

<https://doi.org/10.1038/s41746-025-01885-8>

Next-generation AI framework for comprehensive oral leukoplakia evaluation and management



JingWen Li^{1,5}, YaFang Zhou^{2,5}, MengJing Zhang², John Adeoye³, Jane JingYa Pu¹, MiMi Zhou⁴, ChuanXia Liu⁴, LiJie Fan⁴, Colman McGrath³, Dian Zhang²✉ & LiWu Zheng¹✉

Oral potentially malignant disorder poses a significant risk of malignant transformation, particularly in cases with epithelial dysplasia (OED). Current OED assessment methods are invasive and lack reliable decision-support tools for cancer risk evaluation and follow-up optimization. This study developed and validated OMMT-PredNet, a fully automated multimodal deep learning framework requiring no manual ROI annotation, for non-invasive OED identification and time-dependent cancer risk prediction. Utilizing data from 649 histopathologically confirmed leukoplakia cases across multiple institutions (2003–2024), including 598 cases in the primary cohort and 51 in the external validation set, the model integrated paired high-resolution clinical images and medical records. OMMT-PredNet achieved an AUC of 0.9592 (95% CI: 0.9491–0.9693) for cancer risk prediction and 0.9219 (95% CI: 0.9088–0.9349) for OED identification, with high specificity (MT: 0.9490; OED: 0.9182) and precision (MT: 0.9442; OED: 0.9303). Calibration and decision curve analyses confirmed clinical applicability, while external validation demonstrated robustness. This multidimensional model effectively predicts OED and cancer risk, highlighting its global applicability in enhancing oral cancer screening and improving patient outcomes.

Oral cancer is the most prevalent type of head and neck cancer worldwide, with a 5-year survival rate ranging from 86.3% during the localized stage to 39.3% at the distant stage^{1,2}. Early prediction, detection, and diagnosis of oral cancer are crucial for enabling clinicians to effectively manage the disease and implement timely interventions^{3,4}. The cancer progression features an extended pre-pathologic phase, spanning from initial exposure to risk factors to the onset of overt malignancy. This phase is often linked to the development of oral potentially malignant disorders (OPMDs), which carry an increased risk of malignant transformation (MT)⁵.

Within this collection of OPMD, oral leukoplakia (OLK) is widely recognized as the most frequently encountered condition in clinical practice. It is defined as “a predominantly white plaque of questionable risk having excluded (other) known diseases or disorders that carry no increased risk for cancer” by the World Health Organization^{6,7}. The overall MT rate exhibited considerable variation, spanning from 1.1% to 40.8%, while the annualized MT rate ranged between 1.56% and 4.9%^{8–10}. Notably, oral epithelial dysplasia (OED) present in 15% to 80% of OLK cases, significantly elevates the

risk of malignancy, with the MT rate rising sharply to 40% in dysplastic lesions^{11,12}. The identification and classification of OED in OLK are among the most important indicators for monitoring disease progression, guiding treatment decisions, and assessing the malignant potential¹³. However, the assessment of OED typically requires histopathological analysis through surgical biopsy. While current guidelines recommend clinical follow-up of OLK every 3–12 months, the requirement for repeated biopsies when monitoring suspicious changes—particularly in high-risk cases—may impact patient compliance, delay timely intervention, and increase health-care burdens. This highlights the necessity for alternative, less-invasive methods for assessing OED, meanwhile carrying out time-to-event malignant-transformation-free survival prediction to enhance patient compliance and refining clinical outcomes.

Artificial intelligence (AI) represents a cutting-edge technology for developing accurate, unbiased decision-making systems in detecting OPMD and oral cancer¹⁴. Machine learning (ML) and deep learning (DL) have been effectively applied to diverse data sources, including electronic

¹Division of Oral & Maxillofacial Surgery, Faculty of Dentistry, University of Hong Kong, Hong Kong SAR, China. ²Department of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. ³Division of Applied Oral Sciences & Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China. ⁴Stomatology Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. ⁵These authors contributed equally: JingWen Li, YaFang Zhou. ✉ e-mail: zhangd@szu.edu.cn; lwzheng@hku.hk

health records, histopathological images, clinical photographs, and molecular biomarkers^{15,16}. Recent meta-analyses indicate that among various clinical imaging modalities, clinical photography exhibits the highest diagnostic odds ratio¹⁷. By leveraging diverse algorithms, AI can adaptively extract essential low- and high-level features from images—such as contours, edges, shapes, colors, textures, and patterns—thereby eliminating clinician subjectivity, and more importantly, investigating key characteristics that are challenging to perceive with the naked eye¹⁸. Although well-established computer vision algorithms, such as standard CNNs and ResNet50, perform effectively in general image recognition, they exhibit limitations in medical applications. Their restricted adaptability to domain-specific features, dependence on extensive annotated data, and lack of interpretability hinder clinical utility. To address these challenges, targeted architectural innovations are essential for improving diagnostic reliability, practical applicability, and transparent decision-making in medical imaging analysis. Meanwhile, despite recent advancements in AI applications for healthcare, existing models remain constrained to a single-dimensional analysis of image or textual features and continue to provide single-task outputs focused solely on either disease identification or cancer prediction^{19,20}. This fundamental limitation significantly restricts their clinical applicability, as real-world diagnostic scenarios require comprehensive, multi-faceted evaluations. As we all know, cancer prognostication represents a multimodal issue influenced by various aspects, including demographic data, clinical manifestations, visual morphology, and histopathological evaluations. This critical gap has necessitated the development of multimodal AI frameworks that integrate diverse data sources to enhance diagnostic accuracy and clinical relevance. Moreover, the majority of models define the MT risk as a purely binary classification (likely or unlikely), rather than incorporating “real-world data” and generating patient outcomes as probabilities of transformation over time, which would further enhance clinical applicability²¹.

In this context, we propose a customized, fully automated end-to-end deep learning framework that operates without manual ROI annotation, named OLK multi-modal multi-task prediction network (OMMT-PredNet). This framework is specifically designed to achieve two key objectives related to OLK utilizing multicenter real-world data: the non-invasive identification of OED, as well as the prediction of cancer risk over time.

Results

Baseline characteristics

This study included a sample of 598 OLK patients (296 females, 303 males) over an average follow-up duration of 7.51 years, ranging from 1.6 to 21.57 years. The proportion of OLK cases exhibiting dysplasia in this cohort was 23.41%. The overall MT rate was 12.21%, with significantly higher incidence in females (15.20%) versus males (9.24%). OLK was particularly prevalent among the elderly, with 65.22% of patients aged over fifty. The MT rate for individuals aged 50 and above was notably elevated at 14.62% (57 cases), nearly double that of patients under 50, who had a rate of 7.69% (16 cases). Additionally, 104 patients (17.39%) reported a history of smoking and/or alcohol consumption, or continued these habits post-diagnosis. In terms of clinical subtypes, the prevalence of homogeneous and non-homogeneous cases was equal, with each comprising 299 cases. However, the MT rate for non-homogeneous lesions was strikingly higher, approximately 11 times that of homogeneous lesions, with 67 cases compared to just 6 cases. Furthermore, the most common lesion site was the buccal mucosa (272 cases, 45.48%), followed by the tongue (173 cases, 28.93%) and gingiva (98 cases, 16.39%). The baseline characteristics of the study population are detailed in Table 1.

Overall performance of the multimodal deep learning system on multi-tasks

The OMMT-PredNet system achieved an overall C-index of 0.7947 (95% CI = 0.7789, 0.8105), indicating strong predictive accuracy. To further assess the performance of our multi-modal deep learning system across various tasks, the receiver operating characteristic (ROC) curve analysis was

Table 1 | Baseline characteristics of the study population

Characteristics	Total n = 598	Patients with MT N (%) = 73 (12.2)	Patients without MT N (%) = 525 (87.8)	P value
Mean follow-up duration: 7.51 years				
<i>Gender</i>				0.018
Female	296 (49.5)	45 (7.5)	251 (42.0)	
Male	302 (50.5)	28 (4.7)	274 (45.8)	
<i>Age</i>				0.008
<50	208 (34.8)	16 (2.7)	192 (32.1)	
≥50	390 (65.2)	57 (9.5)	333 (55.7)	
<i>Alcohol</i>				0.318
Current or former	131 (21.9)	18 (3.0)	113 (18.9)	
Never	467 (78.1)	55 (9.2)	412 (68.9)	
<i>Tobacco</i>				<0.001
Current or former	104 (17.4)	31 (5.2)	73 (12.2)	
Never	494 (82.6)	42 (7.0)	452 (75.6)	
<i>Clinical sub-type</i>				<0.001
Homogeneous	299 (50.0)	6 (1.0)	293 (49.0)	
Non-homogeneous	299 (50.0)	67 (11.2)	232 (38.8)	
<i>Lesion site</i>				<0.001
Buccal mucosa	272 (45.5)	16 (2.7)	256 (42.8)	
Tongue	173 (28.9)	35 (5.9)	138 (23.1)	
Gingiva	98 (16.4)	10 (1.7)	88 (14.7)	
Floor of the mouth	23 (3.8)	9 (1.5)	14 (2.3)	
Palate	22 (3.7)	1 (0.2)	21 (3.5)	
Lip	10 (1.7)	2 (0.3)	8 (1.3)	
<i>Diameter</i>				0.043
<2 cm = 1	570 (95.3)	66 (11.0)	504 (84.3)	
≥2 cm = 2	28 (4.7)	7 (1.2)	21 (3.5)	
<i>Lesion size</i>				<0.001
<2 cm	482 (80.6)	47 (7.9)	435 (72.7)	
≥2 cm ² , <5 cm ²	109 (18.2)	21 (3.5)	88 (14.7)	
≥5 cm ²	7 (1.2)	5 (0.9)	2 (0.3)	
<i>Treatment</i>				<0.001
Drug only	461 (77.1)	45 (7.5)	416 (69.6)	
Photodynamic	30 (5.0)	2 (0.3)	8 (4.7)	
Excision	73 (12.2)	21 (3.5)	52 (8.7)	
No treatment	34 (5.7)	5 (0.9)	29 (4.8)	
<i>OED status</i>				<0.001
Class I	123 (20.6)	57 (9.5)	66 (11.1)	
Class II	17 (2.8)	15 (2.5)	2 (0.3)	
No OED	458 (76.6)	1 (0.2)	457 (76.4)	

conducted. The area under the curve (AUC) was calculated to evaluate the model's accuracy, yielding values of 0.9592 (95% CI = 0.9491, 0.9693) for time-to-event cancer risk prediction and 0.9219 (95% CI = 0.9088, 0.9349) for OED identification (Fig. 1A, B). These ROC curves demonstrate that the classifier has a robust ability to discriminate between positive and negative classes. Ablation studies comparing model performance with and without the CBAM module revealed that OMMT-PredNet demonstrates consistently superior performance across all evaluated metrics (Supplemental Note 1).

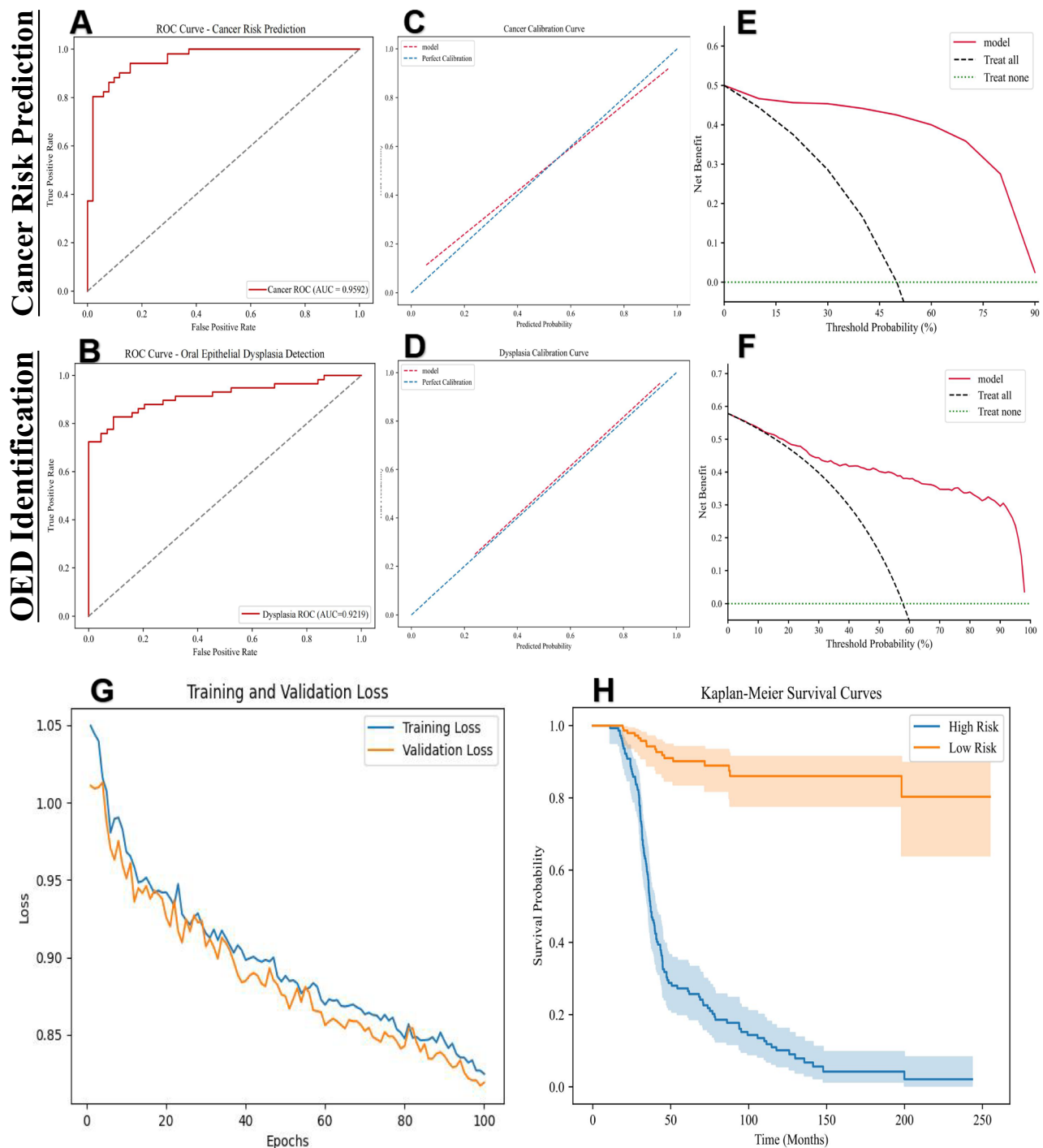


Fig. 1 | Performance evaluation of OMMT-PredNet model for OED identification and cancer risk assessment. **A** and **B** ROC curve for OED [AUC = 0.9219 (95% CI = 0.9088, 0.9349)] and malignant transformation [AUC = 0.9592 (95% CI = 0.9491, 0.9693)] prediction, illustrating the model's sensitivity and specificity at various thresholds. **C** and **D** Calibration plot for multi-tasks, showing the agreement between predicted probabilities and actual outcomes. The plots feature a diagonal line representing perfect calibration, where predicted probabilities match observed frequencies. Points above the line indicate overestimation of risk, while points below suggest underestimation. **E** and **F** Decision curve analysis (DCA) for multi-tasks, demonstrates the net benefits of using the model across different probability thresholds, comparing it

to the strategies of treating all patients or treating none. DCA revealed that OMMT-PredNet model offers a net benefit compared to both the treat-all and treat-none strategies within clinically relevant thresholds, highlighting its potential utility in clinical decision-making. **G** The training and validation loss curves. The training loss consistently declined from approximately 1.05 to 0.85 over 100 epochs, indicating effective learning. The validation loss showed a similar trend, with both losses closely aligned, suggesting good generalization and no overfitting. The stability of the loss curves also indicates effective convergence, demonstrating the model's efficacy and potential for reliable performance. **H** Kaplan-Meier analysis is performed to visualize patient stratification of low- and high-risk patients for individuals.

Calibration and decision curve analysis

The calibration curves indicated a strong concordance between predicted probabilities and observed outcomes, suggesting that the model reliably estimates the probabilities for both MT and OED (Fig. 1C, D). The Brier

scores were recorded at 0.0911 (95% CI = 0.0757, 0.1065) for MT prediction and 0.1080 (95% CI = 0.1018, 0.1142) for OED evaluation, both significantly below the random guessing threshold of 0.25. Furthermore, in terms of the model's net benefit, the results of the decision curve analysis (DCA)

Table 2 | Performance Metrics of OMMT-PredNet for Multi-Task Analysis in Oral Leukoplakia

Parameters	AUC	Brier score	Youden's index	F1 score	Precision	Balanced accuracy	Sensitivity	Specificity	PPV	NPV	
Multi-Task 1: Oral epithelial dysplasia identification	Training & Internal Validation (Hong Kong Cohort = 598)	0.9219 (95% CI = 0.9088, 0.9349)	0.1080 (95% CI = 0.1018, 0.1142)	0.7113 (95% CI = 0.6794, 0.7432)	0.8550 (95% CI = 0.8456, 0.8644)	0.9303 (95% CI = 0.8786, 0.9820)	0.8471 (95% CI = 0.8362, 0.8580)	0.7931 (95% CI = 0.7503, 0.8359)	0.9182 (95% CI = 0.8502, 0.9862)	0.9303 (95% CI = 0.8786, 0.9820)	0.7722 (95% CI = 0.7474, 0.7970)
	External Validation (ZheJiang Cohort = 51)	0.8786 (95% CI = 0.8471, 0.9101)	0.1379 (95% CI = 0.1189, 0.1569)	0.6571 (95% CI = 0.5829, 0.7313)	0.8089 (95% CI = 0.7646, 0.8532)	0.9218 (95% CI = 0.8185, 1.0251)	0.8286 (95% CI = 0.7915, 0.8657)	0.7286 (95% CI = 0.6314, 0.8258)	0.9286 (95% CI = 0.8200, 1.0372)	0.9218 (95% CI = 0.8185, 1.0251)	0.7777 (95% CI = 0.7224, 0.8330)
Multi-Task 2: Time-to-event cancer risk prediction	Training & Internal Validation (Hong Kong cohort = 598)	0.9592 (95% CI = 0.9491, 0.9693)	0.0911 (95% CI = 0.0757, 0.1065)	0.8039 (95% CI = 0.7584, 0.8494)	0.8972 (95% CI = 0.8741, 0.9203)	0.9442 (95% CI = 0.9094, 0.9790)	0.9020 (95% CI = 0.8792, 0.9248)	0.8549 (95% CI = 0.8331, 0.8767)	0.9490 (95% CI = 0.9163, 0.9817)	0.9442 (95% CI = 0.9094, 0.9790)	0.8674 (95% CI = 0.8481, 0.8867)
	External validation (ZheJiang cohort = 51)	0.9469 (95% CI = 0.9145, 0.9793)	0.1230 (95% CI = 0.0565, 0.1895)	0.6234 (95% CI = 0.2934, 0.9534)	0.7455 (95% CI = 0.7329, 0.7581)	0.7333 (95% CI = 0.5482, 0.9184)	0.8643 (95% CI = 0.8445, 0.8841)	0.8000 (95% CI = 0.6413, 0.9587)	0.8857 (95% CI = 0.8064, 0.9650)	0.7333 (95% CI = 0.5482, 0.9184)	0.9329 (95% CI = 0.8927, 0.9731)

OMMT-PredNet C-index = 0.7947 (95% CI = 0.7789, 0.8105).

demonstrated that our OMMT-PredNet model offers a net benefit compared to both the treat-all and treat-none strategies within clinically relevant thresholds, highlighting its potential utility in clinical decision-making (Fig. 1E, F).

Model stability

Moreover, Fig. 1G illustrates the training and validation loss curves. The training loss consistently declined from approximately 1.05 to 0.85 over 100 epochs, indicating effective learning. The validation loss showed a similar trend, with both losses closely aligned, suggesting good generalization and no overfitting. The stability of the loss curves also indicates effective convergence, demonstrating the model's efficacy and potential for reliable performance. Besides, a set of metrics was also calculated to further evaluate the predictive performance of the multi-task model. The precision for MT and OED prediction was 0.9442 (95% CI = 0.9094, 0.9790) and 0.9303 (95% CI = 0.8786, 0.9820), respectively. The specificity for these tasks was 0.9490 (95% CI = 0.9163, 0.9817) and 0.9182 (95% CI = 0.8502, 0.9862). Likewise, the balanced accuracy, F1 score, PPV, and NPV were 0.9020 (95% CI = 0.8792, 0.9248), 0.8972 (95% CI = 0.8741, 0.9203), 0.9442 (95% CI = 0.9094, 0.9790), and 0.8674 (95% CI = 0.8481, 0.8867) at MT prediction task and 0.8471 (95% CI = 0.8362, 0.8580), 0.8550 (95% CI = 0.8456, 0.8644), 0.9303 (95% CI = 0.8786, 0.9820), and 0.7722 (95% CI = 0.7474, 0.7970) at OED identification task. The predictive performance metrics were presented in Table 2.

Multimodal interpretability and external validation

The Kaplan–Meier curves were employed to visualize patient stratification into low- and high-risk populations based on the ROC cutoff value of 0.8078. Then, log-rank test indicated a statistically significant difference in survival distributions between these groups (*P*-value < 2.64e−29, Fig. 1H). Meanwhile, we achieved a visual interpretation of the relative importance of different tissue regions. As shown in Fig. 2A, the sample explainability heat maps for correctly classified patients with OLK were presented, emphasizing the most discriminative features of the oral images during model evaluation. Expert evaluations by oral medicine specialists, the generated heat maps demonstrate that our end-to-end automated framework primarily concentrates on lesion texture and homogeneity features, with particular attention to erosive and atrophic regions.

In the external validation phase, the overall performance metrics of the OMMT-PredNet showed a slight decline compared to the testing cohort. The model achieved AUCs of 0.9469 (95% CI: 0.9145, 0.9793) for cancer risk prediction and 0.8786 (95% CI: 0.8471, 0.9101) for OED classification. Sensitivity and specificity were 80.00% and 88.57%, respectively, for cancer risk prediction, while precision for OED identification was 92.18%. These results underscore the model's robustness, generalizability, and its potential for integration into clinical workflows (Table 2, Fig. 2B).

Discussion

Accurate prediction of OLK malignancy risk is crucial for the prevention and early diagnosis of oral cancer. Within this context, the identification of OED status in OLK is the key indicator for assessing this malignant potential²². Currently, there is a paucity of reliable decision-support models to aid clinicians in the rational assessment of MT risk and the formulation of appropriate patient follow-up and long-term monitoring strategies²³. Typically, Oral Medicine specialists provide clinical care to patients with OPMD, while general dental practitioners play a crucial role in initial screening and long-term monitoring. However, the complex diagnostic process and criteria necessitate not only dental expertise but also a comprehensive understanding of medical knowledge, posing challenges for general dental practitioners in fulfilling this mission. As artificial intelligence is increasingly applied in precision medicine and outcome prediction, this study sought to utilize real-world data to guide clinical decision-making from multiple dimensions and explore the potential for non-invasive alternative diagnostic methods. The results demonstrated that our customized OMMT-PredNet framework achieved notable performance

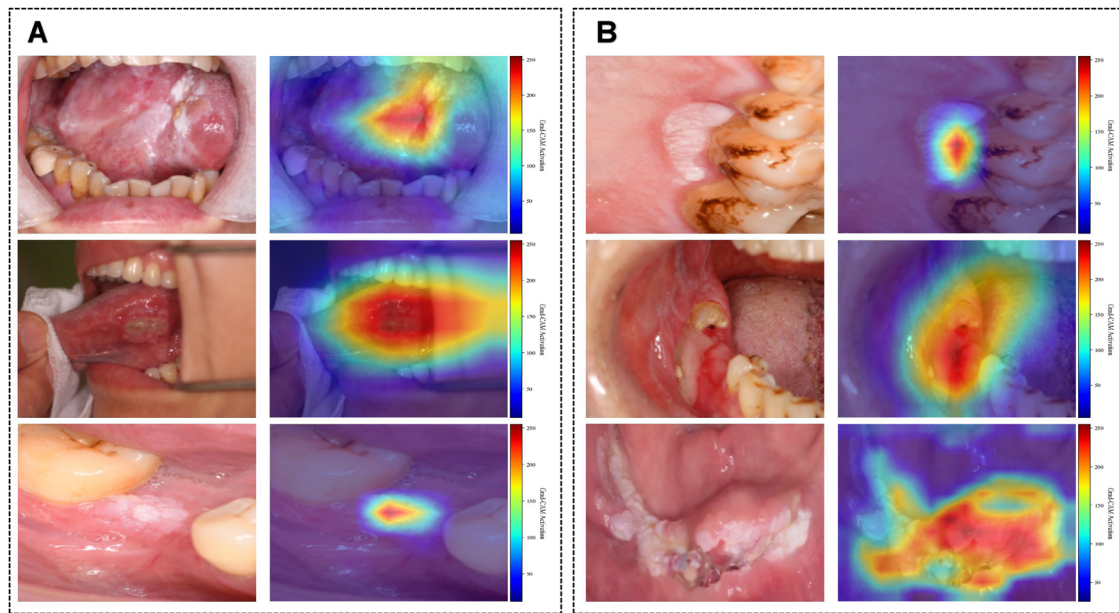


Fig. 2 | Heat maps illustrating the regions of interest identified by the multi-deep learning model. A Three examples of correctly classified oral leukoplakia patients in the testing dataset. **B** Three examples from the external validation dataset. These heat maps highlight areas that significantly contribute to the model's predicted

probabilities. The brightly colored regions are emphasized relative to other areas, as indicated by the scale bar on the right. The accuracy of the heat maps in the external validation dataset reflects the model's strong generalizability.

metrics across multiple tasks, specifically an AUC of 0.9592 for MT prediction and 0.9219 for OED identification, with precision values of 0.9442 and 0.9303, balanced accuracy of 0.9020 and 0.8471, and F1 scores of 0.8972 and 0.8550, respectively.

To the best of our knowledge, this study presents the first and most accurate deep learning-based multi-modal tool designed to construct a robust AI-enabled approach for the above-mentioned multi-task evaluation. The proposed OMMT-PredNet model produced calibrated predicted probabilities of multitasks with good Brier scores and had good stability, discrimination, fairness, and potential net benefit across all evaluation stages. The F1-scores across multiple tasks demonstrated robust performance in balancing precision and recall—a critical requirement for clinical applications where both false positives and false negatives carry significant consequences. The model's NPV is currently suboptimal, which could be enhanced through future integration of non-invasive molecular markers of OED into the diagnostic framework²⁴. Furthermore, we performed fine-grained stratification of patients into low- and high-risk groups based on multi-modal fusion, and incorporated integrated interpretability heat maps into the model's workflow to elucidate the factors influencing the predicted probabilities. Besides, we incorporated data from different institutions as an external validation phase to ensure the model's generalizability.

Ferrer-Sánchez et al. developed a multitask convolutional neural network (CNN) model for OLK risk assessment, achieving a specificity of 0.692 for MT and 0.740 for dysplasia detection, with moderate segmentation performance (Dice = 0.561). A recent systematic review analyzed ML models for OPMD cancer risk prediction, identifying top-performing algorithms such as Random Forest for EMRs (specificity up to 100%) and demonstrating robust accuracy (74–95.4%). However, most studies were based on unimodal textual data from electronic medical records. While these results underscore the potential of deep learning in risk prediction, the clinical translation of this research faces significant limitations, including restricted sample sizes and a lack of external validation for clinical applicability. More importantly, most existing models frame the outcome as a simple binary classification task, significantly limiting their clinical utility due to the oversimplification of complex, multifactorial malignant progression^{21,25–27}. The prediction of time-to-event outcomes in cancer risk is crucial; even experts often struggle to provide precise probabilities of

malignant transformation for patients. In contrast, AI can effectively quantify various dimensions of data, thereby enhancing predictive accuracy. This highlights the fundamental purpose of “predictive models”—forecasting the potential future tendency of disease progression—which is also the primary significance of our study.

Grading of OED is a contentious issue, and reaching a consensus is difficult as different regions of the world apply different grading criteria^{13,28}. In our study, the fundamental value of dysplasia grading lies in its ability to inform cancer risk assessment—a principle that shaped our model's multitask architecture. While multiple classification systems exist, they all ultimately serve the same clinical purpose: accurate malignant transformation risk stratification in OLK. After careful deliberation, we maintained the original diagnostic reports of OED for the following reasons: The original diagnoses reflect the actual clinical decisions that guided patient management at the time of diagnosis. Preserving this real-world context is critical for evaluating long-term outcomes; Re-evaluating all slides under contemporary standards would be influenced by technical artifacts (e.g., fading, sectioning variations) in older specimens could introduce additional bias rather than mitigate it. Our model processes OED as textual labels rather than histopathological images as direct input. The study does not seek to analyze tissue architecture or pathological features per se. From a technical perspective, the proposed class I/II serves as a preprocessing framework rather than proposing a new pathological classification. Comparative studies have also demonstrated comparable predictive efficacy among different grading systems in malignant transformation prediction²⁹. Our external validation using post-2017 data further confirmed its clinical applicability and generalizability.

Manual ROI annotations for clinical photos are time-consuming, subjective, error-prone, costly, lack standardization, and can lead to fatigue and inconsistency among annotators. These challenges can impact the efficiency and accuracy of AI models trained on annotated data, especially for large datasets or real-time applications. Notably, we are the first to implement an end-to-end automated training process without manual ROI annotations of the lesions, allowing the model to learn from the entire clinical image autonomously and achieve strong predictive performance. This novel framework harnesses the ResNet50 architecture, renowned for its

residual learning capabilities, which mitigate the vanishing gradient problem and enable the training of deeper networks. Prior research has demonstrated the robustness and potential of ResNet architectures in unsupervised settings, highlighting their efficacy in achieving competitive performance even without manual labels, thus reinforcing the validity of our approach³⁰. The integration of convolutional block attention module (CBAM) further enhances the model's ability to focus on salient features by dynamically recalibrating channel and spatial importance. This dual attention mechanism allows the model to effectively filter out noise and prioritize critical regions relevant to OED identification. As a result, our approach not only streamlines the training pipeline by eliminating the need for labor-intensive annotations but also enhances predictive accuracy, establishing a new benchmark for automated disease assessment, applicable not only to OLK or other types of OPMD but also provides evidence for future medical image analysis.

Acquiring real-world data poses significant challenges, particularly when it involves obtaining comprehensive clinical images of lesions alongside extensive longitudinal follow-up records that encompass demographic information, clinical manifestations, visual morphology, and histopathological evaluations. Despite we have made concerted efforts to collect the largest possible number of eligible cases with the longest follow-up periods, the sample size remains limited, which is particularly challenging to overcome in the context of retrospective studies. Though notable advancements in AI research across various medical domains, investigations into the application of AI for the detection, prediction, and prognosis of oral precancerous and cancerous lesions are still relatively sparse. Our experiences from the current study underscore the need for more in-depth exploration and research in this specific area in the future. Future multicenter studies incorporating diverse patient populations and imaging protocols are strongly recommended to rigorously evaluate the framework's robustness and clinical applicability across different healthcare settings.

In summary, this study effectively proposed and validated a multimodal framework for the accurate prediction of OED status and associated cancer risk over time in OLK patients using multidimensional data. By employing AI-driven analysis of "real-world data", we generated "real-world evidence" that supports precision medicine and clinical practice. The model demonstrated robust performance in stratifying risk, identifying high-risk visual features linked to histopathological progression, and predicting cancer development over time. As a triage tool, it can guide biopsy decisions while reducing unnecessary procedures; as a monitoring system, it enables risk-adapted follow-up strategies, particularly where specialist access is constrained. This framework addresses a critical need in OLK management by transforming routinely collected data into precise risk assessments without requiring additional infrastructure: a paradigm especially valuable for low- and middle-income regions. Future implementation could bridge diagnostic gaps in primary care while optimizing early intervention and surveillance protocols.

Methods

The retrospective analysis comprised 598 high-resolution oral photographs of consecutive patients diagnosed with OLK, accompanied by their corresponding clinical data. All patients were diagnosed clinically and histopathologically in accordance with the World Health Organization (WHO) criteria^{7,10}, following a standardized protocol: (1) Clinical assessment to exclude other clinically recognizable white or white/red lesions. A provisional diagnosis of OLK is made when a white lesion or white/red lesion is observed on the oral mucosa that cannot be definitively diagnosed as another condition; (2) Biopsy of the lesion to exclude other pathologies that may present as white or white/red lesions. The histopathological findings associated with oral leukoplakia range from simple epithelial hyperplasia with hyperparakeratosis or hyperorthokeratosis to the presence or absence and degree of epithelial dysplasia. In addition, all patients diagnosed with OLK were explicitly advised to abstain from alcohol and tobacco use. Data were collected from the Hospital Authority Clinical Management System (HACMS) of Queen Mary Hospital and Prince Philip Dental Hospital, covering

the period from January 2003 to December 2024 (follow-up endpoint). Sample inclusion criteria in dataset collection were defined by: (a) Both clinical and histopathological assessments were employed to confirm the OLK in all patients. (b) A minimum follow-up duration of 6 months was maintained for all patients. (c) The pathological results were reviewed and confirmed by two experienced specialists to ensure the accuracy of the diagnosis. Regarding the exclusion criteria, patients diagnosed with OLK alongside Head and neck squamous cell carcinoma (HNSCC) either before their initial visit or within 6 months after the first visit were excluded from the study. Additionally, cases with less than 6 months of follow-up and those diagnosed as proliferative verrucous leukoplakia (PVL) were also excluded.

This study adheres to both the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline and the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline. This study was conducted in accordance with the principles outlined in the Declaration of Helsinki. Approval for this study was obtained from the Institutional Review Board of the University of Hong Kong/Hong Kong West Cluster (Reference number: UW23-094). To safeguard patient privacy and confidentiality, the researchers anonymized all clinical data, removing any potential identifiers before proceeding with data analysis.

The presence or absence of OED was assessed through histopathological examination by certified pathologists. Given the extended time span of our real-world data, OED classification followed three WHO systems: the 1978 system (mild/moderate/severe dysplasia), the 2005 system (hyperplasia/mild/moderate/severe dysplasia and carcinoma *in situ*), and the 2017 system (mild/moderate/severe dysplasia)^{31–33}. To maintain fidelity to real-world clinical practice, including diagnostically ambiguous cases (e.g., mild-moderate or moderate-severe), we preserved original diagnostic classifications. Notably, hyperplasia cases were explicitly excluded from our dysplasia classification framework. This approach ensured that our AI-driven approach accurately reflected clinical decision-making patterns. To address datasets' heterogeneity while retaining clinical relevance, we adopted a binary stratification system (Class I/II): Class I included mild, mild-moderate, and moderate cases, whereas Class II comprised moderate-severe, severe, and carcinoma *in situ* cases. The timing of the transformation of OLK to OSCC is defined as the interval between the biopsy date of the original OLK lesion and the first histological diagnosis of malignancy.

The following textual data were collected for each corresponding case: (1) Personal characteristics (age and gender, smoking habits, and alcohol consumption); (2) Clinical manifestations (clinical subtype, lesion location, diameter, and size); (3) Presence of OED at initial diagnosis or during follow-up, along with its grading; (4) Treatment options; and (5) Occurrence of malignant transformation (time-to-event). For discrete variables, label encoding is used to convert them into integer labels, thereby avoiding the high sparsity caused by one-hot encoding. Label encoding maps each value of a feature to a unique integer, transforming clinical text data into information that can be understood by the model.

$$\text{Label Encoding: } c_i \begin{cases} 0 & \text{Category 0} \\ 1 & \text{Category 1} \\ \vdots & \vdots \\ n-1 & \text{Category } n-1 \end{cases}$$

Here, c_i represents the encoded value for the i th categorical feature, and n denotes the total number of categories for that feature.

High-resolution clinical images were obtained for each patient, effectively illustrating the OLK lesions. For the image processing phase, no regions of interest (ROIs) were defined. To address the inconsistent image dimensions in the original clinical photos, all input images were resized to a fixed spatial resolution prior to being fed into the neural network. Specifically, a uniform resizing operation was applied to standardize all images to 224×224 pixels, ensuring consistency in network input and enhancing the reliability of feature extraction during model training. Besides, data augmentation techniques were employed to expand the dataset artificially. This

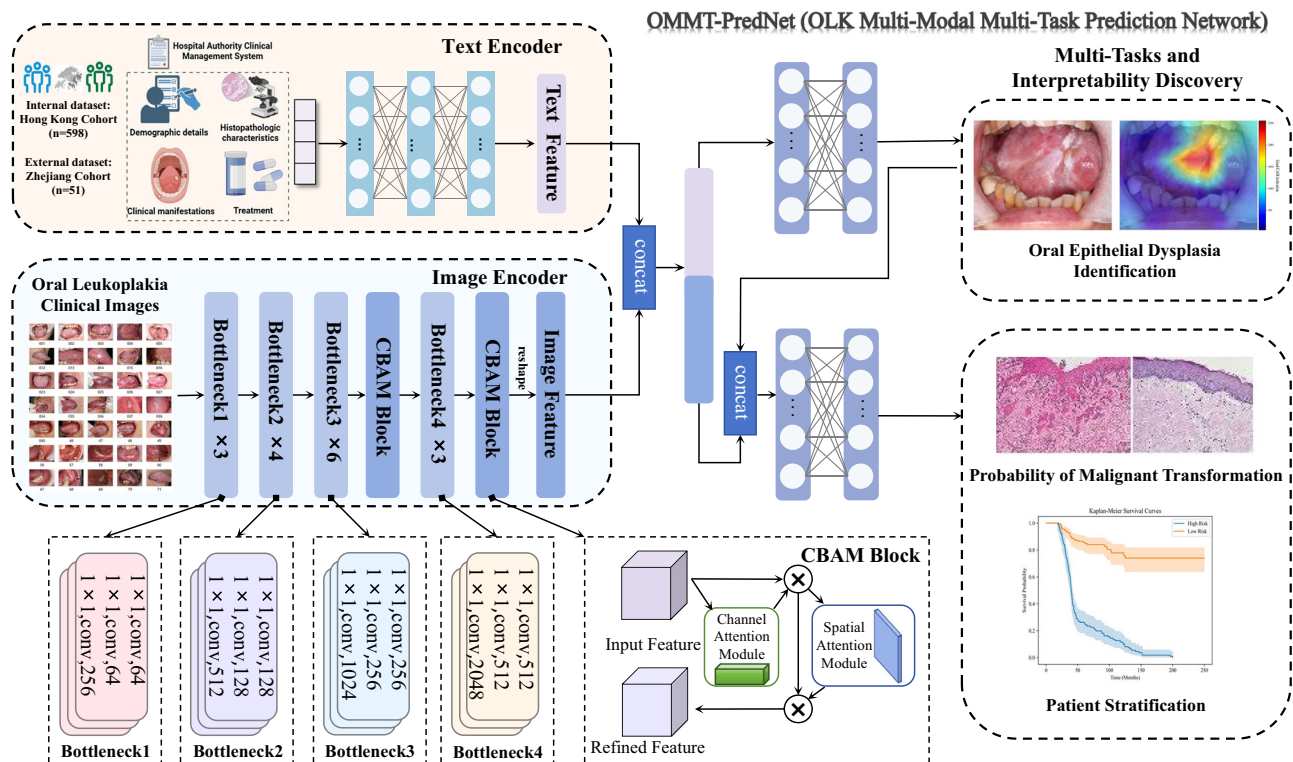


Fig. 3 | Illustration depicts the architecture of the end-to-end automated multi-model framework OMMT-PredNet (OLK multi-modal multi-task prediction network) for multiple tasks. ResNet50 Residual Network 50, CBAM convolutional block attention module.

approach introduces minor variations or utilizes advanced methods to generate additional examples, enhancing the model's efficiency. The applied data augmentation techniques include random vertical flip, random rotation, center crop, random affine transformation, gaussian blur, random noise addition, and color Jitter, all of which contribute to the model's robustness and adaptability. Specific hyperparameter values are detailed in Supplemental Note 2.

In this study, we present OLK multi-modal multi-task prediction network (OMMT-PredNet), a customized end-to-end deep learning framework that achieves full automation for multitasks prediction. The fundamental principles and architectural foundations of the proposed system are as follows:

We designed a novel model architecture that combines ResNet50 with the convolutional block attention module (CBAM) (ResCBAM) to evaluate the presence of OED in clinical images of OLK. Description of model construction and deployment phases undergone in this study is shown in Fig. 3. ResNet50 is a deep convolutional neural network architecture comprising 50 layers, specifically engineered to overcome the challenges associated with training very deep networks. It consists of 48 convolutional layers, along with a MaxPooling layer and an Average Pooling layer, resulting in over 23 million trainable parameters. This complexity enables ResNet50 to capture intricate patterns present in the clinical images.

A key innovation of ResNet50 is the implementation of residual learning, which mitigates the issue of vanishing gradients commonly encountered in deep networks. The architecture employs skip connections that allow gradients to propagate through the network without degradation, facilitating the training of deeper models. The residual block can be mathematically expressed as Eq. (1):

$$R(x) = \text{Output} - \text{Input} = H(x) - x \quad (1)$$

where x represents the input to the neural network block and $H(x)$ represents the true underlying function that the network aims to approximate. By

rearranging this equation, we express the output as Eq. (2):

$$\text{Output} = H(x) + x \quad (2)$$

This formulation indicates that the residual block learns the residual mapping $R(x)$ enabling the network to focus on learning the differences between the input and the desired output, thereby simplifying the learning process and enhancing generalization. A detailed explanation of model principles was described in Supplemental Note 3.

While ResNet50 demonstrates strong performance in various image classification tasks, it has limitations. One notable drawback is its inability to selectively focus on the most informative features within the feature maps, which can lead to suboptimal performance on datasets with high variability, such as those with OLK lesions that present OED. The model may struggle to distinguish between subtle differences in dysplastic changes due to the presence of noise and irrelevant information. Moreover, our model was trained without any manually annotated ROI. This presents a unique challenge, as the model must autonomously identify and assess relevant features across the entire image. In such scenarios, the lack of explicit guidance can further complicate the learning process. To address these limitations, we introduced CBAM after both layers 3 and 4 in the ResNet50 architecture. This attention mechanism enhances feature representation by adaptively refining channel-wise and spatial feature responses, effectively amplifying diagnostically relevant features while suppressing noise and less informative regions. This is particularly beneficial given the characteristics of our dataset, which includes images with varying degrees of dysplasia and noise. By incorporating CBAM, we enable the model to dynamically adjust its attention based on the importance of each feature and spatial region, thereby improving its classification performance even in the absence of manually defined ROIs.

CBAM operates as a lightweight attention mechanism designed to improve the representational capacity of convolutional neural networks. CBAM functions in two sequential stages: channel attention and spatial

attention. A detailed explanation of model principles was described in Supplemental Note 4.

Channel attention module: This module computes attention scores for each channel, allowing the network to emphasize informative features while suppressing less relevant ones. The channel attention map is generated using global average pooling and global max pooling, followed by a shared feed-forward network, represented mathematically as Eq. (3):

$$M_C = \sigma(\text{MLP}([F_{\text{avg}}, F_{\text{max}}])) \quad (3)$$

where F_{avg} and F_{max} denote the pooled feature maps, and σ is the sigmoid activation function.

Spatial attention module: Following channel attention, the spatial attention module further refines the feature maps by focusing on pertinent spatial regions. It generates a spatial attention map using the output from the channel attention module, combining global average pooling and global max pooling across the channel dimension (Eq. (4)):

$$M_s = \sigma(f([F_{\text{avg}}^c, F_{\text{max}}^c])) \quad (4)$$

where f is a convolutional operation that results in a 2D attention map.

The final output of the combined model is obtained by multiplying the original feature map with both the channel and spatial attention maps (Eq. (5)):

$$F_{\text{refined}} = M_C \odot (M_s \odot F) \quad (5)$$

where F is the original feature map, and F_{refined} is the enhanced representation.

The architecture's ability to automatically highlight essential features and regions significantly improves its performance in OED detection, thereby facilitating accurate diagnosis.

To analyze the relevant risk factors, we employed the textual feature encoder (TFE), which leverages multiple fully connected layers to capture complex interactions and higher-order abstractions. The architecture of the TFE is structured as follows:

Input layer: Accepts pre-processed clinical information.

Hidden layers: (a) l th Fully connected layer: Comprises N_l neurons with ReLU activation functions; (b) $(l + 1)$ th fully connected layer: Consists of N_{l+1} neurons with ReLU activation functions. (c) **Dropout layer:** Applied after each fully connected layer to reduce overfitting by randomly deactivating a fraction of neurons during training. Mathematically, the transformation at each layer l can be expressed as Eq. (6):

$$h^{(l)} = \text{ReLU}(W^{(l)}h^{(l-1)} + b^{(l)}) \quad (6)$$

where $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector for layer l , respectively, and $h^{(l-1)}$ is the input from the previous layer.

Output layer: Generates the final encoded clinical feature vector F_{Textual} , which serves as an input to subsequent stages of the multimodal model. The output layer may use a linear activation function or another suitable activation depending on the specific requirements of the model (Eq. (7)).

$$F_{\text{Textual}} = W^{\text{out}} + h^{(L)} + b^{\text{out}} \quad (7)$$

where L denotes the final hidden layer.

The OLK image is processed by the ResCBAM image encoder module to obtain the feature representation F_{Image} . The feature representation F_{Textual} is derived from the TFE module. Image features extracted via ResNet50 + CBAM are dimensionally aligned with clinical text embeddings. Subsequently, the image features and text features are fused through concatenation (concat). The combined feature representation F is obtained

through the following calculation (Eq. (8)):

$$F = \text{concat}(F_{\text{Image}}, F_{\text{Textual}}) \quad (8)$$

After feature fusion, we employed an OED-specific classifier to predict grades, followed by survival analysis incorporating multi-tasks. Our framework enables fully automated end-to-end training, where each module operates independently. Therefore, the final loss consists of three components: the OED classification loss, the MT prediction loss, and the survival loss, which together facilitate multi-task learning. Specifically, we utilize Cross Entropy Loss for the classification task, BCE With Logits Loss for scoring trajectories, and Cox proportional hazards loss for survival analysis. The details of losses are as follows:

OED classification loss: The cross-entropy loss is utilized for the classification of OED, defined as follows (Eq. (9)):

$$\text{Loss}(y, \hat{y}) = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (9)$$

where y_i represents the true label, the \hat{y}_i denotes the predicted probability for the positive class. This loss function has been implemented to optimize the model's performance in the classification of OED.

MT prediction loss: For scoring trajectories, we utilize the BCE with logits loss, denoted as L_{BCE} , which combines a Sigmoid layer and the binary cross-entropy loss in one single class. It is given by (Eqs. (10) and (11)):

$$L_{\text{BCE}} = -y \cdot \log(\sigma(x)) - (1 - y) \cdot \log(1 - \sigma(x)) \quad (10)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

where y is the true label indicating the presence of MT, and x denotes the raw output (logits) from the model prior to applying the sigmoid function $\sigma(x)$. This loss function is crucial for optimizing the model's ability to accurately predict malignancy potential.

Survival loss: For survival analysis, we employ the Cox proportional hazards loss, which is designed to handle censored data effectively³⁴. It is expressed as Eq. (12):

$$l(\theta) := \sum_{i: E_i=1} (\hat{h}_{\theta}(x_i) - \log \sum_{j \in R(T_i)} e^{\hat{h}_{\theta}(x_j)}) \quad (12)$$

$\hat{h}_{\theta}(x_i)$ represents the predicted hazard function for the i th subject, while $R(T_i)$ denotes the cohort at risk at time T_i . The goal is to assess the malignancy potential by modeling the time-to-event data, allowing for the identification of factors influencing malignancy progression. The Cox-PH model's ability to handle censorship and its reliance on hazard ratios provide valuable insights into malignancy potential over time, enabling better risk stratification and clinical decision-making.

Total loss (automatic weighted loss): Automatic weighted loss is a method for automatically determining the weights of individual loss functions in multi-task learning (MTL) models that involve multiple loss functions. The primary objective is to optimize the weight distribution among different tasks or loss functions in an automated manner, thereby enhancing the overall performance of the model without the need for manual weight adjustment. To effectively balance these tasks, we employ a multi-task learning approach, as demonstrated in Eq. (13)^{35,36}.

$$L = \frac{1}{\sigma_1^2} L_{\text{reg}} + \frac{1}{\sigma_2^2} L_{\text{conf}} + \frac{1}{\sigma_3^2} L_{\text{cls}} + \sum_{i=1}^3 \log(\sigma_i + 1) \quad (13)$$

where σ_i , $i \in \{1, 2, 3\}$ are learnable loss weights. Through backpropagation of the model, continuously minimizing the total loss L , the model is able to

optimize the weight distribution of the neural network, thereby enhancing overall performance.

The patient cases were randomly divided into non-overlapping training (80%) and test (20%) sets, which were used to train the models and evaluate their performance. These sets were constructed at the patient case level, meaning that all slides corresponding to a given patient case were assigned exclusively to either the training or test set; slides from the same case were never included in both sets simultaneously. We conducted experiments for each disease model using a five-fold cross-validation approach, reassessing patient cases into non-overlapping training and testing cohorts five times. The model hyperparameters were tuned using only the training folds, while the model's generalization capability was evaluated on the held-out test folds.

Additionally, for external validation of the intelligent model, we utilized a cohort of 51 clinical images and patients' corresponding textual data from Zhejiang University School of Medicine (Baseline characteristics were shown in Supplemental Note 5). While the sample size was moderate, the dataset demonstrated strong representativeness with parameters well-aligned to our model input requirements, featuring a mean follow-up duration of 2.06 years (range: 0.59–6.41 years) based on contemporary diagnostic standards.

The performance of the cross-validated concordance index (c-Index) is reported as the average c-Index across the five folds. To estimate 95% confidence intervals during cross-validation, we utilized non-parametric bootstrapping with 1000 replicates on the out-of-sample predictions from the validation folds^{37,38}. In addition to the c-Index, we also report the cumulative/dynamic AUC (referred to as Survival AUC), which is a time-dependent measure of model performance that evaluates the model's ability to stratify patient risk at various time points, while also correcting for optimistic bias due to censoring through inverse probability of censoring weighting. To construct the Kaplan-Meier curves, we aggregated out-of-sample risk predictions from the validation folds and plotted them against survival time. The cutoff value was determined through ROC curve analysis on the validation set by calculating the Youden index for each potential threshold and selecting the value that maximized this metric. For assessing the statistical significance of patient stratification in the Kaplan-Meier analysis, we employed the log-rank test to determine whether the differences between two survival distributions were statistically significant (P -value < 0.05)³⁹.

To assess the reliability of the model's predicted probabilities, we performed calibration analysis using calibration curves, evaluating the agreement between predicted risks and observed outcomes. Clinical utility was further evaluated using DCA, quantifying the net benefit of the model across different threshold probabilities to inform its practical applicability in decision-making. Ablation studies were conducted to evaluate the contribution of the CBAM module by systematically comparing metrics with and without this component. Besides, a comprehensive set of metrics, including sensitivity, specificity, accuracy, precision, positive and negative predictive value, F1 scores, Brier score loss, and Youden's J statistic, was evaluated. Mathematical calculations for the performance metrics are given below (Eqs. (14)–(20)):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (15)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (16)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (17)$$

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (18)$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (19)$$

$$\text{F1 score} = 2 * \frac{\text{precision} * \text{sensitivity}}{\text{precision} + \text{sensitivity}} \quad (20)$$

All results were analyzed using Python version 3.10 (PyCharm 2023.2.8, Professional Edition, Wilmington, DE, USA).

Data availability

The data that support the findings of this study are not publicly available due to privacy restrictions, but it may be available on request from the corresponding author.

Code availability

The underlying code for this study will be made available through a public GitHub repository upon publication. All analyses were conducted using Python version 3.10 in PyCharm 2023.2.8 (Professional Edition, Wilmington, DE, USA).

Received: 19 February 2025; Accepted: 13 July 2025;

Published online: 10 August 2025

References

- Organization, W. H. *Global Oral Health Status Report: Towards Universal Health Coverage for Oral Health by 2030* (World Health Organization, 2022).
- Races, A. & Males, M. SEER cancer statistics review 1975–2017 (National Cancer Institute, Bethesda, MD, USA, 2020).
- Ford, P. & Farah, C. Early detection and diagnosis of oral cancer: strategies for improvement. *J. Cancer Policy* **1**, e2–e7 (2013).
- Awan, K. Oral cancer: early detection is crucial. *J. Int. Oral Health* **6**, i–ii (2014).
- Iocca, O. 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* **42**, 539–555 (2020).
- Warnakulasuriya, S. 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.* **27**, 1862–1880 (2021).
- Warnakulasuriya, S., Johnson, N. W. & van der Waal, I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. *J. Oral Pathol. Med.* **36**, 575–580 (2007).
- Aguirre-Urizar, J. M., Lafuente-Ibanez de Mendoza, I. & Warnakulasuriya, S. Malignant transformation of oral leukoplakia: Systematic review and meta-analysis of the last 5 years. *Oral Dis.* **27**, 1881–1895 (2021).
- Evren, I. et al. Annual malignant transformation rate of oral leukoplakia remains consistent: a long-term follow-up study. *Oral Oncol.* **110**, 105014 (2020).
- Warnakulasuriya, S. Oral potentially malignant disorders: a comprehensive review on clinical aspects and management. *Oral Oncol.* **102**, 104550 (2020).
- Ho, M. W. et al. The clinical determinants of malignant transformation in oral epithelial dysplasia. *Oral Oncol.* **48**, 969–976 (2012).
- Pimenta-Barros, L. A., Ramos-Garcia, P., Gonzalez-Moles, M. A., Aguirre-Urizar, J. M. & Warnakulasuriya, S. Malignant transformation of oral leukoplakia: systematic review and comprehensive meta-analysis. *Oral Dis.* **31**, 69–80 (2025).
- Odell, E., Kujan, O., Warnakulasuriya, S. & Sloan, P. Oral epithelial dysplasia: recognition, grading and clinical significance. *Oral Dis.* **27**, 1947–1976 (2021).
- Hegde, S., Ajila, V., Zhu, W. & Zeng, C. Artificial intelligence in early diagnosis and prevention of oral cancer. *Asia Pac. J. Oncol. Nurs.* **9**, 100133 (2022).

15. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
16. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
17. Li, J. et al. Diagnostic accuracy of artificial intelligence assisted clinical imaging in the detection of oral potentially malignant disorders and oral cancer: a systematic review and meta-analysis. *Int. J. Surg.* **110**, 5034–5046 (2024).
18. Topol, E. J. As artificial intelligence goes multimodal, medical applications multiply. *Science* **381**, adk6139 (2023).
19. Adeoye, J. et al. A deep learning system to predict epithelial dysplasia in oral leukoplakia. *J. Dent Res.* **103**, 1218–1226 (2024).
20. Peng, J. et al. Oral epithelial dysplasia detection and grading in oral leukoplakia using deep learning. *BMC Oral Health* **24**, 434 (2024).
21. Wang, X. 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.* **49**, 417–426 (2020).
22. Chaturvedi, A. K. et al. Oral leukoplakia and risk of progression to oral cancer: a population-based cohort study. *J. Natl Cancer Inst.* **112**, 1047–1054 (2020).
23. Adeoye, J., Tan, J. Y., Choi, S. W. & Thomson, P. Prediction models applying machine learning to oral cavity cancer outcomes: a systematic review. *Int. J. Med. Inform.* **154**, 104557 (2021).
24. Alajaji, S. A., Khoury, Z. H., Jessri, M., Sciubba, J. J. & Sultan, A. S. An update on the use of artificial intelligence in digital pathology for oral epithelial dysplasia research. *Head Neck Pathol.* **18**, 38 (2024).
25. Ferrer-Sanchez, A., Bagan, J., Vila-Frances, J., Magdalena-Benedito, R. & Bagan-Debon, L. Prediction of the risk of cancer and the grade of dysplasia in leukoplakia lesions using deep learning. *Oral Oncol.* **132**, 105967 (2022).
26. Liu, Y. et al. Quantitative prediction of oral cancer risk in patients with oral leukoplakia. *Oncotarget* **8**, 46057–46064 (2017).
27. Uppal, S., Kumar Shrivastava, P., Khan, A., Sharma, A. & Kumar Shrivastav, A. Machine learning methods in predicting the risk of malignant transformation of oral potentially malignant disorders: a systematic review. *Int. J. Med. Inform.* **186**, 105421 (2024).
28. Khoury, Z. H., Sultan, M. & Sultan, A. S. Oral epithelial dysplasia grading systems: a systematic review & meta-analysis. *Int. J. Surg. Pathol.* **30**, 499–511 (2022).
29. Ellonen, R. et al. Binary- and three-tiered oral epithelial dysplasia grading system and malignant transformation. *Oral Dis.* <https://doi.org/10.1111/odi.15334> (2025).
30. Tomasev, N. et al. Pushing the limits of self-supervised resnets: can we outperform supervised learning without labels on imagenet? arXiv preprint arXiv:2201.05119 (2022).
31. Barnes, L. *Pathology and Genetics of Head and Neck Tumours* Vol. 9 (IARC, 2005).
32. El-Naggar, A. K., Chan, J. K., Rubin Grandis, J. & Slootweg, P. J. *WHO Classification of Head and Neck Tumours* (WHO, 2017).
33. Tabor, M. P. et al. Comparative molecular and histological grading of epithelial dysplasia of the oral cavity and the oropharynx. *J. Pathol.* **199**, 354–360 (2003).
34. Nagpal, C., Potosnak, W. & Dubrawski, A. Auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. In *Proc. Machine Learning for Healthcare Conference*, 585–608 (2022).
35. Kendall, A., Gal, Y., Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. IEEE conference on computer vision and pattern recognition*, 7482–7491 (2018).
36. Liebel, L. & Körner, M. Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334 (2018).
37. LeDell, E., Petersen, M. & van der Laan, M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.* **9**, 1583–1607 (2015).
38. Tsamardinos, I., Greasidou, E. & Borboudakis, G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. Learn.* **107**, 1895–1922 (2018).
39. Bland, J. M. & Altman, D. G. The logrank test. *BMJ* **328**, 1073 (2004).

Acknowledgements

We would like to express our sincere gratitude to Yiming Peng from the Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, for his statistical technical support. This study was supported by Stable Support Project of Shenzhen (No. 20231122145548001); Natural Science Foundation of Shenzhen Municipality (No. JCYJ20220531091407016); Futian Healthcare Research Project (No. FTWS069, FTWS055); Key Research and Development Program of Zhejiang Province (Grant number:2025C02100); University of Hong Kong Seed Fund for PI Research (Grant number: 2402101364).

Author contributions

L.W.Z. and J.W.L. had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. J.W.L., Y.F.Z., L.W.Z., and D.Z. conceived and designed the study. J.W.L. and L.W.Z. obtained the data, and D.Z., Y.F.Z., and M.J.Z. developed the analytic codes for analysis. J.A., J.Y.P., M.J.Z. provided material support. L.J.F., M.M.Z., and C.X.L. provided external validation support. C.M, J.W.L., D.Z., and L.W.Z. provided clinical insights into the study, critically reviewed, and conducted several rounds of revisions. L.W.Z. and D.Z. supervised the study. All authors read and approved the final version of the manuscript and agreed to submit it for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01885-8>.

Correspondence and requests for materials should be addressed to Dian Zhang or LiWu Zheng.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025