solely based on ML algorithms, and we did not conduct tests using deep learning models such as convolutional neural networks and recurrent neural networks (RNNs). Therefore, conducting a comparative analysis encompassing a broader range of algorithms on a prospective cohort may be worthwhile. Nevertheless, our study demonstrated the usefulness of ML algorithms in analyzing 5-year survival rates during the treatment decision-making process. We believe that with further research, these ML models can be used to predict survival in clinical settings.

CONCLUSION

This study highlighted the efficacy of ML algorithms in predicting 5-year survival in patients with H&N cancer. Notably, among these algorithms, the gradient-boosting model, specifically Light GBM, delivered the best performance. Critical features driving the decision-making process included age, BMI, clinical T stage, and N stage. We anticipate that ML algorithms could be integrated into clinical practice in the future.

Table 1. Feature sets (n=1287)

Feature sets	n (%)
Demographic	
Age, mean±SD (Min-Max)	57.5±12.3 (18-89)
Sex (male/female)	970/317 (75.4/24.6)
BMI, mean±SD (Min-Max)	23.6±3.4 (15.1-41.5)
Past and Social history	
HTN	333 (25.9)
DM	167 (13.0)
Hepatitis	33 (2.6)
TB	65 (5.1)
Smoking (Non/Ex/Current)	492/321/474 (38.2/24.9/36.9)
Alcohol (Non/Current)	612/675 (47.5/52.5)
Primary sites	
Oral cavity	291 (22.6)
Oropharynx	179 (13.9)
Larynx	353 (27.4)
Hypopharynx	92 (7.2)
Nasopharynx	138 (10.7)
Salivary gland	146 (11.3)
Nasal cavity	88 (6.9)
Clinical and histopathologic attributes	
T-stage (T1/T2/T3/T4)	457/326/216/288 (35.5/25.3/16.8/22.4)
N-stage (N0/N1/N2/N3)	768/151/337/31 (59.7/11.7/26.2/2.4)
M-stage (M0/M1)	1261/26 (98.0/2.0)
Differentiation (unknown/well/moderate/poorly)	423/301/441/122 (32.8/23.4/34.3/9.5)
Overall stage (I/II/III/IV)	369/181/229/508 (28.7/14.1/17.8/39.4)
Initial treatment modality (Surgery/RT/CCRT/Palliative CTx)	814/149/274/50 (63.2/11.6/21.3/3.9)
Status (Survival/Dead)	951/336 (73.9/26.1)

BMI, body mass index; HTN, hypertension; DM, diabetes mellitus; TB, tuberculosis; RT, radiation therapy; CCRT, concurrent chemoradiation therapy; CTx, chemotheraphy

Table 2. Best hyperparameter

Algorithms	Hyperparameters	Parameter Range	Best Hyperparameters
SVM	C	[0.1 – 1000]	457.206
	gamma	[0.0001 – 1]	0.0039
	degree	[1 – 5]	4
	kernel	linear or RBF	RBF
RF	n_estimators	[10 – 1000]	999
	max_depth	[1 – 50]	49
	min_sample_split	[2 – 20]	2
	min_sample_leaf	[1 – 10]	2
NB	var_smoothing	$[1x10^{-10}\text{-}1x10^{-1}]$	0.0176
XGboost	n_estimators	[10 – 1000]	942
	max_dept	[1 – 50]	14
	gamma	[0 – 1]	0.3083
	min_child_weight	[1 – 10]	2
	subsample	[0.1 – 1]	0.5149
	colsample_bytree	[0.1 – 1]	0.5741
	learning_rate	[0.01 – 0.3]	0.0248
LightGBM	n_estimators	[10 – 1000]	490
	num_leaves	[2 – 200]	102
	max_depth	[1 – 50]	25
	min_child_samples	[1 – 100]	2
	Subsample	[0.1 – 1]	0.1267
	colsample_bytree	[0.1 – 1]	0.6185
	learning_rate	[0.01 – 0.3]	0.1626

SVM, support vector machine; RBF, radial basis function; RF, random forest; NB, naïve Bayes.

Table 3. Model performance.

	Sensitivity	**Accuracy**	**F1 score**	**AUC**
SVM	0.702	0.737	0.727	0.816
RF	0.774	0.811	0.805	0.900
NB	0.611	0.698	0.665	0.769
XGboost	0.790	0.814	0.812	0.890
LightGBM	0.807	0.822	0.820	0.905

AUC, area under the receiver operating characteristic curve; SVM, support vector machine; RF, random forest; NB, naïve Bayes

Table 4. Performance of each primary site using the Light GBM model

	Sensitivity	**Accuracy**	**F1 score**	**AUC**
Oral cavity	0.709	0.720	0.714	0.775
Oropharynx	0.827	0.844	0.844	0.895
Larynx	0.884	0.888	0.891	0.937
Hypopharynx	0.537	0.571	0.555	0.637
Salivary gland	0.951	0.968	0.967	0.988
Nasal cavity	0.667	0.691	0.674	0.700
Nasopharynx	0.771	0.793	0.788	0.849

AUC, area under the receiver operating characteristic curve

Supplement Table 1. Missing data (n=1287)

Feature sets	Missing data (%)
Demographic	
Age	18 (1.4)
Sex	10 (0.8)
BMI	31 (2.4)
Past and Social history	
HTN	0
DM	0
Hepatitis	0
TB	0
Smoking	25 (1.9)
Alcohol	24 (1.8)
Primary sites	
Oral cavity	0
Oropharynx	0
Larynx	0
Hypopharynx	0
Nasopharynx	0
Salivary gland	0
Nasal cavity	0
Clinical and histopathologic attributes	
T-stage	132 (10.3)
N-stage	131 (10.2)
M-stage	133 (10.3)
Differentiation	0
Overall stage	135 (10.5)
Treatment modality	1 (0.1)

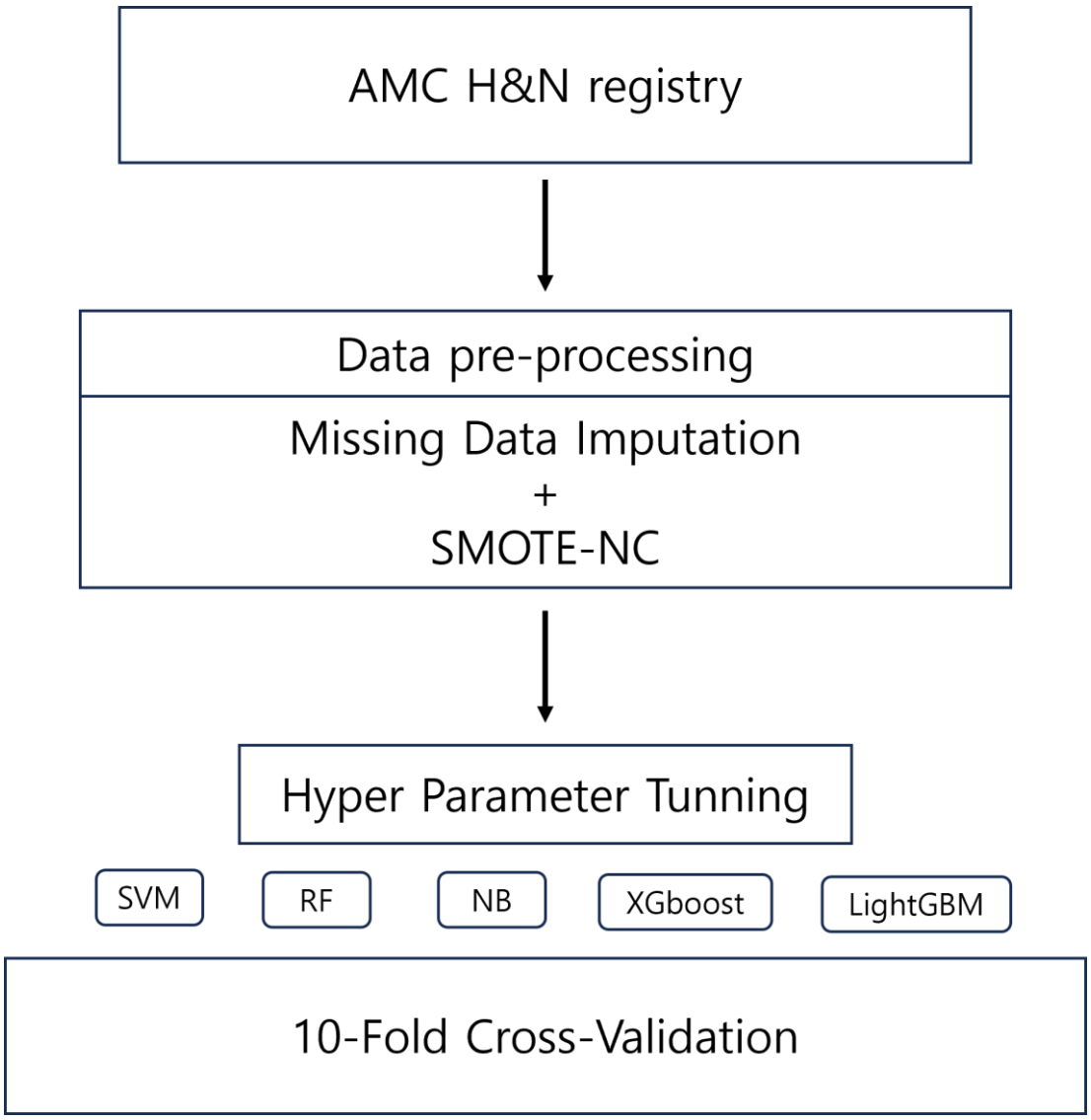
BMI, body mass index; HTN, hypertension; DM, diabetes mellitus; TB, tuberculosis

Supplement Table 2. Five-Year Survival Rates for primary site

	5-year survival rates
Oral cavity	0.66
Oropharynx	0.80
Larynx	0.83
Hypopharynx	0.57
Nasopharynx	0.66
Salivary gland	0.84
Nasal cavity	0.62

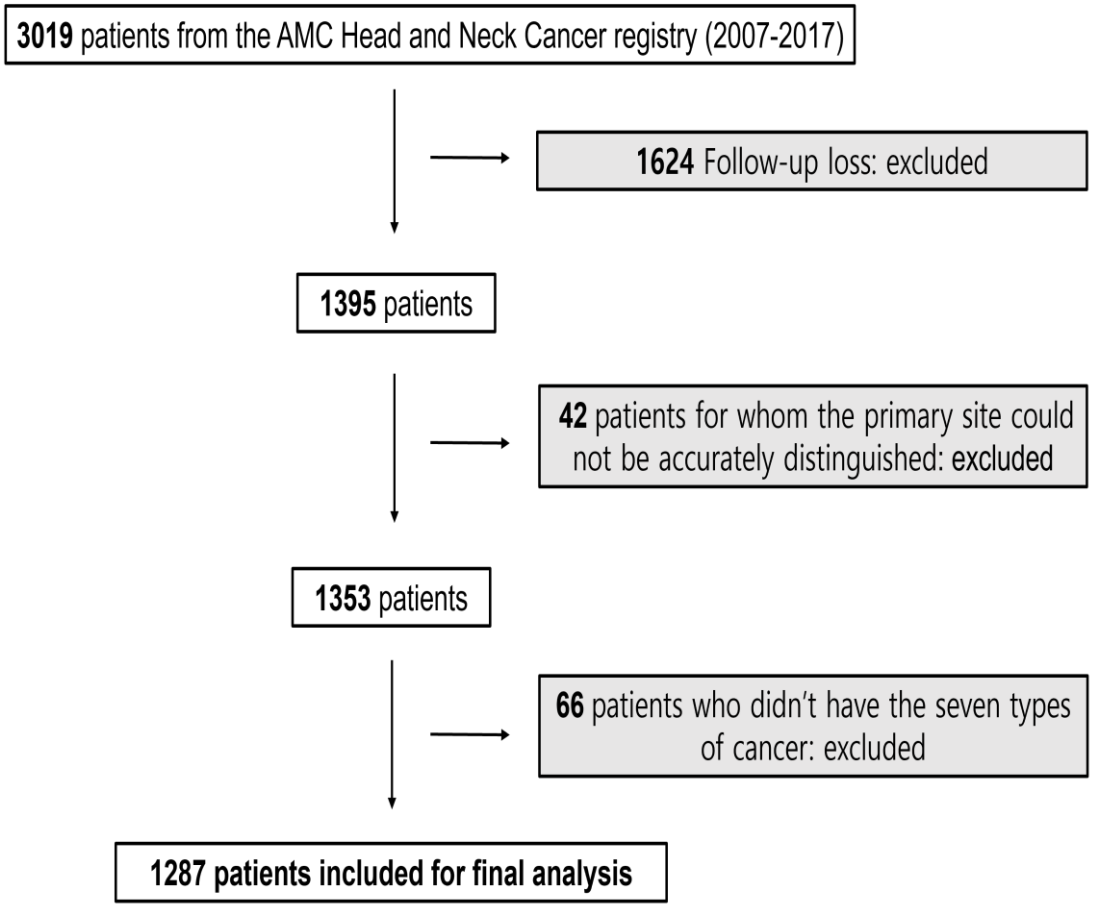
\\

Figure 1. Summary of modeling methodology



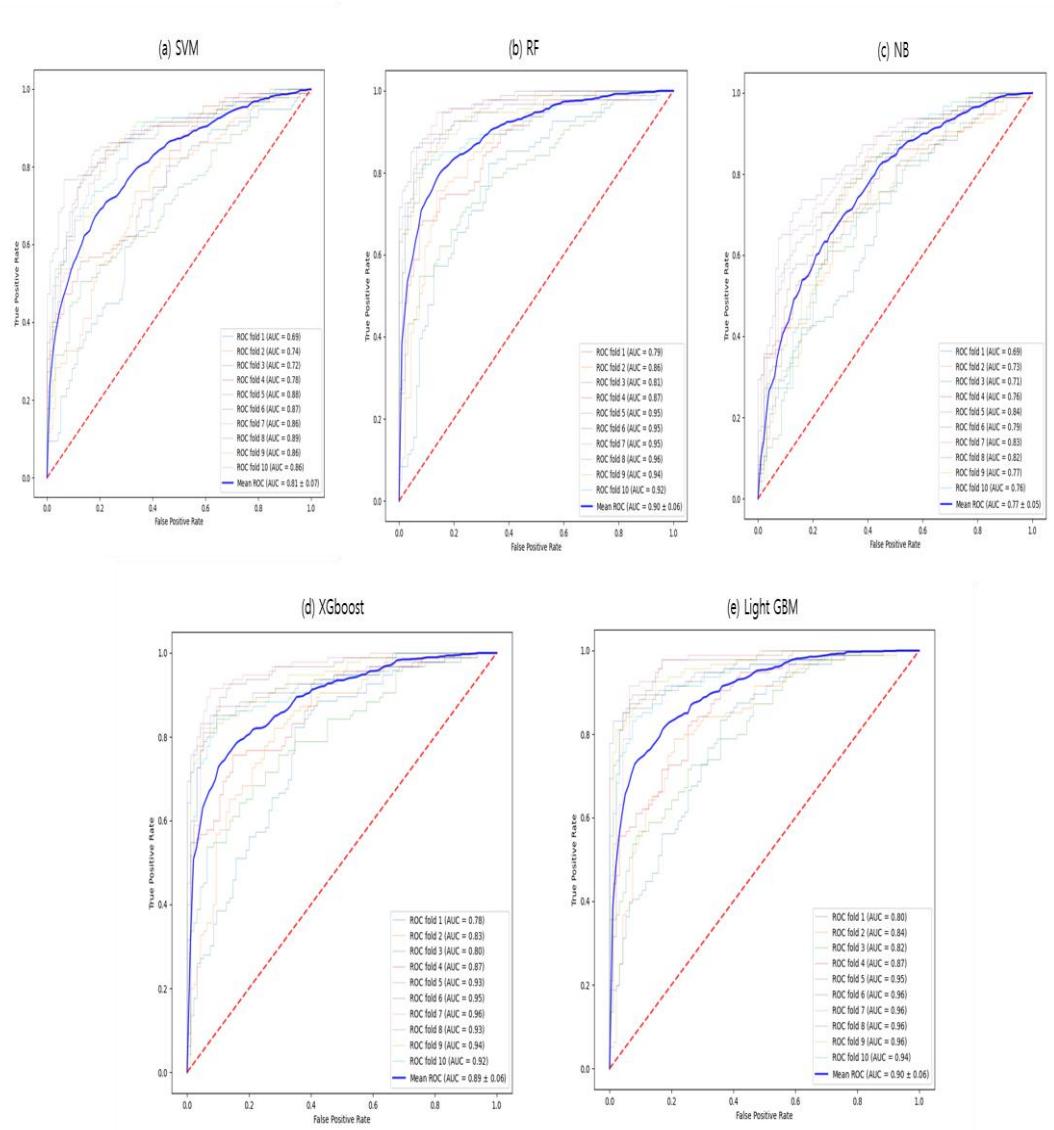
AMC, Asan medical center; H&N, head & neck; SMOTE-NC, synthetic minority oversampling technique-nominal continuous; SVM, support vector machine; RF, random forest; NB, naïve Bayes

Figure 2. Flow chart for patient inclusion



AMC, Asan medical center.

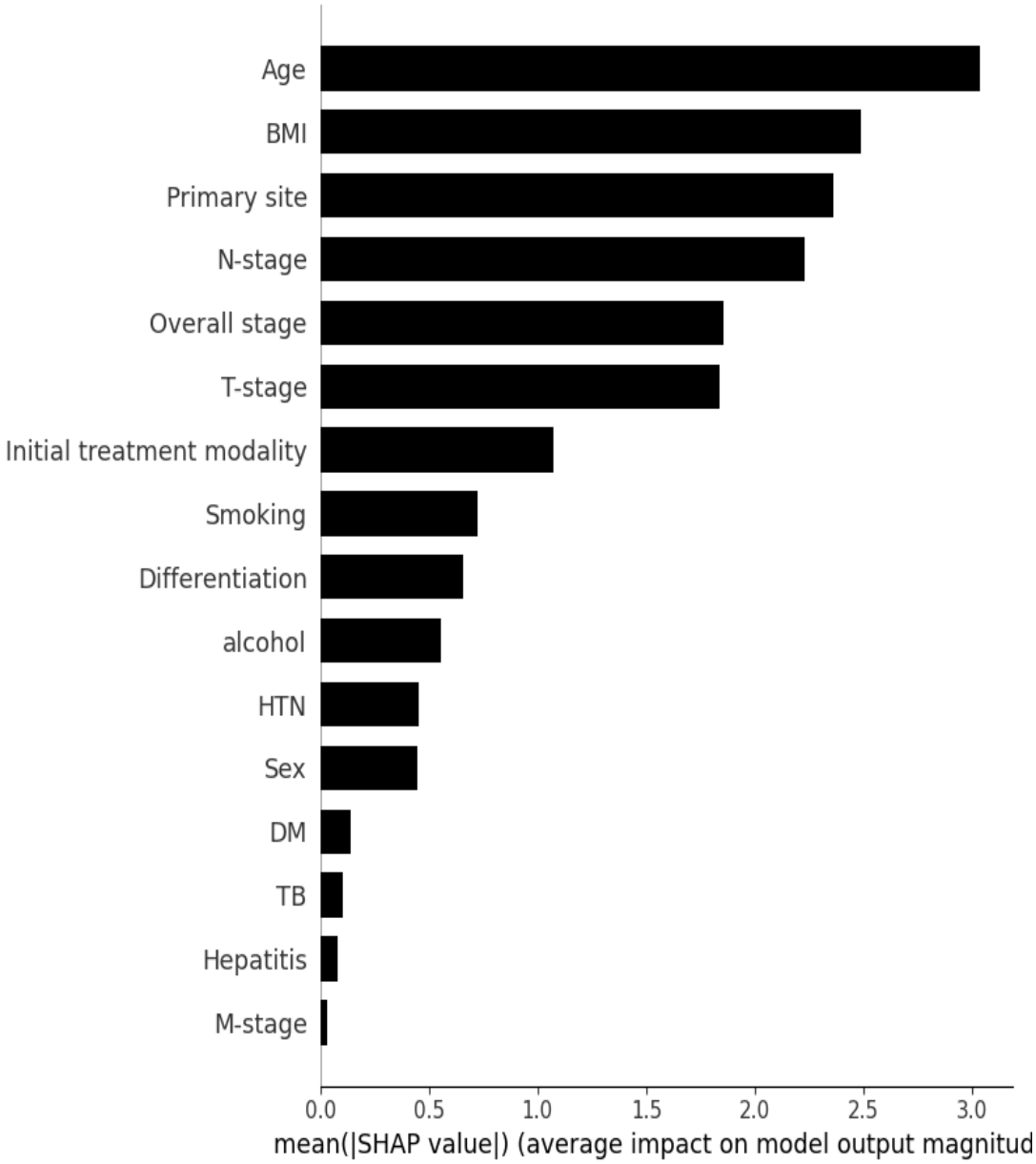
Figure 3. ROC curve based on five machine learning algorithms using 10-fold cross validation



(a) SVM, (b) NB, (c) RF, (d) XGboost, (e) Light GBM

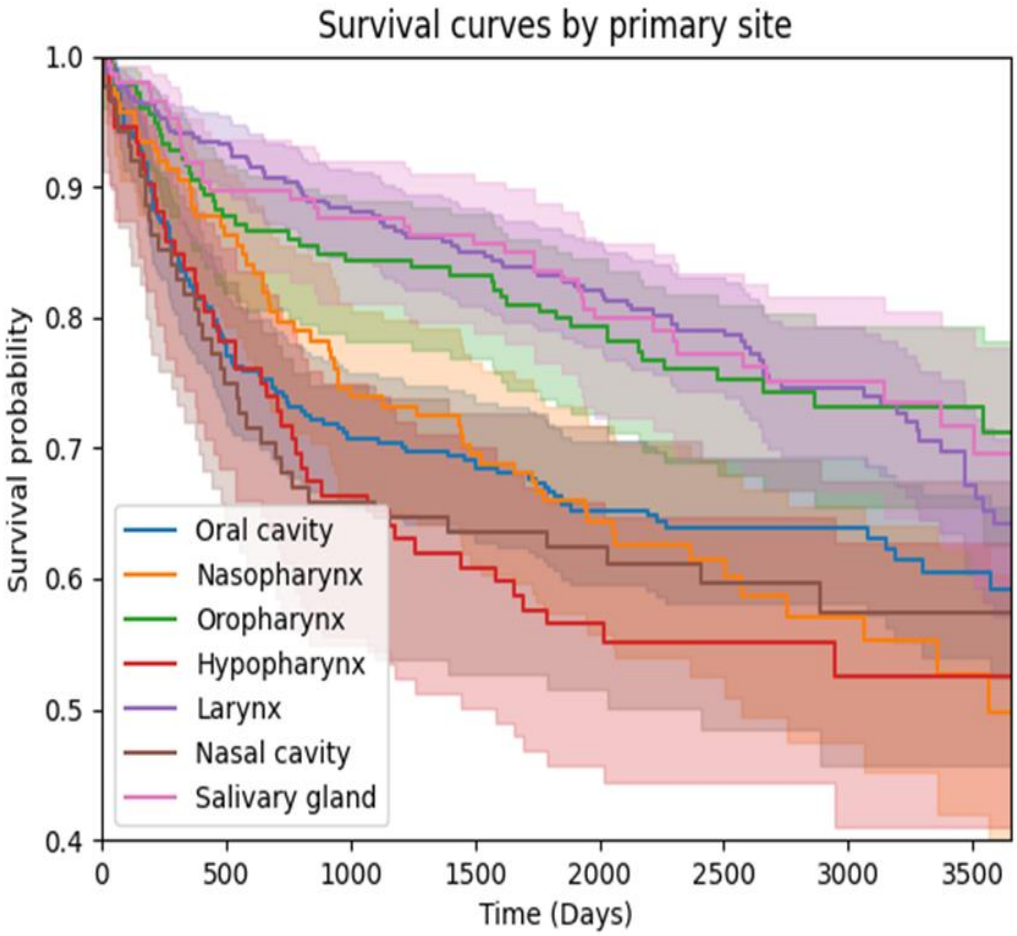
ROC, receiver operating characteristic; SVM, support vector machine; RF, random forest; NB, naïve Bayes.

Figure 4. SHAP feature importance

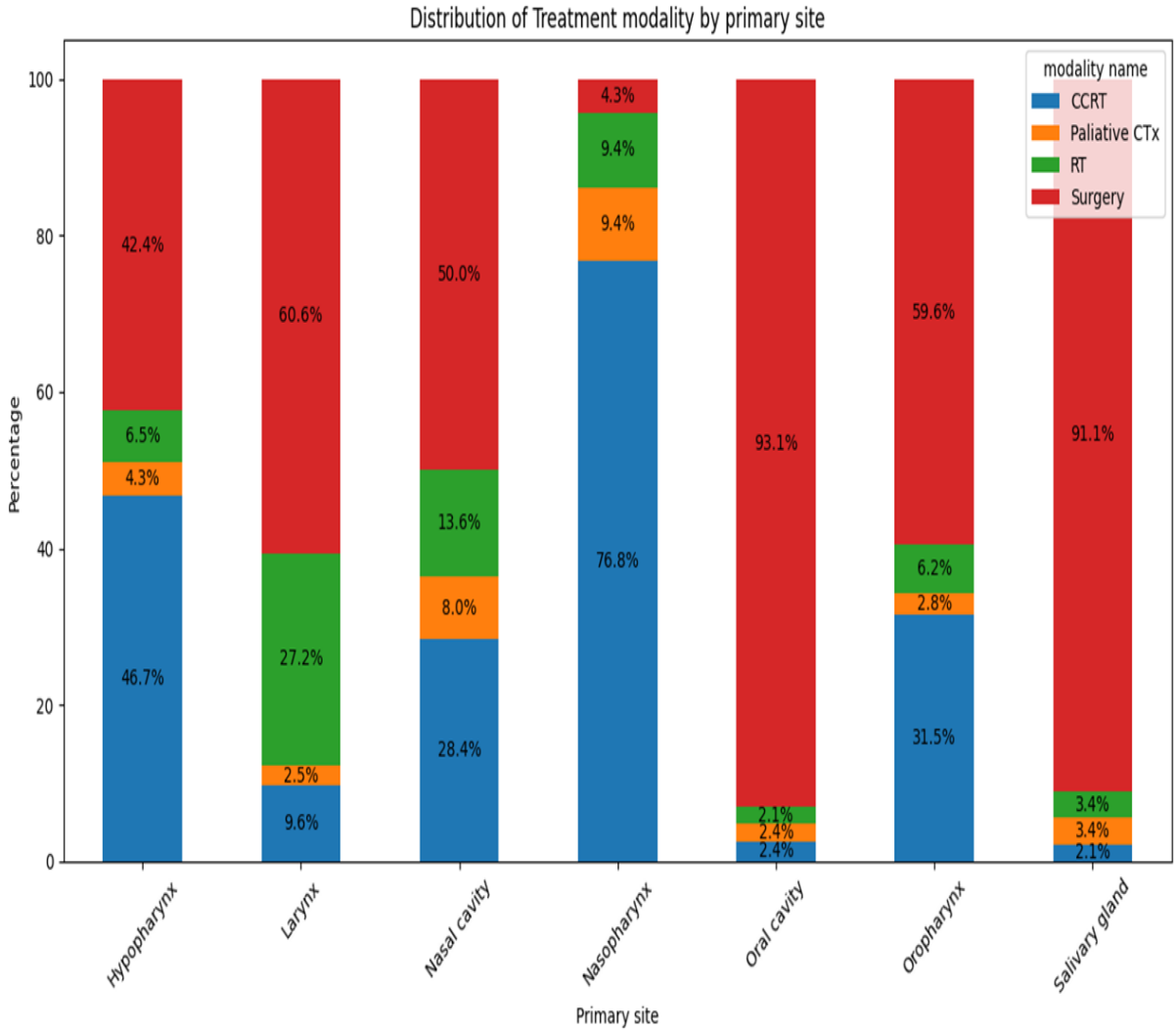


SHAP, SHapley Additive exPlanations.

Supplement figure 1. Survival Curves by Primary Site



Supplement figure 2. Treatment Modalities for Each Primary Site



REFERENCES

1. Gormley M, Creaney G, Schache A, Ingarfield K, Conway DI. Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. *British Dental Journal*. 2022;233(9):780-786.

2. Salahuddin Z, Chen Y, Zhong X, et al. From Head and Neck Tumour and Lymph Node Segmentation to Survival Prediction on PET/CT: An End-to-End Framework Featuring Uncertainty, Fairness, and Multi-Region Multi-Modal Radiomics. *Cancers*. 2023;15(7):1932.

3. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930.

4. Gong X, Zheng B, Xu G, Chen H, Chen C. Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer. *Journal of Thoracic Disease*. 2021;13(11):6240.

5. Alabi RO, Mäkitie AA, Pirinen M, Elmusrati M, Leivo I, Almangush A. Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. *International journal of medical informatics*. 2021;145:104313.

6. Adeoye J, Hui L, Koohi-Moghadam M, Tan JY, Choi S-W, Thomson P. Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis. *International journal of medical informatics*. 2022;157:104635.

7. Kotevski DP, Smee RI, Vajdic CM, Field M. Machine learning and nomogram prognostic modeling for 2-year head and neck cancer–specific survival using electronic health record data: a multisite study. *JCO Clinical Cancer Informatics*. 2023;7:e2200128.

8. Kaur I, Doja M, Ahmad T. Data mining and machine learning in cancer survival research: an overview and future recommendations. *Journal of Biomedical Informatics*. 2022;128:104026.

9. Amin MB, Edge SB, Greene FL, et al. *AJCC cancer staging manual*. vol 1024. Springer; 2017.

10. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE transactions on information theory*. 1967;13(1):21-27.

11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 2002;16:321-357.

12. Cortes C, Vapnik V. Support vector machine. *Machine learning*. 1995;20(3):273-297.

13. Raikwal J, Saxena K. Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. *International Journal of Computer Applications*. 2012;50(14)

14. Breiman L. Random forests. *Machine learning*. 2001;45:5-32.

15. Rish I. An empirical study of the naive Bayes classifier. 2001:41-46.

16. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 2016:785-794.

17. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*. 2017;30

18. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*. 2020;415:295-316.

19. Syarif I, Prugel-Bennett A, Wills G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 2016;14(4):1502-1509.

20. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *Journal of machine learning research*. 2012;13(2)

21. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30

22. Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*. 2023;

23. Wang K-J, Makond B, Wang K-M. An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC medical informatics and decision making*. 2013;13(1):1-14.

24. Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*. 2015;58:49-59.

25. Doja M, Kaur I, Ahmad T. Age-specific survival in prostate cancer using machine learning. *Data Technologies and Applications*. 2020;54(2):215-234.

26. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*. 2019;17(1):26-40.

27. Pulte D, Brenner H. Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. *The oncologist*. 2010;15(9):994-1001.

28. Pruegsanusak K, Peeravut S, Leelamanit V, et al. Survival and prognostic factors of different sites of head and neck cancer: an analysis from Thailand. *Asian Pacific Journal of Cancer Prevention*. 2012;13(3):885-890.

29. Takenaka Y, Takemoto N, Nakahara S, et al. Prognostic significance of body mass index before treatment for head and neck cancer. *Head & neck*. 2015;37(10):1518-1523.

30. Cho J-K, Hyun SH, Choi N, et al. Significance of lymph node metastasis in cancer dissemination of head and neck cancer. *Translational oncology*. 2015;8(2):119-125.

31. Divi V, Chen MM, Nussenbaum B, et al. Lymph node count from neck dissection predicts mortality in head and neck cancer. *Journal of Clinical Oncology*. 2016;34(32):3892-3897.

32. Kang MJ, Jung K-W, Bang SH, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2020. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*. 2023;55(2):385-399.

33. Kordzińska-Cisek I, Grzybowska-Szatkowska L. Salivary gland cancer− epidemiology. *Nowotwory Journal of Oncology*. 2018;68(1):22-27.

국문 요약

배경

두경부암 환자에 대한 정확한 예후 예측은 임상에서 매우 중요하다. 이 연구의 목적은 두경부암 환자의 생존 예측의 기계 학습 (머신러닝) 알고리즘의 유용성을 입증하고 효과적인 모델을 찾아내는 것이다.

방법

2007 년부터 2017 년까지 수집된 서울아산병원 두경부암 레지스트리에 등록된 3019 명의 환자를 대상으로 연구를 진행하였다. 다양한 기계 학습 알고리즘의 성능을 비교하기 위해 사용된 데이터 세트에는 인구 통계학적 특성, 과거 및 사회적 특성, 주요 장소의 임상 및 조직 병리학적 특성 및 치료 방법이 포함되었다. 우리는 두경부암을 5 년 생존 상태를 기준으로 분류하기 위해 데이터세에 총 다섯 가지 기계 학습 모델을 적용하였다. 이 모델들은 최근 개발된 두 가지 그래디언트 부스팅 모델 (XGBoost 및 LightGBM), 일반적으로 사용되는 트리 기반 모델인 Random Forest (RF), Support Vector Machines (SVM) 및 Naive Bayes (NB)를 포함 한다. 모델 성능을 평가하기 위하여 10 겹-교차 검증을 실시하였다.

결과

본 연구는 최종적으로 선정된 1,287 명의 환자로 구성되었다. 다섯 모델 중

LightGBM이 가장 뛰어난 성능을 보였다. 모델 성능을 평가하기 위해 민감도, 정확도, F1 점수, 및 AUC-ROC 네 가지 지표를 사용하였다. 모델 성능의 평균 점수는 다음과 같았다: SVM - 0.745, RF - 0.823, NB - 0.686, XGBoost - 0.827, 및 LightGBM - 0.839.LightGBM 모델에서는 SHapley Additive exPlanations (SHAP) 값이 계산되었고, 결과에서 16개의 특징 중 중요도 측면에서 상위 다섯 가지는 다음과 같이 내림차순으로 나타났다: 나이,체질량 지수 (BMI), 원발 부위, N-스테이지, 병기, T-스테이지

결론

본 연구에서는 나이, 체질량 지수 및 스테이지에 관련된 특성이 두경부암의 생존 가능성을 예측하는 데 가장 중요하다는 것을 발견하였다. 또한 LightGBM 모델은 적절한 데이터셋과 결합하여 두경부 암으로 진단된 환자들을 위한 정확한 예후 모델을 구축하는 데 사용될 수 있었다. 향후 기계 학습 모델들을 임상에 적용될 수 있을거라 기대 한다.