Journal of Big Data

RESEARCH Open Access

Predicting oral cancer risk in patients with oral leukoplakia and oral lichenoid mucositis using machine learning

John Adeoye[1], Mohamad Koohi-Moghadam[2], Siu-Wai Choi[3], Li-Wu Zheng[1], Anthony Wing Ip Lo[4], Raymond King-Yin Tsang[5], Velda Ling Yu Chow[6], Abdulwarith Akinshipo[7], Peter Thomson[8*] and Yu-Xiong Su[1*]

*Correspondence: peter.thomson1@jcu.edu.au; richsu@hku.hk

[1] Division of Oral and Maxillofacial Surgery, Faculty of Dentistry, The University of Hong Kong, Hong Kong, SAR, China
[2] Division of Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong, SAR, China
[3] Department of Orthopedics and Traumatology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China
[4] Department of Pathology, Queen Mary Hospital, Hong Kong, SAR, China
[5] Division of Otorhinolaryngology, Department of Surgery, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China
[6] Division of Head and Neck Surgery, Department of Surgery, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China
[7] Department of Oral and Maxillofacial Pathology and Biology, Faculty of Dental Sciences, University of Lagos, Lagos, Nigeria
[8] College of Medicine and Dentistry, James Cook University, Cairns, QLD, Australia

Abstract

Oral cancer may arise from oral leukoplakia and oral lichenoid mucositis (oral lichen planus and oral lichenoid lesions) subtypes of oral potentially malignant disorders. As not all patients will develop oral cancer in their lifetime, the availability of malignant transformation predictive platforms would assist in the individualized treatment planning and formulation of optimal follow-up regimens for these patients. Therefore, this study aims to compare and select optimal machine learning (ML)-based models for stratifying the malignant transformation status of patients with oral leukoplakia and oral lichenoid mucositis. One thousand one hundred and eighty-seven patients with oral leukoplakia and oral lichenoid mucositis treated at three tertiary health institutions in Hong Kong, Newcastle UK, and Lagos Nigeria were included in the study. Demographic, clinical, pathological, and treatment-based factors obtained at diagnosis and during follow-up were used to populate and compare forty-six machine learning-based models. These were implemented as a set of twenty-six predictors for centers with substantial data quantity and fifteen predictors for centers with insufficient data. Two best models were selected according to the number of variables. We found that the optimal ML-based risk models with twenty-six and fifteen predictors achieved an accuracy of 97% and 94% respectively following model testing. Upon external validation, both models achieved a sensitivity, specificity, and F1-score of 1, 0.88, and 0.67 on consecutive patients treated after the construction of the models. Furthermore, the 15-predictor ML model for centers with reduced data achieved a higher sensitivity for identifying oral leukoplakia and oral lichenoid mucositis patients that developed malignancies in other treatment settings compared to the binary oral epithelial dysplasia system for risk stratification (0.96 vs 0.82). These findings suggest that machine learning-based models could be useful potentially to stratify patients with oral leukoplakia and oral lichenoid mucositis according to their risk of malignant transformation in different settings.

Keywords: Artificial intelligence, Machine learning, Oral leukoplakia, Oral lichen planus, Oral lichenoid lesions, Oral potentially malignant disorders, Oral cancer

Introduction

Background of the study

Oral cancer is the most prevalent malignancy in the head and neck region [1, 2]. Commonly, tumors are of the squamous cell carcinoma subtype histologically and related to the tissue-damaging sequelae of behavioral influences such as tobacco smoking/chewing, areca nut consumption, and heavy alcohol drinking in addition to certain genetic and autoimmune predisposition [3, 4]. In other patients, tumors arise in the absence of known risk factors [5].

The natural history of oral cancer can be diverse. Some patients present with malignant ulcers or exophytic growths while others exhibit precursor mucosal disease with an inherently high risk of malignant transformation; these are collectively known as oral potentially malignant disorders (OPMD) [6, 7]. About 50–62% of oral cavity cancers arise from precancerous lesions such as leukoplakia (white patch), proliferative verrucous leukoplakia (white patch with verrucous appearance), erythroplakia (red patch), erythroleukoplakia (mixed white and red patch), oral lichen planus, and oral lichenoid lesions [6–8]. More widespread conditions include oral submucous fibrosis, oral graft vs host disease (OGVHD), long-term immunosuppressive treatment, Plummer-Vinson syndrome, chronic discoid lupus erythematosus, and dyskeratosis congenita [7, 9, 10]. Since not all patients with oral potentially malignant disorders develop cancer in their lifetime, appropriate recognition and prediction of malignant transformation risk in these disorders could be imperative to obtaining impactful oral cancer prevention, early diagnosis, and disease-specific prognosis.

There are, however, no consistent malignant transformation data in the literature to guide the management of common precursor lesions such as oral leukoplakia and oral lichenoid mucositis (oral lichen planus and oral lichenoid lesions) [7, 8, 11]. The average malignant transformation proportion of oral leukoplakia, oral lichen planus, and oral lichenoid lesions ranges from 5.9 to 14%, 0.9 to 2.3%, and 1.6 to 7% [12–15]. As such, practitioners, especially those working in non-specialist services, are unable to reliably inform or reassure patients about the risk of oral cancer development. Presently, grading oral epithelial dysplasia following histopathology appears to be the most common method of estimating the risk of malignant transformation, although, this is fraught with subpar accuracy, poor precision, and unclear reproducibility [16, 17]. Also, molecular markers suggested for outcome prediction have only been associated with certain risk groups without overt predictive optimization for use in clinical practice [8, 18].

Nomograms based on statistical inferences have been proposed to serve as predictive adjuncts in clinical practice [19]. Nomograms constructed to predict malignant transformation have shown promising accuracy in different populations [20–22], but their implementation may be inelegant, technical, and unautomated which may prolong patient consultation time. Also, such platforms do not have a threshold score for risk stratification and none of the tools have been externally validated at this time [3].

Aims and significance

The application of artificial intelligence and machine learning in clinical prediction and decision-making has been found to be relatively superior to traditional statistical

methods in many fields including oncology [23–26]. Learning models based on these classifiers have outperformed alternate methods of prediction with intrinsic merits that make them attractive for potential clinical application compared to some clinical nomograms [23]. However, little information is available from comprehensive studies that have utilized a comparative modeling approach to support the consideration of machine learning models to predict malignant transformation in oral leukoplakia and oral lichenoid mucositis using variables available from medical records. Therefore, this study aimed to compare and validate machine learning-based models to predict the malignant transformation risk of oral leukoplakia and oral lichenoid mucositis using demographic, clinicopathological, and treatment features obtainable from patient medical records in different treatment centers.

Research hypothesis

This study hypothesizes that machine learning models based on health records information would predict oral cancer risk in oral leukoplakia and oral lichenoid mucositis with satisfactory to good accuracy, sensitivity, and specificity. Additionally, the performance of the machine learning models will be equivalent to or even better than the sole application of the binary oral epithelial dysplasia grading system for risk stratification in oral leukoplakia and oral lichenoid mucositis.

Materials and methods

Data description

Retrospective patients with clinical diagnoses of oral leukoplakia, oral lichen planus, and oral lichenoid lesion) who had biopsy and histopathology for definitive diagnosis were included in this study. The rationale for selecting these diseases was due to their global prevalence and variability of their malignant transformation proportions/rates which necessitate individualized prediction in these specific disorders [8, 27, 28]. As such, patients with erythroplakia or proliferative verrucous leukoplakia at other oral cavity sites with coexisting diseases of interest were excluded.

Patient cohorts from three institutions in Hong Kong, Newcastle Upon Tyne, and Lagos Nigeria were used (Fig. 1). The Hong Kong cohort included two patient groups. The first group comprised 716 patients managed at the Departments of Oral and Maxillofacial Surgery, Otorhinolaryngology, and Head and Neck Surgery of the Queen Mary Hospital, Hong Kong between January 2003 and December 2019 (collected for machine-learning-based model development). This cohort has also been described in a previous report by our group [29]. Further, the second group of Hong Kong patients (n = 58) encountered from January 1 to December 31, 2020, as well as the Newcastle and Lagos patient cohorts (n = 413) were sourced following model construction for external validation. Newcastle patients were treated at the Maxillofacial Surgery Unit of the Newcastle Dental Hospital and the Royal Victoria Infirmary, Newcastle-Upon-Tyne, the United Kingdom [30–32]. Likewise, the Lagos cohort was treated at the Dental clinics of the Lagos University Teaching Hospital, Lagos Nigeria between January 2013, and September 2019. None of the patients had inherited diseases that predisposed them to oral cancer such as Fanconi anemia, xeroderma pigmentosum, and Li-Fraumeni syndrome.

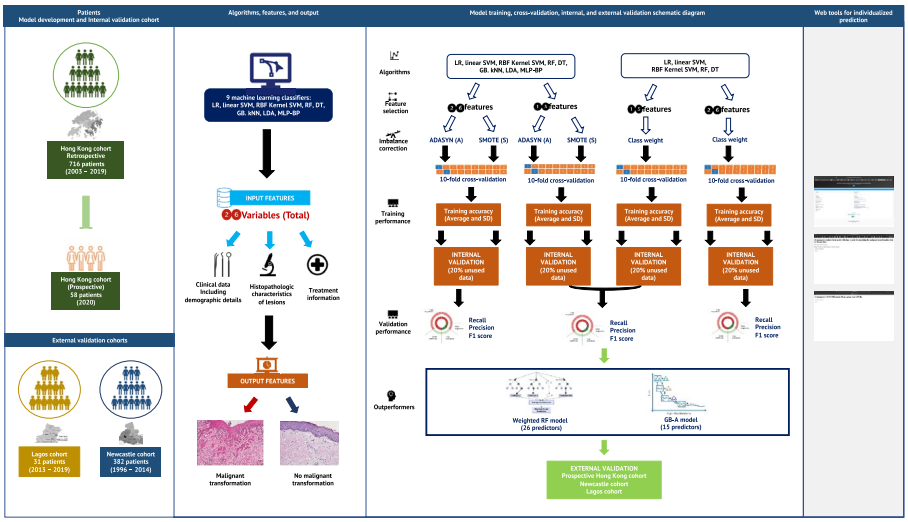

Fig. 1 Description of model construction and deployment phases underwent in this study

Variable selection and rationale

In total, 27 variables were considered which cuts across sociodemographic, clinical, pathological, and treatment information (Table 1). These variables have been presented in several reports as independent risk predictors for the malignant transformation of these precancerous lesions [21, 33–35]. Treatment for these conditions ranges from routine observation, pharmacological management, and scalpel or laser surgical excision/ablation and this was factored into model development [36]. A binary outcome indicating the presence or absence of malignant transformation was obtained for patients in all cohorts in this study. This was determined as of December 31, 2021, for the Hong Kong and Lagos OPMD cohorts and December 31, 2014, for the Newcastle cohort. Irrespective of the follow-up dates, all patients included had a minimum follow-up of 12 months. No statistical assessment of the effect of the predictors on this outcome specifically for these cohorts was conducted in this study. Also, data was not restricted only to patients with oral epithelial dysplasia as oral cancer may also occur in oral leukoplakia and oral lichenoid mucositis without this histologic feature [37]. Only squamous cell carcinomas arising from the precancer lesion, recurrences, or continued progression to adjacent sites were considered in this study.

Model development

Data preprocessing and feature engineering

Electronic entry spreadsheet templates (Microsoft Excel for Mac v 16.52) were used for data collection and column filtering was done to ensure correlation between pertinent variables. Input features for training were continuous (n = 2), Boolean (n = 16), and categorical (n = 9). As OPMDs represent disparate subtypes/entities, the machine learning models in this study were trained to account for the specific clinical subtype (oral leukoplakia vs oral lichenoid mucositis) as this variable was included as a predictor. Both tobacco smoking and alcohol consumption status had missing data that were

Table 1 List of input parameters obtained for machine learning and missing instances

Input feature	Type	Missing instances (Hong Kong Cohort only)	Handling technique for missing data
Age	Continuous	0	NA
Sex	Boolean	0	NA
Tobacco smoking	Boolean	2	Binarization of variables during feature engineering
Alcohol drinking	Categorical (nominal)	33	
Risk habit indulgence following diagnosis	Categorical (nominal)	0	NA
Previous malignancy	Categorical (nominal)	0	NA
Charlson Comorbidity Index (CCI)	Continuous	0	NA
Hypertension status	Boolean	0	NA
Diabetes Mellitus status	Boolean	0	NA
Hyperlipidemia status	Boolean	0	NA
Autoimmune disease status	Boolean	0	NA
Viral hepatitis status	Boolean	0	NA
Type of lesion	Boolean	0	NA
Clinical subtype of lichenoid lesion	Categorical (nominal)	0	NA
Tongue/FOM involved	Boolean	0	NA
Labial/buccal mucosa involved	Boolean	0	NA
Retromolar area involved	Boolean	0	NA
Gingiva involved	Boolean	0	NA
Palate involved	Boolean	0	NA
Number of lesions	Categorical (ordinal)	0	NA
Presence of ulcers or erosions	Boolean	0	NA
Presence of induration	Boolean	0	NA
Treatment at diagnosis	Categorical (nominal)	0	NA
Recurrence after surgical excision	Boolean	0	NA
Number of recurrences	Categorical (ordinal)	0	NA
Oral epithelial dysplasia at diagnosis	Categorical (nominal)	0	NA
Oral epithelial dysplasia detected during follow-up	Categorical (nominal)	0	NA

All predictors were used for modeling and the variables in *bold* were the predictors included in the 15 feature models in this study. Note that 'Tobacco smoking' and 'Alcohol consumption' were binarized into a single variable

handled through the binarization of both variables into two categories which left no missing instances (Table 1). This new variable – patient risk habit category, was based on classifying ever-smokers and ever-drinkers as smoking and alcohol-drinking (SD) patients and non-smokers and non-drinkers as non-smoking and non-alcohol-drinking (NSND) patients. This stratification has been used previously to propose that both groups are distinct entities [5, 38]. Altogether, this left 26 variables available to model malignant transformation risk using machine learning.

We envisaged that all 26 variables (Table 1) obtained following feature engineering may not be easily obtained by specialists at different treatment centers, especially in remote areas where electronic patient documentation was not practiced. Therefore, 15 variables (Table 1 in bold) more likely to be available in manual records or obtained from patients were selected for the construction of another set of models to be applied in this scenario [21–23].

Machine learning algorithms considered

To select the optimal models useful for prediction, the performance of nine popular machine learning classifiers was compared (Fig. 1). These were logistic regression, linear support vector machines (SVM), radial basis function (RBF) kernel SVM, random forest, decision tree, gradient boosting, k-nearest neighbor (kNN), linear discriminant analysis, and multilayer perceptron with backpropagation (MLP-BP). The choice of these algorithms was per previously described models for the prediction of other oncological outcomes [23, 24, 39]. In this study, logistic regression, linear discriminant analysis, and linear SVM were the parametric models used to optimize the function being learned to a known form as a linear combination of the input parameters. Conversely, all the other algorithms use the nonparametric method which does not make a specific assumption of the mapping function from the training data and possesses adjustable parameters.

Model training

The first group of the Hong Kong cohort (n = 716) was used to train the algorithms. Data splitting into training and test sets was done according to the 80:20 rule. Two sets of models were trained per algorithm based on the number of input variables (Fig. 1). One set was trained with all 26 features (following binarization of alcohol drinking and tobacco smoking status) while the other set of models included the 15 predictors that would be obtained easily in different centers. Ten-fold cross-validation was employed with the training data to obtain performance measures across each fold for tuning hyperparameters after each training session and to evaluate the model stability. Model stability is defined as the sensitivity of a machine learning model to variations in the training dataset and it is considered a measure of the reliability of the trained models [40]. In this study, the model stability was evaluated using the coefficient of variation (COV) as a ratio of the standard deviation (s) to the mean ($\bar{x}$) of the training accuracy for each model to provide the variability of this measure. In this study, the COV is inversely proportional to the model stability such that models with better stability were expected to have lower COV values compared to less-stable models. Mathematically, the coefficient of variation is defined as:

$$COV = \frac{s}{\bar{X}} * 100$$

As the occurrence of malignant transformation can be described as a rare event relative to the entire cohort (event to censored ratio of 1:10), we also compared the performance of two synthetic oversampling techniques and class weight assignment for parametric and tree-based models (i.e., logistic regression, SVM, decision tree, and random forest). Oversampling class imbalance techniques used were synthetic minority oversampling technique (SMOTE for categorical and continuous variables) and adaptive synthetic sampling approach (ADASYN). Both methods create synthetic data for the minority class using the k-nearest neighbor algorithm. While both methods are similar, ADASYN uses a density distribution to decide the number of synthetic samples to be generated with more samples for lower-density minority regions (instances that are hard to learn) as opposed to an equal number of synthetic samples for SMOTE [41]. Class

weight assignment for parametric and tree-based algorithms was implemented by automatically adjusting the different cost function weights given to the minority and majority outcome classes in an inversely proportional fashion relative to their frequencies in the training data. Overall, this meant that six models were each trained for logistic regression, SVM, decision tree, and random forest algorithms, and four models for others based on feature selection and the type of class imbalance. In total, 46 models were trained in this study.

Manual hyperparameter tuning according to the cross-validation performance was implemented in gradient boosting, decision tree, random forest, kNN, and MLP-BP to select optimal models while automated hyperparameter optimization was employed in other models. Manual categorical encoding was also used with categorical data to allow the implementation of SVM and kNN during training. The MLP-BP architecture had 64 hidden layers and was trained with rectified linear unit (ReLU) activation, adaptive moment estimation (ADAM) optimizer, and a learning rate of 0.001. Also, for MLP-BP, early stopping regularization was implemented to stop model performance when there was no improvement on a randomly preselected 10% validation set.

Model validation

The unseen 20% dataset of the first Hong Kong patient group was used for testing the trained models. Testing performance measures were generated for all models trained per algorithm i.e., 46 models. These metrics were used to compare the overall performance and to select the best-performing model based on 26 predictors and 15 predictors in this study. Furthermore, two rounds of external validation were performed. First, patients in the second Hong Kong cohort group (n = 58) were used to independently validate the two best-performing models based on 26 features and 15 features. Second, external validation of the 15-feature model (intended for general use due to its common input features) was performed using the Newcastle and Lagos cohorts (n = 413). The performance of these models in the latter external validation round was compared to the binary dysplasia grading system currently used for risk stratification in clinical practice. In this system, atypia extending below and to the middle of the squamous epithelium on histology is classified as low grade/low risk and atypia above this level is deemed high grade/high risk [42, 43].

Performance measures

Six measures were obtained for each of the models to assess all-around performance. This included accuracy, sensitivity, specificity, precision, negative predictive value, and F1-score (defined as the harmonic mean between the precision and sensitivity). Mathematical calculations for the performance metrics are given below:

$$Sensitivity = \frac{TP}{TP + FN} \quad Specificity = \frac{TN}{TN + FP}$$
$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad Precision = \frac{TP}{TP + FP}$$
$$NPV = \frac{TN}{TN + FN} \quad F1\,score = 2 * \frac{precision * sensitivity}{precision + sensitivity}$$

The sensitivity and F1 score were the main measures for comparing model performance as the development of malignant lesions is time-dependent and credible prediction models are expected to identify majority of the cases that will have malignancies to enable surgical management and closer routine follow-up [44]. However, greater preference was given to the sensitivity than precision for models with similar F1-scores ($\pm$0.05).

Explainability and net benefit analyses

To explain the rationale behind the risk prediction of the best-performing models as a measure of the input features, the Shapley additive explanations (SHAP) framework was used in this study. This is a model-agnostic approach to machine learning interpretability that is based on competitive game theory [45]. Global SHAP values were obtained for the predictions of the models using the Hong Kong external validation dataset.

Furthermore, decision curve analysis [46] was performed to estimate the net benefit of the models based on 26 and 15 features for selecting patients with oral leukoplakia and oral lichenoid mucositis that may benefit from surgical intervention and close monitoring. Net benefit was determined for all probability thresholds at or below 50% and these were used to plot the decision curves of both models. Decision curves were compared to standard references which included proffering treatment for all patients and treating none of the patients. Predicted output from the Hong Kong external validation dataset was also used for decision curve analysis.

Web-based application for future beta testing

To encourage further validation of trained models, we generated web platforms (based on 26 and 15 input variables) deployed on the backend of the best-performing models (Fig. 1). These tools were developed using the Flask module in Python and deployed using the Heroku cloud platform. Patient data with blinded outputs may be uploaded or inputted to this platform to automatically generate predictions without any need to repeat the model development process. We created two types of web-based graphical user interfaces i.e., data frame and interactive user-defined platforms. For the data frame web tool, the performance of the algorithms can be determined by uploading institutional EHR datasets in form of spreadsheets (CSV format) without the need to develop the model internally. Likewise, the interactive platform can be used to determine the risk of malignant transformation for single instances of patients with oral leukoplakia or oral lichenoid mucositis.

Computation

Descriptive statistics were performed using SPSS v 26 (IBM Corp., Armonk, NY, USA). Training, testing, and validation of all supervised learning algorithms as well as model deployment were performed with Python v 3.8.7 using scikit-learn [47] and Flask [48] (Python Software Foundation, Wilmington, DE, USA) [23]. McNemar's test was used to compare the sensitivity values of the best 15-feature model and binary epithelial dysplasia grading during external validation. Probability values below 5% was used to denote statistical significance. The Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist [49] was adhered to during the study.

Results

Patients' description

One thousand one hundred and eighty-seven patients were included in this study with 716 patients used for model construction and 471 patients used for external validation. For the training cohort, the average follow-up time was 90.9 months and the median age (IQR) of patients was 58 (49–67) years. Majority of them were females (56%) and non-smokers/non-drinkers (65.5%). Most of the lesions were of the oral leukoplakia subtype than oral lichenoid mucositis (54.3% vs 45.7%, Table 2). Lesions involving the buccal or labial mucosa were more common than other anatomic sites (56.8%) with only 6.6% of patients presenting with indurated lesions. Most patients received no treatment or were placed on periodic observation (41.9%) while 30.9% of them received surgical intervention for their mucosal disease. Post-excision recurrence was observed among 19% of those that received surgical treatment with 13.5% experiencing only one recurrence as of the censoring date. Epithelial dysplasia was found on histology in 9.5% of patients at diagnosis and 7% of patients during follow-up biopsies. Approximately, one-tenth of the patients experienced malignant transformation of their lesions. Clinical and histologic characteristics of tumors are also shown in Table 2.

A detailed description of the external validation cohorts is presented in Table 2 and Additional file 1: Table S1. This cohort had an average follow-up time of 77.3 months. Overall, the prevalence of malignant transformation in the external validation cohort was 6% and the Newcastle cohort comprised patients with a clinically high risk of malignant transformation as 89% of the patients had epithelial dysplasia of various grades at diagnosis.

Predictive performance of classifiers

In total, 46 models were trained and the distribution of their accuracies in each cross-validation fold are shown in Figs. 2, 3. Mean accuracies during training ranged from 0.81 to 0.93 and the respective comparison of models developed within each algorithm are detailed below.

Table 2 Demographic, clinicopathologic, and outcome information of 774 Hong Kong patients used for training and validation

Variables	First patient group (2003 – 2019) N = 716 N (%)	Second patient group (2020) N = 58 N (%)
Median age at diagnosis (IQR)	58 (49–67)	61.5 (53.8–68.3)
Gender		
Female	401 (56.0)	33 (56.9)
Male	315 (44.0)	25 (43.1)
Patient category		
NSND	469 (65.5)	41 (70.7)
SD	247 (34.5)	17 (29.3)
Continued risk habits following diagnosis		
Yes	14 (2.0)	11 (19.0)
No	167 (23.3)	1 (1.7)
Not applicable	535 (74.7)	46 (79.3)
Previous malignancy		
Head and neck tumors	21 (2.9)	0
Other tumors	46 (6.4)	3 (5.2)
Hematologic malignancies	23 (3.2)	6 (10.3)
No malignancy	626 (87.4)	49 (84.5)
Charlson comorbidity index–mean (SD)	0.72 (1.01)	0.64 (1.02)
Hypertension	22 (37.9)	211 (29.5)
Diabetes mellitus	9 (15.5)	111 (15.5)
Hyperlipidemia	21 (36.2)	122 (17.0)
Autoimmune disease	3 (5.2)	42 (5.9)
Viral hepatitis infection	3 (5.2)	69 (9.6)
Lesion		
Oral leukoplakia	389 (54.3)	41 (70.7)
Oral lichen planus/lichenoid lesion	327 (45.7)	17 (29.3)
Clinical subtype of lichenoid lesion		
Reticular/Papular	100 (14.0)	4 (6.9)
Erosive/Atrophic	142 (19.8)	6 (10.3)
Plaque	85 (11.9)	7 (12.1)
Tongue/FOM	245 (34.2)	25 (43.1)
Buccal/Labial mucosa	407 (56.8)	27 (46.6)
Retromolar area	26 (3.6)	3 (5.2)
Gingiva	88 (12.3)	2 (3.4)
Palate	23 (3.2)	3 (5.2)
Number of lesions		
Single	469 (65.5)	44 (75.9)
Bilateral or double	210 (29.3)	10 (17.2)
Multiple	37 (5.2)	4 (6.9)
Presence of ulcers or erosions	228 (31.8)	19 (32.8)
Induration	47 (6.6)	5 (8.6)
Treatment		
Surgical intervention	221 (30.9)	20 (34.5)
Pharmacological treatment	195 (27.2)	7 (12.1)
No treatment	300 (41.9)	31 (53.4)
Post-excision recurrence	42 (19.0)	2 (3.4)

Table 2 (continued)

Variables	First patient group (2003 – 2019) N = 716 N (%)	Second patient group (2020) N = 58 N (%)
Number of recurrences		
1	30 (13.5)	2 (3.4)
2	7 (3.2)	0
3	4 (1.8)	0
4	1 (0.5)	0
Oral epithelial dysplasia at diagnosis		
Absent	641 (89.5)	48 (82.8)
Mild	34 (4.7)	6 (10.3)
Moderate	27 (3.8)	0
Severe	7 (1.0)	0
Unknown (defaulted biopsy at diagnosis)	7 (1.0)	4 (6.9)
Oral epithelial dysplasia at follow-up		
Absent	658 (91.9)	48 (82.8)
Mild	11 (1.5)	0
Moderate	15 (2.1)	1 (1.7)
Severe	24 (3.4)	7 (12.1)
Unknown (defaulted biopsy during follow-up)	8 (1.1)	2 (3.4)
Malignant transformation	76 (10.6)	6 (10.3)
AJCC TNM stage		
Stage I	47 (6.6)	3 (5.2)
Stage II	9 (1.3)	2 (3.4)
Stage III	6 (0.8)	0
Stage IV	12 (1.7)	0
Tumor grade		
Well differentiated	23 (3.2)	NA
Moderate differentiated	30 (4.2)	
Poorly differentiated	3 (0.4)	
Tumor prognosis		
Remission	58 (8.1)	4 (6.9)
Recurrence	6 (0.8)	2 (3.4)
Cancer-related death	6 (0.8)	0
Second primary tumor	6 (0.8)	0

[a] *NA* not available

Logistic regression

The average accuracies of all models were similar during training, although, the models trained following oversampling with SMOTE were more stable with a lower coefficient of variation across the cross-validation folds. Upon validation, the models with balanced class weights performed better than those resampled with ADASYN or SMOTE based on their F1 scores (Tables 3, 4, 5). Nonetheless, the models with 26 features performed better than the 15-feature models, and the 26-parameter balanced class weight assignment model had the best accuracy, and F1 score of 0.92 and 0.60 respectively as well as a sensitivity of 0.75 following testing.

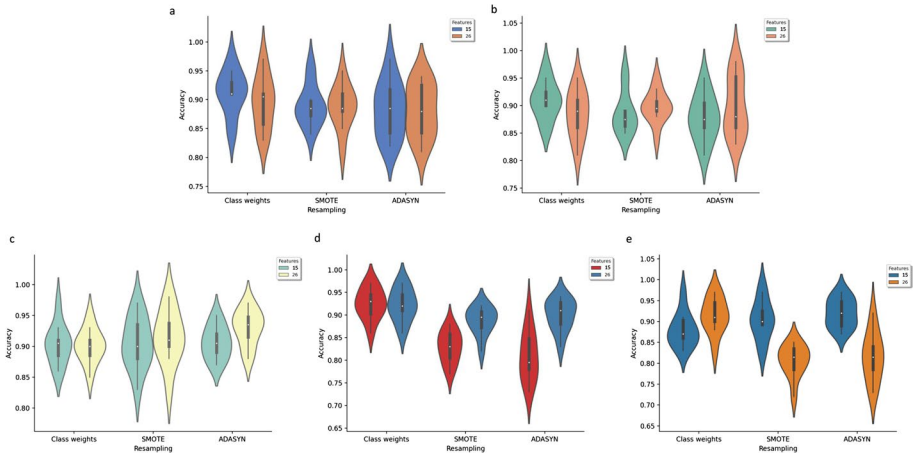

Fig. 2 Violin plots showing the distribution of accuracy estimates obtained across the cross-validation folds during model training of five algorithms using three class imbalance correction methods—(**a**) Logistic regression (**b**) Linear SVM (**c**) Radial basis function SVM (**d**) Random Forest (**e**) Decision tree

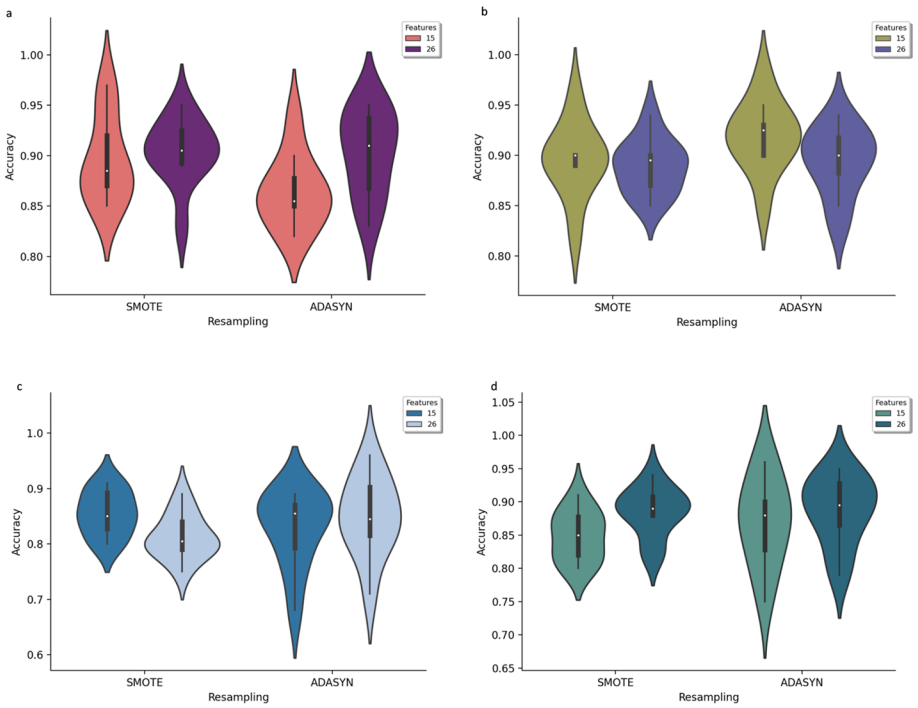

Fig. 3 Violin plots showing the distribution of accuracy estimates obtained across the cross-validation folds during model training of four algorithms using two synthetic class imbalance correction methods—(**a**) gradient boosting (**b**) k-nearest neighbor (**c**) multilayer perceptron (**d**) linear discriminant analysis

Linear SVM

Similar mean accuracies were obtained for all models with the 26-parameter SMOTE model being the most consistent across the ten cross-validation folds (A detailed description of the external validation cohorts is presented in Table 2 and Additional file 1: Table S2). The 26-parameter models had better accuracy and F1 scores than

Table 3 Performance of nine machine learning classifiers on the training and testing datasets using 26 input features and synthetic oversampling

Algorithms	Imbalanced class technique	SMOTE									ADASYN								
	Dataset	Training			Testing						Training			Testing					
	Performance measures	Mean accuracy	SD	Range	Accuracy	Sensitivity	Precision	F1-score	SP	NPV	Mean accuracy	SD	Range	Accuracy	Sensitivity	Precision	F1-score	SP	NPV
Logistic regression		0.89	0.036	0.81–0.95	0.88	0.75	0.39	0.51	0.89	0.98	0.88	0.043	0.81–0.93	0.92	0.67	0.53	0.59	0.95	0.97
Linear SVM		0.90	0.027	0.84–0.95	0.87	0.75	0.38	0.50	0.89	0.97	0.90	0.051	0.83–0.98	0.95	0.67	0.73	0.70	0.98	0.97
RBF-Kernel SVM		0.92	0.041	0.83–0.98	0.92	0.50	0.55	0.52	0.96	0.95	0.93	0.027	0.88–0.97	0.92	0.33	0.57	0.42	0.98	0.94
Random forest		0.89	0.029	0.82–0.92	0.87	0.67	0.35	0.46	0.89	0.97	0.90	0.033	0.83–0.94	0.91	0.67	0.47	0.55	0.93	0.97
Decision tree		0.81	0.038	0.72–0.85	0.71	0.75	0.19	0.31	0.71	0.97	0.82	0.056	0.73–0.92	**0.95**	**0.75**	**0.69**	**0.72**	**0.97**	**0.98**
Gradient boosting		0.91	0.030	0.83–0.95	0.90	0.75	0.43	0.56	0.91	0.98	0.90	0.04	0.83–0.95	0.95	0.67	0.73	0.70	0.98	0.97
kNN		0.89	0.025	0.85–0.94	0.87	0.42	0.29	0.35	0.91	0.94	0.90	0.032	0.83–0.94	0.83	0.42	0.23	0.29	0.87	0.94
MLP-BP		0.82	0.039	0.75–0.89	**0.94**	**0.75**	**0.60**	**0.67**	**0.95**	**0.98**	0.85	0.066	0.71–0.96	0.76	0.21	0.21	0.32	0.77	0.96
LDA		0.89	0.034	0.82–0.94	0.87	0.67	0.35	0.46	0.89	0.97	0.89	0.049	0.79–0.95	0.93	0.58	0.58	0.58	0.96	0.96

SMOTE synthetic minority oversampling technique, *ADASYN* adaptive synthetic sampling, *SD* standard deviation, *SP* specificity, *NPV* negative predictive value, *SVM* support vector machines, *RBF* radial basis function, *kNN* *k*-nearest neighbor, *MLP-BP* multilayer perceptron with backpropagation, *LDA* inear discriminant analysis

Values in *bold* represent the best-performing algorithm in each group

Table 4 Performance of nine machine learning classifiers on the training and testing datasets using 15 input features and synthetic oversampling

Algorithms	Imbalanced class technique	SMOTE									ADASYN								
	Dataset	Training			Testing						Training			Testing					
	Performance measures	Mean accuracy	SD	Range	Accuracy	Sensitivity	Precision	F1-score	SP	NPV	Mean accuracy	SD	Range	Accuracy	Sensitivity	Precision	F1-score	SP	NPV
Logistic regression		0.89	0.033	0.84–0.96	0.84	0.67	0.30	0.41	0.85	0.97	0.88	0.047	0.82–0.97	0.91	0.67	0.47	0.55	0.93	0.97
Linear SVM		0.89	0.036	0.85–0.96	0.82	0.83	0.29	0.44	0.82	0.98	0.88	0.04	0.81–0.95	0.94	0.67	0.62	0.64	0.96	0.97
RBF-Kernel SVM		0.91	0.039	0.83–0.97	0.87	0.50	0.33	0.40	0.91	0.95	0.91	0.025	0.87–0.95	0.90	0.42	0.42	0.42	0.95	0.95
Random forest		0.83	0.034	0.77–0.88	**0.96**	**0.58**	**0.88**	**0.70**	**0.99**	**0.96**	0.81	0.055	0.75–0.91	0.88	0.67	0.38	0.49	0.90	0.97
Decision tree		0.91	0.045	0.83–0.98	0.91	0.50	0.46	0.48	0.95	0.95	0.92	0.032	0.87–0.97	0.90	0.67	0.42	0.52	0.92	0.97
Gradient boosting		0.90	0.040	0.85–0.97	0.86	0.83	0.36	0.50	0.86	0.98	0.87	0.035	0.82–0.94	**0.94**	**0.75**	**0.64**	**0.69**	**0.96**	**0.98**
kNN		0.90	0.035	0.82–0.96	0.87	0.42	0.29	0.35	0.91	0.94	0.92	0.032	0.85–0.98	0.89	0.33	0.33	0.33	0.94	0.94
MLP-BP		0.86	0.038	0.80–0.91	0.90	0.75	0.43	0.55	0.91	0.98	0.83	0.067	0.68–0.89	0.90	0.75	0.42	0.55	0.91	0.98
LDA		0.85	0.036	0.80–0.91	0.83	0.67	0.28	0.39	0.84	0.96	0.87	0.066	0.75–0.96	0.89	0.58	0.39	0.47	0.92	0.96

SMOTE synthetic minority oversampling technique, *ADASYN* adaptive synthetic sampling, *SD* standard deviation, *SP* specificity, *NPV* negative predictive value, *SVM* support vector machines, *RBF*—radial basis function, *kNN* k-nearest neighbor, *MLP-BP* multilayer perceptron with backpropagation, *LDA* linear discriminant analysis

Values in *bold* represent the best-performing algorithm in each group

Table 5 Performance of parametric and tree-based classifiers on the training and testing datasets using class weights

Algorithms	Number of features	26 features									15 features								
	Dataset	Training			Testing						Training			Testing					
	Performance measures	Mean accuracy	SD	Range	Accuracy	Sensitivity	Precision	F1-score	SP	NPV	Mean accuracy	SD	Range	Accuracy	Sensitivity	Precision	F1-score	SP	NPV
Logistic regression		0.89	0.044	0.83–0.95	0.92	0.75	0.50	0.60	0.93	0.98	0.91	0.035	0.84–0.97	0.92	0.75	0.53	0.62	0.94	0.98
Linear SVM		0.89	0.040	0.81–0.95	0.93	0.75	0.56	0.64	0.95	0.98	0.91	0.032	0.86–0.97	**0.94**	**0.67**	**0.67**	**0.67**	**0.97**	**0.97**
RBF-Kernel SVM		0.90	0.028	0.85–0.95	0.92	0.33	0.50	0.40	0.97	0.94	0.90	0.030	0.86–0.97	0.94	0.33	0.80	0.47	0.99	0.94
Random forest		0.92	0.032	0.86–0.97	**0.97**	**0.75**	**0.90**	**0.81**	**0.99**	**0.98**	0.92	0.032	0.86–0.97	0.94	0.42	0.83	0.56	0.99	0.95
Decision tree		0.91	0.040	0.83–0.97	0.95	0.75	0.69	0.72	0.97	0.98	0.88	0.038	0.83–0.97	0.92	0.42	0.56	0.48	0.97	0.95

SD standard deviation, *SP* specificity, *NPV* negative predictive value, *SVM* support vector machines, *RBF*—radial basis function

Values in *bold* represent the best-performing algorithm in each group

models trained with 15 features, and the 26-parameter ADASYN model had the best accuracy and F1 score of 0.95 and 0.70, and a sensitivity of 0.67 following testing.

RBF-Kernel SVM

All six models had excellent average accuracy on training, however, the 15-parameter ADASYN model was the most stable (Additional file 1: Table S2). Models with 26 features were slightly better than those with 15 features and lower performances were observed irrespective of the class imbalance technique implemented (Tables 3, 4, 5). Overall, the 26-parameter SMOTE model had the best accuracy and F1-score of 0.92 and 0.52 upon testing; although, the sensitivity of this model on the test set was 0.50.

Random forest

Irrespective of the class imbalance resampling used, the 26-feature model had equal or better mean training accuracy and was more stable than the models with 15 features (Tables 2, 3, 4). However, the 26-parameter balanced class weight model outperformed all other algorithms on the test dataset with accuracy, F1 score, and sensitivity of 0.97, 0.81, and 0.75 respectively.

Decision tree

The mean accuracies of the 15-feature models were higher than the 26-feature models for models trained with synthetic oversampling (Tables 3, 4). However, the 26-feature model had better mean training accuracy when class weight assignment was used (Table 5). The 15-feature ADASYN model was the most stable across the validation folds (Additional file 1: Table S2). Following testing, the 26-parameter models based on ADASYN and class weight distribution both outperformed other models with similar accuracy, F1 score, and sensitivity of 0.95, 0.72, and 0.75 respectively.

Gradient boosting

Mean accuracies on training followed the same trend based on the type of imbalance technique, although, the accuracy of the 26-parameter models was higher and stable generally (Tables 3, 4). Upon evaluation with the test dataset, the 26-parameter model with ADASYN technique had the best accuracy and F1 score of 0.95 and 0.70 with a sensitivity of 0.67.

kNN

While the 15-parameter ADASYN had the highest mean accuracy for the models during training, the 26-parameter SMOTE model was most stable across the folds (Additional file 1: Table S2). Models with 26 and 15 features based on SMOTE-resampled training data had similar accuracy, F1 score, and sensitivity of 0.87, 0.35, and 0.42 which was better than both ADASYN-resampled models upon evaluation with the test dataset (Tables 3, 4).

MLP-BP

The 15-parameter SMOTE model had the highest accuracy on the training dataset and the most stable performance estimates across the cross-validation folds (Tables 3,

4). On evaluation with the test dataset, the 26-parameter SMOTE model had the best accuracy, F1 score, and sensitivity of 0.94, 0.67, and 0.75.

Linear discriminant analysis

Irrespective of the class imbalance method, the 26-parameter model had a higher mean accuracy than the 15-parameter model on the training dataset and the SMOTE-resampled models were generally more stable (Tables 3, 4). The 26-parameter ADASYN model performed best with an accuracy, F1 score, and sensitivity of 0.93, 0.58, and 0.58 respectively following model testing.

Comparison of model performance upon testing

Testing performances of these trained models are also presented in Tables 3, 4, 5. For all algorithms, the 26-parameter SMOTE models had better stability across the cross-validation folds (Additional file 1: Table S2). However, the 15-parameter RBF-kernel SVM model based on ADASYN resampling had the lowest coefficient of variation and best stability of any model upon comparing the mean training accuracies. Also, according to the F1 scores, for models using synthetic class imbalance correction, those based on ADASYN had better mean F1 scores than those based on SMOTE (Additional file 1: Table S3). Likewise, irrespective of class imbalance, the 26-parameter models had better mean F1 scores than the 15-parameter models.

Overall, for the models with 26 predictors, the weighted random forest model outperformed other models with an F1 score, sensitivity, precision, and accuracy of 0.81, 0.75, 0.90, and 0.97 respectively (Table 5). Also, this model was better than all others irrespective of the number of variables or class resampling technique. For the 15-parameter models, both random forest with SMOTE (F1-score: 0.70, sensitivity: 0.58, accuracy: 0.96) and gradient boosting with ADASYN (F1-score: 0.69, sensitivity: 0.75, accuracy:

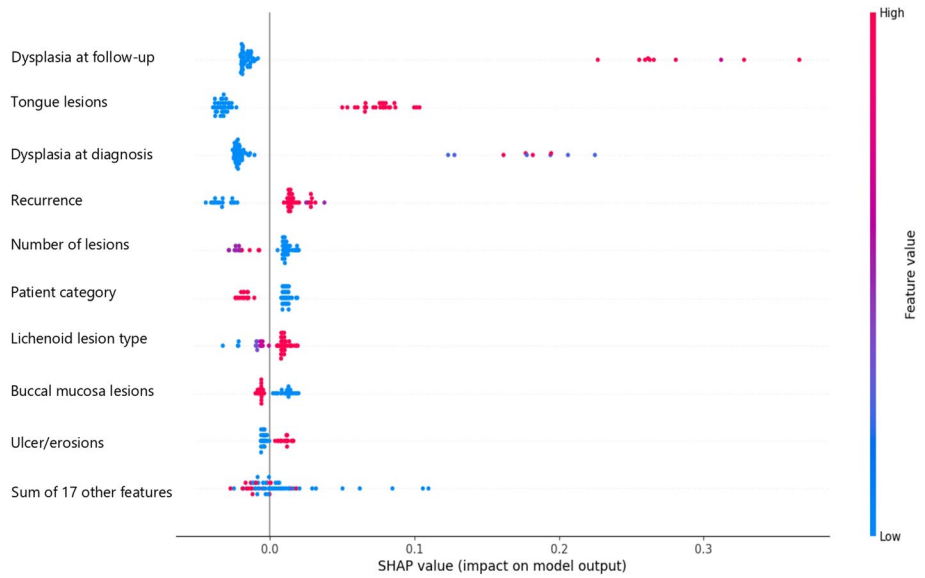

Fig. 4 Global SHAP summary plot to explain the importance of the features to the model predictions of the weighted random forest model (with 26 input parameters)

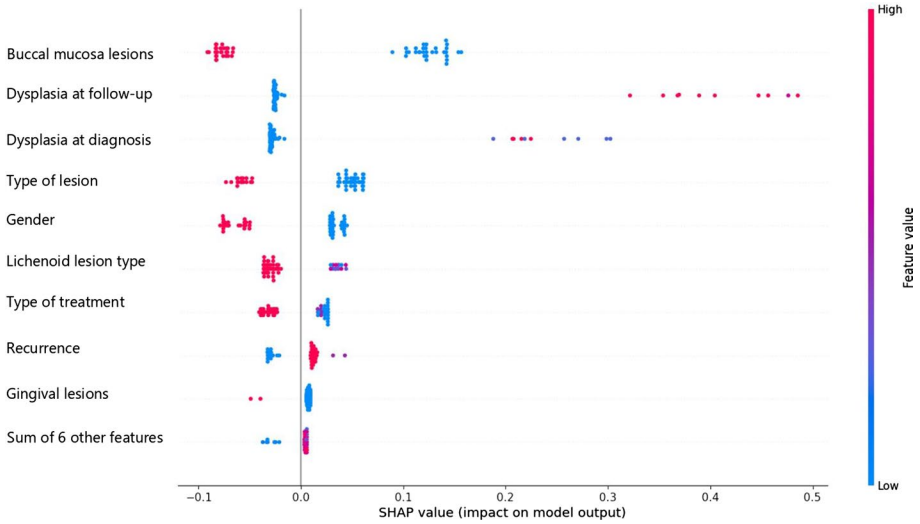

Fig. 5 Global SHAP summary plot to explain the importance of the features to the model predictions of the gradient boosting-ADASYN model (with 15 input parameters)

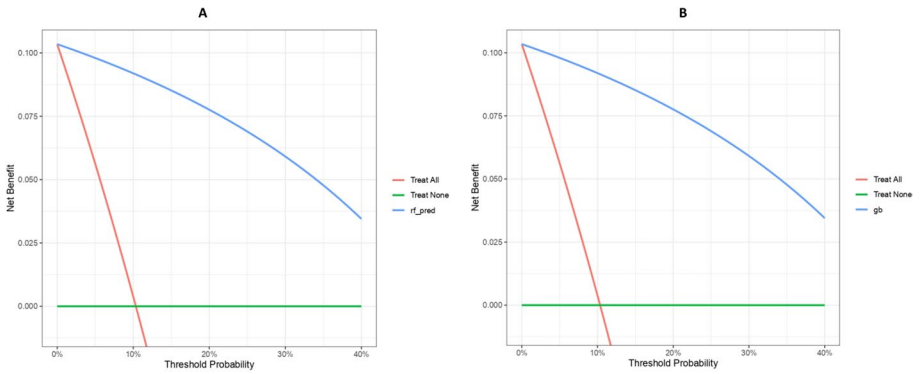

Fig. 6 Decision curves indicating the net benefits of the weighted random forest model (**A**) and gradient boosting-ADASYN model (**B**) in selecting patients for intervention and close monitoring

0.94) had similar performances that bested other models (Table 4). However, based on the prioritization of sensitivity over precision for models with similar F1 scores, the gradient boosting with ADASYN resampling had the best performance for the 15 feature models in this study.

External validation of the two best models using prospective Hong Kong patients

We assessed the performance of the two outperforming models (weighted random forest for 26 predictors and gradient boosting-ADASYN model for 15 variables) using 58 consecutive Hong Kong patients with oral leukoplakia and oral lichenoid mucositis. For all patients, both models had perfect sensitivity (1.00) for predicting malignant transformation as of the last date of follow-up (December 31, 2021) and both displayed satisfactory yet comparable predictive accuracies. Both risk models for oral leukoplakia and oral lichenoid mucositis achieved similar AUC, accuracy, specificity, and F1 score of 0.94, 0.90, 0.88, and 0.67 respectively.

Explainability using global SHAP values for both 26 and 15 feature models is presented in Figs. 4, 5. For the weighted random forest model, epithelial dysplasia at diagnosis and follow-up, tongue lesions, number of lesions, and occurrence of recurrent lesions were the five most important features contributing to the predicted output. Likewise, for the gradient boosting-ADASYN model, buccal mucosa lesions, epithelial dysplasia at diagnosis and follow-up, type of lesions (leukoplakia vs lichenoid mucositis), and the patient's gender were covariates that were pertinent to predicted outputs in the external validation cohort. Decision curve analysis to determine the net benefit of the predicted outputs of the models is plotted in Fig. 6. Overall, both models offered higher benefits when used for risk stratification at different threshold probabilities than if surgical intervention is performed in all patients in the external validation cohort.

External validation of the 15-feature model using patients from different settings

Since the malignant transformation prediction model based on 15 variables was specifically developed to be used in other areas according to the ease of data availability, we further assessed its generalizability using patients managed in Newcastle, UK (high-risk cohort) and Lagos, Nigeria for which the same variables were available. The 15-parameter gradient boosting-ADASYN risk model achieved a higher sensitivity for identifying patients with malignant transformation in these cohorts than the binary epithelial dysplasia grading system (0.96 vs 0.82). However, the difference in sensitivity values between the ML model and dysplasia grading did not achieve statistical significance ($p = 0.250$).

Web deployment

The data frame web-based deployment of both risk models according to 26 predictors (weighted random forest) and 15 predictors (gradient boosting-ADASYN can be found at https://wrf-26.herokuapp.com and https://mtp-gb-a-15.herokuapp.com. The templates for input parameters and coding of variables have been included in Additional file 2 (for 26 features) and Additional file 3 (for 15 features). Predicted outputs are given as high risk (1) and low risk (0) for malignant transformation. Further, the interactive web applications of the models to enable day-to-day validation are available at https://opmd-predict-facdent-hku.herokuapp.com and https://opmd-predict-facdent-hku-26.herokuapp.com which also includes the predicted class probabilities.

Discussion

Predicting the malignant transformation of oral leukoplakia and oral lichenoid mucositis often pose a conundrum to clinicians especially in non-specialist centers and resource-limited settings [10]. Adjunctive decision-making and risk stratification platforms that could predict cancer occurrence in OPMDs with good sensitivity and precision would prove to be substantive assets in contemporary clinical practice. Machine learning represents one of the most advanced and sophisticated methods for developing such tools leading to the objective of this study to compare different models and select the most promising ones for further optimization.

Generally, this study found that machine learning models with 26 variables performed better than those with 15 variables in sensitivity and precision. This alludes to the ability of clinical factors such as risk factor category, clinical history of comorbidities, viral

hepatitis status, number of lesions, and presence of induration which were included in the 26-feature models to better delineate high and low-risks patients with oral leukoplakia and oral lichenoid mucositis when they are available. Notably, the performance of the machine learning models was found to vary with the method of class imbalance correction as those employing ADASYN obtained better F1-scores than SMOTE irrespective of the number of variables or the type of machine learning algorithm employed. This supports the notion of a potential increase in model performance that may be offered by ADASYN over SMOTE since it focuses on instances that are difficult to classify [50]. Nonetheless, we maintain that this is likely to vary with the type of machine learning classifier used [51].

This study developed two models based random forest and gradient boosting algorithms which were the outperformers to predict the malignant transformation of oral leukoplakia and oral lichenoid mucositis. In this study, these models were selected for achieving a satisfactory sensitivity (all > 75%) for patients with high risk of malignancy in the different datasets used during internal and external validation irrespective of the number of predictors available for the task. Random forest and gradient boosting are robust ensemble learning algorithms that seek to reduce the generalization error of predictions by considering decisions from different weak learners resulting in better overall performance [52, 53]. The performance of these algorithms for producing better predictive models for oral cavity cancer outcomes (such as locoregional recurrence) has been previously reported by our group and others which supports the findings of this study [23, 39, 54–56].

To date, three models for predicting the malignant transformation of OPMDs have been proposed [55, 57, 58] none of which have resulted from a comprehensive model comparison experiments, internal/external validation, or utilized predictive features available at different point-of-care centers [57]. Comparing all tools proposed, the best-performing models in this study had better sensitivity on internal and even external validation. Furthermore, we obtained better overall performance for the developed models in this study when compared to the few nomograms proposed to stratify the risk of cancer occurrence in patients with OPMDs [21, 22]. One of such nomograms (Newcastle nomogram) was developed with a similar dataset with that used for external validation in this study and our models displayed better sensitivity for identifying patients with high risk of malignant transformation [22]. Also, the outperforming machine learning models had better sensitivity than reported for the WHO dysplasia grading system or binary epithelial dysplasia grading system in predicting the risk of malignant transformation in different OPMDs [16, 17, 42]

Limitations

The main limitation in the implementation of the models developed and validated in this study is their moderate precision. However, the high sensitivity achieved on external validation indicates that they are indeed better suited to risk stratification for better clinical judgment currently than the incontrovertible prediction that some white lesions will develop malignancies. To improve the precision, we suggest the inclusion of other cytologic or molecular-based features which may further indicate malignant outcomes among clinically high-risk patients with high-grade dysplasia in future endeavors [18,

59, 60]. Furthermore, the models were developed using a retrospective dataset that was not purposeful for training the algorithms. However, the data was collected from electronic records linked across different specialties and clinics involved in the management of oral precancer and correlated across multiple platforms and visits to ensure correct imputation. Likewise, our generalizability assessments of these models are only incipient and solely to assess the potential utility of these models in cases from other centers. Hopefully, the availability of web platforms based on the malignant transformation risk models constructed in this study will encourage further external validation and impact assessment in other populations using randomized controlled trial designs [61] before routine implementation at various general and oncological practices in specialist and remote centers.

Conclusions

Overall, this study found that machine learning-based models have satisfactory sensitivity and accuracy in identifying patients with oral leukoplakia and oral lichenoid mucositis that are at risk of malignant transformation. Tree-based algorithms (gradient boosting and random forest) performed best and were used to develop two promising models based on 15 and 26 features respectively. Although not statistically significant, the 15-parameter gradient boosting model had a higher sensitivity than the binary epithelial dysplasia grading system. Furthermore, this study provided proof-of-concept that demographic, clinicopathological, and treatment information obtainable from electronic health records are useful for predicting the risk of malignant transformation of oral leukoplakia and oral lichenoid mucositis as a binary outcome. The models were deployed as web-based platforms to encourage further external validation and serve as tools for future research to determine their impact in the management of patients with oral leukoplakia and oral lichenoid mucositis.

Abbreviations

ADASYN	Adaptive synthetic sampling
kNN	K-nearest neighbors
MLP-BP	Multilayer perceptron and back propagation
OPMDs	Oral potentially malignant disorders
RBF	Radial basis function
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machines

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-023-00714-7.

Additional file 1: Table S1. Demographic, clinicopathologic, and outcome information of 413 patients for external validation. **Table S2.** Coefficient of variation for the models to assess their stability around the average accuracy. **Table S3.** Mean F1 scores for trained models.

Additional file 2: Format of the CSV file to be uploaded to the 26-feature model for generating malignant transformation risk status.

Additional file 3: Format of the CSV file to be uploaded to the 15-feature model for generating malignant transformation risk status.

Acknowledgements
Not applicable.

Author contributions
JA, PT, and Y-XS conceived the study. AWIL, VLYC, RK-YT, AA, PT, L-WZ, and Y-XS were involved in the diagnosis and management of patient cohorts. JA, AWIL, and AA performed data curation. JA and MK-M performed the machine learning experiments and deployment of web-based tools. S-WC, L-WZ, PT, and Y-XS provided supervision and validation. JA wrote the original draft of the manuscript. JA, MK-M, AWIL, VLYC, RK-YT, AA, L-WZ, PT, and Y-XS critically revised the manuscript. All authors read and approved the final manuscript.

Funding
No external funding.

Availability of data and materials
The datasets generated and/or analysed during the current study are not publicly available due to the need to maintain patient confidentiality but are available from the corresponding authors on reasonable request. Deployment of outperforming models was done using Flask module on python software and codes are available in these GitHub repos: https://github.com/jaadeoye/interactive_mtp_15 and https://github.com/jaadeoye/interactive-mtp-26.

Declarations

Ethics approval and consent to participate
Approval to conduct this study was granted by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (Reference Number UW-21-495). Informed consent for patients was waved due to the retrospective nature of this study. All clinical data were anonymized by the researchers, and all potential patient identifiers were removed before data analysis.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 17 May 2022 Accepted: 13 March 2023
Published online: 30 March 2023

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. Oral Oncol. 2009;45(4–5):309–16.
3. Adeoye J, Thomson P. Strategies to improve diagnosis and risk assessment for oral cancer patients. Fac Dental J. 2020;11(3):122–7.
4. Ram H, Sarkar J, Kumar H, Konwar R, Bhatt MLB, Mohammad S. Oral cancer: risk factors and molecular pathogenesis. J Maxill Oral Surg. 2011;10(2):132–7.
5. Adeoye J, Tan JY, Ip CM, Choi S-W, Thomson P. "Fact or fiction?": Oral cavity cancer in nonsmoking, nonalcohol drinking patients as a distinct entity—scoping review. Head Neck. 2021;43:3662.
6. Speight PM, Morgan PR. The natural history and pathology of oral cancer and precancer. Community Dent Health. 1993;10(Suppl 1):31–41.
7. Warnakulasuriya S, Johnson NW, van der Waal I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. J Oral Pathol Med. 2007;36(10):575–80.
8. Warnakulasuriya S, Kujan O, Aguirre-Urizar JM, Bagan JV, González-Moles M, Kerr AR, et al. Oral potentially malignant disorders: a consensus report from an international seminar on nomenclature and classification, convened by the WHO collaborating centre for oral cancer. Oral Dis. 2021;27(8):1862–80.
9. Shearston K, Fateh B, Tai S, Hove D, Farah CS. Oral lichenoid dysplasia and not oral lichen planus undergoes malignant transformation at high rates. J Oral Pathol Med. 2019;48(7):538–45.
10. van der Waal I. Oral potentially malignant disorders: is malignant transformation predictable and preventable? Med Oral Patol Oral Cir Bucal. 2014;19(4):e386–90.
11. Speight PM, Khurram SA, Kujan O. Oral potentially malignant disorders: risk of progression to malignancy. Oral Surg Oral Med Oral Pathol Oral Radiol. 2018;125(6):612–27.
12. Iocca O, Sollecito TP, Alawi F, Weinstein GS, Newman JG, De Virgilio A, et al. Potentially malignant disorders of the oral cavity and oral dysplasia: a systematic review and meta-analysis of malignant transformation rate by subtype. Head Neck. 2020;42(3):539–55.
13. Aguirre-Urizar JM, Lafuente-Ibáñez de Mendoza I, Warnakulasuriya S. Malignant transformation of oral leukoplakia: systematic review and meta-analysis of the last 5 years. Oral Diseases. 2021;27(8):1881–95.
14. Warnakulasuriya S, Ariyawardana A. Malignant transformation of oral leukoplakia: a systematic review of observational studies. J Oral Pathol Med. 2016;45(3):155–66.
15. González-Moles M, Ramos-García P, Warnakulasuriya S. An appraisal of highest quality studies reporting malignant transformation of oral lichen planus based on a systematic review. Oral Dis. 2021;27(8):1908–18.
16. Sathasivam HP, Sloan P, Thomson PJ, Robinson M. The clinical utility of contemporary oral epithelial dysplasia grading systems. J Oral Pathol Med. 2021. https://doi.org/10.1111/jop.13262.

17. de Freitas Silva BS, Batista DCR, de Souza Roriz CF, Silva LR, Normando AGC, dos Santos Silva AR, et al. Binary and WHO dysplasia grading systems for the prediction of malignant transformation of oral leukoplakia and erythroplakia: a systematic review and meta-analysis. Clin Oral Invest. 2021;25(7):4329–40.
18. Sathasivam HP, Kist R, Sloan P, Thomson P, Nugent M, Alexander J, et al. Predicting the clinical outcome of oral potentially malignant disorders using transcriptomic-based molecular pathology. Br J Cancer. 2021;125(3):413–21.
19. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. Lancet Oncol. 2015;16(4):e173–80.
20. Chen F, Lin L, Yan L, Liu F, Qiu Y, Wang J, et al. Nomograms and risk scores for predicting the risk of oral cancer in different sexes: a large-scale case-control study. J Cancer. 2018;9(14):2543–8.
21. Wang T, Wang L, Yang H, Lu H, Zhang J, Li N, et al. Development and validation of nomogram for prediction of malignant transformation in oral leukoplakia: a large-scale cohort study. J Oral Pathol Med. 2019;48(6):491–8.
22. Goodson ML, Smith DR, Thomson PJ. The"newcastle nomogram"-statistical modelling predicts malignant transformation in potentially malignant disorders. J Oral Pathol Med. 2019;48(8):662–8.
23. Adeoye J, Tan JY, Choi S-W, Thomson P. Prediction models applying machine learning to oral cavity cancer outcomes: a systematic review. Int J Med Informatics. 2021;154: 104557.
24. Alabi RO, Youssef O, Pirinen M, Elmusrati M, Mäkitie AA, Leivo I, et al. Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future—a systematic review. Artif Intell Med. 2021;115: 102060.
25. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8.
26. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
27. Aguirre-Urizar JM, Lafuente-Ibáñez de Mendoza I, Warnakulasuriya S. Malignant transformation of oral leukoplakia: systematic review and meta-analysis of the last 5 years. Oral Dis. 2021;27(8):1881–95.
28. Ramos-García P, González-Moles M, Warnakulasuriya S. Oral cancer development in lichen planus and related conditions-3.0 evidence level: a systematic review of systematic reviews. Oral Dis. 2021;27(8):1919–35.
29. Adeoye J, Koohi-Moghadam M, Lo AWI, Tsang RK-Y, Chow VLY, Zheng L-W, et al. Deep learning predicts the malignant-transformation-free survival of oral potentially malignant disorders. Cancers. 2021;13(23):6054.
30. Thomson PJ, Goodson ML, Smith DR. Profiling cancer risk in oral potentially malignant disorders—a patient cohort study. J Oral Pathol Med. 2017;46(10):888–95.
31. Thomson PJ, Goodson ML, Smith DR. Potentially malignant disorders revisited-The lichenoid lesion/proliferative verrucous leukoplakia conundrum. J Oral Pathol Med. 2018;47(6):557–65.
32. Thomson PJ, Goodson ML, Cocks K, Turner JE. Interventional laser surgery for oral potentially malignant disorders: a longitudinal patient cohort study. Int J Oral Maxillofac Surg. 2017;46(3):337–42.
33. Warnakulasuriya S, Ariyawardana A. Malignant transformation of oral leukoplakia: a systematic review of observational studies. J Oral Pathol Med. 2016;45(3):155–66.
34. Idrees M, Kujan O, Shearston K, Farah CS. Oral lichen planus has a very low malignant transformation rate: a systematic review and meta-analysis using strict diagnostic and inclusion criteria. J Oral Pathol Med. 2021;50(3):287–98.
35. Giuliani M, Troiano G, Cordaro M, Corsalini M, Gioco G, Lo Muzio L, et al. Rate of malignant transformation of oral lichen planus: a systematic review. Oral Dis. 2019;25(3):693–709.
36. Lodi G, Franchini R, Warnakulasuriya S, Varoni EM, Sardella A, Kerr AR, et al. Interventions for treating oral leukoplakia to prevent oral cancer. Cochrane Database Syst Rev. 2016;7(7):CD001829.
37. Chaturvedi AK, Udaltsova N, Engels EA, Katzel JA, Yanik EL, Katki HA, et al. Oral leukoplakia and risk of progression to oral cancer: a population-based cohort study. J Natl Cancer Inst. 2020;112(10):1047–54.
38. Adeoye J, Thomson P. A call for an established oral cancer classification by etiology and revision of related terminology. Oral Diseases. 2021. https://doi.org/10.1111/odi.13784.
39. Chu CS, Lee NP, Adeoye J, Thomson P, Choi S-W. Machine learning and treatment outcome prediction for oral cancer. J Oral Pathol Med. 2020;49(10):977–85.
40. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst. 2007;12:95–116.
41. Gnip P, Vokorokos L, Drotár P. Selective oversampling approach for strongly imbalanced data. PeerJ Comput Sci. 2021. https://doi.org/10.7717/peerj-cs.604.
42. Kujan O, Oliver RJ, Khattab A, Roberts SA, Thakker N, Sloan P. Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. Oral Oncol. 2006;42(10):987–93.
43. Khoury ZH, Sultan M, Sultan AS. Oral epithelial dysplasia grading systems: a systematic review & meta-analysis. Int J Surg Pathol. 2022;30:499.
44. Jäwert F, Nyman J, Olsson E, Adok C, Helmersson M, Öhman J. Regular clinical follow-up of oral potentially malignant disorders results in improved survival for patients who develop oral cancer. Oral Oncol. 2021;121: 105469.
45. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017. 30.
46. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
47. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn machine learning in python. J Machine Learn Res. 2011;12:2825–30.
48. Grinberg M. Flask web development: developing web applications with python: "O'Reilly Media, Inc.". 2018.
49. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Cancer. 2015;112(2):251–9.
50. He H, Bai Y, Garcia EA, Li S, editors. ADASYN. Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence); 2008. IEEE.
51. Danquah RA. Handling imbalanced data: a case study for binary class problems. arXiv Preprint arXiv. 2020;2010:201004326.

52. Opitz D, Maclin R. Popular ensemble methods: An empirical study. J Artif Intell Res. 1999;11:169–98.
53. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. Front Comp Sci. 2020;14:241–58.
54. 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.
55. Wang X, Yang J, Wei C, Zhou G, Wu L, Gao Q, et al. A personalized computational model predicts cancer risk level of oral potentially malignant disorders and its web application for promotion of non-invasive screening. J Oral Pathol Med. 2020;49(5):417–26.
56. Adeoye J, Zheng L-W, Thomson P, Choi S-W, Su Y-X. Explainable ensemble learning model improves identification of candidates for oral cancer screening. Oral Oncol. 2023;136: 106278.
57. Liu Y, Li Y, Fu Y, Liu T, Liu X, Zhang X, et al. Quantitative prediction of oral cancer risk in patients with oral leukoplakia. Oncotarget. 2017;8(28):46057–64.
58. Shams WK, Htike ZZ. Oral cancer prediction using gene expression profiling and machine learning. Int J Appl Eng Res. 2017;12:4893–8.
59. Odell EW. Aneuploidy and loss of heterozygosity as risk markers for malignant transformation in oral mucosa. Oral Dis. 2021;27(8):1993–2007.
60. Monteiro L, Mello FW, Warnakulasuriya S. Tissue biomarkers for predicting the risk of oral cancer in patients diagnosed with oral leukoplakia: a systematic review. Oral Dis. 2021;27(8):1977–92.
61. Zhou Q, Chen Z-h, Cao Y-h, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. Digit Med. 2021;4(1):154.

Publisher's Note