GingiCasc: A Tooth-ROI Framework for Ordinal MGI Gingivitis Severity Grading

Linyi Zhang
Seoul National University
2025-28627
zhanglinyi1225@snu.ac.kr

Kyungjin Kim
Seoul National University
2024-28748
kkj_james@snu.ac.kr

Ungyu Lee
Seoul National University
2024-29390
unyui24@snu.ac.kr

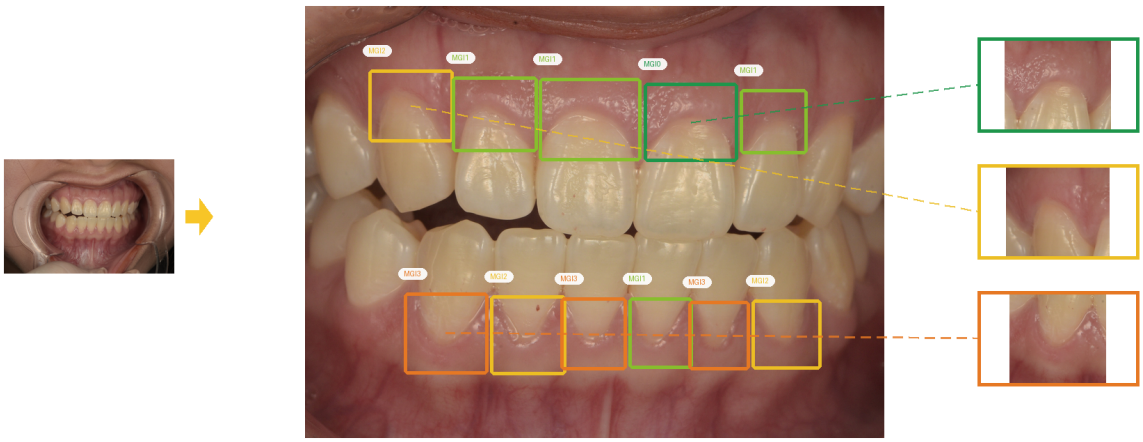

Fig. 1: GingiCasc performs tooth-level gingivitis assessment from intraoral images. It detects tooth ROIs and predicts ordinal MGI0–MGI4 grades in one framework. Selected matched teeth are enlarged on the right for visual inspection. No lesion masks or lesion-level annotations are used.

Abstract

Gingivitis is prevalent and its clinical assessment relies on tooth-by-tooth visual inspection, which consumes dental resources and limits scalable screening. Intraoral photographs provide a low-cost imaging source, but existing image-based methods often report image-level diagnoses or coarse localisation outputs that do not directly match tooth-level clinical records. Many studies also use private data, unreleased code, and heterogeneous metrics, which limits reproducibility and fair comparison. This paper formulates gingivitis assessment from intraoral images as tooth ROI detection and ordinal Modified Gingival Index (MGI) grading, and proposes GingiCasc, a Cascade R-CNN-based framework for detecting tooth ROIs and predicting MGI0–MGI4 severity grades.

GingiCasc uses a ResNet-50-FPN backbone and three cascade RoI stages to refine tooth proposals, and predicts tooth boxes, MGI grades, and grade probabilities from aligned RoI representations. To address long-tailed MGI labels and ordinal severity errors, the training objective combines class-balanced weighting, additional MGI0/MGI1 class weight multipliers, and a severity-distance penalty. Experiments on the public Duy 2024 provided split show that GingiCasc improves MGI grading metrics over two R50 baselines and achieves the highest mAP among the compared methods. These results support tooth-level ROI detection with matched-tooth ordinal grading as a reproducible evaluation setting for intraoral gingivitis analysis.

1 Introduction

Gingivitis is among the most common inflammatory oral diseases, and its prevalence can exceed 50% in several adult populations. When not controlled in time, gingivitis increases the risk of progression to periodontitis[29]. Clinical assessment usually depends on local visual signs, including gingival colour, contour, swelling, and bleeding[27, 39]. These signs appear around specific teeth, and clinical records therefore require tooth-level descriptions of disease severity rather than a single judgement for an entire intraoral image[26, 27, 39]. Tooth-by-tooth visual assessment preserves the local information needed for diagnosis, but it also consumes dental resources and is difficult to scale directly to population-level screening[32, 39].

Deep learning models have recently been used to represent colour, boundary, and local morphology in medical images, providing a methodological basis for automated gingivitis analysis from intraoral RGB images[18–20, 25, 35, 36]. Gingivitis severity, however, is not a single visual attribute

of the whole image[26, 39]. A model that only outputs an image-level diagnosis or a coarse abnormal region still does not correspond to the clinical process of recording severity for individual teeth[20, 39].

Existing intraoral-image studies on gingivitis can be roughly grouped by output granularity and supervision form into image-level classification, regional localisation or object detection, and segmentation-style analysis[1, 6, 20, 28]. Image-level methods directly predict the inflammatory state or severity of a full image, and show that intraoral photographs contain learnable visual patterns related to gingival inflammation[20, 28]. Regional localisation and object-detection methods further provide local regions or boxes, moving model outputs from whole-image decisions to local structures[1, 15, 20, 33]. Segmentation-style studies annotate and predict pixel-level health or inflammation regions along the gingival margin[6]. Nevertheless, image-level methods do not directly produce tooth-level severity records[20, 26, 39]. Across studies, task boundaries, supervision forms, and reported metrics differ substantially, and localisation, detection, or segmentation outputs are not necessarily aligned with tooth-level severity labels[1, 6, 20, 28, 35]. In a systematic review of 23 deep-learning studies on dental plaque and gingivitis images, Moharrami et al. reported lower certainty of evidence for gingivitis than for plaque, and noted remaining limitations in external testing, multi-centre data, and consistency of reporting[28]. Gingivitis image analysis therefore still lacks a task definition whose inputs and outputs are standardised around tooth-level clinical recording[26, 28, 39].

This paper defines intraoral-image gingivitis analysis as tooth ROI detection and tooth-level Modified Gingival Index (MGI) grading[11, 26]. Given an intraoral image, the model must localise tooth boxes and output an MGI0–MGI4 grade with its probability distribution for each tooth ROI[11, 26]. This setting differs from image-level gingivitis classification and from lesion segmentation: both the training supervision and the model outputs are defined by tooth ROIs and MGI severity labels, not by gingivitis lesion masks or lesion-level boxes[6, 11, 20].

The MGI is a non-invasive visual scoring system for gingival severity[26]. It describes the progression from healthy gingiva to severe inflammation. MGI0–MGI4 has a natural order, and confusion between adjacent grades and confusion across distant grades have different clinical meanings[5, 10, 26, 30]. A standard multiclass objective only checks whether the predicted class equals the ground-truth class, and does not directly encode the distance between grades. Under imbalanced grade distributions, infrequent grades can also be absorbed by frequent grades[5, 7, 10, 24, 26, 30]. Tooth-level MGI grading therefore requires the model to handle tooth ROI localisation, class-frequency imbalance, and severity-distance errors together[7, 11, 26].

Motivated by the above analysis, this paper proposes GingiCasc, a framework that performs tooth ROI detection and ordinal MGI grading in a single model[4, 15, 23, 26, 33]. GingiCasc first extracts multi-scale features with a ResNet-50-FPN backbone, and then generates and updates tooth ROI candidates through three cascade RoI stages[4, 15, 23, 33]. The three RoI stages use increasing IoU thresholds (0.5/0.6/0.7), so candidate boxes move from a looser matching criterion to a stricter localisation target during training[4]. At the RoI level, the model outputs tooth boxes, MGI class logits, and MGI probability distributions, keeping localisation results aligned with tooth-level severity predictions[26, 39].

The training objective explicitly incorporates severity information to address class imbalance and distant ordinal errors[5, 7, 10, 24, 26]. MGI classification uses class-balanced weighting based on the effective number of samples[7], with additional class weight multipliers for MGI0 and MGI1. A severity-distance penalty uses the distance between the predicted probability distribution and the ground-truth grade to penalise distant grade errors, so that the objective reflects the order of MGI grades[5, 10, 26]. Object detection studies commonly use IoU-based AP and mAP to quantify localisation performance[31]. For evaluation, this paper further adopts class-agnostic one-to-one tooth matching, and reports tooth localisation metrics separately from matched-tooth MGI grading metrics. This protocol avoids forcing missed teeth into an MGI grade and allows detection errors and grading errors to be analysed separately.

The main contributions of this paper are summarised as follows.

1. We define a tooth ROI detection and MGI0–MGI4 tooth-level grading task, so that model outputs correspond to the tooth boxes and severity grades required by tooth-by-tooth clinical recording[11, 26].

2. We propose GingiCasc, which integrates MGI severity prediction into three-stage cascade-style RoI detection and trains the grading head with class-balanced weighting, MGI0/MGI1 class weight multipliers, and a severity-distance penalty[4, 7, 23, 26].

3. We use a class-agnostic tooth-matching evaluation protocol that reports tooth localisation and matched-tooth MGI grading separately, thereby distinguishing detection errors from grading errors.

4. We evaluate three random seeds on the test split and show that GingiCasc improves the main MGI grading metrics over two R50 baselines, while reporting detection and grading metrics separately.

The remainder of this paper is organised as follows. Section 2 reviews intraoral-image gingivitis analysis, MGI grading, ROI detection, and class-imbalance handling. Section 3 describes the GingiCasc architecture and training objective. Section 4 presents the dataset, evaluation protocol, and experimental results. Section 5 summarises the scope and conclusions of the paper.

2 Related Work

Intraoral-image gingivitis analysis involves image acquisition, local-structure localisation, severity recording, and model evaluation. To clarify the relation between this study and prior

work, this section reviews four groups of studies: gingivitis analysis from intraoral images, tooth-level MGI severity grading and ordinal learning, ROI-based detection frameworks, and class-imbalance and severity-aware losses.

2.1 Gingivitis Analysis From Intraoral Images

Intraoral images provide a practical data source for non-invasive gingivitis screening. Early photographic and image-analysis studies used intraoral images to record gingivitis and gingival-margin inflammation, and examined agreement between image-based assessment and clinical examination[2, 37]. Later intraoral-scanner studies further explored the feasibility of non-invasive imaging for gingival-inflammation assessment[8]. With the adoption of deep learning in medical and dental image analysis, automatic analysis of intraoral photographs gained stronger representation tools[18, 19, 25, 34, 35]. In gingivitis tasks, existing studies have used intraoral RGB images for screening, photo-based detection, local localisation, and inflammation-grade analysis[1, 6, 20, 38, 40]. According to output form, current methods can be grouped into image-level classification, local region localisation or object detection, and gingival-region analysis.

Image-level classification methods directly predict the presence of gingivitis or chronic gingivitis from a whole intraoral photograph. Li et al. studied automatic gingivitis screening from RGB intraoral images and used classification with localisation to show learnable visual patterns associated with gingival inflammation[20]. A later transfer-ensemble study by Li et al. identified chronic gingivitis, further supporting intraoral photographs as inputs for binary screening[21]. Chau et al. evaluated the accuracy of photo-based artificial-intelligence gingivitis detection[6]. Tobias and Spanier used dental selfies for MGI classification, indicating that patient-side intraoral images may also contain gingival-severity cues[38]. These methods usually treat the whole image or a cropped image as the prediction unit, and therefore do not directly bind predictions to tooth-level severity records.

Local-region localisation and object-detection methods move the model output from image-level classes to local boxes or response regions. Alalharith et al. used Faster R-CNN to detect early signs of gingivitis in orthodontic patients[1]; the localisation results of Li et al. also highlighted image regions related to gingivitis[20]. These methods are closer to the local visual judgements made in clinical inspection, but the target regions and output protocols differ across studies, and the resulting regions do not necessarily align with the MGI severity label of each tooth.

Gingival-region analysis targets local conditions of the gingival margin or gingival tissue. Wen et al. used a gingival-removal strategy to build an automatic gingival-inflammation grading model, shifting both input and analysis towards gingival tissue regions[40]. Such studies emphasise fine-grained visual changes in gingival tissue, but their target differs from assigning an MGI label to each detected tooth ROI. In a systematic review of deep-learning studies on dental plaque and gingivitis in RGB intraoral photographs, Moharrami et al. reported limited certainty of evidence for gingivitis, as well as limitations in external testing, multi-centre data, and reporting consistency[28]. Existing studies therefore support the feasibility of automated gingivitis analysis from intraoral images, but their task boundaries, supervision objects, and evaluation metrics remain weakly aligned with tooth-by-tooth MGI recording.

2.2 Tooth-level MGI Severity Grading and Ordinal Learning

Clinical gingivitis assessment relies on local visual signs. The Gingival Index and Modified Gingival Index use observations of colour, contour, and bleeding-related changes to describe gingival-margin inflammation[26, 27]; clinical diagnosis also depends on case definitions and local signs of plaque-induced gingivitis[39]. The MGI records local gingival status on an ordered scale from MGI0 to MGI4[26]. For an intraoral-image model, tooth-level MGI grading is therefore not just binary screening: the model must establish correspondence between a predicted region and a specific tooth, and then assign the grade of the adjacent gingiva. Image-level accuracy alone cannot show whether the model localises the correct tooth or separate missed detections, mismatches, and grade errors.

Ordinal-learning studies distinguish ordered labels from nominal classes. Frank and Hall transformed ordinal classification into a set of ordered binary classification problems[13]; Gutiérrez et al. reviewed ordinal-regression methods and noted that ordered labels require modelling and evaluation objectives that differ from unordered multiclass classification[16]. In deep models, Niu et al. used a multiple-output CNN for ordinal regression in age estimation[30]; Diaz and Marathe used soft labels to represent relationships between adjacent grades[10]; and Cao et al. proposed rank-consistent ordinal regression to constrain ordered predictions[5]. These studies show that the distance structure of ordered labels can be incorporated into training objectives or label representations.

Ordinal tasks should not be evaluated only with exact accuracy. Baccianella et al. discussed evaluation measures for ordinal regression and emphasised the distance between predicted and true ranks[3]. Weighted kappa loss further introduced ordinal disagreement into deep classification losses[9]. This paper does not directly adopt these ordinal-regression architectures; instead, it adds a severity-distance penalty to the MGI classification head of a tooth ROI detector. The evaluation reports MAE, quadratic weighted kappa, adjacent accuracy, and balanced accuracy, so that exact agreement, grade distance, and class-balanced behaviour are considered together.

2.3 ROI-based Detection Frameworks and Matching Evaluation

ROI-based detectors organise object localisation and region-level classification in a single framework. R-CNN first used candidate regions and extracted deep features for each region[15]. Fast R-CNN performed region classification and bbox regression on shared feature maps through RoI

pooling[14]. Faster R-CNN introduced a Region Proposal Network, integrating proposal generation and the detection network end to end[33]. Feature Pyramid Networks construct multi-scale feature pyramids with top-down and lateral connections, which benefits detection across object sizes[23]. Cascade R-CNN trains multiple detection heads with progressively higher IoU thresholds, matching proposal refinement with higher-quality localisation targets[4].

Proposal generation, RoI feature extraction, bbox regression, and region-level classification in region-based detection provide a reusable structure for tooth ROI detection. In medical images, however, the object definition must remain aligned with clinical labels. The detection target in this paper is not a generic object or a gingivitis region, but a tooth ROI with an MGI severity label. GingiCasc uses a ResNet-50-FPN backbone and three cascade RoI stages, and outputs both tooth boxes and MGI class logits at the RoI level. This design uses the local representation capability of region-based detection while restricting the predicted region to tooth ROIs, so that regional predictions correspond to tooth-level MGI grading.

Detection evaluation usually depends on IoU-based matching and AP/mAP. The PASCAL VOC and COCO benchmarks popularised AP/mAP as standard object-detection measures[12, 22]; Padilla et al. compared common object-detection metrics and summarised different AP and mAP calculations[31]. For tooth-level grading, detection AP alone is insufficient because the model must also assign the correct MGI grade on matched teeth. This paper uses class-agnostic one-to-one tooth matching at IoU ≥ 0.5, and reports tooth-localisation metrics separately from matched-tooth grading metrics.

2.4 Class Imbalance and Severity-aware Losses

Class imbalance is common in medical image classification and detection. He and Garcia reviewed imbalanced-data learning and described how high-frequency classes can affect classifier training and evaluation interpretation[17]. In deep detection and classification, focal loss handles foreground/background and difficulty imbalance by down-weighting easy samples[24]; class-balanced loss uses the effective number of samples to construct class weights and reduce training bias caused by sample-count differences[7]. For MGI grading, when low-severity samples are less frequent, ordinary cross-entropy can bias the model towards more frequent grades.

Ordered labels introduce another error structure: adjacent grade errors and errors across several grades should not receive identical penalties. Ordinal soft labels, rank-consistent ordinal regression, and weighted kappa loss each use grade-distance information in different ways[5, 9, 10]. GingiCasc uses class-balanced MGI classification, MGI0/MGI1 class weight multipliers, and a severity-distance penalty in the MGI classification head. The first two components address frequency differences among tooth-level MGI labels, whereas the last penalises distant grade errors according to the distance between the predicted distribution and the ground-truth grade. In contrast to general ordinal-regression or imbalance-learning methods, the aim here is to incorporate class frequency and severity distance into the MGI grading head of a tooth ROI detector, and to report grading performance on matched teeth.

3 Method

3.1 Problem Formulation

Given an intraoral image I, this paper formulates gingivitis assessment as a joint problem of tooth ROI detection and ordinal MGI grading. The training annotations are written as

$$\mathcal{G}(I) = \{(b_j, y_j)\}_{j=1}^{N}, \qquad b_j \in \mathbb{R}^4, \quad y_j \in \mathcal{Y} = \{0, 1, 2, 3, 4\}, \tag{1}$$

where $b_j = (u_j, v_j, w_j, h_j)$ is the bounding box of the j-th tooth ROI, and y_j is the MGI grade of the gingiva adjacent to the corresponding tooth. The raw tooth classes 2–6 are mapped to MGI0–MGI4 by

$$y = r - 2, \qquad r \in \{2, 3, 4, 5, 6\}. \tag{2}$$

Raw classes 0 and 1 indicate maxilla and mandible jaw ROI metadata. They are excluded from the main training target and are not used as tooth-detection classes.

The model output is

$$\widehat{\mathcal{G}}(I) = \{(\hat{b}_i, \hat{y}_i, \hat{s}_i, \hat{p}_i)\}_{i=1}^{M}, \qquad \hat{p}_i \in [0, 1]^5, \quad \sum_{c=0}^{4} \hat{p}_{i,c} = 1, \tag{3}$$

where $\hat{b}_i$ is a predicted tooth box, $\hat{s}_i$ is the tooth-detection score, $\hat{p}_i$ is the probability distribution over the five MGI grades, and $\hat{y}_i = \arg\max_c \hat{p}_{i,c}$. This definition fixes each prediction unit as a tooth ROI and requires the model to output both tooth location and the corresponding MGI severity label. Training and evaluation are therefore organised around tooth-level detection and matched-tooth grading.

3.2 Cascade-style Tooth ROI Detector

GingiCasc uses a region-based cascade detector. An input image first passes through a ResNet-50-FPN backbone to produce multi-scale features

$$\mathcal{F} = B_\theta(I), \tag{4}$$

and a region proposal network then generates initial proposals $\mathcal{R}^{(0)}$. These proposals enter $T = 3$ RoI stages sequentially. Stage t uses the IoU threshold

$$\tau = (\tau_1, \tau_2, \tau_3) = (0.5, 0.6, 0.7) \tag{5}$$

for positive/negative assignment, with bbox regression weights $(10, 10, 5, 5)$, $(20, 20, 10, 10)$, and $(30, 30, 15, 15)$. The increasing τ_t moves training from looser tooth-proposal matching towards stricter tooth localisation.

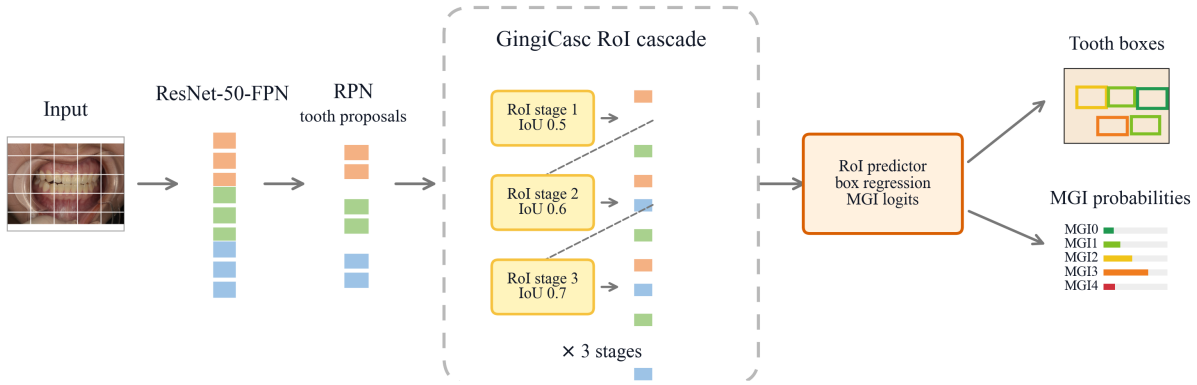

Fig. 2: GingiCasc method overview. The model detects tooth ROIs, predicts MGI grades, and uses severity-aware training terms for tooth-level ordinal grading. Evaluation uses class-agnostic one-to-one tooth matching to separate detection metrics from matched-tooth grading metrics.

For stage t, RoI feature extraction and the box head produce the proposal representation $h_i^{(t)}$. The predictor outputs background-plus-MGI class logits and class-specific box offsets:

$$z_i^{(t)} \in \mathbb{R}^6, \qquad \Delta_i^{(t)} \in \mathbb{R}^{6\times 4}. \tag{6}$$

Class 0 is background, and classes 1–5 correspond to MGI0–MGI4. Let $\ell_i^{(t)} \in \{0,1,2,3,4,5\}$ denote the detector label assigned by stage matching. A foreground proposal satisfies $\ell_i^{(t)} > 0$, with MGI label $y_i^{(t)} = \ell_i^{(t)} - 1$.

During training, each stage computes a classification loss and a box-regression loss independently, and the cascade stages are weighted by

$$\alpha = (\alpha_1, \alpha_2, \alpha_3) = (1.0, 0.5, 0.25). \tag{7}$$

The first two stages update proposals with the predicted boxes from the current stage:

$$\mathcal{R}^{(t)} = \mathrm{Refine}\left(\mathcal{R}^{(t-1)}, \Delta^{(t)}, z^{(t)}\right), \qquad t = 1, 2. \tag{8}$$

The final stage produces the final box regression. During inference, the class logits from the three stages are averaged:

$$\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_i^{(t)}. \tag{9}$$

The MGI probability vector is obtained by normalising the foreground logits:

$$\hat{p}_{i,c} = \frac{\exp(\bar{z}_{i,c+1})}{\sum_{k=0}^{4}\exp(\bar{z}_{i,k+1})}, \qquad c \in \{0,1,2,3,4\}. \tag{10}$$

Post-processing removes background predictions, applies score filtering, small-box filtering, and class-wise non-maximum suppression, and preserves the MGI probability vector aligned with each tooth box.

3.3 Severity-aware Training Objective

The GingiCasc training objective consists of a cascade detection objective and a severity-aware MGI objective. The latter combines class-balanced weighting, MGI0/MGI1 class weight multipliers, and a severity-distance penalty to handle class imbalance and grade-distance errors.

Let n_c be the number of samples for MGI class c in the training set. Class-balanced weighting constructs the initial weight from the effective number of samples[7]:

$$\tilde{w}_c = \frac{1-\beta}{1-\beta^{n_c}}, \qquad \beta = 0.999. \tag{11}$$

To preserve the loss scale, the weights are mean-normalised:

$$w_c = \frac{\tilde{w}_c}{\frac{1}{5}\sum_{k=0}^{4}\tilde{w}_k}. \tag{12}$$

Low-severity MGI samples use additional class weight multipliers

$$m = (1.5, 2.0, 1.0, 1.0, 1.0), \tag{13}$$

and are normalised again to obtain the final foreground class weight:

$$v_c = \frac{w_c m_c}{\frac{1}{5}\sum_{k=0}^{4} w_k m_k}. \tag{14}$$

For detector labels, the background weight is $\eta_0 = 1$, and the MGI foreground weight is $\eta_{c+1} = v_c$.

For stage t, let Q_t be the number of sampled proposals, and define $D_t = \sum_{i=1}^{Q_t}\eta_{\ell_i^{(t)}}$. The weighted classification loss is

$$\mathcal{L}_{\mathrm{cls}}^{(t)} = -\frac{1}{D_t}\sum_{i=1}^{Q_t}\eta_{\ell_i^{(t)}}\log\frac{\exp(z_{i,\ell_i^{(t)}}^{(t)})}{\sum_{k=0}^{5}\exp(z_{i,k}^{(t)})}. \tag{15}$$

Box regression is applied only to foreground proposals:

$$\mathcal{P}_t = \{i \mid \ell_i^{(t)} > 0\}, \tag{16}$$

$$\mathcal{L}_{\mathrm{box}}^{(t)} = \frac{1}{Q_t}\sum_{i\in\mathcal{P}_t}\mathrm{smooth}_{L_1}^{\beta_{\mathrm{box}}}\left(\Delta_{i,\ell_i^{(t)}}^{(t)} - \Delta_i^{*(t)}\right), \tag{17}$$

$$\beta_{\mathrm{box}} = \frac{1}{9}. \tag{18}$$

Here $\Delta_i^{*(t)}$ is the regression target encoded from the matched ground-truth box. When $\mathcal{P}_t = \emptyset$, the summation is empty and the box loss is 0.

MGI grades are ordered, and distant grade errors are more severe than adjacent grade errors. For foreground proposal $i \in \mathcal{P}_t$, foreground logits are first converted to MGI probabilities:

$$p_{i,c}^{(t)} = \frac{\exp(z_{i,c+1}^{(t)})}{\sum_{k=0}^{4} \exp(z_{i,k+1}^{(t)})}. \tag{19}$$

The severity-distance penalty is then defined as the expected absolute distance from the predicted distribution to the true MGI class:

$$\mathcal{L}_{\text{dist}}^{(t)} = \frac{1}{|\mathcal{P}_t|} \sum_{i \in \mathcal{P}_t} \frac{1}{4} \sum_{c=0}^{4} p_{i,c}^{(t)} \left| c - y_i^{(t)} \right|. \tag{20}$$

When $\mathcal{P}_t = \emptyset$, $\mathcal{L}_{\text{dist}}^{(t)}$ is also set to 0. This term does not change the MGI label space; it explicitly penalises cross-grade errors in the training objective.

The stage loss and total loss are

$$\mathcal{L}^{(t)} = \mathcal{L}_{\text{cls}}^{(t)} + \mathcal{L}_{\text{box}}^{(t)} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}^{(t)}, \tag{21}$$

$$\mathcal{L}_{\text{total}} = \sum_{t=1}^{3} \alpha_t \mathcal{L}^{(t)}, \tag{22}$$

where $\lambda_{\text{dist}} = 0.5$. GingiCasc therefore incorporates the ordered structure of MGI0–MGI4 through class weights and a severity-distance penalty within the same detection objective.

3.4 Prediction Interface

For a single image, inference returns tooth boxes, MGI labels, detection scores, and MGI probability vectors:

$$\widehat{\mathcal{G}}(I) = \{(\hat{b}_i, \hat{y}_i, \hat{s}_i, \hat{p}_i)\}_{i=1}^{M}. \tag{23}$$

Here $\hat{s}_i$ is the tooth foreground score, $\hat{p}_i$ is obtained by normalising the averaged foreground logits from the three cascade stages, and $\hat{y}_i = \arg\max_c \hat{p}_{i,c}$. This interface defines only the prediction record passed to the evaluator. Class-agnostic tooth matching, missed-tooth statistics, and matched-tooth grading metrics are defined in Section 4 as part of the experimental protocol.

4 Experiments

4.1 Dataset and Split

The model was evaluated on the public intraoral-image dataset released by Duy et al.[11], using the provided split. In the raw annotations, classes 2–6 denote tooth ROIs with MGI grades and are mapped to MGI0–MGI4. Classes 0 and 1 denote maxilla and mandible jaw ROI metadata, and are used only for data-consistency checks. They are excluded from the main training target and are not used as tooth-detection classes. The converted COCO-style annotations are used for training, validation, and testing.

Table 1: Dataset distribution after remapping raw tooth labels to MGI0–MGI4. Counts are tooth-level annotations in the COCO-style files; jaw ROI metadata boxes are excluded from this table.

Split	Images	Tooth ROIs	MGI0	MGI1	MGI2	MGI3	MGI4
Train	732	8,790	244	1,218	2,178	3,399	1,751
Validation	182	2,185	30	306	774	755	320
Test	182	2,175	35	307	899	713	221

Table 1 summarises the remapped tooth-level annotations. The training, validation, and test splits contain 732, 182, and 182 images, with 8,790, 2,185, and 2,175 tooth ROI annotations, respectively. MGI3 and MGI2 are the dominant grades, whereas MGI0 and MGI1 have fewer supporting samples. This distribution motivates evaluation with both class-aggregated grading metrics and ordinal error measures.

The data audit covered 1,096 images. No missing image, missing label, or invalid annotation row was found. SHA-256 image hashing identified 14 duplicate image groups, including 8 groups crossing different splits. The data files do not include patient identifiers, and patient-independent splitting cannot be verified. This paper keeps the provided split without split repair, and reports duplicate images and missing patient identifiers as dataset limitations.

4.2 Evaluation Protocol

The evaluation protocol separates tooth localisation from matched-tooth MGI grading. For each image, the model outputs a prediction record $(\hat{b}_i, \hat{y}_i, \hat{s}_i, \hat{p}_i)$, where $\hat{b}_i$ is the tooth box, $\hat{s}_i$ is the tooth-detection score, $\hat{y}_i$ is the predicted MGI label, and $\hat{p}_i$ is the MGI probability vector. Predictions are first sorted by tooth score in descending order, and class-agnostic one-to-one tooth matching is then performed. Let $\mathcal{U}$ be the set of unmatched ground-truth tooth ROIs. For prediction i, the candidate set in the same image is

$$\mathcal{U}_i = \{j \in \mathcal{U} \mid \text{img}(j) = \text{img}(i)\}. \tag{24}$$

If $\mathcal{U}_i = \emptyset$, the prediction remains unmatched. Otherwise, the ground truth with the highest IoU is selected:

$$j^*(i) = \arg\max_{j \in \mathcal{U}_i} \text{IoU}(\hat{b}_i, b_j). \tag{25}$$

When $\text{IoU}(\hat{b}_i, b_{j^*(i)}) \geq 0.5$, the pair $(i, j^*(i))$ is added to the matched set $\mathcal{M}$, and $j^*(i)$ is removed from $\mathcal{U}$. Otherwise, the prediction is treated as unmatched. This matching step does not use the predicted MGI label, so whether a tooth ROI is detected and whether its MGI grade is correct are counted separately.

Detection performance is measured from tooth boxes and tooth scores, including AP50, AP75, detection recall at IoU 0.5, and mAP over COCO-style IoU thresholds $\Gamma = \{0.50, 0.55, \ldots, 0.95\}$:

$$\text{mAP} = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \text{AP}_\gamma \,. \tag{26}$$

Here AP_γ is computed from the precision–recall curve sorted by tooth score with 101-point interpolation. Unmatched ground-truth tooth ROIs are counted as missed teeth, and the missed-tooth rate is recorded.

MGI grading metrics are computed only on the matched set $\mathcal{M}$. The confusion matrix $C \in \mathbb{N}^{5\times5}$ is defined as

$$C_{ab} = \sum_{(i,j)\in\mathcal{M}} \mathbf{1}[y_j = a]\mathbf{1}[\hat{y}_i = b], \qquad a, b \in \{0, 1, 2, 3, 4\}. \tag{27}$$

Macro-F1, balanced accuracy, and quadratic weighted kappa (QWK) are computed from C. Ordinal grading error is further measured by grade distance on matched teeth:

$$\mathrm{MAE} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} |\hat{y}_i - y_j|, \tag{28}$$

$$\mathrm{AdjAcc} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \mathbf{1}\left[|\hat{y}_i - y_j| \le 1\right]. \tag{29}$$

When $\mathcal{M} = \emptyset$, the matched-tooth metrics are set to 0 under the evaluator's safe-division rule. This protocol avoids forcing missed tooth ROIs into an MGI grade and allows detection failure, missed teeth, and matched-tooth grading error to be analysed separately.

4.3 Experimental Setup

The comparative experiment includes Faster R-CNN R50 CE, Cascade R-CNN R50 CE, and GingiCasc. Both baselines use a ResNet-50-FPN backbone and a cross-entropy classification objective. GingiCasc uses the three-stage cascade tooth ROI detector with class-balanced weighting, MGI0/MGI1 class weight multipliers, and a severity-distance penalty. All test results are summarised over three random seeds, 20260613, 20260614, and 20260615.

For a metric value x_1, x_2, x_3 from the three random seeds, mean$\pm$std is defined as

$$\bar{x} = \frac{1}{3}\sum_{k=1}^{3} x_k, \qquad s = \sqrt{\frac{1}{2}\sum_{k=1}^{3}(x_k - \bar{x})^2}. \tag{30}$$

mAP and AP50 summarise tooth detection, whereas Macro-F1, QWK, MAE, AdjAcc, and BalAcc summarise matched-tooth grading.

The reference experiment compares GingiCasc with three same-family training configurations. The class-balanced and severity-distance settings remove corresponding training terms, whereas the ordinal-auxiliary setting is a reference configuration rather than a one-factor removal from the final reported loss. This design characterises the relation between severity-aware training choices and grading metrics without treating all rows as controlled single-component ablations.

4.4 Comparative Experiment Results

Table 2 reports the comparative experiment. GingiCasc reaches a Macro-F1 of 0.32 ± 0.01, higher than 0.20 ± 0.01 for Faster R-CNN R50 CE and 0.24 ± 0.01 for Cascade R-CNN R50 CE. For ordinal consistency, GingiCasc obtains a QWK of 0.39 ± 0.02, compared with 0.16 ± 0.01 and 0.23 ± 0.03 for the two baselines. Its MAE is 0.65 ± 0.06, lower than 0.89 ± 0.01 and 0.80 ± 0.05 for Faster R-CNN R50 CE and Cascade R-CNN R50 CE, respectively. After class-agnostic matching, GingiCasc improves class-aggregated grading metrics and ordinal agreement, while reducing mean grade-distance error.

Detection metrics differ across IoU thresholds. GingiCasc obtains an mAP of 0.39 ± 0.02, higher than 0.34 ± 0.04 for Faster R-CNN R50 CE and 0.29 ± 0.01 for Cascade R-CNN R50 CE. At AP50, Faster R-CNN R50 CE obtains 0.87 ± 0.03, GingiCasc obtains 0.86 ± 0.02, and Cascade R-CNN R50 CE obtains 0.80 ± 0.08. Thus, GingiCasc achieves the highest mAP across IoU thresholds, whereas its AP50 remains below that of Faster R-CNN R50 CE.

4.5 Reference Configuration Results

Table 3 reports the reference configuration experiment. Among the listed same-family configurations, GingiCasc obtains the best matched-tooth grading results for Macro-F1, QWK, MAE, AdjAcc, and BalAcc. The Macro-F1 and QWK values are 0.22 ± 0.01 and 0.18 ± 0.01 for w/o class-balanced loss, 0.24 ± 0.02 and 0.23 ± 0.04 for w/o severity-distance penalty, and 0.25 ± 0.01 and 0.28 ± 0.03 for the ordinal-auxiliary reference configuration. GingiCasc reaches 0.32 ± 0.01 and 0.39 ± 0.02 for the same two metrics, while reducing MAE to 0.65 ± 0.06.

4.6 Limitations

The comparative experiment and reference configuration experiment show that GingiCasc improves matched-tooth grading metrics over three random seeds, mainly in Macro-F1, QWK, MAE, and AdjAcc. Detection results depend more strongly on the IoU threshold: GingiCasc has the highest mAP, but its AP50 is lower than that of Faster R-CNN R50 CE. These results indicate that ordinal grading under class imbalance should be interpreted with class-aggregated metrics and grade-distance metrics together.

Two dataset limitations remain. The provided split contains cross-split duplicate images, which may overestimate real generalisation. Missing patient identifiers also prevent verification of a patient-independent split. This paper does not use external datasets and does not report cross-centre testing. The results should therefore be interpreted as three-seed test results on the Duy 2024 provided split under a unified evaluator, not as patient-independent or external-generalisation evidence.

5 Conclusion

This paper studies tooth ROI detection and MGI0–MGI4 ordinal grading from intraoral images. Unlike image-level gingivitis screening or lesion segmentation, the task output is restricted to tooth boxes and tooth-level MGI grades, matching the tooth-by-tooth structure of clinical severity recording.

Table 2: Comparative test results. Values are mean±std over three random seeds. Higher values are better for mAP, AP50, Macro-F1, QWK, AdjAcc, and BalAcc; lower values are better for MAE.

Method	mAP	AP50	Macro-F1	QWK	MAE	AdjAcc	BalAcc
Faster R-CNN R50 CE	0.34±0.04	**0.87±0.03**	0.20±0.01	0.16±0.01	0.89±0.01	0.80±0.01	0.29±0.01
Cascade R-CNN R50 CE	0.29±0.01	0.80±0.08	0.24±0.01	0.23±0.03	0.80±0.05	0.84±0.02	0.29±0.01
GingiCasc	**0.39±0.02**	0.86±0.02	**0.32±0.01**	**0.39±0.02**	**0.65±0.06**	**0.91±0.03**	**0.33±0.02**

Table 3: Reference configuration test results. Values are mean±std over three random seeds. Higher values are better for all metrics except MAE. The ordinal-auxiliary row is a reference configuration rather than a one-factor removal from the final reported loss.

Method	mAP	AP50	Macro-F1	QWK	MAE	AdjAcc	BalAcc
GingiCasc ordinal-auxiliary reference	0.34±0.03	0.86±0.03	0.25±0.01	0.28±0.03	0.78±0.06	0.85±0.03	0.31±0.01
GingiCasc w/o class-balanced loss	0.33±0.02	**0.88±0.02**	0.22±0.01	0.18±0.01	0.83±0.03	0.83±0.01	0.28±0.01
GingiCasc w/o severity-distance penalty	0.34±0.03	0.85±0.01	0.24±0.02	0.23±0.04	0.82±0.04	0.83±0.01	0.30±0.01
GingiCasc	**0.39±0.02**	0.86±0.02	**0.32±0.01**	**0.39±0.02**	**0.65±0.06**	**0.91±0.03**	**0.33±0.02**

On the Duy 2024 provided split, three-seed test comparisons show that GingiCasc improves Macro-F1, QWK, MAE, adjacent accuracy, and balanced accuracy over Faster R-CNN R50 CE and Cascade R-CNN R50 CE. GingiCasc also obtains the highest mAP, but its AP50 is lower than that of Faster R-CNN R50 CE. The evidence therefore supports tooth-level grading improvements and mAP improvement under the reported evaluation protocol.

The reference configuration experiment shows that GingiCasc has higher Macro-F1, QWK, adjacent accuracy, and balanced accuracy, and lower MAE, than the listed same-family configurations. This result is consistent with the severity-aware training objective used in the reported method, but it does not imply universal performance under other splits or external datasets.

The study remains constrained by the dataset. Image auditing identified duplicate image groups, and the data lack patient identifiers, so patient-independent splitting cannot be verified. No external dataset or cross-centre validation is reported. Future work should test the model under patient-level splits, multi-centre data, larger low-grade samples, and more consistent annotations. Because the task and annotations do not include lesion masks or lesion boxes, the results should not be interpreted as lesion segmentation or lesion-localisation evidence.

References

[1] Dima M. Alalharith, Hajar M. Alharthi, Wejdan M. Alghamdi, Yasmine M. Alsenbel, Nida Aslam, Irfan Ullah Khan, Suliman Y. Shahin, Simona Dianišková, Muhanad S. Alhareky, and Kasumi K. Barouch. A deep learning-based approach for the detection of early signs of gingivitis in orthodontic patients using faster region-based convolutional neural networks. *International Journal of Environmental Research and Public Health*, 17(22):8447, 2020. doi: 10.3390/ijerph17228447.

[2] D. Arnbjerg, S. Poulsen, and J. Heidmann. Evaluation of a photographic method for diagnosis of gingivitis and caries. *Scandinavian Journal of Dental Research*, 100(4):207–210, 1992. doi: 10.1111/j.1600-0722.1992.tb01743.x.

[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Evaluation measures for ordinal regression. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 283–287. IEEE, 2009. doi: 10.1109/ISDA.2009.230.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021. doi: 10.1109/TPAMI.2019.2956516.

[5] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008.

[6] Reinhard Chun Wang Chau, Guan-Hua Li, In Meei Tew, Khaing Myat Thu, Colman McGrath, Wai-Lun Lo, Wing-Kuen Ling, Richard Tai-Chiu Hsung, and Walter Yu Hang Lam. Accuracy of artificial intelligence-based photographic detection of gingivitis. *International Dental Journal*, 73(5):724–730, 2023. doi: 10.1016/j.identj.2023.03.007.

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[8] Sinéad Daly, Joon Seong, Charles Parkinson, Robert Newcombe, Nicholas Claydon, and Nicola West. A proof of concept study to confirm the suitability of an intra oral scanner to record oral images for the non-invasive assessment of gingival inflammation. *Journal of Dentistry*, 105:103579, 2021. doi: 10.1016/j.jdent.2020.103579.

[9] Jordi de la Torre, Domenec Puig, and Aida Valls. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154, 2018. doi: 10.1016/j.patrec.2017.05.018.

[10] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 4738–4747, 2019. doi: 10.1109/CVPR.2019.00487.

[11] Hoang Bao Duy, Tran Thi Hue, Tong Minh Son, Le Long Nghia, Luong Thi Hong Lan, Nguyen Minh Duc, and Le Hoang Son. A dental intraoral image dataset of gingivitis for image captioning. *Data in Brief*, 57:110960, 2024. doi: 10.1016/j.dib.2024.110960.

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

[13] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *Machine Learning: ECML 2001*, volume 2167 of *Lecture Notes in Computer Science*, pages 145–156. Springer, 2001. doi: 10.1007/3-540-44795-4_13.

[14] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. doi: 10.1109/ICCV.2015.169.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. doi: 10.1109/CVPR.2014.81.

[16] Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sánchez-Monedero, Francisco Fernández-Navarro, and César Hervás-Martínez. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016. doi: 10.1109/TKDE.2015.2457911.

[17] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

[19] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584, 2017. doi: 10.3348/kjr.2017.18.4.570.

[20] Wen Li, Yuan Liang, Xuan Zhang, Chao Liu, Lei He, Leiying Miao, and Weibin Sun. A deep learning approach to automatic gingivitis screening based on classification and localization in RGB photos. *Scientific Reports*, 11:16831, 2021. doi: 10.1038/s41598-021-96091-3.

[21] Wen Li, Enting Guo, Hong Zhao, Yuyang Li, Leiying Miao, Chao Liu, and Weibin Sun. Evaluation of transfer ensemble learning-based convolutional neural network models for the identification of chronic gingivitis from oral photographs. *BMC Oral Health*, 24(1):814, 2024. doi: 10.1186/s12903-024-04460-x.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

[25] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi: 10.1016/j.media.2017.07.005.

[26] Ralph R. Lobene, Thornton Weatherford, Norman M. Ross, Rosalyn A. Lamm, and Lewis Menaker. A modified gingival index for use in clinical trials. *Clinical Preventive Dentistry*, 8 (1):3–6, 1986.

[27] Harald Löe. The gingival index, the plaque index and the retention index systems. *Journal of Periodontology*, 38(6 Part II): 610–616, 1967. doi: 10.1902/jop.1967.38.6_part2.610.

[28] Mohammad Moharrami, Elaheh Vahab, Mobina Bagherianlemraski, Ghazal Hemmati, Sonica Singhal, Carlos Quinonez, Falk Schwendicke, and Michael Glogauer. Deep learning for detecting dental plaque and gingivitis from oral photographs: A systematic review. *Community Dentistry and Oral Epidemiology*, 53(6):617–632, 2025. doi: 10.1111/cdoe.70001.

[29] Shinya Murakami, Brian L. Mealey, Angelo Mariotti, and Iain L. C. Chapple. Dental plaque-induced gingival conditions. *Journal of Clinical Periodontology*, 45(S20):S17–S27, 2018. doi: 10.1111/jcpe.12937.

[30] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016. doi: 10.1109/CVPR.2016.532.

[31] Rafael Padilla, Wesley L. Passos, Thadeu L. B. Dias, Sergio L. Netto, and Eduardo A. B. da Silva. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3):279, 2021. doi: 10.3390/electronics10030279.

[32] Marco A. Peres, Lorna M. D. Macpherson, Robert J. Weyant, Blánaid Daly, Renato Venturelli, Manu R. Mathur, Stefan Listl, Roger Keller Celeste, Carol C. Guarnizo-Herreño, Cormac Kearns, Habib Benzian, Paul Allison, and Richard G. Watt. Oral diseases: a global public health challenge. *The Lancet*, 394 (10194):249–260, 2019. doi: 10.1016/S0140-6736(19)31146-8.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 91–99, 2015.

[34] Marta Revilla-León, Miguel Gómez-Polo, Abdul B. Barmak, Wardah Inam, Joseph Y. K. Kan, John C. Kois, and Orhan Akal. Artificial intelligence models for diagnosing gingivitis and periodontal disease: A systematic review. *The Journal of Prosthetic Dentistry*, 130(6):816–824, 2023. doi: 10.1016/j.prosdent.2022.01.026.

[35] Falk Schwendicke, Wojciech Samek, and Joachim Krois. Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, 99(7):769–774, 2020. doi: 10.1177/0022034520915714.

[36] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017. doi: 10.1146/annurev-bioeng-071516-044442.

[37] R. N. Smith, D. L. Lath, A. Rawlinson, M. Karmo, and A. H. Brook. Gingival inflammation assessment by image analysis: Measurement and validation. *International Journal of Dental Hygiene*, 6(2):137–142, 2008. doi: 10.1111/j.1601-5037.2008.00294.x.

[38] Guy Tobias and Assaf B. Spanier. Modified gingival index (MGI) classification using dental selfies. *Applied Sciences*, 10 (24):8923, 2020. doi: 10.3390/app10248923.

[39] Leonardo Trombelli, Roberto Farina, Claudio O. Silva, and Dimitris N. Tatakis. Plaque-induced gingivitis: Case definition and diagnostic considerations. *Journal of Clinical Periodontology*, 45(S20):S44–S67, 2018. doi: 10.1111/jcpe.12939.

[40] Chang Wen, Xueying Bai, Jiaxin Yang, Sihong Li, Xiaoxuan Wang, and Dong Yang. Deep learning based approach: Automated gingival inflammation grading model using gingival removal strategy. *Scientific Reports*, 14(1):19780, 2024. doi: 10.1038/s41598-024-70311-y.