Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

A Risk Prediction Scheme for Secondary Primary Esophageal Squamous Cell Carcinoma in Head and Neck Cancer Survivors

Chun-Chia Chen[1,2], Ming-Yi Lu[3,4*], Chi-Chang Chang[5,6*]

[1] Institute of Medicine, Chung Shan Medical University, Taichung, 40201, Taiwan
[2] Division of Plastic Surgery, Department of Surgery, Chi Mei Medical Center, Tainan, 71004, Taiwan
[3] School of Dentistry, College of Oral Medicine, Chung Shan Medical University, Taichung, 40201, Taiwan
[4] Division of Oral and Maxillofacial Surgery, Department of Dentistry, Chung Shan Medical University Hospital, Taichung, 40201, Taiwan
[5] School of Medical Informatics, Chung Shan Medical University & IT Office, Chung Shan Medical University Hospital, Taichung, 40201, Taiwan
[6] Department of Information Management, Ming Chuan University, Taoyuan, 33300, Taiwan
E-mail: miexyz@gmail.com; changintw@gmail.com

Received: 16 March 2025; **Revised:** 10 April 2025; **Accepted:** 10 July 2025

Abstract: In this study, we developed a machine learning scheme to predict the occurrence of Second Primary Esophageal Squamous cell Carcinoma (SPESC) among patients with primary Head and Neck Cancer (HNC). This study retrospectively collected 2,863 records of patients with HNC, including 65 cases of SPESC. Data on 19 risk factors for SPESC were analyzed from the aforementioned records to identify significant risk factors and protective factors for SPESC. On the basis of gain ratios, the following significant risk factors were identified for the occurrence of SPESC in patients with HNC: age at HNC diagnosis < 65 years, tumor grade/differentiation > 2, smoking behavior, drinking behavior, and existence of a tumor depth measurement in the pathology report. The only significant protective factor for SPESC was Body Mass Index (BMI) ≥ 24. Data on the aforementioned factors were integrated into seven machine learning algorithms to predict the occurrence of SPESC: C4.5, C5.0, support vector machine, random forest, Classification and Regression Trees (CART), linear dynamic analysis, and logistic regression. Among these seven algorithms, CART exhibited the largest area under the receiver operating characteristic curve (0.9240); the highest accuracy (0.8678), recall rate (0.8573), F1 score (0.8871), and Matthews correlation coefficient (0.7305); and the lowest false positive rate (0.1990) for SPESC detection. Thus, CART was selected as the machine learning algorithm of the proposed model. The highest model accuracy of 88.0% was achieved under the following candidate factor conditions: drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: < 24, radiotherapy: yes; and tumor grade/differentiation: > 2. Overall, the developed prediction scheme enables the early and accurate detection of SPESC among patients with HNC, thus improving patient outcomes and ensuring that patients with a low risk of developing SPESC do not need to undergo unnecessary invasive procedures.

***Keywords*:** prediction scheme, secondary primary esophageal squamous cell carcinoma, head and neck cancer, clinical risk factors, machine learning classifiers, CART

MSC: 65L05, 34K06, 34K28

DOI: https://doi.org/10.37256/cm.6520255753

Abbreviation

SPESC	Second Primary Esophageal Squamous cell Carcinoma
HNC	Head and Neck Cancer
CART	Classification And Regression Trees

1. Introduction

Squamous cell carcinoma is one of the most common malignancies affecting the human body. This malignancy often occurs in the head and neck region, including the oral cavity, oropharynx, hypopharynx, and larynx, resulting in Head and Neck Cancer (HNC). HNC is a major disease worldwide, with 931,931 new cases of HNC and 672,129 deaths resulting from HNC being reported in 2020 [1]. Approximately 70% of new cases of HNC occur in low- and middle-income countries, with the male-to-female incidence ratio being 3 : 1 [2]. The consumption of specific carcinogen-containing products is associated with the high prevalence of head and neck squamous cell carcinoma [3]. Patients with HNC tend to have Second Primary Tumors (SPTs), as explained by field cancerization theory [4]. In addition to the head, neck, and lungs, the esophagus is one of the areas most commonly affected by HNC, with the incidence of second primary esophageal cancer with primary HNC varying from 5.9% to 43% [5–9]; the highest prevalence of 43% was observed for primary cancers in the hypopharynx [6]. The mean survival interval between diagnosis and death was estimated to be 22.76 months for primary cancers in the hypopharynx [5]. Furthermore, worse prognoses have been reported when second primary malignancies occur with these cancers [5, 6, 8, 9].

In contrast to head, neck, and lung malignancies, esophageal malignancies are not easily detected in their early stages. Therefore, establishing a clinical risk prediction scheme applicable to patients with primary HNC who may develop Second Primary Esophageal Squamous cell Carcinoma (SPESC) is crucial. If a second primary cancer occurs in the esophagus or elsewhere, early detection would help physicians provide early therapy for improving outcomes. Although positron emission tomography or computed tomography is commonly performed on survivors of HNC during follow-up, these methods have limited effectiveness for detecting early esophageal cancer. Routine esophageal screening is recommended for patients who are suspected to have esophageal cancer. However, esophagoscopy and panendoscopy are invasive procedures and are unsuitable for frequent routine follow-up. Previous studies have not developed suitable schemes for predicting the occurrence of SPESC in survivors of HNC. Therefore, the present study developed a machine learning scheme for this purpose.

2. Materials and methods

2.1 *Sample and candidate risk factors*

This study retrospectively collected 2,863 records from five cancer registries. The study was approved by the Institutional Review Board of the Chung Shan Medical University Hospital (IRB no. CS2-20114) and did not require patient consent. Data on the following 19 candidate risk factors were analyzed on the basis of relevant literature and guidance from clinical experts (Table 1): (1) sex; (2) age at diagnosis; (3) tumor grade/differentiation; (4) tumor size; (5) examined regional lymph nodes; (6) combined stage group; (7) surgery; (8) radiotherapy; (9) chemotherapy; (10) Body Mass Index (BMI; kg/cm^2); (11) smoking behavior; (12) betel nut chewing behavior; (13) drinking behavior; (14) lymph node sizes (mm); (15) level I-III lymph nodes; (16) level IV and V and retropharyngeal lymph nodes; (17) level VI and VII and facial lymph nodes; (18) parapharyngeal, parotid, and suboccipital/retroauricular lymph nodes; and (19) tumor depth measurement in the pathology report (mm).

Table 1. Selected candidate risk factors for SPESC

No	Candidate risk factors	Definition
X1	Sex	Male/Female
X2	Age at diagnosis	< 65 years/≧ 65 years
X3	Grade/Differentiation	≦ 2/> 2
X4	Tumor size (cm)	≦ 4 cm/> 4 cm
X5	Regional lymph nodes examined	No/Yes
X6	Combine stage group	≦ stage II/> stage II
X7	Surgery	No/Yes
X8	Radiotherapy	No/Yes
X9	Chemotherapy	No/Yes
X10	BMI (kg/m^2)	18.5 ≦ BMI < 24/< 18.5/≧ 24
X11	Smoking behavior	No/Yes
X12	Betel nut chewing behavior	No/Yes
X13	Drinking behavior	No/Yes
X14	Size of lymph nodes (mm)	≦ 10 mm/> 10 mm
X15	Levels I-III, lymph nodes for head and heck	Negative/Positive
X16	Levels IV-V and retropharyngeal lymph nodes	Negative/Positive
X17	Levels VI-VII and facial lymph nodes	Negative/Positive
X18	Parapharyngeal, parotid and suboccipital/Retroauricular lymph nodes	Negative/Positive
X19	Measured depth in pathology report (mm)	≦ 49 mm/> 49 mm
Y	SPESC	No/Yes

2.2 *Machine learning classifiers*

T Statistical analysis was conducted to select features for machine learning, and the gain ratio was used to rank the analyzed features by their importance. Data on significant risk factors were then integrated into seven machine learning algorithms for predicting the occurrence of SPESC: the C4.5, C5.0, Support Vector Machine (SVM), Random Forest (RF), Classification and Regression Trees (CART), Linear Discriminant Analysis (LDA), and Logistic Regression (LGR) algorithms. These algorithms were trained through 10-fold cross validation, with 70% and 30% of the adopted data used for training and testing, respectively. Decision trees can be created using the C4.5 algorithm developed by Ross Quinlan. When these decision trees are used for classification, C4.5 is considered a statistical classifier. The information gain ratio is one of the most used indexes to construct the C5.0 decision tree.

Let S represents a set of samples, p_i is the probability that an arbitrary sample belongs to the class B_i, $i = 1, \ldots, n$. Info (S) indicates the information entropy of S,

$$\text{Info}\ (S) = \sum_{i=1}^{n} p_i \log\ (p_i) \tag{1}$$

Assumed that categorical attribute x_i has n different values. We can use attribute x_i to divide S into n subsets $\{s_1, s_2, \ldots, s_n\}$ and s_{in} is the sample number of class B_i in subset s_n. Info (S, x_i) is the needed information entropy for calculate information gain and is as follows.

$$\text{Info}\ (S, x_i) = \sum_{j=1}^{n} \frac{s_{1j} + s_{2j} + \ldots + s_{nj}}{S} \text{Info}\ (x_i) \tag{2}$$

C4.5 is a landmark decision tree generation algorithm and is highly popular for machine learning tasks. By selecting attributes for each node on the basis of information entropy, this algorithm can construct decision trees through a top-down recursive divide-and-conquer approach [10, 11]. The aforementioned approach is also commonly used to predict and classify data in the CART algorithm. A rule based on these variables can then be employed to predict the value of an outcome variable (Y in this study) from known prediction variables (X1-X19 in this study). The prediction variables can be a mixture of categorical and continuous variables [12, 13]. RF is an ensemble learning method used for classification, regression, and other tasks involving aggregation of the results of a given dataset or the generation of a random decision tree [14]. The popular C5.0 decision tree has lower accuracy than do gradient-boosted trees; however, data characteristics can influence its performance. By using a recursive approach, the C5.0 algorithm generates a tree in accordance with the information in a top-down scheme [15]. LGR analysis can be used to analyze the binary outcomes of medicine use and to model the relationships between explanatory variables (X1-X19) and a response variable (Y) [16, 17]. In LDA, two or more classes of events are characterized by a linear combination of features. This method can be used to conduct linear classification as well as dimensionality reduction before classification; such dimensionality reduction can improve the feature extraction process and facilitate data compression [18, 19]. The SVM algorithm analyzes data for regression or classification tasks by using a supervised learning method. This algorithm is one of the most robust prediction methods, and SVM classifiers can be used to separate two classes by using a linear decision boundary [20, 21].

Let $\{(x_i,\ y_i)\}_{i=1}^{N}$, $x_i \in R^d$, $y_i \in \{-1,\ 1\}$ be the training set with input vectors and labels, where y_i is known target, N is the number of sample observations and d is the dimension of each observation. The SVM algorithm is to seek the hyperplane $w \cdot x_i + b = 0$, where w is the vector of hyperplane and b is a bias term, to separate the data from two classes with maximal margin width $2/\|w\|^2$, and the all points under the boundary is named support vector. To optimal the hyperplane, SVM is to solve the following optimization problem.

$$\text{Min } \Phi(x) = \frac{1}{2}\|w\|^2 \tag{3}$$

$$\text{S.t. } y_i(w^T x_i + b) \geq 1,\ i = 1,\ 2,\ ...,\ N \tag{4}$$

By using Lagrangian multipliers α to the Eq. (1), it thus yields the following dual Lagrangian form,

$$\text{Max } \Phi(w,\ b,\ \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1,\ j=1}^{N} \alpha_i \alpha_j y_i y_j {x_i}^T x_j \tag{5}$$

$$\text{S.t. } \sum_{j=1}^{N} \alpha_j y_j = 0,\ 0 \leq \alpha_i \leq C,\ i = 1,\ 2,\ ...,\ N \tag{6}$$

In Eq. (5), C is the penalty factor and used to determine the degree of penalty assigned to an error classification; the value of α must be nonnegative real coefficients. In SVM algorithm, any function that meets Mercer's condition can be used as the kernel function to map the data into a high-dimensional feature space. Although several choices for the kernel function are available, one of the most widely used kernel unction is the Radial Basis Function (RBF) defined as $K(x_i,\ x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2),\ \gamma \geq 0$, where γ denotes the width of the RBF. Thus, the RBF is adopted in this study as kernel function.

In this study, the performance of the aforementioned machine learning classifiers was evaluated in terms of their recall, specificity, accuracy, false positive ratio, F1 score, precision, Matthews correlation coefficient, and Area Under the receiver operating characteristic Curve (AUC). They are

$$\text{Precision} = \frac{TP}{TP+FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP+FN}. \tag{8}$$

3. Results

The 2,863 records collected in this study covered the period from 2000 to 2020 and included 65 SPESC cases. Table 2 presents the clinical data of the 2,863 patients, including the results of chi-square tests and Odds Ratios (ORs) for different risk factors. The following significant risk factors were identified for SPESC in patients with HNC: age at HNC diagnosis of < 65 years [OR = 2.96, 95% Confidence Interval (CI) = 1.07-8.19], tumor grade/differentiation of > 2 (OR = 2.54, 95% CI = 1.44-4.46), smoking behavior (OR = 3.20, 95% CI = 1.16-8.84), drinking behavior (OR = 5.92, 95% CI = 2.37-14.78), and tumor depth (OR = 1.64, 95% CI = 1.00-2.70). The only protective factor for SPESC was BMI $\geqq$ 24 kg/m^2 (OR = 0.30, 95% CI = 0.18-0.52).

A total of 19 candidate risk factors for HNC were used as predictive features in this study, and these factors were ranked by their gain ratio to determine their importance (Table 3). The gain ratios of the aforementioned parameters indicated that drinking behavior was the most important factor, followed by tumor size; sex; smoking behavior; surgery; level IV and V and retropharyngeal lymph nodes; betel nut chewing behavior; level VI and VII and facial lymph nodes; age at diagnosis; BMI; parapharyngeal, parotid, and suboccipital/retroauricular lymph nodes; tumor grade/differentiation; measured tumor depth mentioned in the pathology report; lymph node sizes; radiotherapy; combined stage group; levels I–III lymph nodes; chemotherapy; and examined regional lymph nodes.

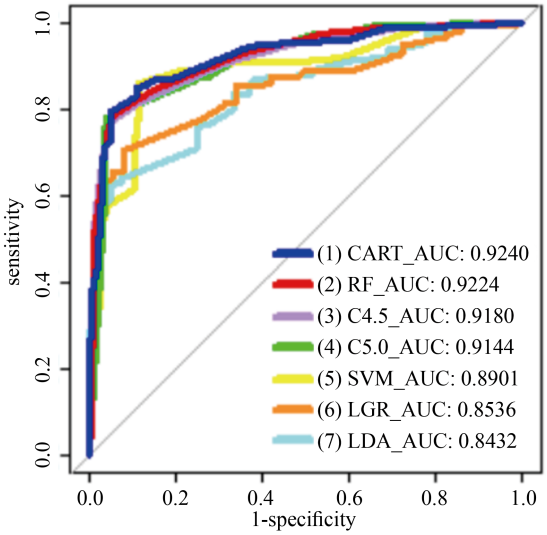

Figure 1. Receiver operating characteristic curves of all classifiers for the test dataset

The prediction performance of seven machine learning classifiers was investigated: the C4.5, C5.0, SVM, RF, CART, LDA, and LGR algorithms (Table 4). Of these classifiers, CART had the highest AUC value (0.9240), accuracy (0.8678), recall rate (0.8573), F1 score (0.8871), and Matthews correlation coefficient (0.7305) and the lowest false positive rate (0.1990) for SPESC occurrence. C5.0 had higher specificity but lower accuracy and F1 score than did CART. CART considerably outperformed the other algorithms, mainly because of its random bootstrap subsampling design and splitting

feature selection. In addition, single CART trees outperformed those arranged in series through a boosting ensemble method. Of all classifiers, CART had the largest AUC (0.9240), which indicated that it had the highest true positive rate and lowest false positive rate (Figure 1). By contrast, the LDA algorithm had the lowest AUC value.

Table 2. Clinical data of all patients with HNC whose data were examined in this study (N = 2,863) (*** $p < 0.001$; ** $p = 0.001$; * $p < 0.05$. The ORs for all listed clinical risk factors were analyzed)

Candidate risk factors	Attribute	p-value	Odds ratio
Sex	Female Male	0.029*	1.00 6.72 [0.92-48.66]
Age at diagnosis	≧ 65 years < 65 years	0.028*	1.00 2.96*[1.07-8.19]
Grade/Differentiation	≦ 2 > 2	0.001**	1.00 2.54*[1.44-4.46]
Tumor size	＞ 4 cm ≦ 4 cm	0.438	1.00 1.3 [0.67-2.49]
Regional lymph nodes examined	No Yes	0.134	1.00 1.47 [0.89-2.42]
Combine stage group	≦ stage II ＞ stage II	0.407	1.00 1.23 [0.75-2.02]
Surgery	Yes No	0.101	1.00 0
Radiotherapy	No Yes	0.393	1.00 1.23 [0.76-2.03]
Chemotherapy	No Yes	0.435	1.00 1.22 [0.74-2.02]
BMI	18.5 ≦ BMI < 24 BMI < 18.5 BMI ≧ 24	< 0.001***	1.00 1.49 [0.57-3.88] 0.30*[0.18-0.52]
Smoking behavior	No Yes	0.018*	1.00 3.20*[1.16-8.84]
Betel nut chewing behavior	No Yes	0.394	1.00 1.30 [0.71-2.35]
Drinking behavior	No Yes	< 0.001***	1.00 5.92*[2.37-14.78]
Size of lymph nodes	＞ 10 mm ≦ 10 mm	0.488	1.00 1.24 [0.67-2.30]

Table 2. (cont.)

Candidate risk factors	Attribute	*p*-value	Odds ratio
Levels I-III, lymph nodes	Positive Negative	0.185	1.00 1.41 [0.85-2.33]
Levels IV-V and retropharyngeal lymph nodes	Positive Negative	0.453	1.00 2.10 [0.29-15.33]
Levels VI-VII and facial lymph nodes	Positive Negative	0.528	1.00 0
Parapharyngeal, parotid, and suboccipital/Retroauricular lymph nodes	Positive Negative	0.733	1.00 0
Measured depth in pathology report (mm)	≦ 049 mm > 049 mm	0.048*	1.00 1.64 [1.00-2.70]

Table 3. Ranking of the examined candidate risk factors for SPESC in terms of their importance

Rank	Candidate risk factors
X13	Drinking behavior
X4	Tumor size (cm)
X1	Sex
X11	Smoking behavior
X7	Surgery
X16	Levels IV-V and retropharyngeal lymph nodes
X12	Betel nut chewing behavior
X17	Levels VI-VII and facial lymph nodes
X2	Age at diagnosis
X10	BMI (kg/m^2)
X18	Parapharyngeal, parotid and suboccipital/Retroauricular lymph nodes
X3	Grade/Differentiation
X19	Measured depth in pathology report (mm)
X14	Size of lymph nodes (mm)
X8	Radiotherapy
X6	Combine stage group
X15	Levels I-III, lymph nodes for head and neck
X9	Chemotherapy
X5	Regional lymph nodes examined

Table 4. Ranking of the examined candidate risk factors for SPESC in terms of their importance

Method	Specificity	Recall	Accuracy	FPR	F1 score	MCC	Precision	AUC
CART	0.8838	0.8573	0.8678	0.1990	0.8871	0.7305	0.9191	0.9240
RF	0.9492	0.7925	0.8542	0.2518	0.8682	0.7247	0.9600	0.9224
C4.5	0.9456	0.7736	0.8413	0.2693	0.8553	0.7029	0.9563	0.9180
C5.0	0.9619	0.7783	0.8506	0.2618	0.8633	0.7236	0.9692	0.9144
SVM	0.9456	0.7736	0.8413	0.2693	0.8553	0.7029	0.9563	0.8901
LGR	0.9220	0.7017	0.7884	0.3325	0.8008	0.6118	0.9326	0.8536
LDA	0.9528	0.6262	0.7548	0.3765	0.7559	0.5779	0.9533	0.8432

We ranked the clinical risk factors by their odds ratios. Drinking behavior was the most important clinical risk factor in this study 5.92 * [2.37-14.78].

Of the seven classifiers, CART had the highest AUC value (0.9240) and Matthews correlation coefficient (0.7305). Therefore, this algorithm was incorporated into the developed scheme for predicting SPESC risk.

In decision tree analysis, all samples passed through 16 subsets of branches from the root node to the leaf node (Figure 2). Of all considered candidate factors, drinking behavior had the strongest influence on SPESC risk; thus, this factor was selected as the root node. Furthermore, age at diagnosis and smoking behavior were selected as nodes for the generation of second- and third-order decision trees, respectively. As presented in Table 5, accuracies of 58.5%-88.0% were achieved with different combinations of candidate factor conditions.

The highest accuracy was achieved with condition combination 4.

Table 5. SPESC prediction accuracies achieved with different combinations of candidate factor conditions in CART model

Combination	Candidate factor	Conditions accuracy
1	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: < 24 kg/m^2, radiotherapy: no, and betel nut chewing behavior: no	84.5%
2	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: < 24 kg/m^2, radiotherapy: no, betel nut chewing behavior: yes, and Measured Depth in Pathology Report: ≤ 49 mm	66.8%
3	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: < 24 kg/m^2, radiotherapy: yes; tumor grade/differentiation: ≤ 2, and Levels I-III, Lymph Nodes for Head and Neck: positive	82.5%
4	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: < 24 kg/m^2, radiotherapy: yes; tumor grade/differentiation: > 2, and SSF3: positive	88.0%
5	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: ≥ 24 kg/m^2, and betel nut chewing behavior: no	58.5%
6	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: ≤ 4 cm, BMI: ≥ 24 kg/m^2, betel nut chewing behavior: yes, and tumor grade/differentiation: > 2	76.5%
7	Drinking behavior: yes, age at diagnosis: < 65 years, smoking behavior: yes, surgery: no, tumor size: > 4 cm, SSF3: positive, and betel nut chewing behavior: no	60.0%

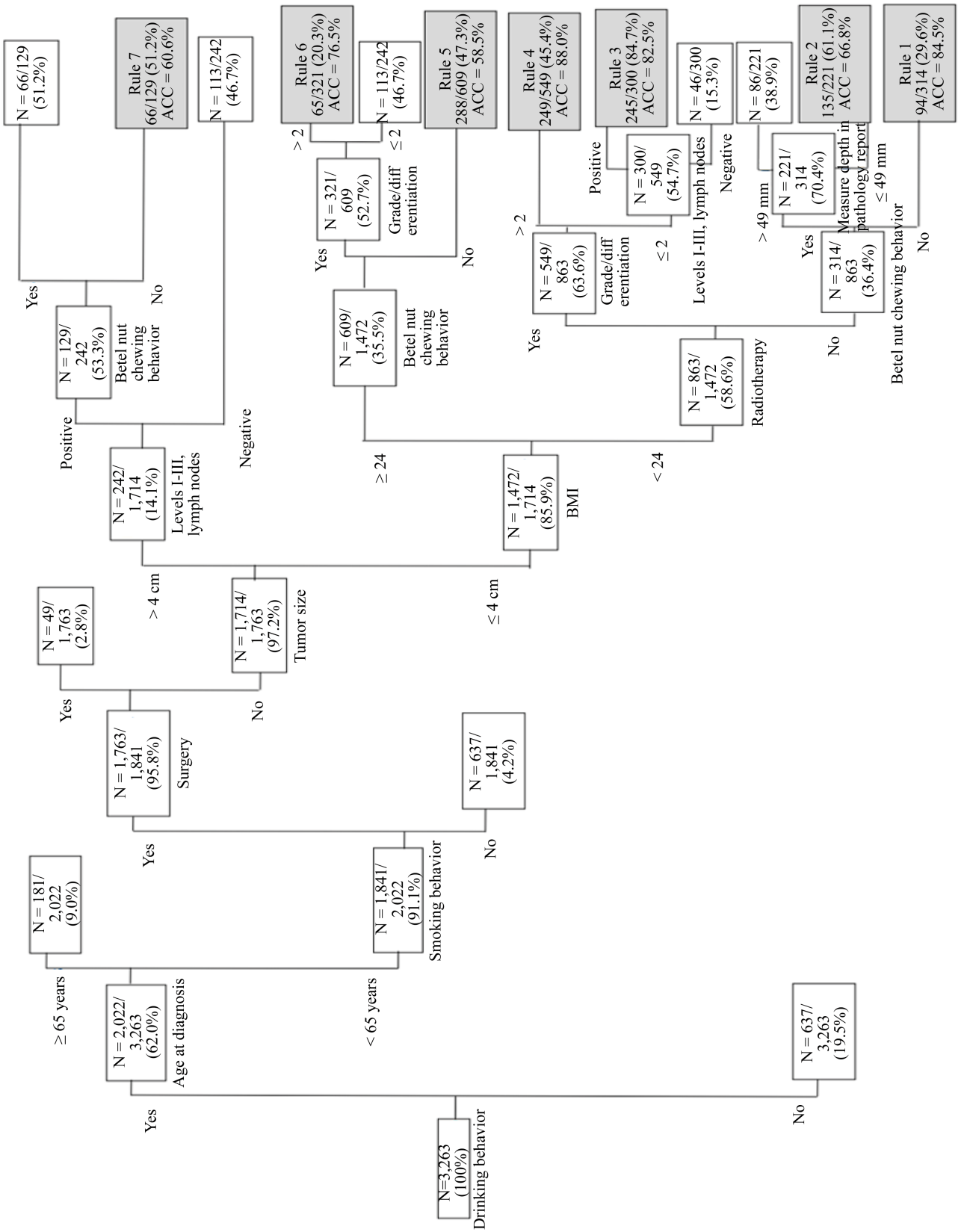

Figure 2. Decision tree for predicting a diagnosis of SPESC by using CART model

4. Discussion

Esophageal squamous cell carcinoma is one of the most common second primary malignancies in survivors of HNC [22]. According to the literature, survivors of oropharynx or hypopharynx malignancies should undergo regular endoscopic examinations for a minimum of 10 years to ensure that they do not develop primary esophageal cancer. However, the early detection of esophageal malignancies is challenging. Bugter et al. [23] examined the clinical data of 246 patients with HNC who developed second primary cancers over a median follow-up period of 4.96 years; they found that 23 of these patients developed esophageal cancer (1.4%). In the present study, 65 out of 2,863 survivors of HNC (2.3%) developed SPESC, with this proportion being similar to those reported in other studies (2% to 17%) [24]. Moreover, these 65 patients received a diagnosis of SPESC during the advanced stage of the cancer.

In one study, panendoscopy and other imaging methods enabled the detection of SPESC in 24% and 15% of all cases, respectively [25]. When combined with positron emission tomography or computed tomography, panendoscopy can improve the detection of unknown primary HNC [25–27]. However, previous studies have neither attempted to predict the occurrence of SPESC nor achieved early diagnosis of SPESC. Patients with HNC are recommended to undergo routine panendoscopy screening or Lugol chromoendoscopy; however, panendoscopy is an invasive procedure, and complications such as esophageal perforation, stricture, bleeding, and respiratory distress have been reported for this procedure. Therefore, panendoscopy is not used in routine clinical practice. Consequently, a reliable prediction scheme must be established for identifying patients with HNC who have high SPESC risk. Such a scheme can not only facilitate the early diagnosis of SPESC but also prevent the life-threatening complications associated with routine panendoscopy screening.

Some research has been conducted in Asia regarding risk factors and prediction models for SPESC associated with primary HNC; however, this research has not been sufficiently comprehensive. Bugter et al. developed a cause-specific Cox model to predict the occurrence of SPTs in 1,581 Western patients with cancer [23]. They identified tobacco and alcohol use, comorbidities, and tumor-containing oral cavity subsite as risk factors for SPTs. The C-index, namely the discriminative accuracy, of their SPT prediction model was 0.65 (95% CI = 0.61-0.68). However, considerably fewer variables were examined in the aforementioned study than in the present study. Bugter et al. used Adult Comorbidity Evaluation-27 to assess comorbidities. In contrast to Bugter et al., we considered BMI as a candidate factor and confirmed that it played a crucial role in our scheme. This factor has been increasingly considered in recent studies on cancer and metabolic diseases. Furthermore, most studies on pathological risk factors (i.e., X14-X19) have analyzed the relationship between image accuracy and the recurrence or progression of primary tumor. To the best of our knowledge, no research has used data on pathological risk factors to predict the occurrence of SPTs. Thus, by developing a clinical scheme for predicting the risk of SPESC with primary HNC, the present study makes a crucial contribution to the relevant literature.

A previous study developed a stacking-ensemble-based classification model to predict the occurrence of SPTs in patients with HNC [28]. By stacking ensembles of classifiers, we achieved a larger AUC and higher accuracy than achieved using any single classifier; however, the specificity and sensitivity were lower. Of all classifiers used in the present study, C4.5 had the highest AUC value of 0.9104 and thus was selected as the final classifier. Several clinical characteristics can be used to identify patients with HNC who have high SPESC risk. In the present study, the risk factors identified for the development of SPESC in the aforementioned patients included alcohol consumption, betel nut chewing, and smoking. This finding is in line with those of other studies. Alcohol consumption behavior was the main risk factor in our prediction scheme. Tseng et al. reported that the incidence of SPESC in patients with HNC in Taiwan increased from 1% in 1999 to 2.4% in 2009 [29]. Moreover, no drinking behavior, an age at diagnosis of $\geq$ 65 years, no smoking behavior, and surgery are positive clinical characteristics among survivors of HNC. Our prediction scheme revealed that the presence of these four protective clinical characteristics resulted in very low SPESC risk. This scheme can be used by clinicians to determine which patients should receive frequent SPESC screening.

Because of their significant ORs, an age at diagnosis of $<$ 65 years, tumor grade/differentiation of $>$ 2, smoking behavior, drinking behavior, and tumor depth were identified as significant clinical risk factors for SPESC in the present study. By contrast, BMI was the only protective clinical factor in this study. The highest model accuracy of 88.0% was achieved under the following candidate factor conditions: drinking behavior: yes; age at diagnosis: $<$ 65 years; smoking

behavior: yes; surgery: no; tumor size: ≤ 4 cm; BMI: < 24 kg/m^2; radiotherapy: yes; and tumor grade/differentiation: > 2. In conclusion, the results of statistical analysis and machine learning indicate that alcohol consumption is the strongest risk factor for SPESC. In addition to alcohol consumption behavior, an age of < 65 years at HNC diagnosis, smoking behavior, and surgery are positively associated with SPESC.

The present study contributes to the clinical literature and society by developing a noninvasive and efficient approach for identifying survivors of HNC who have high SPESC risk. The developed prediction scheme enables the early detection of SPESC, the development of personalized screening strategies, and the avoidance of unnecessary procedures for patients with a low risk of developing SPESC; thus, this scheme can contribute to improving quality of life and long-term outcomes for patients with cancer [30, 31]. It can also reduce health-care costs associated with overscreening and late-stage cancer treatments as well as guide policy and health-care strategies, particularly for populations with limited access to advanced health-care facilities [32]. A drawback of this study is that it collected data from a single institution, which may have limited the generalizability of its results. Moreover, this study's retrospective design may have introduced biases related to data collection and patient selection. The developed prediction scheme was not validated with independent datasets, which is crucial for confirming its applicability in different clinical settings. Future research can focus on addressing the aforementioned gaps to validate and refine the developed scheme, thereby ensuring its robustness and broad applicability [33, 34].

5. Conclusions

This study developed a scheme for predicting the occurrence of SPESC among survivors of HNC, thus addressing a crucial clinical need. A total of 19 clinical risk factors were analyzed, and significant risk factors and protective factors for SPESC were identified on the basis of their gain ratios. Data on the significant factors were then integrated into seven machine learning classifiers to generate predictions for the occurrence of SPESC. The results indicated that of the seven machine learning classifiers, CART exhibited the best performance, with its AUC, accuracy, and F1 score being 0.924, 86.8%, and 0.887, respectively. Thus, CART was selected as the machine learning algorithm for the developed scheme. The significant risk factors identified for SPESC included an age of < 65 years at HNC diagnosis, tumor grade/differentiation of > 2, smoking behavior, alcohol consumption, and tumor depth. Moreover, a BMI of ≥ 24 kg/m^2 was identified as the only significant protective factor for SPESC, which is a unique finding. The developed scheme can integrate data related to diverse clinical variables; thus, it can be used by clinicians to conduct robust identification of patients with HNC who have high SPESC risk. Such identification can ensure that patients with low SPESC risk do not undergo unnecessary invasive procedures. The developed scheme represents a major advancement in predictive oncology, and it can be employed to develop personalized treatment strategies and improve treatment outcomes for survivors of HNC. The results of this study highlight the potential of using machine learning approaches in advancing cancer surveillance and prevention efforts. Future studies should validate the developed scheme across broader populations and health-care settings to validate its utility and scalability.

Author contributions

Data curation, Chi-Chang Chang and Chun-Chia Chen; formal analysis, Chi-Chang Chang and Ming-Yi Lu; investigation, Chi-Chang Chang and Chun-Chia Chen; methodology, Chi-Chang Chang; validation, Chi-Chang Chang; writing-original draft, Chun-Chia Chen, and Ming-Yi Lu; writing-review and editing, Chi-Chang Chang and Ming-Yi Lu. All authors have read and agreed to the published version of the manuscript.

Funding

This study was funded by the Ministry of Science and Technology of Taiwan (grant number: MOST 110-2321-B-040-003), jointly funded by Chi Mei Medical Center and Kaohsiung Medical University Chung-Ho Memorial Hospital (grant number: 113CM-KMU-02), and the Chung Shan Medical University Hospital Foundation (grant number: CSH-2022-C-035).

Institutional review board statement

The study was conducted in accordance with the guidelines of the Declaration of Helsinki and was approved by the Institutional Review Board of the Chung Shan Medical University Hospital (protocol code CSMUH No: CS2-20114, 2020.6.24).

Data availability statement

Researchers who meet relevant access criteria can obtain the data examined in this study from the Institutional Review Board of the Chung Shan Medical University Hospital, Taichung City, Taiwan. Requests for these data may be sent to the Chung Shan Medical University Hospital (email: irb@csh.org.tw).

Conflict of interest

The authors declare that they have no conflicts of interest with respect to the work reported in this paper.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2021; 71(3): 209-249. Available from: https://doi.org/10.3322/caac.21660.

[2] Wild CP, Weiderpass E, Stewart BW. *World cancer report: cancer research for cancer prevention*. Lyon (FR): international agency for research on cancer. 2020. Available from: https://www.iccp-portal.org/system/files/resources/IARC%20World%20Cancer%20Report%202020.pdf [Accessed 15th August 2025].

[3] Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*. 2020; 6: 92. Available from: https://doi.org/10.1038/s41572-020-00224-3.

[4] Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*. 1953; 6(5): 963-968. Available from: https://doi.org/10.1002/1097-0142.

[5] Pan D, Xu W, Gao X, Yiyang F, Wei S, Zhu G. Survival outcomes in esophageal cancer patients with a prior cancer. *Medicine*. 2021; 100(7): e24798. Available from: https://doi.org/10.1097/MD.0000000000024798.

[6] Huang YW, Wang YP, Lee TL, Chang CF, Hou MC, Tai SK, et al. Image-enhanced endoscopy for detection of second primary esophageal neoplasms in patients with hypopharyngeal cancer: prevalence, risk factors, and characteristics. *Journal of the Chinese Medical Association*. 2021; 84(10): 963-968. Available from: https://doi.org/10.1097/JCMA.0000000000000592.

[7] Sheikh M, Roshandel G, McCormack V, Malekzadeh R. Current status and future prospects for esophageal cancer. *Cancers*. 2023; 15(3): 765. Available from: https://doi.org/10.3390/cancers15030765.

[8] Moon PK, Ma Y, Megwalu UC. Head and neck cancer stage at presentation and survival outcomes among Native Hawaiian and other Pacific Islander patients compared with Asian and White patients. *JAMA Otolaryngology-Head & Neck Surgery*. 2022; 148(7): 636-645. Available from: https://doi.org/10.1001/jamaoto.2022.1086.

[9] van de Ven SEM, de Graaf W, Bugter O, Spaander MCW, Nikkessen S, de Jonge PJF, et al. Screening for synchronous esophageal second primary tumors in patients with head and neck cancer. *Diseases of the Esophagus*. 2021; 34(10): doab037. Available from: https://doi.org/10.1093/dote/doab037.

[10] Salzberg SL. C4.5: Programs for machine learning by J. Ross Quinlan. *Machine Learning*. 1994; 16: 235-240. Available from: https://doi.org/10.1007/BF00993309.

[11] Leng J, Qiu H, Huang Q, Zhang J, Zhou H. Recommendations for broadening eligibility criteria in esophagus cancer clinical trials: the mortality disparity of esophagus cancer as a first or second primary malignancy. *Journal of Thoracic Disease*. 2024; 16(6): 3882-3896. Available from: https://doi.org/10.21037/jtd-23-1881.

[12] Kuhn M. A short introduction to the caret package. *R Foundation for Statistical Computing*. 2015; 1: 1-10.

[13] Deane-Mayer Z, Knowles J. *CaretEnsemble: ensembles of caret models*. Version 2.0.0. 2016. Available from: https://cran.r-project.org/package=caretEnsemble [Accessed 15th August 2025].

[14] Breiman L. Random forest. *Machine Learning*. 2001; 45: 5-32. Available from: https://doi.org/10.1023/A:1010933404324.

[15] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*. 2006; 15(3): 651-674. Available from: https://doi.org/10.1198/106186006X133933.

[16] Landwehr N, Hall M, Frank E. Logistic model trees. *Machine Learning*. 2005; 59(1-2): 161-205. Available from: https://doi.org/10.1007/s10994-005-0466-3.

[17] Friedman JH. Multivariate adaptive regression splines. *Annals of Statistics*. 1991; 19(1): 1-67. Available from: https://doi.org/10.1214/aos/1176347963.

[18] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986; 323(6088): 533-536. Available from: https://doi.org/10.1038/323533a0.

[19] Cai Q, Hong Y, Huang X, Chen T, Chen C. Current status and prospects of diagnosis and treatment for esophageal cancer with supraclavicular lymph node metastasis. *Frontiers in Oncology*. 2024; 14: 1431507. Available from: https://doi.org/10.3389/fonc.2024.1431507.

[20] Grubinger T, Zeileis A, Pfeiffer KP. Evtree: evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software*. 2014; 61(1): 1-29. Available from: https://doi.org/10.18637/jss.v061.i01.

[21] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995; 20(3): 273-297. Available from: https://doi.org/10.1007/BF00994018.

[22] Oshima K, Tsushima T, Ito Y, Kato K. Recent progress in chemoradiotherapy for oesophageal squamous cell carcinoma. *Japanese Journal of Clinical Oncology*. 2024; 54(4): 395-402. Available from: https://doi.org/10.1093/jjco/hyae005.

[23] Bugter O, van Iwaarden DLP, van Leeuwen N, Nieboer D, Dronkers EAC, Hardillo JAU, et al. A cause-specific Cox model for second primary tumors in patients with head and neck cancer: a RONCDOC study. *Head & Neck*. 2021; 43(6): 1881-1889. Available from: https://doi.org/10.1002/hed.26666.

[24] Vogt A, Schmid S, Heinimann K, Frick H, Herrmann C, Cerny T, et al. Multiple primary tumours: challenges and approaches, a review. *ESMO Open*. 2017; 2(2): e000172. Available from: https://doi.org/10.1136/esmoopen-2017-000172.

[25] Wan M, Yang X, He L, Meng H. Elucidating the clonal relationship of esophageal second primary tumors in patients with laryngeal squamous cell carcinoma. *Infectious Agents and Cancer*. 2023; 18(1): 75. Available from: https://doi.org/10.1186/s13027-023-00540-5.

[26] Du J, Bao Z, Liang T, Zhao H, Zhao J, Xu R, et al. Risk factors for metachronous esophageal squamous cell carcinoma after endoscopic or surgical resection of esophageal carcinoma: a systematic review and meta-analysis. *Frontiers in Oncology*. 2023; 13: 1241572. Available from: https://doi.org/10.3389/fonc.2023.1241572.

[27] Guo QQ, Ma SZ, Zhao Y, Beeraka NM, Gu H, Zheng YF, et al. Association of definitive radiotherapy for esophageal cancer and the incidence of secondary head and neck cancers: a SEER population-based study. *World Journal of Oncology*. 2024; 15(4): 598-611. Available from: https://doi.org/10.14740/wjon1787.

[28] Chang CC, Huang TH, Shueng PW, Chen SH, Chen CC, Lu CJ, et al. Developing a stacked ensemble-based classification scheme to predict second primary cancers in head and neck cancer survivors. *International Journal of Environmental Research and Public Health*. 2021; 18(23): 12499. Available from: https://doi.org/10.3390/ijerph182312499.

[29] Tseng CM, Wang HH, Lee CT, Tai CM, Tseng CH, Chen CC, et al. A nationwide population-based study to assess the risk of metachronous esophageal cancers in head and neck cancer survivors. *Scientific Reports*. 2020; 10: 884. Available from: https://doi.org/10.1038/s41598-020-57630-6.

[30] Leng J, Qiu H, Huang Q, Zhang J, Zhou H. Recommendations for broadening eligibility criteria in esophagus cancer clinical trials: the mortality disparity of esophagus cancer as a first or second primary malignancy. *Journal of Thoracic Disease*. 2024; 16(6): 3882-3896. Available from: https://doi.org/10.21037/jtd-23-1881.

[31] Cai Q, Hong Y, Huang X, Chen T, Chen C. Current status and prospects of diagnosis and treatment for esophageal cancer with supraclavicular lymph node metastasis. *Frontiers in Oncology*. 2024; 14: 1431507. Available from: https://doi.org/10.3389/fonc.2024.1431507.

[32] Oshima K, Tsushima T, Ito Y, Kato K. Recent progress in chemoradiotherapy for oesophageal squamous cell carcinoma. *Japanese Journal of Clinical Oncology*. 2024; 54(4): 395-402. Available from: https://doi.org/10.1093/jjco/hyae005.

[33] Sheikh M, Roshandel G, McCormack V, Malekzadeh R. Current status and future prospects for esophageal cancer. *Cancers*. 2023; 15(3): 765. Available from: https://doi.org/10.3390/cancers15030765.

[34] Moon PK, Ma Y, Megwalu UC. Head and neck cancer stage at presentation and survival outcomes among Native Hawaiian and other Pacific Islander patients compared with Asian and White patients. *JAMA Otolaryngology-Head & Neck Surgery*. 2022; 148(7): 636-645. Available from: https://doi.org/10.1001/jamaoto.2022.1086.