POTENTIAL PREDICTORS FOR ORAL SQUAMOUS CELL CARCINOMA STAGING USING DECISION TREE ANALYSIS

Wan Muhamad Amir W Ahmad[1]*, Muhammad Azeem Yaqoob[2], Hazik Bin Shahzad[3], Mohamad Nasarudin Adnan[4], Farah Muna Mohamad Ghazali[5], Noraini Mohamad[6], Norhayati Yusop[7], Nor Azlida Aleng[8], Nor Farid Mohd Noor[9]

Article History: **Received:** 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Introduction: Oral cancer is the sixth most common cancer worldwide, with a mortality rate of up to 50%. According to the report in 2012, there were 8.2 million cancer deaths and 14.1 million new cases of oral squamous cell carcinoma (OSCC). These malignancies are still frequently not discovered and their survival rate has remained essentially unchanged over the past three decades. **Objective:** This paper aims to determine the potential predictors which contribute to the TNM staging of OSCC using decision tree analysis of 57 patients who attended Hospital USM, Kelantan. **Method**: Two methods of statistical analysis were used which were decision tree analysis and ordinal logistic regression analysis. **Results:** Using decision tree analysis, three factors were related to the TNM staging which are T Classifications, N Classifications, and Surgical Margin. From the ordinal logistic regression point of view, T Classifications, N Classifications, and Surgical Margin are contributing to the TNM staging. **Conclusion:** From all the analysis, it can be concluded that the proposed method produces excellent outcomes. The new approach in methodology delivers an accurate estimation of the final model's fit. The model's enhanced methodology leads to better outcomes and efficient decision-making.
The method approach provides an accurate evaluation of the final model's fit. The model's superior performance resulted in better outcomes and more effective decision-making.

Keywords: Oral Squamous Cell Carcinomas (OSCC), Decision Tree Analysis, Ordinal Logistic Regression

[1*,2,3,4,5,6,7]School of Dental Sciences, Health Campus. Universiti Sains Malaysia (USM),
[8]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia.
[9]Faculty of Medicine, Universiti Sultan Zainal Abidin (UniSZA), Medical Campus, Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia.

DOI: 10.31838/ecb/2023.12.s2.077

1. Introduction

The sixth most common tumor that is also malignant is oral squamous cell carcinoma (OSCC) and it accounts for up to 50% of oral cavity deaths [1, 2]. The highest prevalence of these cancers was found in Asia, which accounts for 2 to 5% of all cancer cases worldwide. The global survival rate has continued unchanged over the past few decades despite recent advances in therapeutic approaches [3-5]. These cancers have a multifactorial etiology which includes tobacco use, smoking, paan, betel quid, alcohol consumption, and viral stimuli [6-8]. According to a systematic review of the lip, oral cavity, and pharyngeal cancers in 2012, it was estimated that there were 263,900 new cases worldwide and that there were also 128,000 deaths from oral cavity cancer, which is an increase from previous years [4]. According to the WHO, the age-standardized rate (ASR) for the occurrence of oral cancer in the Malaysian population is 3.0 for every 100,000 people with the highest incidence of oral cancer among female populations in India at 10.2 per 100,000 [9]. Data from the Malaysian Oral Cancer Research and Coordinating Center (OCRCC) published in April 2011 showed that there were 1587 oral cancer deaths in Malaysia or 1.5% of all fatalities. Malaysia stands in 14th place worldwide having an ASR death rate of 7.72 for every 100000 people [10]. According to a prior study, in Kelantan, oral cancer patients had a median survival time of 9 months and an 18.0% 5-year survival rate [11]. The factors that led to poor survival included being elderly, male, having an advanced stage at diagnosis, and not receiving treatment.

In many developing nations, using tobacco, and drinking alcohol have all been linked to an increase in the risk of oral cancer [12]. Recent studies have demonstrated that the human papillomavirus (HPV) plays a significant causal role in OSCC [8]. All over the world, the human papillomavirus is a significant public health risk. Tonsils and the base of the tongue are two locations frequently affected by HPV-related cancers [6]. The estimated prevalence of HPV in normal oral mucosa ranges from 0.6% to 81.0% [13]. Numerous studies have documented the presence of HPV in oropharyngeal cancers. However, HPV is less prevalent in the oral cavity than oropharyngeal cancer. According to a previous assessment by the International Agency for Research on Cancer (IARC) on HPV carcinogenicity, there is ample evidence of HPV carcinogenic activity in oral cavity and oropharynx, yet little evidence for laryngeal cancer [8]. According to a recent report published by the WHO on the classification of head and neck tumors with HPV carcinogenesis, 3% of OSCCs are associated with HPV [12]. The role of HPV in the oral cavity has been well-documented, but it varies considerably based on anatomical site, ethnicity, detection method, and geographical variation in prevalence. There are limited studies looking at HPV and OSCC together in the Malaysian population [14].

According to Chaturvedi et al (2012), there are 170 sub-types of HPV that are classified as low-risk and high-risk [6]. The benign oral papillomatous lesions oral squamous papilloma, oral condyloma accuminatum, oral verruca Vulgaris, and focal epithelial hyperplasia appear to be strongly correlated with low-risk HPV such as HPV-6 and 11 [6]. On the other hand, oral potentially malignant disorders and OSCC are both linked to high-risk HPV, specifically HPV 16 and 18 [15]. It is well known that these high-risk HPV subtypes are frequently linked to cervical and other anogenital cancers [2, 8, 16]. The most useful prognostic tool for tumor survival in use is the tumor-node-metastasis (TNM) staging system [13]. Additionally, the clinical characteristics of the patient, along with their gender, age, and smoking behaviours, are taken into account when deciding on a therapeutic approach to minimize the risk of complications and maximize the prognosis for the many different types of cancer [8, 17]. This raises a crucial issue in the management of OSCC because many of the identification factors are linked to a poor prognosis. Numerous different clinicopathological factors, including age, smoking habits, TNM staging, lymph node involvement, and microvascular involvement have been examined as prognostic factors for OSCC in earlier studies [18].

2. Material, Methods, And Data

Decision Tree

Supervised machine learning with partitioning data on specific parameters gives us decision trees. The decision tree results are represented by leaves, and data are divided at the decision nodes. It is a good and useful tool for categorization, decision-making, and predictions in chronological choice problems [19]. This approach has been extensively employed in the medical field. For example, decision-makers commonly face sequential decisions with options depending on chance, but resulting in different results. Decision trees are the optimal approach to display this type of information graphically. It aids patrons in understanding issues naturally and making clearer judgments, as well as offering guidance on decision-making processes and potential outcomes. Three different node types make up a decision tree, as shown in Fig. 1: (a) choice node (b) chance node (c) endpoint node/terminal node [20]. A node is split when the global null hypothesis is rejected, and the covariate with the strongest link to the outcome is chosen for splitting. If the minimal p-value exceeds the multiplicity-

adjusted significance threshold, then the node is not split and is labeled as the terminal node.

A data point x in this study is a vector of d attribute with optional class label y. The set of attributes is denoted by A = $A1, A2,..., Ad$. We can characterize x as $\{x_1,x_2,...,x_d\}$, where $x_1 \in A1$, $x_2 \in A2$, ..., $x_d \in A_d$. Let $Y = \{y_1,y_2,...,y_m\}$ be the set of class labels. Each training item x is mapped to a class value y where $y \in Y$. The complete set of training data is X. A partitioning rule S (splitting rule) subdivides data set X into a set of subsets collectively known as X_S; that is, $X_S=\{X_1, X_2,...,X_k\}$ where $Ui\ X_i = \boldsymbol{X}$. A decision has each set of parent nodes corresponds to an X_S partitioning of the parent's data set, with the entire data set connected with the root. The number of items in X_i that belong to class y_j is $|\boldsymbol{X}_{ij}|$. The probability that a randomly selected member of X_i is of class y_j is $p_{ij} = \left|X_{ij}\right| / \left|X_i\right|$

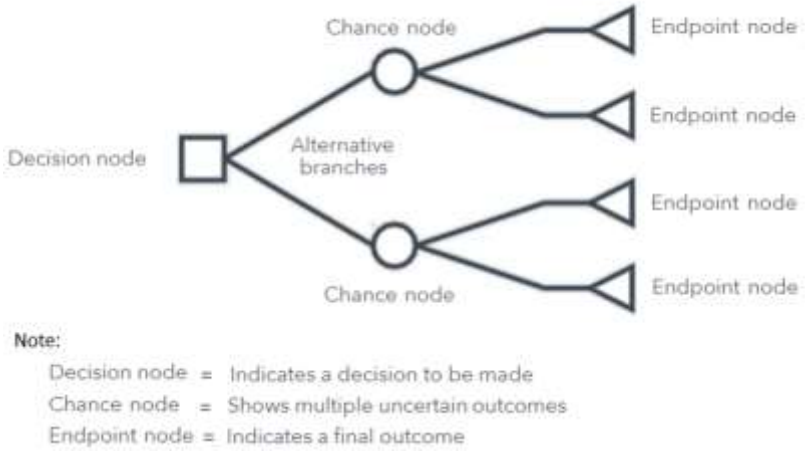

Figure 1: Decision trees are graphical models for describing sequential decision problems

Ordinal Regression

Ordinal logistic regression (OLR) is ideal for modeling a categorical variable (variable with more than two categories). Regarding our situation, the OSCC has already been graded as an ordinal variable with four different levels. Therefore the next step is to model the ordinal regression. The value of the regression parameter will be calculated using the maximum likelihood technique. The ordinal model is provided by

$$y_i^* = x_i\beta + \varepsilon_i \quad (1)$$

The dependent variable, however, is categorized, so we must use $C_x(x) = \ln\left[\frac{P(Y \le j \mid x)}{P(Y > j \mid x)}\right]$ and

$$\ln\left(\frac{\sum pr(event)}{1-\sum pr(event)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots + \beta_k X_k$$

. This may be summed up as

$$\ln\left(\frac{P(Y \le j \mid x)}{1-P(Y \le j \mid x)}\right) = \alpha_j + \beta_i X_k$$

$$i = 1...k,\ \ j = 1,2,..., p-1$$

(2)

where

α_j = called threshold or intercept

β_i = Parameter in the model

X_{i1} = Set of factors or independent variables.

Equation (2) above is an ordinal logistic model for k predictors with the p-1 levels response variable [18].

The Data

The research was conducted at Hospital Universiti Sains Malaysia (USM) in Kubang Kerian, Kelantan, Malaysia. A total of 57 patients took part in this study. The data summary for the selected variable in the analysis is described in Table 1.

Table 1: Data description of research variables

Code-variables	**Explanation of user Variables**
TNM Staging	1 = Stage 1; 2 = Stage 2; 3=Stage 3; 4=Stage 4a, 4b, 4c
T-Classification	1 = T1; 2=T2 ; 3 = T3 ; 4 = T4a,4b,4c
N-Classification	1 = N0; 2= N1; 3 =N2a; 4 = N3

Surgical Margin	1 = Not involve; 2 = Yes; 3 = Not applicable

The IBM SPSS program was used to do the statistical analysis. To get the best research outcomes, this study uses two separate statistical approaches which were decision tree analysis and ordinal regression analysis. In this instance, four variables were selected based on the importance of each variable to TNM staging. The decision tree analysis was used to select the variables for the ordinal regression modeling. It was discovered that T classification, N classification, and surgical margin were related to the TNM staging.

3. Results

The decision tree result was constructed based on the recommendation of IBM SPSS. To ensure that this model fits the data, the fitting of the model will be examined through the statistical procedure. Figure 2 shows the decision plot. It was found that in the first split, the *p*-value for the split is significant $[\chi^2(df) = 40.240(2); p < 0.05]$ and was assigned to the T classification.

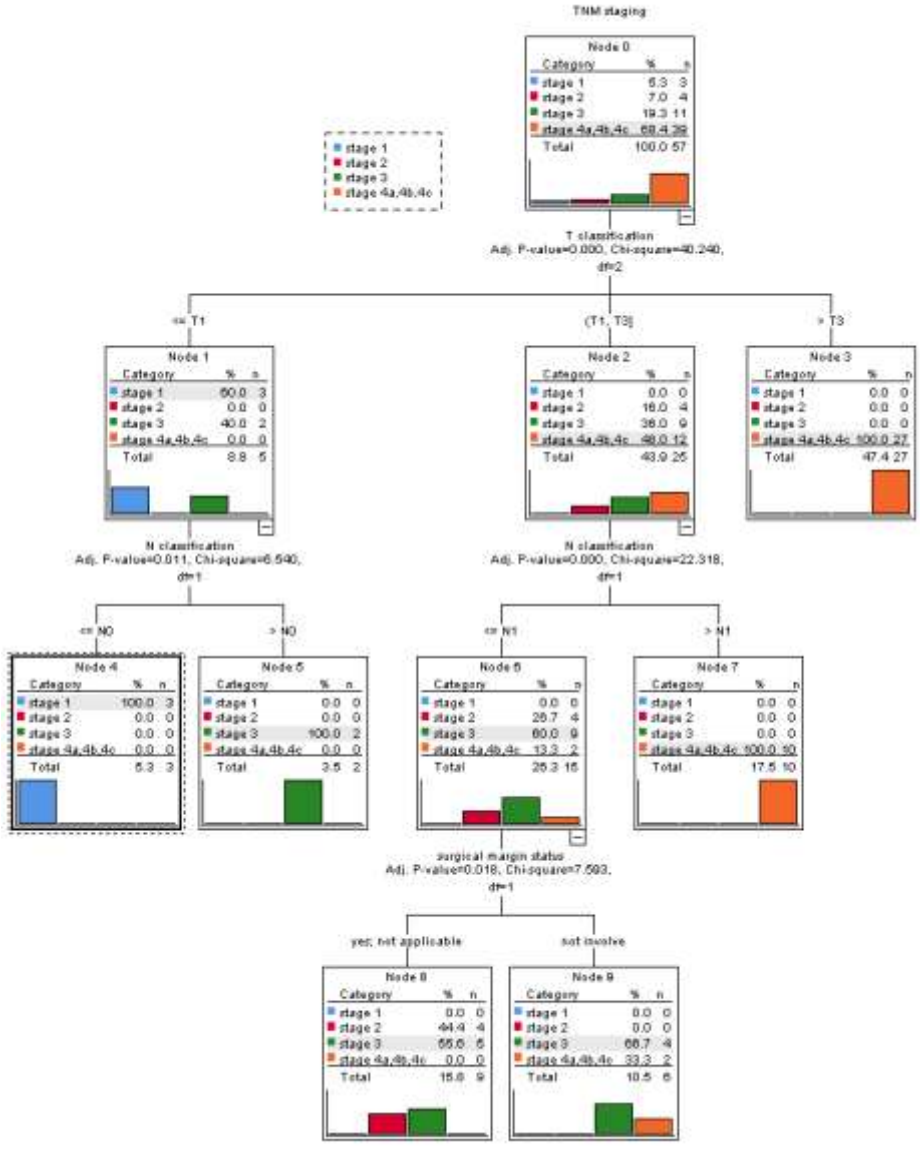

Figure 2: The Decision Tree Analysis

The T1 Classification is the starting point for the discussion. Three nodes fall under the T1 classification. Node 1 contains three cases at stage 1 and two cases at stage 3. There are 25 cases of the T2 to T3 classification at node 2, consisting of four cases (16%) at stage 2, Nine cases(36%) at stage 3,

and twelve cases (48%) at stage 4. At Node 3 (T4/T4a/T4b classification), 27 cases were reported. Furthermore, Node 4 and Node 5 fell under the N classification. The split value was given as $[\chi^2(df) = 6.540(1); p < 0.05]$. Node 4 shows that there are 3 cases found in the stage of no lymph node palpable. While with the presence of the lymph Node (Node 5), two cases were found. Node 2 was divided into Node 6, and Node 7 with the splitting value given as $[\chi^2(df) = 22.318(1); p < 0.05]$. At Node 6 with splitting $\leq$N1(metastasis in a single ipsilateral lymph node), the TNM stage 2, Stage 3, and Stage 4a, 4b, 4c were given as 4(26.7%), 9(60.0%), 2 (13.2). while Node 7 is referring to the Metastasis as specified in N2a, 2b, 2c. There are ten cases for stages 4a, 4b, and 4c, according to Figure 2. The third splitting was referred to as the surgical margin status $[\chi^2(df) = 7.593(1); p < 0.05]$ under Node 6. Under the surgical marginal Status "Yes and Not Applicable", which is referred to the Node 8, four cases were reported for stage 2 and five for stage 3. While the surgical marginal status was "Not involved" four cases were reported for stage 3 and two for stages 4a, 4b, and 4c. Using this statistical tool for the classification, three factors have a high potential for the TNM staging classification. The TNM staging scenario can be determined based on these three: T Classifications, N Classification, and surgical marginal status. This can be seen through the significant result of the splitting classification with the assistance of the statistical point of view. These variables will be a feeder for the ordinal regression modeling.

Table 2: Parameter estimate on the ordinal logistic regression model.

							95% Confidence Interval	
		Estimate	**Std. Error**	**Wald**	**df**	**p-value**	**Lower Bound**	**Upper Bound**
Threshold	[TNMstaging = 2]	7.777	2.364	10.822	1	0.001	3.143	12.410
	[TNMstaging = 3]	9.719	2.708	12.882	1	0.000	4.412	15.027
	[TNMstaging = 4]	11.823	2.908	16.527	1	0.000	6.123	17.523
Location	T Classification	2.442	0.588	17.213	1	0.000	1.288	3.5950
	N Classification	1.128	0.656	2.953	1	0.086	-0.158	2.4140
	Surgical Margins	0.790	0.638	1.533	1	0.216	-0.460	2.0390

Ordered logistic regression was applied with the simplifying the SPSS syntax; Significant at the level of 0.25
Model fitting information : $[\chi^2$ (df) =42.312(3); p < 0.05]
Goodness of fit test (Pearson) $[\chi^2$ (df) =34.66(30); p < 0.05]
Pseudo R-Square (Nagelkerke) : 62.3%

The results of an ordinal regression model incorporating ordinal regression and decision tree analysis are summarised in Table 2. The threshold is utilized to compute the predicted probabilities that a patient with a given set of characteristics belongs to a particular category. The proposed ordered logistic regression models for the different cut-off points shall be different and represented by a separate equation, so the formulations for the second, third, and fourth category becomes (the estimated model):

Ordinal logistic regression for TNMstaging = 2

$$\text{Logit}(P(Y \leq 2) = 7.777 + (2.442 \times \text{T Classification}) + (1.128 \times \text{N Classification}) + (0.790 \times \text{Surgical Margin}) \quad (3)$$

Ordinal logistic regression for TNMstaging = 3

$$\text{Logit}(P(Y \leq 3) = 9.719 + (2.442 \times \text{T Classification}) + (1.128 \times \text{N Classification}) + (\times \text{Surgical Margin}) \quad (4)$$

Ordinal logistic regression for TNMstaging = 4

$$\text{Logit}(P(Y\ 4) = 11.823 + (2.442 \times \text{T Classification}) + (1.128 \times \text{N Classification}) + (0.790 \times \text{Surgical Margin}) \quad (5)$$

The ordinal model considers three factors that adjust the class odds ratio, including T-Classifications, N-Classification, and Surgical Marginal status. Its shows that the T-Classifications [$\beta_{TClassifications}$ =2.442; p < 0.25; 95% CI (1.288, 3.595)], N-Classification [$\beta_{NClassifications}$ =1.128; p < 0.25; 95% CI (-0.158, 2.414)], and Surgical Margin Status. [$\beta_{Surgical\ Marginal\ status}$= 0.790; p < 0.25; 95% CI (-0.460, 2.039)]. T-Classifications, N-Classification, Surgical Margin and show that when either increase, there is a higher TNM staging.

4. Discussion

The discussion of this proposed model can be divided into two phases: the first phase focuses on the decision tree technique used to identify the factor, while the second phase examines the results

obtained. The methodology suggested for an ordered logistic regression model with decision tree analysis requires that the response variable is ordinal. The proposed method provides an alternative to ordinal regression analysis. This study helps health workers to identify individual TNM staging based on the high potential parameter. This paper's goal was to discuss the suggested idea of combining the ordinal regression methodologies and the results from a statistical perspective. The study's main objective was to evaluate a decision tree and ordinal regression methodology combination and apply biostatistical techniques. The decision tree method effectively assesses the variables that should be carefully chosen for the final model in this particular research project. Alternatively, decision trees can be used to select the most pertinent input variables that will be used to build the models, which can then be used to formulate a clinical hypothesis and inform future researchers. The decision tree-derived variable selection shows that such prediction models can be created successfully; this method is very helpful for the planning of health services. The most significant variables, in this case, were T-Classification, N-Classification, and surgical margin. This method can be utilized as an alternative to ordered regression modeling in situations where the selection of appropriate variables was based on a computational analysis that predicted the significance of the independent variable that should be selected for the final model. The findings assisted the decision-maker in achieving the best possible outcomes. Integrating a decision tree with ordered regression analysis allows us to create a more robust clinical application of risk factors. Using decision tree analysis, more conclusive, detailed, and reliable results can be obtained. The proposed method and obtained outcomes prove the superiority of the modeling approach taken. The purpose of this article is to create a new technique that combines a decision tree with ordered logistic regression. This methodology can be used by policymakers and health professionals to determine the correlation between TNM staging and its effects.

5. Conclusion

The factor that caused the TNM staging was successfully identified by the proposed method. The decision tree analysis plays as a tool for factor determination, and the obtained result would enhance the level of TNM staging and the diagnosis. The result shows that TNM staging was closely correlated with three factors: T-classification, N-classification, and surgical margin. Besides that, the methodology proposed provides an accurate evaluation of the final model's fit. The model's superior performance led to improved results, effective decision-making management, and the TNM staging prediction level. It is hoped that the proposed methodology will help clinicians manage and predict TNM staging in OSCC patients. The proposed idea of combining two statistical analyses can later be implemented as a single-user tool for statistical data analysis. The idea of combining ordered logistic regression and decision tree analysis produces very good results and this procedure can be an alternative for regression modeling with high accuracy. In conclusion, successful methodology development is a result of both the concept of combining statistical tools and the ability to obtain results using the suggested approach. This concept may be a useful advancement in methodology development research in the future.

Ethical approval

The study was approved by the Universiti Sains Malaysia Research Ethics and Human Research Committee (USM/JEPeM/16050184).

Conflict of Interest

The authors declare no conflict of interest.

Informed Consent

Informed consent was obtained from all individuals in this study

Research Funding

This study was funded by the Ministry of Higher Education (MoHE) Fundamental Research Grant Scheme (FRGS) FRGS/1/2022/STG06/USM/02/10

Acknowledgments

The authors would like to thank Universiti Sains Malaysia (U.S.M.). and to the Ministry of Higher Education for Fundamental Research Grant Scheme (FRGS).

6. References

Ajila V, Shetty H, Babu S, Shetty V, Hegde S. Human papilloma virus associated squamous cell carcinoma of the head and neck. Journal of sexually transmitted diseases. 2015;2015.

Petry KU. HPV and cervical cancer. Scandinavian Journal of Clinical and Laboratory Investigation. 2014;74(sup244):59-62.

Mehrotra R, Yadav S. Oral squamous cell carcinoma: etiology, pathogenesis and prognostic value of genomic alterations. Indian journal of cancer. 2006;43(2):60-6.

Shield KD, Ferlay J, Jemal A, Sankaranarayanan R, Chaturvedi AK, Bray F, et al. The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. CA: a cancer journal for clinicians. 2017;67(1):51-64.

Siddiqui IA, Farooq MU, Siddiqui RA, Rafi ST. Role of toluidine blue in early detection of oral cancer. Pakistan Journal of Medical Sciences. 2006;22(2):184.

Chaturvedi AK. Epidemiology and clinical aspects of HPV in head and neck cancers. Head and neck pathology. 2012;6:16-24.

Rashid S, Manzar S, Kazmi F, Shahzad HB, Jan ZA, Minam M. Influence of Risk Habits on Demographic Factors and its Impact on Oral Submucous Fibrosis. Pakistan Journal of Medical Sciences. 2021.

Vargas-Ferreira F, Nedel F, Etges A, Gomes APN, Furuse C, Tarquinio SBC. Etiologic factors associated with oral squamous cell carcinoma in non-smokers and non-alcoholic drinkers: a brief approach. Brazilian dental journal. 2012;23:586-90.

Cheong SC, Vatanasapt P, Yi-Hsin Y, Zain RB, Kerr AR, Johnson NW. Oral cancer in South East Asia: Current status and future directions. Translational Research in Oral Oncology. 2017;2:2057178X17702921.

Ahmad WMAW, Yaqoob MA, Noor NFM, Ghazali FMM, Rahman NA, Tang L, et al. The predictive model of oral squamous cell survival carcinoma: a methodology of validation. BioMed research international. 2021;2021.

Razak AA, Saddki N, Naing NN, Abdullah N. Oral cancer survival among Malay patients in Hospital Universiti Sains Malaysia, Kelantan. Asian Pac J Cancer Prev. 2010;11(1):187-91.

Seethala RR, Stenman G. Update from the 4th edition of the World Health Organization classification of head and neck tumours: tumors of the salivary gland. Head and neck pathology. 2017;11:55-67.

Fan X, Liu Y, Heilman SA, Chen JJ. Human papillomavirus E7 induces rereplication in response to DNA damage. Journal of virology. 2013;87(2):1200-10.

Goot-Heah K, Kwai-Lin T, Froemming GRA, Abraham MT, Rosdy NMMNM, Zain RB. Human papilloma virus 18 detection in oral squamous cell carcinoma and potentially malignant lesions using saliva samples. Asian Pacific Journal of Cancer Prevention. 2012;13(12):6109-13.

Prabhu S, Wilson D. Human papillomavirus and oral disease–emerging evidence: a review. Australian dental journal. 2013;58(1):2-10.

Wakeham K, Kavanagh K. The burden of HPV-associated anogenital cancers. Current oncology reports. 2014;16:1-11.

Lipworth L, Rossi M, McLaughlin J, Negri E, Talamini R, Levi F, et al. Dietary vitamin D and cancers of the oral cavity and esophagus. Annals of oncology. 2009;20(9):1576-81.

Grimm M, Cetindis M, Biegner T, Lehman M, Munz A, Teriete P, et al. Serum vitamin D levels of patients with oral squamous cell carcinoma (OSCC) and expression of vitamin D receptor in oral precancerous lesions and OSCC. Medicina oral, patologia oral y cirugia bucal. 2015;20(2):e188.

Mesaric J, Sebalj D. Decision trees for predicting the academic success of Students. Croatian Operational Research Review, 7 (2), 367-388. doi. 2016.

Alpaydin E. Decision Trees. 2014.Deena, S. R., Vickram, S., Manikandan, S., Subbaiya, R., Karmegam, N., Ravindran, B., ... & Awasthi, M. K. (2022). Enhanced biogas production from food waste and activated sludge using advanced techniques–a review. Bioresource Technology, 127234.