Machine Learning and Treatment Outcome Prediction for Oral Cancer

Chui Shan CHU, Nikki P. LEE, John ADEOYE, Peter THOMSON, Siu-Wai CHOI

Oral & Maxillofacial Surgery, Faculty of Dentistry, The University of Hong Kong

Correspondence:

Peter Thomson, Oral & Maxillofacial Surgery, Faculty of Dentistry, The University of Hong Kong

E-mail: thomsonp@hku.hk

Abstract

Background: The natural history of oral squamous cell carcinoma (OSCC) is complicated by progressive disease including loco-regional tumour recurrence and development of distant metastases. Accurate prediction of tumour behaviour is crucial in delivering individualized treatment plans and developing optimal patient follow-up and surveillance strategies. Machine learning algorithms may be employed in oncology research to improve clinical outcome prediction.

Methods: Retrospective review of 467 OSCC patients treated over a 19-year period facilitated construction of a detailed clinico-pathological database. 34 prognostic features from the database were used to populate 4 machine learning algorithms, linear regression (LR), decision tree (DT), support vector machine (SVM) and k-nearest neighbours (KNN) models, to attempt progressive disease outcome prediction. Principal component analysis (PCA) and bivariate analysis were used to reduce data dimensionality and highlight correlated variables. Models were validated for accuracy, sensitivity and specificity, with predictive ability assessed by Receiver Operating Characteristic (ROC) and Area under the Curve (AUC) calculation.

Results: Out of 408 fully characterized OSCC patients, 151 (37%) had died and 131 (32%) exhibited progressive disease at the time of data retrieval. The DT model with 34 prognostic features was most successful in identifying ‘true positive’ progressive disease, achieving 70.59% accuracy (AUC 0.67), 41.98% sensitivity and a high specificity of 84.12%.

Conclusion: Machine learning models assist clinicians in accessing digitized health information and appear promising in predicting progressive disease outcomes. The future will see increasing emphasis on the use of artificial intelligence to enhance understanding of aggressive tumour behaviour, recurrence and disease progression.

1. INTRODUCTION

Oral cancer, principally squamous cell carcinoma arising from the oral mucosal lining (OSCC), accounts for half the annual global mortality attributed to head and neck malignancy, with cancer-related deaths associated with aggressive primary tumours and advanced stage disease at clinical presentation[1]. Post-diagnosis, the natural history in any given patient may be complicated by progressive disease in terms of loco-regional recurrence of the primary tumour and/or development of distant blood-borne metastases. Accurate classification of risk and prediction of tumour behaviour at the time of initial diagnosis and intervention is therefore crucial in delivering individualized treatment plans and developing optimal patient follow-up and surveillance strategies[2,3].

In recent years, demographic, clinico-pathological, therapeutic and bio-molecular data have all been used to populate clinical decision-making tools, including statistical regression models and prognostic nomograms, in an attempt to predict poor clinical outcome post-OSCC treatment. Unfortunately, such methods have gained limited acceptance in contemporary clinical practice due to data validity concerns and little demonstrable predictive accuracy[3-12].

Within the last decade, machine learning algorithms that automate analytical model building have been employed in oncology research to improve prediction and attempt more reliable forecasts of clinical outcome. Their popularity is based upon a presumed ability to sequentially detect patterns, garner information and undergo automated training based on data input, especially complex non-homogenous data, ultimately making clinical predictions with minimal human intervention[13,14]. Whilst a degree of predictive accuracy for algorithms has been reported, in particular the use of support vector machines, boosted decision trees, decision forest and artificial neural networks, there is a need to validate the predictive power of machine learning by analysing disease progression within well-defined OSCC patient cohorts prior to widespread translation to clinical practice[15-19].

In a recent publication, we reported upon post-treatment outcomes for a 467 OSCC patient cohort in Hong Kong and observed that histopathological features of invasive tumour behaviour such as perineural invasion

(PNI), bone invasion (BNI), lymphovascular invasion (LVI) and extra-nodal extension (ENE), especially in combination, showed potential application as prognostic markers of rapid disease progression and poor clinical outcome[20]. The aim of this study, therefore, was to revisit this well-characterised patient cohort to evaluate the ability of supervised machine learning models to predict disease outcome.

2. METHODS

2.1 Study Population and Clinico-Pathological Data

A retrospective review of OSCC patients treated over a 19-year period, between 1st October 2000 and 1st October 2019, at the Queen Mary Hospital in Hong Kong was performed using records from the Hospital Authority Clinical Management System (HA CMS)[20]. Consecutively treated adult patients with clinical subtypes corresponding to ICD-10 C00-C06, C09 and C10 were retrieved from the database. Patient demographic information included age, sex, date of diagnosis, status at time of data retrieval (alive or dead), previous cancer history, and smoking, alcohol, human papillomavirus (HPV) and Epstein-Barr virus (EBV) status. Clinico-pathological data recorded tumour site, grading, histopathological characteristics of tumour invasiveness, resection margin status, pTNM classification, disease staging and, where appropriate, use of cervical lymph node dissection and/or adjuvant chemo-radiotherapy regimes. Outcome was recorded as either disease-free or progressive disease, defined by loco-regional tumour recurrence and/or development of distant metastases. Overall survival was determined from the date of primary diagnosis until death or most recent clinic follow-up.

2.2 Prognostic Features

A series of 34 demographic, clinico-pathological and lifestyle factors, extracted from the database in view of their association with progressive disease risk, were selected as prognostic features to populate the prediction models[21]. These are listed in Table 1, which also summarizes the manner in which each feature was classified in the model.

2.3 Prediction Models

MATLAB R2020a (MathWorks, Inc., Natick, MA, USA), a mathematical programming platform facilitating data plotting and analysis, was employed to build linear regression (LR), decision tree (DT), support vector machine (SVM) and k-nearest neighbours (KNN) models, 4 frequently used models for outcome prediction[22,23]. The 34 prognostic features (predictors) and the presence of progressive disease (outcome) were used to develop the models, which were evaluated by 15-fold cross validation to avoid overfitting (a model with too many variables which may just be 'noise'). Two methods were used before building the models to investigate whether data reduction could enhance performance: principal component analysis (PCA) to reduce data dimensionality and highlight correlated variables, and bivariate analysis (IBM SPSS for Windows 10 version 25) to identify prognostic features positively correlated with outcome ($p<0.05$). Models were validated for accuracy, sensitivity (true positive) and specificity (true negative), with diagnostic ability assessed via Receiver Operating Characteristic (ROC) and Area under the Curve (AUC) calculation. For each of the models, predictive performance using all 34 prognostic features was compared with those derived from PCA and bivariate analysis.

2.4 Ethical Approval

Approval to conduct this retrospective study was granted by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (Reference number UW-19-704). All clinical data were anonymised by the researchers, and all potential patient identifiers were removed prior to data analysis.

3. RESULTS

3.1 Patients, Outcome and Predictions

In total, 467 OSCC patients were identified from the HA CMS database and their full demographic and clinico-pathological data are summarized in Table 2. Ultimately, 59 patients were excluded from study analysis due to unavailability of clinico-pathological data (43) or because patients documented as alive failed to return for their most recent clinic assessments (16). Data from 408 OSCC patients (244 males and 164 females) were thus used to populate the machine learning models; at the time of data retrieval, 151 (37%) had died, and 131 (32%) exhibited progressive disease.

Bivariate analysis identified 13 prognostic features positively predictive of progressive disease development, and these are listed in Table 3, whilst PCA utilised between 16 to 34 components dependent upon the model, as summarized in Table 4. By listing predicted and actual progressive and non-progressive disease outcomes, Table 4 provides a performance comparison of LR, DT, SVM and KNN models with PCA and bivariate analyses.

3.2 Linear Regression

Compared to using all 34 predictive features or 13 selected by bivariate analysis, the LR model reduced to 18 components by PCA achieved the highest accuracy of 70.83% (AUC 0.68). PCA also performed best in terms of specificity (88.81%), although sensitivity was higher in the 34 feature model (35.11%).

3.3 Decision Tree

The DT model with 34 features attained 70.59% accuracy (AUC 0.67), with a sensitivity of 41.98% and specificity 84.12%, superior to both bivariate analysis and 19-component PCA.

3.4 Support Vector Machine

The SVM model with 34 features shared identical accuracy with 34-component PCA (69.85%, AUC 0.68). Whilst sensitivity at 24.43% increased to 25.95% after PCA application, specificity fell by 0.73%. Bivariate analysis reduced accuracy and sensitivity, to 68.63% and 15.27% respectively, although specificity reached 93.86%. In general, all 3 SMV models performed well in terms of specificity (greater than 90%).

3.5 K-Nearest Neighbour

Using bivariate analysis, the KNN model achieved 69.36% accuracy (AUC 0.71), with 35.11% sensitivity and 85.56% specificity. PCA (16 components) performed better than the 34 feature model for both accuracy (68.38% vs 66.42%) and specificity (88.45% vs 85.56%), although sensitivity was identical at 25.95%.

3.6 Comparison of Model Performance

Overall, the DT model using 34 prognostic features appeared most successful in identifying 'true positive' progressive disease, achieving 70.59% accuracy, 41.98% sensitivity, and a high specificity of 84.12%. In general, specificity was much higher (ranging from 79.42 to 93.86%) than sensitivity (15.27 to 41.98%) in all models.

4. DISCUSSION

4.1 Machine Learning Models

Machine learning is an increasingly popular application of artificial intelligence using computers to acquire and analyse complex data sets, identify patterns and develop predictive, decision-making algorithms that improve automatically with experience. Models are built from large, representative sets of 'training data' and progress to additional data processing to facilitate prediction. For this study, 4 models were populated with clinico-pathological data from a cohort of previously treated OSCC patients: LR to estimate relationships between dependent variables and their associated features, DT based upon individual observations (branches) and their perceived value (leaves), SVM with non-probabilistic, binary linear classification to predict categorization, and KNN non-parametric classification and regression. Machine learning differs from conventional statistics which, requiring prior knowledge of methods necessary to meet study objectives, test specific hypotheses and draw inference from study samples[24].

4.2 Predicting OSCC Outcome

It is frustrating that our ability to predict clinical outcome for OSCC patients in contemporary clinical practice remains limited. As a general observation, the incidence of progressive disease increases with length of patient follow-up. Whilst it is possible to attempt characterization of 'high-risk' patients using clinico-pathological features, there is inevitable cohort bias. It seems reasonable, therefore, to utilize artificial intelligence to improve accuracy of predictive diagnoses and facilitate targeted treatment intervention[2,20,25]. All 12 models in this study performed reasonably, although DT using 34 prognostic features was best at predicting OSCC progression, achieving 71% accuracy but only 42% sensitivity. There are few comparable data in the literature, although a recent systematic review reported SVM accuracy between 56.7 to 99.4%[26].

In a study of 311 early-stage tongue SCCs, an artificial neural network (ANN) was used to characterise invasive histopathology and achieved 88% accuracy and 71% sensitivity for locoregional recurrence prediction[16], whilst a decision forest algorithm to predict occult nodal metastasis in 71 T1/T2 OSCC patients reported an AUC of 0.84, with 91.7% sensitivity and 57.6% specificity[17]. Predictive ability of these models may have been improved by the measurement of specific disease outcomes in better defined patient cohorts with same stage disease.

4.3 Study Limitations

This was a retrospective study of clinico-pathological data retrieved from pre-existent HA CMS records. Consecutive OSCC patients were recruited from a number of HA facilities and exhibited heterogeneity of presenting disease. Machine learning algorithms are dependent upon the quality and precision of inputted data. It may be that conventional medical record information, which currently lacks genetic profiling, biomarker analyses and advanced histopathological imaging, are ultimately inadequate for predictive analyses. Deep neural networks, which facilitate multiple layer extraction of increasingly complex data and mimic human decision making, may be better applied in the future to study the inherently complex nature of tumour biology.

5. CONCLUSIONS

Machine learning models in this study have shown promise in predicting progressive OSCC disease outcomes. The future will see increasing emphasis on artificial intelligence to assist clinicians in utilizing digitized health information to predict outcome, inform personalized treatment decisions and rationalize intervention. It is hoped this will enhance understanding of biological mechanisms driving aggressive tumour behaviour and identify progressive disease at the earliest possible stage.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 68:394-424.

2. Thomson PJ. Perspectives on Oral Squamous Cell Carcinoma Prevention – Proliferation, Position, Progression and Prediction. Journal of Oral Pathology & Medicine 2018 47: 803-807.

3. Jerjes W, Upile T, Petrie A, Riskalla A, Hamdoon Z, Vourvachis M, et al. Clinicopathological parameters, recurrence, locoregional and distant metastasis in 115 T1-T2 oral squamous cell carcinoma patients. Head Neck Oncol. 2010 2:9.

4. Waldram R, Taylor AE, Whittam S, Iyizoba-Ebozue Z, Murray L, Frood R, et al. Evaluation of Locoregional Recurrence Patterns Following Adjuvant (Chemo)Radiotherapy for Oral Cavity Carcinoma. Clin Oncol (R Coll Radiol). 2020 32:228-37.

5. Larson AR, Kemmer J, Formeister E, El-Sayed I, Ha P, George J, et al. Beyond Depth of Invasion: Adverse Pathologic Tumor Features in Early Oral Tongue Squamous Cell Carcinoma. Laryngoscope. 2020 130:1715-1720.

6. Bhattasali O, Ryoo JJ, Thompson LDR, Abdalla IA, Chen J, Iganej S. Impact of chemotherapy regimen on treatment outcomes in patients with HPV-associated oropharyngeal cancer with T4 disease treated with definitive concurrent chemoradiation. Oral Oncol. 2019 95:74-78.

7. Kibe Y, Nakamura N, Kuno H, Hiyama T, Hayashi R, Zenda S, et al. Frequency and predictors of detecting early locoregional recurrence/disease progression of oral squamous cell carcinoma with high-risk factors on imaging tests before postoperative adjuvant radiotherapy. Int J Clin Oncol. 2019 24:1182-1189.

8. Sawant S, Ahire C, Dongre H, Joshi S, Jamghare S, Rane P, et al. Prognostic significance of elevated serum CD44 levels in patients with oral squamous cell carcinoma. J Oral Pathol Med. 2018 47:665-73.

9. Zaman SU, Aqil S, Sulaiman MA. Predictors of locoregional recurrence in early stage buccal cancer with pathologically clear surgical margins and negative neck. Acta Otorrinolaringológica Esp. 2018 69:226-30.

10. Sridharan S, Thompson LDR, Purgina B, Sturgis CD, Shah AA, Burkey B, et al. Early squamous cell carcinoma of the oral tongue with histologically benign lymph nodes: A model predicting local control and vetting of the eighth edition of the American Joint Committee on Cancer pathologic T stage. Cancer. 2019 125:3198-3207.

11. Montero PH, Yu C, Palmer FL, Patel PD, Ganly I, Shah JP, et al. Nomograms for preoperative prediction of prognosis in patients with oral cavity squamous cell carcinoma. Cancer. 2014 120:214-221.

12. Wang SJ, Patel SG, Shah JP, Goldstein DP, Irish JC, Carvalho AL, et al. An oral cavity carcinoma nomogram to predict benefit of adjuvant radiotherapy. JAMA Otolaryngol Head Neck Surg. 2013 139:554-559.

13. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nature Methods. 2018 15:233-234.

14. Zurada J, Levitan AS, Guan J. A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. The Journal of Real Estate Research. 2011 33:349-388.

15. Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, et al. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. Int J Med Inform. 2020;136:104068.

16. Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, et al. Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. Virchows Arch. 2019 475:489-497.

17. Bur AM, Holcomb A, Goodwin S, Woodroof J, Karadaghy O, Shnayder Y, et al. Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. Oral Oncol. 2019 92:20-25.

18. Chang SW, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics. 2013 14:170.

19. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019 9:6994.

20. Adeoye J, Thomson PJ, Choi S-W. Prognostic Significance of Multi-Positive Invasive Histopathology in Oral Cancer. Journal of Oral Pathology & Medicine 2020 *(In Press).*

21. Warnakulasuriya S. Prognostic and predictive markers for oral squamous cell carcinoma: the importance of clinical, pathological and molecular markers. Saudi Journal of Medicine and Medical Sciences. 2014 2:12.

22. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002 35:352-359.

23. Al-Aidaroos K, Bakar A, Othman Z. Medical Data Classification with Naive Bayes Approach. Information Technology Journal. 2012 11:1166-1174.

24. Bzdok D, Altman N, Krzywinski M. Points of significance: statistics versus machine learning. Nature Publishing Group; 2018.

25. Thomson PJ. Minimal intervention in oral cancer management: idealistic or realistic? Faculty Dental Journal 2018 9: 151-154.

26. Patil S, Awan KH, Arakeri G, Seneviratne CJ, Muddur N, Malik S, Ferrari M, Rahimi S, Brennan PA. Machine learning and its potential applications to the genomic study of head and neck cancer - A systematic review. Journal of Oral Pathology & Medicine 2019 48: 773-779.

Table 1: Individual Prognostic Features and Variable Classification Used in Machine Learning

	Prognostic Features	Variable Classification
1	Sex	Binary (Male/Female)
2	Age	Numerical
3	Smoking History	Categorical (unknown/non-/past/current smoker)
4	Alcohol Drinking	Categorical (unknown/non-/past/current drinker)
5	HPV Status	Categorical (unknown/negative/positive)
6	EBV Status	Categorical (unknown/negative/positive)
7	Past Cancer History	Categorical (unknown/no/yes)
8	Anterior Tongue	Binary (not involved/involved)
9	Posterior Tongue	Binary (not involved/involved)
10	Buccal Mucosa	Binary (not involved/involved)
11	Lips	Binary (not involved/involved)
12	Hard Palate	Binary (not involved/involved)
13	Soft Palate / Oropharynx	Binary (not involved/involved)
14	Maxillary Gingiva	Binary (not involved/involved)
15	Mandibular Gingiva	Binary (not involved/involved)
16	Tonsil	Binary (not involved/involved)
17	Floor of Mouth	Binary (not involved/involved)
18	Retromolar Region	Binary (not involved/involved)
19	Neck Dissection	Binary (no/yes)
20	Second Primary Tumour*	Binary (no/yes)
21	T Classification	Binary (smaller than 4cm/equal and larger than 4cm)
22	N Classification	Categorical (no lymph nodes/smaller than 6cm/equal or larger than 6cm)
23	Disease Staging	Categorical (stage I/II/III/IV)
24	Frozen Section Results	Categorical (no assessment/margin negative/dysplasia or in-situ tumor/margin positive)
25	Resection Margin Status	Categorical (unknown/negative/positive)
26	Tumor Grading	Categorical (unknown/well/moderately/poorly differentiated)
27	Cervical Lymph Node Metastasis	Categorical (unknown/no/yes)
28	DOI	Categorical (unknown/less than 1cm/equal or deeper than 1cm)
29	BNI	Categorical (unknown/negative/positive)
30	LVI	Categorical (unknown/negative/positive)

31	PNI	Categorical (unknown/negative/positive)
32	ENE	Categorical (unknown/negative/positive)
33	Radiotherapy	Categorical (no/neo-adjuvant/adjuvant)
34	Chemo-Radiotherapy	Categorical (no/neo-adjuvant/adjuvant)

* Second primary tumour was defined as the presence of two malignant tumours, at least 2 cm apart or detected 6 months or more after primary tumour diagnosis.

HPV Human papillomavirus; EBV Epstein-Barr virus; DOI Depth of invasion; BNI Bone invasion; LVI Lymphovascular invasion; PNI Perineural invasion; ENE Extranodal extension

Table 2: Patient Demographics and Clinico-Pathological Tumour Data

Variable	All (n= 467)	Male (n= 275)	Female (n= 192)
Age in Years at Diagnosis, mean (SD)	61.4 (14.1)	61.2 (13.4)	61.8 (15.1)
Current Status (n)			
Alive	282	159	123
Dead	181	113	68
Missing	3	2	1
Age at Death in Years, mean (SD)	68.3 (13.6)	68.9 (12.0)	67.3 (16.0)
History of Non-Head & Neck Cancer (n)			
Yes	85	56	29
No	381	218	163
Missing	1	1	
Tobacco Smoking (at time of diagnosis) (n)			
Non-smoker	262	94	168
Past smoker	89	84	5
Current smoker	86	79	7
Unknown	30	18	12
Alcohol Drinking (at time of diagnosis) (n)			
Non-drinker	248	98	150
Past drinker	36	34	2
Current drinker	109	99	10
Unknown	74	44	30
HPV status (n)			
Positive	24	20	4
Negative	50	36	14
Unknown	393	219	174
EBV status (n)			
Positive	9	7	2
Negative	18	15	3
Unknown	440	253	187
Tumour site, number (%)			
Tongue (anterior)	201 (43.0)	111 (40.3)	90 (46.8)
Tongue (base/posterior)	28 (6.0)	23 (8.4)	5 (2.6)
Buccal mucosa	69 (14.8)	32 (11.6)	37 (19.3)
Floor of mouth	26 (5.6)	23 (8.4)	3 (1.6)
Lips	3 (0.6)	3 (1.1)	0 (0.0)
Gingiva (mandibular)	57 (12.2)	30 (10.9)	27 (14.1)
Gingiva (maxillary)	18 (3.9)	8 (2.9)	10 (5.2)
Soft palate / Oropharynx	6 (1.3)	5 (1.9)	1 (0.5)
Retromolar Region	12 (2.6)	5 (1.8)	7 (3.6)
Hard Palate	11 (2.4)	4 (1.5)	7 (3.6)

Tonsil	36 (7.7)	31 (11.3)	5 (2.6)
pTNM Classification, number (%)			
pT			
T1	146 (31.3)	85 (30.9)	61 (31.8)
T2	130 (27.8)	70 (25.5)	60 (31.3)
T3	39 (8.4)	26 (9.5)	13 (6.8)
T4a	122 (26.1)	74 (26.9)	48 (1.6)
T4b	5 (1.1)	5 (1.8)	0 (0.0)
Missing	25 (5.4)	15 (5.5)	10 (5.2)
pN			
Nx	18 (3.9)	5 (1.8)	13 (6.8)
N0	249 (53.3)	149 (54.2)	100 (52.1)
N1	56 (12.0)	30 (10.9)	26 (13.5)
N2a	11 (2.4)	6 (2.2)	5 (2.6)
N2b	66 (14.1)	40 (14.5)	26 (13.5)
N2c	29 (6.2)	20 (7.3)	9 (4.7)
N3	13 (2.8)	10 (3.6)	3 (1.6)
Missing	25 (5.4)	15 (5.5)	10 (5.2)
pM			
M0	440 (94.32)	260 (94.5)	180 (93.8)
M1	2 (0.4)	0 (0.0)	2 (1.0)
Missing	25 (5.4)	15 (5.5)	10 (5.2)
Disease Staging			
Stage 1	118 (25.3)	66 (24.0)	52 (27.1)
Stage 2	75 (16.1)	43 (15.6)	32 (16.7)
Stage 3	56 (12.0)	30 (10.9)	26 (13.5)
Stage 4A	176 (37.7)	109 (39.6)	67 (34.9)
Stage 4B	15 (3.2)	12 (4.4)	3 (1.6)
Stage 4C	2 (0.4)	0 (0.0)	2 (1.0)
Missing	25 (5.4)	15 (5.5)	10 (5.2)
Neck Dissection			
No	63 (13.5)	33 (12.0)	30 (15.6)
Yes	393 (84.2)	235 (85.5)	158 (82.3)
Unknown	11 (2.4)	7 (2.5)	4 (2.1)
Tumour Grading			
Well differentiated	132 (28.2)	73 (26.6)	59 (30.8)
Moderately differentiated	248 (53.1)	145 (52.7)	103 (53.6)
Poorly differentiated	54 (11.6)	37 (13.5)	17 (8.9)
Missing	33 (7.1)	20 (7.3)	13 (5.2)
Use of Adjuvant Chemo-Radiotherapy			
Combination Chemo-Radiotherapy	107 (22.9)	73 (26.5)	34 (17.7)
Radiotherapy	113 (24.2)	62 (22.5)	51 (26.6)
None	246 (52.7)	140 (50.9)	106 (55.2)
Missing	1 (0.2)	0 (0.0)	1 (0.5)

Frozen Section Margins			
Negative	314 (67.2)	185 (67.3)	129 (67.2)
Positive	52 (11.1)	33 (12.0)	19 (10.0)
Missing	101 (21.6)	57 (20.7)	44 (22.9)
Tumour Resection Margin Status			
Negative	406 (86.9)	234 (85.1)	172 (89.6)
Positive	36 (7.7)	25 (9.1)	11 (5.7)
Missing	25 (5.4)	16 (5.8)	9 (4.7)
Tumour Invasiveness, number (%) positive			
Bony invasion	81 (17.3)	46 (16.7)	35 (18.2)
Perineural invasion	93 (20.0)	55 (20.0)	38 (19.8)
Lymphovascular invasion	91 (19.5)	65 (23.6)	26 (13.5)
Extra-nodal extension	79 (16.9)	51 (18.5)	28 (14.6)
Depth of Invasion (cm)			
<1cm	96 (20.6)	61 (22.2)	35 (18.2)
≥1cm	67 (14.3)	44 (16.0)	23 (12.0)
Missing	304 (65.1)	170 (61.8)	134 (69.8)

HPV – Human papillomavirus; EBV – Epstein-Barr virus

Table 3: Features Positively Predictive of Progressive Disease (Bivariate Analysis)

Selected features	Chi-square value	*p-value*
HPV	13.44	0.001
Anterior tongue	5.90	0.015
Buccal mucosa	8.22	0.004
Tonsil	3.93	0.047
T stage	13.04	0.0003
Overall stage	9.88	0.02
Neck dissection	11.84	0.001
Frozen section positivity	16.02	0.001
Resection margin positivity	10.08	0.006
Presence of metastatic nodules	23.20	0.000009
DOI	16.85	0.0002
PNI	16.39	0.0003
ENE	10.42	0.005

Table 4: Comparative Performance of Machine Learning Models in Identifying Progressive Disease

Linear Regression Models							
34 features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	46	46	67.89	0.68	35.11	83.39
	Non-progressive (Predicted)	85	231				
PCA-18 components		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	43	31	70.83	0.68	32.82	88.81
	Non-progressive (Predicted)	88	246				
Bivariate analysis-13 selected features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	36	39	67.16	0.7	27.48	85.92
	Non-progressive (Predicted)	95	238				
DT models							
34 features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	55	44	70.59	0.67	41.98	84.12
	Non-progressive (Predicted)	76	233				
PCA-19 components		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	47	57	65.44	0.6	35.88	79.42
	Non-progressive (Predicted)	84	220				
Bivariate analysis-13 selected features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	52	50	68.38	0.66	39.69	81.95
	Non-progressive (Predicted)	79	227				

Support Vector Machine models

34 features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	32	24	69.85	0.68	24.43	91.34
	Non-progressive (Predicted)	99	253				
PCA-34 components		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	34	26	69.85	0.68	25.95	90.61
	Non-progressive (Predicted)	97	251				
Bivariate analysis-13 selected features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	20	17	68.63	0.62	15.27	93.86
	Non-progressive (Predicted)	111	260				

K-Nearest Neighbors models

34 features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	34	40	66.42	0.69	25.95	85.56
	Non-progressive (Predicted)	97	237				
PCA-16 components		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	34	32	68.38	0.67	25.95	88.45
	Non-progressive (Predicted)	97	245				
Bivariate analysis-13 selected features		Progressive (Actual)	Non-progressive (Actual)	Accuracy%	AUC	Sensitivity%	Specificity%
	Progressive (Predicted)	46	40	69.36	0.71	35.11	85.56
	Non-progressive (Predicted)	85	237				