International Journal of Medical Informatics 168 (2022) 104896

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

# An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer

Rasheed Omobolaji Alabi [a,b,*], Alhadi Almangush [a,c,d], Mohammed Elmusrati [b], Ilmo Leivo [d], Antti A. Mäkitie [a,e,f]

[a] Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[b] Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland
[c] Department of Pathology, University of Helsinki, Helsinki, Finland
[d] University of Turku, Institute of Biomedicine, Pathology, Turku, Finland
[e] Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[f] Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institute and Karolinska University Hospital, Stockholm, Sweden

ARTICLE INFO

*Keywords:*
Machine Learning
Oropharyngeal Cancer
Prognostication
Precision Medicine
Human Papillomavirus

ABSTRACT

*Background:* The optimal management of oropharyngeal squamous cell carcinoma (OPSCC) includes both surgical and non-surgical, that is, (chemo)radiotherapy treatment options and their combinations. These approaches carry a risk of specific treatment-related side effects. HPV-positive OPSCC has been reported to be more sensitive to (chemo)radiotherapy-based treatment modalities. **Objectives**: This study aims to demonstrate how machine learning can aid in classifying OPSCC patients into risk groups (low-chance or high-chance) for overall survival. We examined the input variables using permutation feature importance. Furthermore, we provided explanations and interpretations using the Local Interpretable Model Agnostic Explanations (LIME) and SHapley Additive Explanation (SHAP) frameworks. **Methods**: The machine learning model for 3164 OPSCC patients was built using data obtained from the Surveillance, Epidemiology, and End Results (SEER) program database. A total of five variants of tree-based machine learning algorithms (voting ensemble, light GBM, XGBoost, Random Forest, and Extreme Random Trees) were used to divide the patients into risk groups. The developed model with the best predictive performance was temporally validated with a different cohort. **Results**: The voting ensemble machine learning algorithm showed an accuracy of 88.3%, Mathews' correlation coefficient of 0.72, and weighted area under curve of 0.93, when temporally validated. Human papillomavirus (HPV) status, age of the patients, T stage, marital status, N stage, and the treatment modality (surgery with postoperative radiotherapy) were found to have the most significant effects on the ability of the machine learning model to predict overall survival. Similarly, for the individual patients with SHAP framework, HPV status, gender, and treatment modality (surgery with postoperative radiotherapy) were the input features that improved the model's prediction. **Conclusion**: The proposed stratification of OPSCC patients into risk groups by machine learning techniques can provide accurate predictions and thus aid clinicians in administering early and personalized interventions. Clinicians could utilize the predicted risk with the explanations offered by the SHAP and LIME frameworks to understand previously undetected relationships between prognostic variables to make informed clinical decisions and effective interventions.

## 1. Introduction

Oropharyngeal squamous cell carcinoma (OPSCC) is one of the most common head and neck carcinomas [1,2]. The incidence of OPSCC has increased in recent years – especially in developed countries, where over 100,000 new cases are diagnosed yearly [3–5]. The current modifiable risk factors include human papillomavirus (HPV), in addition to the traditional heavy alcohol use and smoking [3,6]. The frequently encountered HPV-associated OPSCC affects the younger population in particular, compared with the conventional type of upper respiratory

* Corresponding author at: Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland.
*E-mail address:* rasheed.alabi@helsinki.fi (R. Omobolaji Alabi).

https://doi.org/10.1016/j.ijmedinf.2022.104896
Received 25 June 2022; Received in revised form 27 September 2022; Accepted 7 October 2022
Available online 13 October 2022

airway squamous cell cancer [7–10]. Consequently, the landscape of the head and neck carcinoma has been transformed despite the initially reported decline in the incidence of head and neck cancers [11].

OPSCC and its treatment may have devastating effects on the quality of life of these patients, especially the HPV-related tumors typically present at advanced stage, i.e., with neck metastases, while the primary tumor may still be nonvisible. Considering the associated sequelae of the management of OPSCC, such as dysphagia and xerostomia, the proper individualized planning of the treatment and management of these patients is of utmost importance [7].

Patients treated for HPV-associated OPSCC are usually confronted with decades of survival with an adversely impacted quality of life [12]. This is largely due to the side-effects of aggressive treatment with a combination of surgery and radiotherapy and with concurrent chemotherapy. Therefore, it is imperative to stratify the patients into risk groups for individualized treatment interventions that can positively enhance their quality of life. For instance, suggesting a treatment de-intensification for low-risk HPV-positive OPSCC patients and a more aggressive treatment for high-risk HPV-positive subgroups would lead to more individualized outcomes.

Of note, the increasing costs of cancer care and its treatment are expected to result in a significant burden in the form of greater loss of economic resources (financial loss) and opportunities for patients (morbidity, reduced quality of life, and decreased life expectancy), families, employers, and society at large [13]. The healthcare cost for OPSCC varies from one country, cancer site, and center to another [13]. For example, in the United States, the average overall medical costs for treatment of oropharyngeal cancer were estimated at $77116 for OPSCC patients who received surgery (Sx) only, $88895 for radiotherapy only (RT), $102910 for Sx + RT, and $115779 for chemoradiotherapy (CRT) [14].

Therefore, it is important for clinicians to have an insightful treatment approach that can improve the quality of cancer care and prevent further economic and financial losses. An example of an approach that can enhance decision-making regarding an effective treatment alternative is to provide an artificial intelligence-based second opinion to the clinicians. The subfield of artificial intelligence, that is, machine learning (ML) can be used to classify the OPSCC patients into risk groups in relation to the patients' chance of overall survival.

This study presents an insightful approach that can assist in the effective stratification of OPSCC patients into risk groups using machine learning (ML) techniques. These techniques are able to analyze the hidden and complicated interactions that exist between variables [15]. The same approach has been reported to show significant contributions in the stratification of patients into risk groups in the prediction of locoregional recurrences [15,16]. Therefore, in this study, we aim to explore the potential of machine learning techniques in the prediction of survival of OPSCC patients.

The contribution of this study is fourfold. First, it develops an implementable machine-learning-based overall survival risk stratification model for OPSCC patients. Having a prior knowledge of the risk of the patients in terms of the chance of overall survival (high chance or low chance of overall survival) can enhance early and fitted personalized medical care and interventions. Furthermore, this will prevent increasing financial costs of cancer treatment, improve quality of cancer care, and enable healthcare organizations to efficiently deliver population-based health management interventions. Second, it provides a global explanation for the developed predictive model through the use of permutation feature importance. Third, explanations and interpretations for the individual predictions (local explanations) were provided using the Local Interpretable Model Agnostic Explanations (LIME) and SHapley Additive Explanation (SHAP) frameworks. The essence of this framework is to provide explainable and interpretable ML models that are transparent and trusted, facilitate human-AI model understanding, and aid model adoption. Finally, the combination of a predictive model with explanations and interpretations using LIME & SHAP frameworks is poised to allow clinicians to make informed decisions regarding treatment options rather than trusting the prediction by a ML-based model. The overall survival risk stratification model is expected to demonstrate reasonable predictive ability while offering explanations and interpretations of the predictions. This will allow clinicians to understand previously undetected relationships between prognostic variables to make informed clinical decisions and perform effective interventions.

This paper provides background information about OPSCC and its economic implications. Then, it describes the objectives of the study and the proposed method, our results (using arrays of performance metrics after the training and temporal validations of the model), and the findings' practical implications. We conclude with the limitations of the study and suggested avenues for further research.

## 2. Materials and methods

### 2.1. Collection of data

In this study, we retrieved data from the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH). This database was used because it is a publicly available database with a high-quality, significant number of cases and non-identifiable information on patients with various cancers [17,18]. These important characteristics are aimed at ensuring large-scale outcome analysis research.

#### 2.1.1. Ethical permission

The ethical permission to use the SEER database was granted with the identification number: 17247-Nov2020 (alabir). The access to the human papillomavirus status of the patients was granted with the same identification number.

#### 2.1.2. Selection of patient's attributes

The specialized database of the SEER program of the NCI was searched for Nov 2020 submission [2010– 2015] (Fig. 1). The consideration for oropharyngeal cancer as contained in the SEER database include the base (posterior one-third) of the tongue, oropharynx, tonsil, vallecula, and soft palate (Fig. 1). The inclusion criteria included that all the cases have known diagnostic information. The included known clinical and pathologic characteristics were race, gender, age at diagnosis, marital status, TNM status according to the American Joint Committee on Cancer (AJCC) 7th edition, grade, human papillomavirus status, treatment modalities (surgery, and radiotherapy). The disease-free survival (in months) and overall survival of the patients were also included. Disease-free survival is the time period from the beginning (or end) of treatment until the patient is diagnosed of recurrence while overall survival refers to the time period from the beginning (or end) of treatment until the patients die of any cause. These parameters were considered in a similar study on oropharyngeal cancer [19,20] and in a study on prognostic markers [21].

#### 2.1.3. Extracted cases of oropharyngeal cancer

A total of 3284 cases of oropharyngeal cancer were found eligible for inclusion in this study (Table 1). The detailed extraction process for these cases is presented in Fig. 1. Additionally, a detailed explanation of the clinical and pathologic parameters used in this study is presented in Table 1. The data used in the training of the machine learning-based model included 3164 patients while 120 patients were reserved to temporally validate the model.

### 2.2. Machine learning algorithms

Five variants of ensemble machine learning algorithms were examined in this study. Ensemble learning paradigm, also known as multiple classification systems, has the potential to produce a learner that is

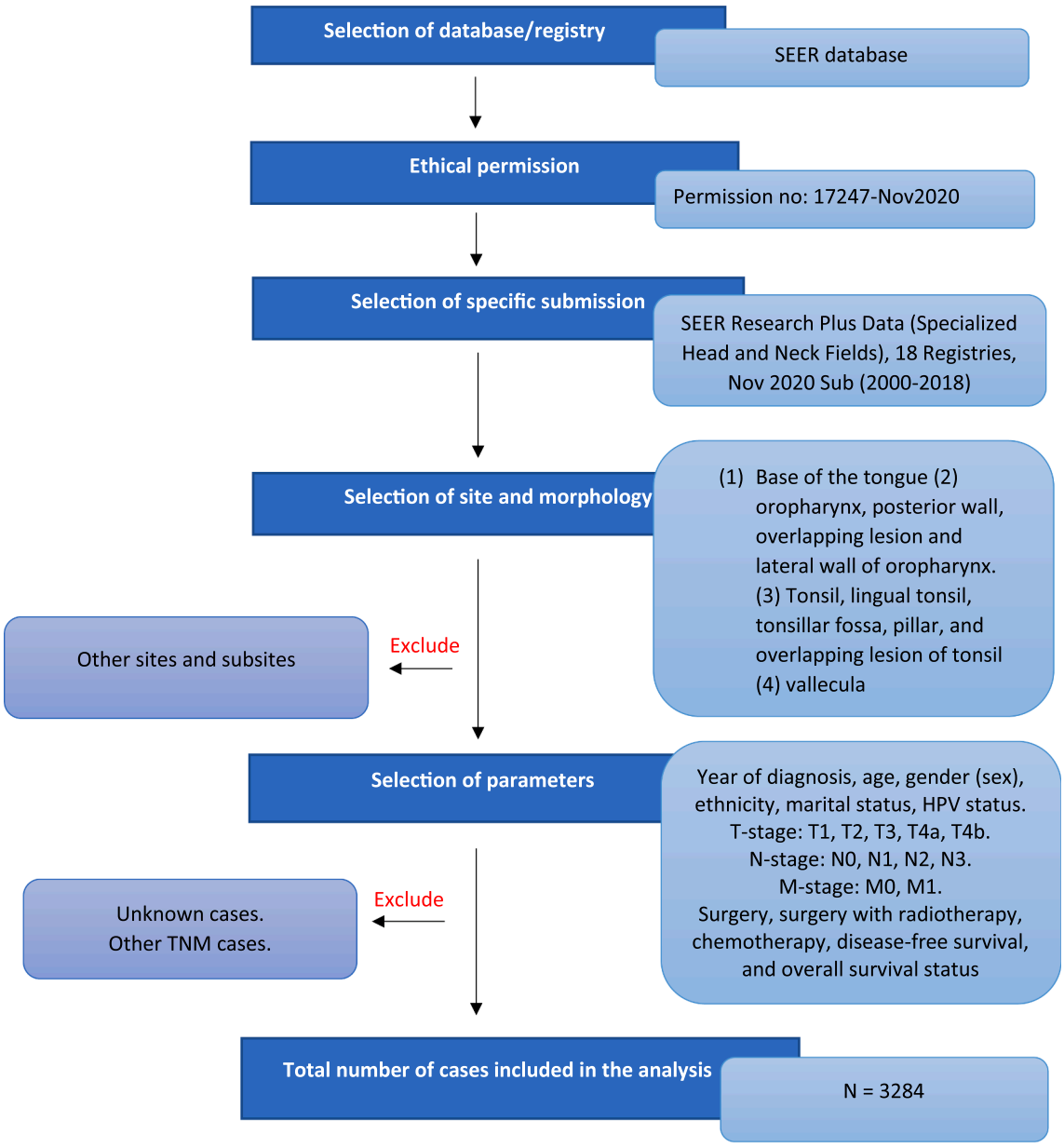


**Fig. 1.** A schematic of data extraction from the Surveillance, Epidemiology, and End Results (SEER).

generalizable [22]. Technically, an ensemble paradigm ensures that multiple versions of the same machine learning model (weak or strong) are trained in such a way that each ensemble member is different (i.e., the decision trees are fit on different subsamples of the training dataset) [23]. Then, this process is followed by a selective combination of member classifiers into a better classification using any of the several appropriate and efficient ensemble methods, such as voting, averaging, bagging, stacking boosting, or boosting [22,24,25]. The combination method largely depends on the problem to be solved [26]. The five variants examined in this study are voting ensemble, light gradient boosting machine (Light GBM), extreme gradient boosting machine (XGBM), random forest, and extreme random trees.

i. Voting ensemble' also known as a *meta-model* or *model of models* because it combines the prediction from multiple other models [27]. The Azure machine learning studio uses the soft voting methodology, whereby it sums the predicted probabilities for each class label. The predicted class label with the largest sum probability is given as the final prediction. Hence, it is known as a majority voting ensemble [22].
ii. Light Gradient Boosting (LightGBM) uses the boosting methodology where many moderately accurate weak learners are integrated (boosted) to form strong learning [22]. The implementation of LightGBM introduces two novel techniques. These are gradient-based one-side sampling and exclusive feature bundling [28]. These two ideas resulted in the training speed and improved predictive performance of LightGBM [28].
iii. Extreme Gradient Boosting (XGBM or XGBoost), like the LightGBM, also uses the boosting methodology. The main difference between LightGBM and XGBM lies in how they weigh samples and hypotheses for training [22]. For example, in XGBM, there is a level-wise (horizontal) growth of the trees [29]. Thus, making it more robust than LightGBM that uses leaf-wise (vertical) growth of trees which makes it prone to overfitting [29,30]. In terms of the hypothesis for training, XGBoost uses a pre-sorted and histogram-based approach for computing the best split, thus, making it less fast compared to LightGBM that utilizes one of the

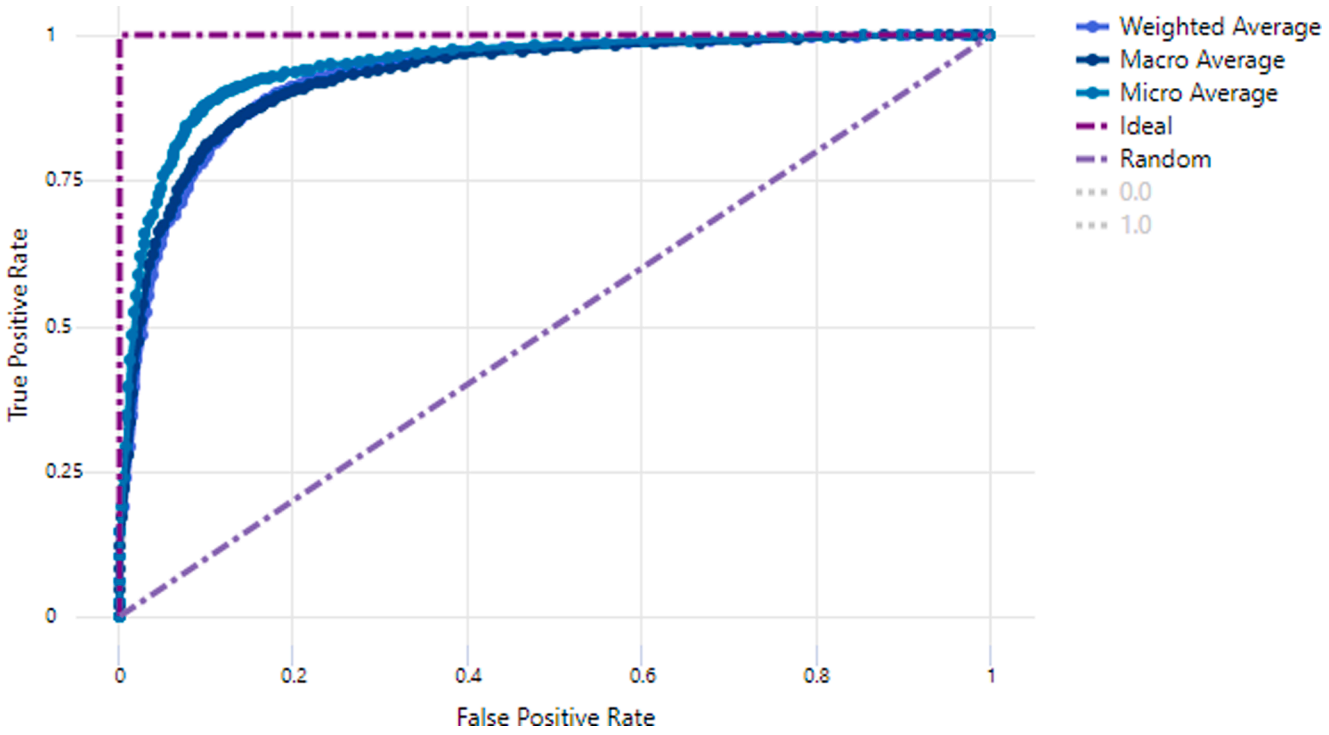


**Fig. 2.** The area under receiving operating characteristics curve for voting ensemble method.

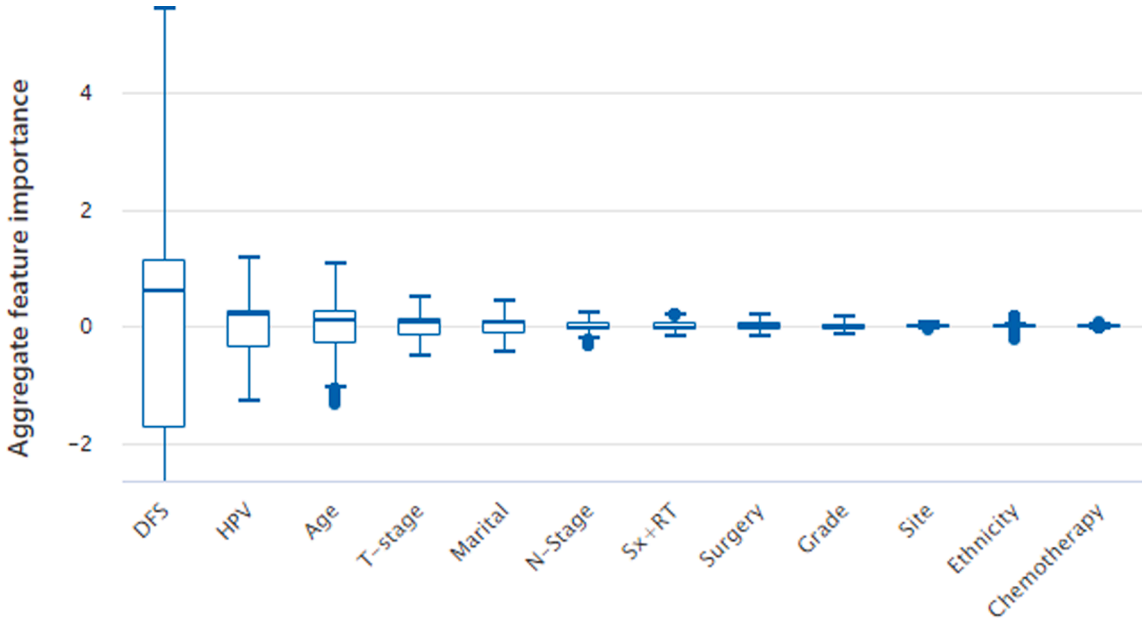


**Fig. 3.** Permutation feature importance of the input parameters on the model.

novel techniques (gradient-based one-side sampling) to compute its best split [30]. Depending on size of data and resources, both LightGBM and XGBM have showed promising model performance for various classification and regression tasks [23,30].

iv. Random forest (RF) samples subsets of the entire dataset with replacements (bootstrapping). Multiple decision trees (a forest) are constructed over different subsets [16]. The final prediction is made using majority voting or averaging from these trees.

v. Extreme random trees, also known as extra trees differs from the RF in that it samples the entire dataset randomly [16,31]. During training, extreme random trees construct trees over every observation in the dataset but with different subsets of features. When constructing each decision tree, the extreme random tree splits nodes randomly. This makes it faster since the node splits are randomized. Thus, it produces a low variance compared to RF [31].

### 2.3. Machine learning training

The machine learning training process started with the preprocessing of data, which is necessary to ensure high data quality for the training process [32–34]. All forms of incomplete data, such as

**Table 1**
Extracted baseline demographic and tumor characteristics of oropharyngeal cancer patients from the SEER database (N = 3284 for both training and temporal validation).

| Parameters | Total Number of cases for training (n = 3164) | Categorization for machine learning training | Total Number of cases for external validation of the web-based tool (n = 120) |
|---|---|---|---|
| Year: | | | |
| 2010 | 267 (8.4 %) | | – |
| 2011 | 388 (12.3 %) | No categorization (Not used in training) | 1 (0.8 %) |
| 2012 | 519 (16.4 %) | | – |
| 2013 | 614 (19.4 %) | | 8 (6.7 %) |
| 2014 | 690 (21.8 %) | | 51 (42.5 %) |
| 2015 | 686 (21.7 %) | | 60 (50.0 %) |
| Gender: | | | |
| Male | 2526 (79.8 %) | 0 = Male | 98 (81.7 %) |
| Female | 638 (20.2 %) | 1 = Female. | 22 (18.3 %) |
| Ethnicity: | | | |
| White | 2825 (89.3 %) | 0 = White | 108 (90.0 %) |
| Black | 233 (7.4 %) | 1 = Black | 11 (9.2 %) |
| Other | 106 (3.4 %) | 2 = Others (American Indian /AK Native, Asian pacific | 1 (0.8 %) |
| Marital Status: | | | |
| Married* | 1900 (60.1 %) | 1 = Married | 76 (63.3 %) |
| Single** | 1264 (39.9 %) | 0 = Single | 44 (36.7 %) |
| Grade: | | | |
| Grade I: Well differentiated | 174 (5.5 %) | Grade I = 1 | 7 (5.8 %) |
| Grade II: Moderately differentiated | 1305 (41.2 %) | Grade II = 2 | 55 (45.8 %) |
| Grade III: Poorly differentiated | 1650 (52.1 %) | Grade III = 3 | 53 (44.2 %) |
| Grade IV: Undifferentiated | 35 (1.1 %) | Grade IV = 4 | 5 (4.2 %) |
| HPV Status: | | | |
| Negative | 1143 (36.1 %) | 0 = HPV-negative | 36 (30.0 %) |
| Positive | 2021 (63.9 %) | 1 = HPV-positive | 84 (70.0 %) |
| Site: | | | |
| Base of tongue | 1178 (37.2 %) | 1 = Base of tongue | 44 (36.6 %) |
| Oropharynx [+] | 218 (6.9 %) | 2 = Oropharynx | 11 (9.2 %) |
| Tonsil [++] | 1742 (55.1 %) | 3 = Tonsil | 62 (51.6 %) |
| Valecullar | 26 (0.8) | 4 = Valecular | 3 (2.5 %) |
| Tumor (*T*-stage) | | | |
| T1 | 875 (27.6 %) | 1 = T1 | 34 (28.3 %) |
| T2 | 1279 (40.4 %) | 2 = T2 | 47 (39.2 %) |
| T3 | 580 (18.3 %) | 3 = T3 | 19 (15.8 %) |
| T4 | 430 (13.6 %) | 4 = T4 | 20 (16.7 %) |
| Nodal (*N*-stage) | | | |
| N0; No regional lymph node metastasis | 1417 (44.8 %) | 0 = N0 | 50 (41.7 %) |
| N1; Single node regional lymph node metastasis | 1424 (45.0 %) | 1 = N1 | 60 (50.0 %) |
| | 323 (10.2 %) | 3 = N3 | 10 (8.3 %) |
| N3; Cancer has spread to one or more lymph node | | | |
| Metastases (M−stage) | | | |
| AJCC M0; No distant metastasis | 3085 (97.5 %) | 0 = M0 | 116 (96.6 %) |
| AJCC M1; Presence of distant metastasis | 79 (2.5 %) | 1 = M1 | 4 (3.3 %) |
| Treatment parameters | | | |
| Surgery with postoperative radiotherapy (Sx + RT) | 1270 (40.1 %) | 1 = Sx + RT | 57 (47.5 %) |
| Surgery with chemoradiotherapy (Sx + CRT) | 579 (18.3 %) | 1 = Sx + CRT | 28 (23.3 %) |
| Definitive chemoradiotherapy | 413 (13.1 %) | 1 = CRT | 11 (9.2 %) |
| Surgery alone | 518 (18.3 %) | 1 = Surgery | 6 (5.0 %) |
| No treatment given | 384 (12.1 %) | 0 = No treatment given | 18 (15.0 %) |
| Overall Status | | | |
| Alive | 2117 (66.9 %) | 0 = Alive | 84 (70.0 %) |
| Dead | 1047 (33.1 %) | 1 = Dead | 36 (30.0 %) |

HPV: Human papillomavirus; *Married including common law; ** Single includes never married, widowed, divorced, unmarried/domestic partner, and separated: [+]Oropharynx includes posterior wall of oropharynx, overlapping lesion of oropharynx, and lateral wall of oropharynx; [++]Tonsil includes lingual tonsil, overlapping lesion of tonsil, and tonsillar pillar.

missing values, incorrect input values, and incomplete entries were removed from the onset of the extraction process (Fig. 1). The resulting data from the pre-processing phase were further checked to ensure that it had been correctly preprocessed. The data was further categorized to ensure that was in a reliable format for the machine learning training phase (Table 1). From the extracted data, the input parameters for the machine learning training included age at diagnosis, gender, ethnicity, marital status, grade, HPV status, tumor site, TNM-stage, and treatment modalities (surgery, surgery + radiotherapy (Sx + RT), and surgery + chemoradiotherapy (Sx + CRT). The target outcome is the overall survival of the patient. The entire training phase was done using Microsoft Azure Machine Learning Studio (Azure ML 2021) to build the predictive model [35].

The training process was performed using 5-fold cross-validation due to the relatively large amount of dataset. This approach minimizes bias and imitates external validation [36]. Cross-validation was chosen in order to test the predictive ability of the model on new data that were not used in estimating it, hence offering the ability to detect overfitting or selection bias [37]. In addition, it gives insight into how the model will generalize to an independent dataset (i.e., an unknown dataset) as demonstrated in subsection 2.3.2. We used the ensemble method as the training algorithm due to its ability to reduce generalization error. Therefore, we selected 5 variants of ensemble methods as the training algorithm. These variants were stacked ensemble, voting ensemble, light gradient boosting machine (Light GBM), extreme gradient boosting machine (XGBM), random forest, and extreme random trees. The hyperparameter tuning was done where necessary to ensure that a reasonable weighted area under curve (AUC) was achieved. The model with the best performance was then temporally validated (**sub-section 2.3.2**). The result from the external validation was evaluated mainly on weighted AUC. Other performance metrics were examined (**sub-section**

**3.4**).

#### 2.3.1. Performance metrics of the trained model

Apart from accuracy, other performance metrics such as Mathew correlation coefficient (MCC), F1 score, confusion matrix categories, sensitivity, specificity, and weighted area under curve were used to evaluate the performance of the model (Table 2). Notably, Mathews' correlation coefficient has been reported to be a reliable metric for classification tasks [38].

#### 2.3.2. Temporal validation of the model

A temporal external validation method was used where 120 cases that had not been used in the training or testing were used to evaluate the true performance of this model. The performance of the model when temporally validated was considered the gold standard performance (Table 3). Temporal validation approach may be posited as a viable validation process for prediction model reproducibility and generalizability [39]. Therefore, it is considered the simplest form of external validation, which is more robust and stronger than internal validation [40] even though the subsequent cohorts used for temporal validation were recruited from the same data source [41].

#### 2.3.3. Permutation feature importance

We performed permutation feature importance (PFI) to examine the global explanation of the model. PFI works by shuffling the data in such a way that one feature is removed at a time while the corresponding effect of the shuffled feature on the performance metrics of the model is estimated [42]. The larger the change, the more important is the feature to the model's performance in stratifying the patients into risk groups for overall survival.

### 2.4. Interpretability with Local Interpretable model Agnostic explanations (LIME) and SHapley additive explanations (SHAP)

We used the LIME framework to examine the probability of the correctness of the predictions made by our trained model. Furthermore, it gives an overview of how each of the input parameters contributed to the risk stratification results given by our model. This framework uses the *LimeTabularExplainer* to fit the training data. To demonstrate how this framework works, we examined the LIME framework explanations on the stratified risk predictions made by our model for a single patient [Fig. 4]. This approach offers an interpretation regarding the prediction made by the model. The SHAP framework, on the other hand, is a model-agnostic approach that is based on cooperative game theory to explain the prediction of any machine learning model [43]. We used the classic SHAP values from game theory to capture the average marginal contribution of each input parameter (**sub-section 2.1.2**) to the single prediction made by the model [43,44]. That is, we used the SHAP framework to divide the variability of the predictions made by the model between the available covariates. Thus, the contribution of each variable to every single prediction by the model can be assessed regardless of the underlying model (Fig. 5a). Therefore, it is a model-agnostic framework. In this study, we used an extreme gradient boosting-based model with Python version 3.10.4. The motivation of the SHAP approach is to offer some level of explanation and interpretability to the predictions made by the model. We demonstrated the SHAP framework on the stratified

**Table 3**
Temporal validation with cases neither used in training nor testing (N = 120 cases).

| | Performance metrics | Voting Ensemble method |
|---|---|---|
| Confusion matrix parameters | True positive | 78 |
| | False positive | 6 |
| | False negative | 8 |
| | True negative | 28 |
| Predictive value | PPV (Precision) | 0.93 |
| | NPV | 0.78 |
| Other metrics | Sensitivity (recall) | 0.91 |
| | Specificity | 0.82 |
| | F1 score | 0.92 |
| Accuracy | Accuracy | 88.3 % |
| | Balanced accuracy | 85.3 % |
| | Weighted accuracy | 90.5 % |
| Correlation | Mathew's correlation | 0.72 |
| AUC | Weighted AUC | 0.934 |

PPV: Positive predictive value; NPV: Negative predictive value; AUC: Area under curve.

**Table 2**
Machine learning algorithm performance metrics from the training phase (N = 3164 cases for training set).

| | Performance metrics | Voting Ensemble | Light GBM | XGBoost Classifier | Random Forest | Extreme Random Trees |
|---|---|---|---|---|---|---|
| Confusion matrix parameters | True positive | 2044 | 2044 | 2038 | 2043 | 2049 |
| | False positive | 73 | 73 | 79 | 74 | 68 |
| | False negative | 251 | 258 | 253 | 257 | 271 |
| | True negative | 796 | 789 | 794 | 790 | 776 |
| Predictive value | PPV (Precision) | 0.97 | 0.97 | 0.96 | 0.97 | 0.97 |
| | NPV | 0.76 | 0.75 | 0.76 | 0.75 | 0.74 |
| Other metrics | Sensitivity (recall) | 0.89 | 0.89 | 0.89 | 0.89 | 0.88 |
| | Specificity | 0.92 | 0.92 | 0.91 | 0.91 | 0.92 |
| | F1 score | 0.93 | 0.93 | 0.92 | 0.93 | 0.92 |
| Accuracy | Accuracy | 89.8 % | 89.5 % | 89.5 % | 89.5 % | 89.2 % |
| | Balanced accuracy | 86.3 % | 85.9 % | 86.0 % | 85.9 % | 85.5 % |
| | Weighted accuracy | 92.5 % | 92.3 % | 92.2 % | 92.3 % | 92.3 % |
| Correlation | Mathews' correlation coefficient | 0.77 | 0.76 | 0.76 | 0.76 | 0.75 |
| AUC | Weighted AUC | 0.929 | 0.926 | 0.929 | 0.925 | 0.923 |

PPV: Positive predictive value; NPV: Negative predictive value; AUC: Area under curve.

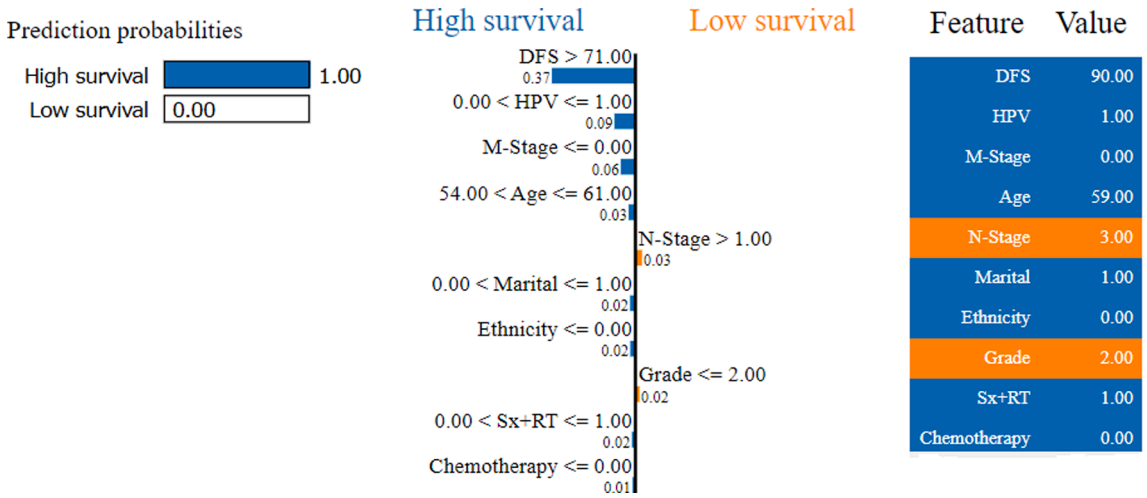


**Fig. 4.** The Local Interpretable Model Agnostic Explanations (LIME) framework for individual predictions.

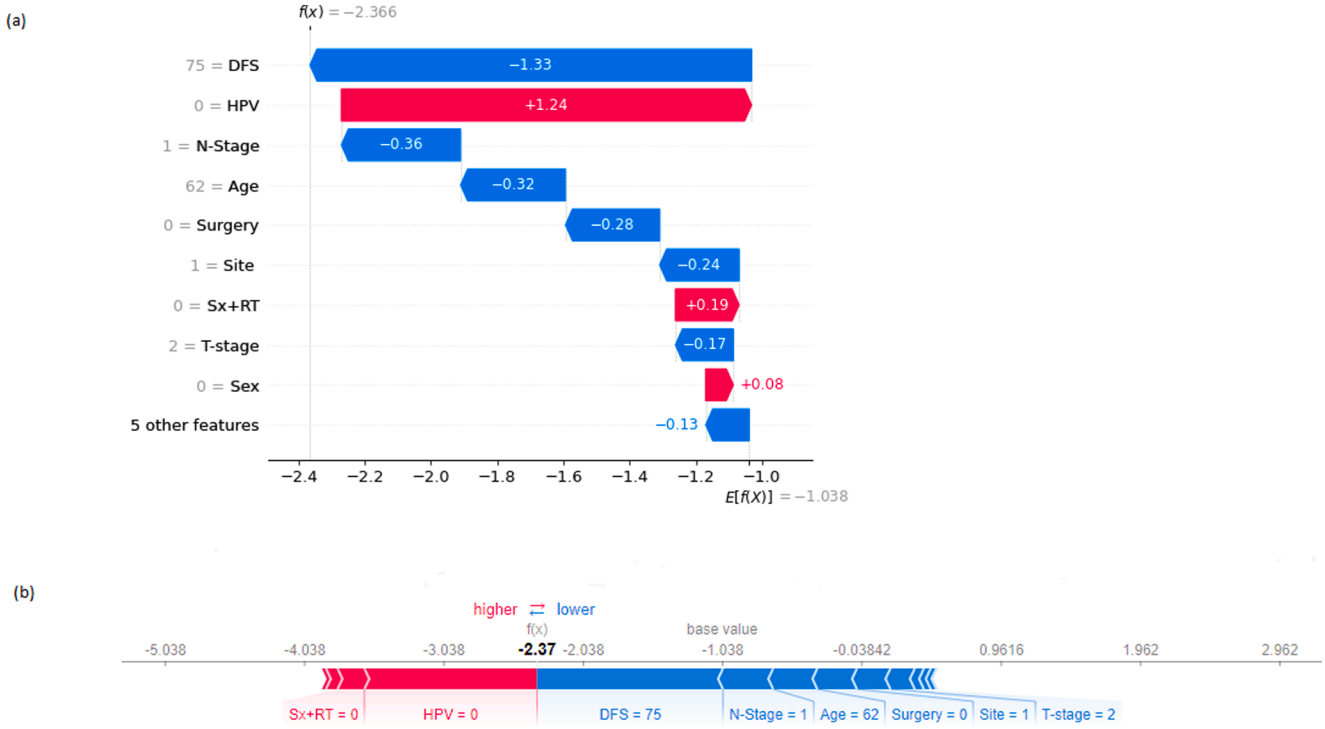


**Fig. 5.** The SHapley Additive Explanation (SHAP) framework for individual predictions.

risk predictions made by our model for any given patient [Fig. 5b].

## 3. Results

### 3.1. Characteristics of the study population

Out of the 3164 cases used for model training and internal validation of this study (Table 1), a total of 267 (8.4 %) cases were reported in the year 2010, 388 (12.3 %) cases in 2011, 519 (16.4 %) in 2012, 614 (19.4 %) cases in 2013, 690 (21.8 %) in 2014, and 686 (21.7 %) in the year 2022 demonstrating increasing occurrence of OPSCC. The median age at diagnosis was 61 (SD ± 10.4: range 20–85: mean age was 61.4 years). A significant amount of the extracted cases were male (2526 [79.8 %] males and 638 [20.2 %] females) in the ratio of 3.96:1 male-to-female.

Considering the ethnicity, 2825 (89.3 %) were of White origin, 233 (7.4 %) were Black, and 106 (3.4 %) were from other origins (American Indian/AK Native, Asian/Pacific Islander). Regarding the marital status of OPSCC patients at the time of diagnosis, the unmarried patients (single [never married], divorced, unmarried or domestic partner, widowed, and separated) comprised a total of 1264 (39.9 %) cases while 1900 (60.1 %) were married (Table 1).

In terms of the HPV status of the patients, 1143 (36.1 %) were HPV negative while 2021 (63.9 %) were HPV positive. In addition, the clinical and pathologic characteristics such as grade showed that 174 (5.5 %) out of the 3164 patients had a tumor with a well-differentiated grade, 1305 (41.2 %) a moderately differentiated grade, 1650 (52.1 %) a poorly differentiated grade, and 35 (1.1 %) an undifferentiated grade. For the staging scheme according to the AJCC TNM classification, 875 (27.6 %) patients had stage T1 tumors, 1279 (40.4 %) stage T2, 580 (18.3 %) stage T3, and 430 stage T4 (13.6 %). Correspondingly, 1417 (44.8 %) had N0, 1424 (45.0 %) had N1, 333 (10.2 %) N3; 3085 (97.5 %) M0, and 79 (2.5 %) M1. The details of the clinicopathologic

characteristics and the corresponding distributions are given in Table 1.

Regarding the treatment modalities, 1270 (40.1 %) had surgery with postoperative radiotherapy, 579 (18.3 %) surgery with chemo-radiotherapy, 413 (13.1 %) definitive (chemo)radiotherapy, 518 (18.3 %) surgery alone and 384 (12.1 %) received none of the available treatments. The follow-up time ranged from 0 to 107 months (Median 49; Mean 49.4; SD ± 27.2). The number of patients who were alive at the last follow-up was 2201 (67.0 %).

### 3.2. Characteristics of the study population for temporal validation

The detailed characteristics of the external validation data (n = 120) are presented in Table 1. The mean age at diagnosis was 59.4 years (Median 59; SD ± 10.8; range 31 – 85 with 98 (81.7 %) male and 22 (18.3 %) female. Considering the ethnicity of the OPSCC patients for external validation, the vast majority were of white origin [108 (90.0 %) White, 11 (9.2 %) Black, and 0.8 % were from another ethnic group]. Additionally, 76 (63.3 %) were married and 44 (36.7 %) were unmarried, which includes single, never married, divorced, unmarried or domestic partner, widowed, and separated. In terms of the clinical and pathologic characteristics such as grade, 7 (5.8 %) out of the 120 OPSCC patients for temporal validation had well-differentiated grade, 55 (45.8 %) moderately differentiated, 53 (44.2 %) poorly differentiated, and 5 (4.2 %) undifferentiated. Regarding HPV status, a total of 36 (30.0 %) had a HPV-negative while 84 (70.0 %) had a HPV-positive tumor.

A total of 44 (36.6 %) originated from the base (posterior one-third) of the tongue, 62 (51.6 %) from the tonsils, and the remaining 14 (11.6 %) cases were from other subsites. For the staging scheme according to the AJCC TNM, 34 (28.3 %) patients had stage T1, 47 (39.2 %) stage T2, 19 (15.8 %) stage T3, and 20 (16.7 %) stage T4. Correspondingly, 50 (41.7 %) had N0, 60 (50.0 %) had N1, 10 (8.3 %) N3; 116 (96.6 %) M0, and 4 (3.3 %) M1. The details of the histopathologic characteristics and the corresponding distributions are given in Table 1.

Considering the treatment modalities in the external validation cohort, 57 (47.5 %) had surgery with postoperative radiotherapy, 28 (23.3 %) surgery with chemoradiotherapy, 11 (9.2 %) definitive (chemo)radiotherapy, 6 (5.0 %) surgery alone and 18 (15.0 %) received none of the available treatments. The follow-up time ranged from 2 to 79 months (Mean 38.2; Median 42.0; SD ± 16.3). The number of patients who were alive at last follow-up was 84 (70.0 %).

### 3.3. Ensemble method performance during training

The performance metrics of the examined algorithms (voting ensemble, stacked ensemble, light GBM, XGBoost, random forest, and extreme random trees) was presented in Table 2. These algorithms showed comparable performance in the risk stratification of OPSCC patients based on the training phase results (Table 2). Remarkably, the voting ensemble variant slightly outperformed the other algorithms, specifically in terms of the weighted area under curve. The accuracy, balanced accuracy, Matthews' correlation coefficient, and weighted area under curve for the voting ensemble were 89.8 %, 86.3 %, 0.77, and 0.929 (Table 2). The area under receiving operating characteristics curve of the voting ensemble model is presented in (Fig. 2). Meanwhile, other performance metrics such as predictive values, confusion matrix parameters, sensitivity, specificity, and F1-score are also presented in Table 2.

### 3.4. Temporal validation of the prediction performance of the model

The trained voting ensemble model gave an accuracy of 88.3 % when temporally validated with new cohorts. The performance from this type validation was considered the gold standard regarding the predictive ability of the model [15,16,45]. Similarly, the sensitivity, specificity, F1-score, and Matthew's correlation were 0.91, 0.82, 0.92, and 0.72 respectively (summarized in Table 3).

### 3.5. Predictive features for the developed model

In terms of the importance of the input features on the ability of the model to stratify the patients into risk groups, the human papillomavirus (HPV) status, age of the patients, T stage, marital status, N stage, and the treatment (surgery followed by radiotherapy) were found to be the most prominent features (Fig. 3).

### 3.6. Explainability and interpretability of the model

The LIME framework explained the degree of correctness of the stratification made by the model for a single prediction (Fig. 4). Additionally, how each of the input parameters contributed to the prediction was given (Fig. 4). As shown in Fig. 4, the model predicted that the patient has a high-risk for survival with 100 % degree of correctness (Fig. 4). Furthermore, HPV status, M−stage, age, ethnicity, Sx + RT, and chemotherapy contributed to the high survival risk prediction made by the model for that particular instance of prediction (Fig. 4). Similarly, SHAP framework showed that HPV status, gender, and treatment modality (surgery with radiotherapy) were the input features that enhanced the model prediction from the base value (the average model output over the training dataset) to the model prediction (shown in red in Fig. 5a&b). As shown in Fig. 6, HPV-positive, young-aged OPSCC patients, early *T-N* stage, married, treatment approach (surgery followed by radiotherapy), and disease-free survival time associated with a high chance of overall survival (Fig. 6). This result is specifically comparable to the feature importance identified in the PFI analysis (Fig. 3).

### 3.7. Comparison of current studies with previous studies

The studies by Dinia et al. and Patel et al. developed a predictive model based on a machine learning paradigm to identify patients at high risk of relapse or death after treatment for HPV-positive OPSCC. However, these studies used a relatively small amount of data (n = 450 for Dinia et al and n = 553 for Patel et al.) and showed a reasonable performance in terms of area under curve (AUC) metrics (0.89 and 0.79) [20,46]. Similarly, the study by Gaebel et al., examined a hybrid approach – expert-based implementation and machine-learning-based model for clinical decision-making using a small clinical dataset (n = 94) [47]. With the introduction of this hybrid approach, the weighted accuracy of the prediction increased (from 52.9 % to 88.3 %). Following the limitations of these studies, we developed and externally validated a machine learning model to identify patients at high risk of overall survival (OS) with a reasonable performance metrics (weighted accuracy: 90.1 %; AUC: 0.92) using population-based registry data. Besides identifying patients at high risk of OS, explanations and interpretations were provided with the prediction using Local Interpretable Model Agnostic Explanations (LIME) and SHapley Additive Explanation (SHAP) frameworks. The LIME and SHAP techniques ensured that how each variable contributed to the predicted outcome was known. In addition, our study also demonstrated a better predictive performance than a similar study that used a national cancer database [19].

### 3.8. Significance of stratifying patients into risk groups

Several studies have emphasized the significance of dividing medical patients into risk groups in recent years [48–51]. For cancer patients, risk stratification becomes pertinent due to the increased risk of cancer recurrence, associated treatment costs, and treatment morbidities that can affect the quality of life of the patients. Thus, the application of a subfield of artificial intelligence, like machine learning techniques, provides opportunities for health care organizations and clinicians to better understand the underlying risks of their large patient populations. Identifying those patients who are likely to be members of high-risk trajectories allows healthcare organizations to stratify patients by level of risk and develop early targeted and personalized interventions to

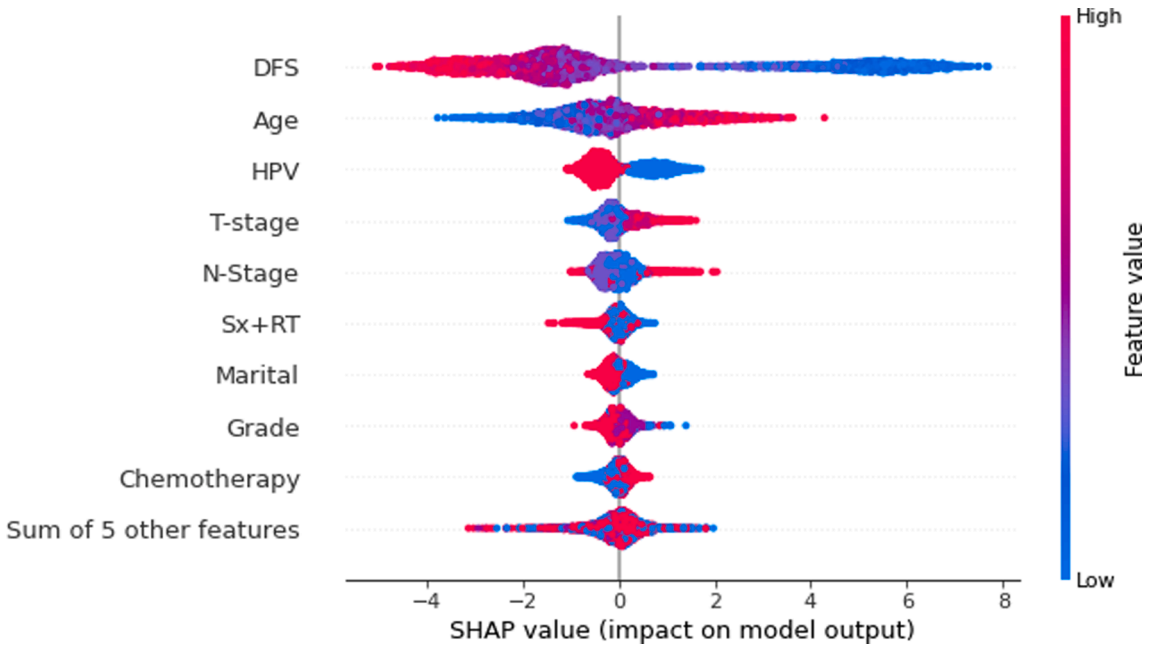


**Fig. 6.** Explainability and interpretability of the SHAP framework for each feature.

improve care quality [48,50]. Clinicians could utilize the explanations and interpretations provided by the LIME and SHAP frameworks in addition to the predicted results to understand previously undetected relationships between these prognostic parameters to allow for more informed clinical decision-making and effective interventions. Most importantly, targeted and personalized interventions can help reduce the incidence of recurrence of OPSCC as cancer relapse leads to further financial loss in terms of costly hospital readmissions and increased treatment costs.

### 3.9. Discussion

The present study examined the potential of ensemble machine learning-based models to stratify oropharyngeal squamous cell carcinoma (OPSCC) patients into favorable overall survival (low-risk) and worse overall survival (high-risk) groups for targeted treatment intervention. This stratification was aimed at guiding treatment decisions to spare patients from possibly ineffective treatment approaches. The model developed showed promising overall risk stratification prediction when externally validated. Additionally, the permutation feature importance (PFI) highlighted that the topmost variables that are most predictive for the overall survival stratification were human papillomavirus status, age, tumor stage, marital status, nodal status, and treatment (surgery followed by radiotherapy). In terms of the explainability of the model for individual prediction, HPV status, gender, and treatment (surgery followed by radiotherapy) were significant. The interpretability of the model with the SHapley additive explanations (SHAP) framework showed that HPV-positive, young-aged OPSCC patients, early *T-N* stage, being married, treatment approach with surgery followed by radiotherapy), and disease-free survival time are associated with a high chance of overall survival. The reliability of the prediction made by the model was provided using the Local Interpretable Model Agnostic Explanations (LIME) framework.

Considering the dearth of evidence suggesting the optimal treatment modality for early-stage OPSCC patients [19,52], it becomes important to carefully select a treatment plan that can enhance the patients' quality of life. It has been reported that the current combined therapies include significant risks of morbidity for the growing group of survivors [9]. Therefore, survivors are faced with potentially severe deterioration in their quality of life marked by xerostomia, dysphagia, and chewing problems [9]. These ailments further emphasize the need to accurately stratify the patients into risk groups for effective targeted treatment planning and improved management of the patients.

Several attempts have been made to optimize treatment planning for OPSCC patients. For example, the study by Karadaghy et al. concluded that the tumor characteristics and the facility type influenced the decision to either undergo primary surgery or primary radiation [19]. In this study, we leveraged the ability of a machine-learning algorithm to unravel the intricate nonlinear interactions among variables (Table 1). The developed model is poised to create additional value in the analysis of clinical data from an international, multi-state, and national cancer registry by assisting in the clinical decision-making process.

The predictive performance showed by the machine learning model presented in this study is capable of assisting clinicians in selecting a treatment approach that contributes to excellent oncologic control while reducing morbidity and enhancing function and quality of life [52]. This is imperative considering that human papillomavirus (HPV)-positive OPSCC has been reported to demonstrate favorable treatment outcomes compared with the traditional HPV-negative OPSCC counterpart, which is usually driven by tobacco and alcohol consumption [52–55].

Because an increase in the incidence of OPSCC is generally found in young and generally healthy cohorts [56–59], it is necessary to minimize treatment-related toxicity. The predictive performance shown by the machine learning algorithms examined in this study may suggest that patients with favorable overall survival (low-risk) may require a single modality treatment while worse overall survival (high-risk) patients may require multimodality treatment.

Permutation feature importance found HPV to be an important factor for the predictive ability of the ML model. This may corroborate why HPV infection has been regarded as the most significant causal factor for OPSCC [6,10,60–62]. Similarly, the age of OPSCC was also considered by the PFI as an important parameter. This observation agrees with another report which emphasized that age was one of the factors associated with highly aggressive HPV-associated OPSCC [63]. This may be attributed to the fact that HPV-related infection might have occurred for many years prior to the development of OPSCC. Additionally, *T*-stage was also considered an important factor. This is evident as it has been reported that oropharyngeal cancers are typically diagnosed late, i.e., at advanced stage with positive regional lymph nodes [6,64,65]. Consequently, the locoregional prognosis and survival of the OPSCC patients

are affected [6].

Understanding the pattern of metastatic spread to the neck lymph nodes in OPSCC is of paramount importance, as highlighted in this study where the PFI analysis identified the nodal stage as an important factor. Nodal status may help to improve the rationale for determining the proper neck treatment approach, indicating the possible treatment (adjuvant therapy or not), and better envisaging the prognosis of the OPSCC patients [66]. Regarding the possible treatment approach, PFI analysis performed in this study equally identified surgery with radiotherapy as an important factor for overall survival in OPSCC patients. This is essentially necessary, especially in the era of HPV-associated OPSCC where there has been a renewed interest in primary surgical management with a less invasive approach such as transoral robotic surgery as well as a refinement of radiation techniques to minimize long-term morbidity and side effects [67].

The marital status of OPSCC patients was also found to be an important parameter of survival in HPV-related OPSCC. This result is consistent with similar studies using SEER data to examine the effects of marital status on survival in patients with head and neck cancer (HNC) [68,69]. A likely reason for this may be due to spousal support, which offers social support and active surveillance of visual and symptomatic head and neck cancer sites to enhance early observation, higher rates of treatment success, and better survival [68,70].

Notably, the afore-mentioned PFI analysis in this study provides the overall behavior of the features on the underlying model. It does not explain the contributions of these features for individual predictions to enhance the explainability and interpretability of the model. Remarkably, explainability has been identified as one of the main concerns hindering the adoption of a machine learning-based model for cancer management [33,71]. According to the SHAP framework to enhance explainable and interpretable machine learning, HPV status, gender, and treatment approach are the main salient features for the individual predictions of the outcome by the model. This is evident as OPSCC is associated with HPV (an independent risk factor) [6], more prevalent in males, and HPV-associated diseases have been found to show significantly better treatment response and prognosis than the non-HPV-associated counterpart [21,72].

To bring the utilization of the ML model closer to reality, we temporally validated the model developed. The motivation for external validation was to ensure that the integrated model addresses possible concerns regarding the generalizability of the tool [16,45]. Several studies have touted the benefits of tree-based (ensemble) machine learning algorithms in the prognostication of outcomes in cancer management [15,16,45]. This is because the algorithm can summarize the impact of input features on the model (i.e., global interpretation) [73]. However, the algorithm is not able to reveal the impact of the input features on individual predictions (i.e., local interpretation), which is needed for explainability and interpretability, despite its potential benefits [73]. Therefore, we used the SHAP framework to ensure both local and global interpretations are presented to enhance interpretable and explainable models [73]. Furthermore, the LIME framework gives the degree of accuracy of the prediction made by the model. This addresses concerns about the trustworthiness of predictions made by the model.

While both the LIME and SHAP frameworks are aimed at providing interpretable and explainable machine learning models, the SHAP framework appears to be more robust as it provides both local and global interpretations of the developed model and the corresponding input parameters. Remarkably, the feature importance produced by the SHAP framework (Fig. 6) is more detailed than the traditional feature importance (Fig. 3) as it not only shows the important features but also how the variables within each parameter contribute to the predictive ability of the model. However, there are growing concerns regarding the interpretations made by the SHAP framework [74]. Despite these, it is hoped that providing explanations and interpretations to promising machine learning models as demonstrated in this study can bring the utilization of these models closer to usage in daily clinical practices.

### 3.10. Practical implications

Our proposed model demonstrated good performance in stratifying patients with different risks of survival by using comprehensive clinicopathologic data from one of the most comprehensive cancer population databases. The model can provide additional personalized information for the postoperative management of patients with OPSCC. Traditionally, clinical decisions have been based on guidelines and accumulated experience [75]. However, this approach may be subjective due to variations in the experience of the clinicians. Thus, our model seeks to provide a second opinion and rigor to clinicians for effective decision making. The model is able to generate individualized predictions by synthesizing data across broad patient bases. Thus, the model will stratify an OPSCC patient to a specific risk group while accounting for the patient's unique characteristics. On an individual level, the individualized predictions made by the model address one of the concerns regarding the traditional American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM), which is the fact that the model considered other tumor- and patient-related risk factors in making the predictions [45]. On a more granular level, understanding the risks of overall survival can help clinicians in achieving a targeted treatment approach. Although analysis of optimal treatment was not performed in this study, as the focus was on the overall survival risk of the OPSCC patients, the model may provide the premises for clinicians to intensify or deintensify treatment approaches (regimens) in order to improve quality of life and prognosis.

## 4. Limitations

The inherent limitations of the SEER database have an impact on the present study. For example, extracting OPSCC patients from this database was challenging because tumors are often reported in aggregate with other pharyngeal or head and neck malignancies. Similarly, the definition of subsite may sometimes be confusing as there are no clear distinctions between the oral cavity and oropharynx. Additionally, there are limitations regarding the recording of the radiotherapy and chemotherapy information in the SEER registry. The model still showed promising performance. It remains important to externally validate our proposed model with new cases to enhance its generalizability and, most importantly, to facilitate clinical application in the future. The predictive model was trained with a dataset that is fairly balanced, thus it had a relatively stable performance. However, data imbalance techniques should be used in the training to ensure a reduction in false positive errors in the model.

### 4.1. Conclusions and future research

This study responds to frequent calls for personalized and precision medicine [76]. It utilized ensemble machine learning algorithms to stratify OPSCC patients into risk groups (as either high-risk or low-risk) based on the chance of overall survival using population-based data. Interpretations and explanations were provided for the predictions made by the model using the LIME and SHAP frameworks. Our findings indicated that the proposed model is able to stratify the patients into risk groups and may help clinicians make informed decisions and facilitate more precise management of OPSCC patients. Ideal and universally accepted global guidelines for the treatment of OPSCC patients are lacking. Therefore, the present approach of the division of OPSCC patients according to their respective chance of overall survival may provide a general recommendation for the treatment of OPSCC patients. In addition, the explanations and interpretation provided alongside the predicted chance of overall survival are posited to provide insights to clinicians in terms of how each prognostic factor contributes to the predicted chance of survival.

In the future, we aim to integrate our model as a web-based prognostic tool to facilitate external validation of the model using a new dataset. It would be interesting to add other important parameters, such as smoking and drinking habits, to the development of the machine-learning model. Considering the increasing application of machine learning for cancer management in various studies, a comprehensive systematic literature review and *meta*-analyses are warranted in a future study to carefully synthesize the published results in these studies. This is posited to define the path to the possible implementation of machine learning models in cancer management.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

*Acknowledgment*

K. Albin Johassons Stiftelse. The Sigrid Jusélius Foundation. The Helsinki University Hospital Research Fund. Turku University Hospital Fund, Helsinki University Hospital Research Fund.

The authors thank Dr. Kenneth Quek for his valuable editing of the English of the manuscript.

**Summary points:**

- A survival risk stratification model was developed by combining a highly accurate machine learning (ML) model with explainable artificial intelligence (xAI).
- We compared five varieties of ensemble algorithms – voting ensemble, light gradient boosting machine (Light GBM), extreme gradient boosting (XGBoost), random forest, and extreme random trees for survival risk stratification in oropharyngeal cancer patients.
- Explainability and interpretability of the model were enhanced using the Local Interpretable Model Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) frameworks.
- The interpretable and explainable machine learning model showed the ability to predictive risk stratification of oropharyngeal cancer patients into distinct risk groups (high-risk and low-risk).
- The human papillomavirus (HPV) status, age of the patients, T stage, marital status, N stage, and the treatment modality (surgery with postoperative radiotherapy) were found to be the most prominent features with significant effects on the ability of the machine learning model to perform overall survival risk stratification in oropharyngeal cancer patients.
- The predictive risk stratification of oropharyngeal cancer patients is important for effective treatment planning care and informed clinical decisions.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104896.

## References

[1] L.A. Koneva, Y. Zhang, S. Virani, P.B. Hall, J.B. McHugh, D.B. Chepeha, et al., HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers, Mol Cancer Res 16 (2018) 90–102, https://doi.org/10.1158/1541-7786.MCR-17-0153.

[2] Z. Gooi, J.Y.K. Chan, C. Fakhry, The epidemiology of the human papillomavirus related to oropharyngeal head and neck cancer: Epidemiology of HPV-Related OSCC, The Laryngoscope 126 (2016) 894–900, https://doi.org/10.1002/lary.25767.

[3] A.C. Chi, T.A. Day, B.W. Neville, Oral cavity and oropharyngeal squamous cell carcinoma-an update: Oral & Oropharyngeal Cancer Update, CA Cancer J. Clin. 65 (2015) 401–421, https://doi.org/10.3322/caac.21293.

[4] S. Warnakulasuriya, Global epidemiology of oral and oropharyngeal cancer, Oral Oncol. 45 (2009) 309–316, https://doi.org/10.1016/j.oraloncology.2008.06.002.

[5] R. Lambert, C. Sauvaget, C.M. de Camargo, R. Sankaranarayanan, Epidemiology of cancer from the oral cavity and oropharynx, Eur. J. Gastroenterol. Hepatol. 23 (2011) 633–641, https://doi.org/10.1097/MEG.0b013e3283484795.

[6] T. Carpén, A. Sjöblom, M. Lundberg, C. Haglund, A. Markkola, S. Syrjänen, et al., Presenting symptoms and clinical findings in HPV-positive and HPV-negative oropharyngeal cancer patients, Acta Otolaryngol. 138 (2018) 513–518, https://doi.org/10.1080/00016489.2017.1405279.

[7] S. Mascharak, B.J. Baird, F.C. Holsinger, Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning: Multispectral Imaging of Oropharynx Cancer, The Laryngoscope 128 (2018) 2514–2520, https://doi.org/10.1002/lary.27159.

[8] Guo T, Eisele D, Fakhry C. The potential impact of prophylactic human papillomavirus vaccination on oropharyngeal cancer. Cancer 2016;1;122(15): 2313–23. https://doi.org/doi: 10.1002/cncr.29992.

[9] S. Høxbroe Michaelsen, C. Grønhøj, J. Høxbroe Michaelsen, J. Friborg, C. von Buchwald, Quality of life in survivors of oropharyngeal cancer: A systematic review and meta-analysis of 1366 patients, Eur. J. Cancer 78 (2017) 91–102, https://doi.org/10.1016/j.ejca.2017.03.006.

[10] A.K. Chaturvedi, E.A. Engels, R.M. Pfeiffer, B.Y. Hernandez, W. Xiao, E. Kim, et al., Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States, J. Clin. Oncol. 29 (2011) 4294–4301, https://doi.org/10.1200/JCO.2011.36.4596.

[11] E.L. You, M. Henry, A.G. Zeitouni, Human Papillomavirus-Associated Oropharyngeal Cancer: Review of Current Evidence and Management, Current Oncology 26 (2019) 119–123, https://doi.org/10.3747/co.26.4819.

[12] Larsen CG, Jensen DH, Carlander A-LF, Kiss K, Andersen L, Olsen CH, et al. Novel nomograms for survival and progression in HPV+ and HPV- oropharyngeal cancer: a population-based study of 1,542 consecutive patients. Oncotarget 2016;7: 71761–72. https://doi.org/10.18632/oncotarget.12335.

[13] K.R. Yabroff, J. Lund, D. Kepka, A. Mariotto, Economic Burden of Cancer in the United States: Estimates, Projections, and Future Research, Cancer Epidemiol. Biomark. Prev. 20 (2011) 2006–2014, https://doi.org/10.1158/1055-9965.EPI-11-0650.

[14] C.G. Gourin, C. Fakhry, H. Quon, H. Kang, A.P. Kiess, R.J. Herbert, et al., Treatment, survival, and costs of oropharyngeal cancer care in the elderly: Oropharyngeal Cancer Care in the Elderly, The Laryngoscope 128 (2018) 1103–1112, https://doi.org/10.1002/lary.26887.

[15] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R. D. Coletta, et al., Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool, Virchows Arch. 475 (2019) 489–497, https://doi.org/10.1007/s00428-019-02642-5.

[16] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R. D. Coletta, et al., Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer, Int. J. Med. Inf. (2019:), 104068, https://doi.org/10.1016/j.ijmedinf.2019.104068.

[17] Y. Li, Z. Zhao, X. Liu, J. Ju, J. Chai, Q. Ni, et al., Nomograms to estimate long-term overall survival and tongue cancer-specific survival of patients with tongue squamous cell carcinoma, Cancer Med 6 (2017) 1002–1013, https://doi.org/10.1002/cam4.1021.

[18] Surveillance, Epidemiology, and End Results (SEER) Program. SEER_Program. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973–2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2018 submission 2018.

[19] O.A. Karadaghy, M. Shew, J. New, A.M. Bur, Machine Learning to Predict Treatment in Oropharyngeal Squamous Cell Carcinoma, ORL (2021) 1–8, https://doi.org/10.1159/000515334.

[20] H. Patel, D.M. Vock, G.E. Marai, C.D. Fuller, A.S.R. Mohamed, G. Canahuate, Oropharyngeal cancer patient stratification using random forest based-learning over high-dimensional radiomic features, Sci Rep 11 (2021) 14057, https://doi.org/10.1038/s41598-021-92072-8.

[21] A. Sjöblom, U.-H. Stenman, J. Hagström, L. Jouhi, C. Haglund, S. Syrjänen, et al., Tumor-Associated Trypsin Inhibitor (TATI) as a Biomarker of Poor Prognosis in Oropharyngeal Squamous Cell Carcinoma Irrespective of HPV Status, Cancers 13 (2021) 2811, https://doi.org/10.3390/cancers13112811.

[22] V.C. Osamor, A.F. Okezie, Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis, Sci Rep 11 (2021) 14806, https://doi.org/10.1038/s41598-021-94347-6.

[23] Brownlee J. A gentle introduction to ensemble learning 2020.

[24] F. Aydın, Z. Aslan, The Construction of a Majority-Voting Ensemble Based on the Interrelation and Amount of Information of Features, The Computer Journal 63 (2020) 1756–1774, https://doi.org/10.1093/comjnl/bxz118.

[25] Y. Zhang, H. Zhang, J. Cai, B. Yang, A Weighted Voting Classifier Based on Differential Evolution, Abstract and Applied Analysis 2014 (2014) 1–6, https://doi.org/10.1155/2014/376950.

[26] S. Karlos, G. Kostopoulos, S. Kotsiantis, A Soft-Voting Ensemble Based Co-Training Scheme Using Static Selection for Binary Classification Problems, Algorithms 13 (2020) 26, https://doi.org/10.3390/a13010026.

[27] Brownlee J. How to develop voting ensembles with Python 2020.

[28] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc.; 2017.

[29] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.
[30] Saha S. XGBoost vs LightGBM: How are they different. Machine Learning Tools 2022. https://neptune.ai/blog/xgboost-vs-lightgbm (accessed September 25, 2022).
[31] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach Learn 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.
[32] R.O. Alabi, A. Almangush, M. Elmusrati, A.A. Mäkitie, Deep Machine Learning for Oral Cancer: From Precise Diagnosis to Precision Medicine, Front Oral Health 2 (2022), 794248, https://doi.org/10.3389/froh.2021.794248.
[33] R.O. Alabi, O. Youssef, M. Pirinen, M. Elmusrati, A.A. Mäkitie, I. Leivo, et al., Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for future—A systematic review, Artif. Intell. Med. 115 (2021), 102060, https://doi.org/10.1016/j.artmed.2021.102060.
[34] Alabi RO, Bello IO, Youssef O, Elmusrati M, Mäkitie AA, Almangush A. Utilizing Deep Machine Learning for Prognostication of Oral Squamous Cell Carcinoma—A Systematic Review. Frontiers in Oral Health 2021;2. https://doi.org/10.3389/froh.2021.686863.
[35] Microsoft Azure Machine Learning Studio. Azure Machine Learning Studio: In Documentation. 2018.
[36] Y.-J. Tseng, H.-Y. Wang, T.-W. Lin, J.-J. Lu, C.-H. Hsieh, C.-T. Liao, Development of a Machine Learning Model for Survival Risk Stratification of Patients With Advanced Oral Cancer, JAMA Network Open 3 (2020) e2011768.
[37] G. Cawley, N. Talbot, On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, Journal of Machine Learning Research 11 (2010) 2079–2107.
[38] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.
[39] C.L. Ramspek, K.J. Jager, F.W. Dekker, C. Zoccali, M. van Diepen, External validation of prognostic models: what, why, how, when and where? Clinical Kidney Journal 14 (2021) 49–58, https://doi.org/10.1093/ckj/sfaa188.
[40] K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E. W. Steyerberg, et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration, Ann. Intern. Med. 162 (2015) W1, https://doi.org/10.7326/M14-0698.
[41] D. Gibertoni, P. Rucci, M. Mandreoli, M. Corradini, D. Martelli, G. Russo, et al., Temporal validation of the CT-PIRP prognostic model for mortality and renal replacement therapy initiation in chronic kidney disease patients, BMC Nephrol 20 (2019) 177, https://doi.org/10.1186/s12882-019-1345-7.
[42] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[43] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, ArXiv:170507874 [Cs, Stat] (2017).
[44] A. Gramegna, P. Giudici, SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk, Front Artif Intell 4 (2021), 752558, https://doi.org/10.3389/frai.2021.752558.
[45] R.O. Alabi, A.A. Mäkitie, M. Pirinen, M. Elmusrati, I. Leivo, A. Almangush, Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer, Int. J. Med. Inf. 145 (2021), 104313, https://doi.org/10.1016/j.ijmedinf.2020.104313.
[46] A. Dinia, S. Ammari, J. Filtes, M. Classe, A. Moya-Plana, F. Bidault, et al., Events prediction after treatment in HPV-driven oropharyngeal carcinoma using machine learning, Eur. J. Cancer 171 (2022) 106–113, https://doi.org/10.1016/j.ejca.2022.05.003.
[47] J. Gaebel, S. Mehlhorn, A. Oeser, A. Dietz, T. Neumuth, M. Stoehr, Clinical decision support models for oropharyngeal cancer treatment: design and evaluation of a multi-stage knowledge abstraction and formalization process, Int J CARS (2022), https://doi.org/10.1007/s11548-022-02675-3.
[48] Ben-Assuli O, Vest JR. Return visits to the emergency department: An analysis using group based curve models. Health Informatics J 2022;28:146045822211054. https://doi.org/10.1177/14604582221105444.
[49] C. Su, Z. Xu, K. Hoffman, P. Goyal, M.M. Safford, J. Lee, et al., Identifying organ dysfunction trajectory-based subphenotypes in critically ill patients with COVID-19, Sci Rep 11 (2021) 15872, https://doi.org/10.1038/s41598-021-95431-7.
[50] R.-G. Roni, H. Tsipi, B.-A. Ofir, S. Nir, K. Robert, Disease evolution and risk-based disease trajectories in congestive heart failure patients, J Biomed Inform 125 (2022), 103949, https://doi.org/10.1016/j.jbi.2021.103949.
[51] O. Ben-Assuli, T. Heart, J.R. Vest, R. Ramon-Gonen, N. Shlomo, R. Klempfner, Profiling Readmissions Using Hidden Markov Model - the Case of Congestive Heart Failure, Null 38 (2021) 237–249, https://doi.org/10.1080/10580530.2020.1847362.
[52] P. Sinha, O.A. Karadaghy, M.M. Doering, M.G. Tuuli, R.S. Jackson, B.H. Haughey, Survival for HPV-positive oropharyngeal squamous cell carcinoma with surgical versus non-surgical treatment approach: A systematic review and meta-analysis, Oral Oncol. 86 (2018) 121–131, https://doi.org/10.1016/j.oraloncology.2018.09.018.
[53] M.B. Wang, I.Y. Liu, J.A. Gornbein, C.T. Nguyen, HPV-Positive Oropharyngeal Carcinoma: A Systematic Review of Treatment and Prognosis, Otolaryngology-Head and Neck Surgery 153 (2015) 758–769, https://doi.org/10.1177/0194599815592157.
[54] M.A. O'Rorke, M.V. Ellison, L.J. Murray, M. Moran, J. James, L.A. Anderson, Human papillomavirus related head and neck cancer survival: A systematic review and meta-analysis, Oral Oncol. 48 (2012) 1191–1201, https://doi.org/10.1016/j.oraloncology.2012.06.019.
[55] P.P. Sedghizadeh, W.D. Billington, D. Paxton, R. Ebeed, S. Mahabady, G.T. Clark, et al., Is p16-positive oropharyngeal squamous cell carcinoma associated with favorable prognosis? A systematic review and meta-analysis, Oral Oncol. 54 (2016) 15–27, https://doi.org/10.1016/j.oraloncology.2016.01.002.
[56] B.S. Chera, R.J. Amdur, Current Status and Future Directions of Treatment Deintensification in Human Papilloma Virus-associated Oropharyngeal Squamous Cell Carcinoma, Seminars in Radiation Oncology 28 (2018) 27–34, https://doi.org/10.1016/j.semradonc.2017.08.001.
[57] S. Cheraghlou, P.K. Yu, M.D. Otremba, H.S. Park, A. Bhatia, C.K. Zogg, et al., Treatment deintensification in human papillomavirus-positive oropharynx cancer: Outcomes from the National Cancer Data Base: HPV-Positive Cancer Treatment Deintensification, Cancer 124 (2018) 717–726, https://doi.org/10.1002/cncr.31104.
[58] N. Gildener-Leapman, J. Kim, S. Abberbock, G.W. Choby, R. Mandal, U. Duvvuri, et al., Utility of up-front transoral robotic surgery in tailoring adjuvant therapy: Up-front transoral robotic surgery in tailoring adjuvant therapy, Head Neck 38 (2016) 1201–1207, https://doi.org/10.1002/hed.24390.
[59] H. Quon, J.D. Richmon, Treatment Deintensification Strategies for HPV-Associated Head and Neck Carcinomas, Otolaryngol. Clin. North Am. 45 (2012) 845–861, https://doi.org/10.1016/j.otc.2012.04.007.
[60] S. Habbous, K.P. Chu, X. Qiu, A. La Delfa, L.T.G. Harland, E. Fadhel, et al., The changing incidence of human papillomavirus-associated oropharyngeal cancer using multiple imputation from 2000 to 2010 at a Comprehensive Cancer Centre, Cancer Epidemiology 37 (2013) 820–829, https://doi.org/10.1016/j.canep.2013.09.011.
[61] H. Mehanna, T. Beech, T. Nicholson, I. El-Hariry, C. McConkey, V. Paleri, et al., Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer-systematic review and meta-analysis of trends by time and region, Head Neck 35 (2013) 747–755, https://doi.org/10.1002/hed.22015.
[62] M. Lundberg, I. Leivo, K. Saarilahti, A.A. Mäkitie, P.S. Mattila, Increased incidence of oropharyngeal cancer and p16 expression, Acta Otolaryngol. 131 (2011) 1008–1011, https://doi.org/10.3109/00016489.2011.575796.
[63] O. Alabi, J.P. O'Neill, 'Good cancer gone bad': a narrative review of HPV oropharyngeal cancer and potential poor outcomes, Eur Arch Otorhinolaryngol 277 (2020) 2185–2191, https://doi.org/10.1007/s00405-020-05991-z.
[64] G. Psychogios, K. Mantsopoulos, C. Bohr, M. Koch, J. Zenk, H. Iro, Incidence of occult cervical metastasis in head and neck carcinomas: Development over time: Occult Cervical Metastasis, J. Surg. Oncol. 107 (2013) 384–387, https://doi.org/10.1002/jso.23221.
[65] S.C. Cantrell, B.W. Peck, G. Li, Q. Wei, E.M. Sturgis, L.E. Ginsberg, Differences in imaging characteristics of HPV-positive and HPV-Negative oropharyngeal cancers: a blinded matched-pair analysis, AJNR Am J Neuroradiol 34 (2013) 2005–2009, https://doi.org/10.3174/ajnr.A3524.
[66] J.G. Vartanian, E. Pontes, I.M.G. Agra, O.D. Campos, J. Gonçalves-Filho, A. L. Carvalho, et al., Distribution of Metastatic Lymph Nodes in Oropharyngeal Carcinoma and Its Implications for the Elective Treatment of the Neck, Arch Otolaryngol Head Neck Surg 129 (2003) 729, https://doi.org/10.1001/archotol.129.7.729.
[67] S.C. Kamran, M.M. Qureshi, S. Jalisi, A. Salama, G. Grillone, M.T. Truong, Primary surgery versus primary radiation-based treatment for locally advanced oropharyngeal cancer, The Laryngoscope 128 (2018) 1353–1364, https://doi.org/10.1002/lary.26903.
[68] G. Inverso, B.A. Mahal, A.A. Aizer, R.B. Donoff, N.G. Chau, R.I. Haddad, Marital status and head and neck cancer outcomes, Cancer 121 (2015) 1273–1278, https://doi.org/10.1002/cncr.29171.
[69] E.W. Schaefer, M.Z. Wilson, D. Goldenberg, H. Mackley, W. Koch, C.S. Hollenbeak, Effect of marriage on outcomes for elderly patients with head and neck cancer: Marriage effect in head and neck cancer, Head Neck 37 (2015) 735–742, https://doi.org/10.1002/hed.23657.
[70] A.A. Aizer, M.-H. Chen, E.P. McCarthy, M.L. Mendu, S. Koo, T.J. Wilhite, et al., Marital Status and Survival in Patients With Cancer, JCO 31 (2013) 3869–3876, https://doi.org/10.1200/JCO.2013.49.6489.
[71] R.O. Alabi, V. Tero, E. Mohammed, Machine learning for prognosis of oral cancer: What are the ethical challenges? CEUR-Workshop Proceedings (2020).
[72] K.K. Ang, J. Harris, R. Wheeler, R. Weber, D.I. Rosenthal, P.F. Nguyen-Tân, et al., Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer, N Engl J Med 363 (2010) 24–35, https://doi.org/10.1056/NEJMoa0912217.
[73] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, et al., From Local Explanations to Global Understanding with Explainable AI for Trees, Nat Mach Intell 2 (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.
[74] Chandan D. Using SHAP for Explainability — Understand these Limitations First. Explainability Done Right 2021. https://towardsdatascience.com/using-shap-for-explainability-understand-these-limitations-first-1bed91c9d21 (accessed June 15, 2022).
[75] D. Bertsimas, H. Wiberg, Machine Learning in Oncology: Methods, Applications, and Challenges, JCO Clinical Cancer Informatics (2020) 885–894, https://doi.org/10.1200/CCI.20.00072.
[76] B. Bhinder, C. Gilvary, N.S. Madhukar, O. Elemento, Artificial Intelligence in Cancer Research and Precision Medicine, Cancer Discovery 11 (2021) 900–915, https://doi.org/10.1158/2159-8290.CD-21-0090.