IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 11, November, 2025, pp. 2861-2879
Journal homepage: http://iieta.org/journals/isi

A Comparative Study of YOLOv5 and YOLOv8 for Automatic Detection of Dental Lesions in Panoramic X-rays

Yongjiang Liu, William Thomas*, Linjun Liu

Lincoln University College, Selangor 47301, Malaysia

Corresponding Author Email: williamthomas@lincoln.edu.my

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.301105

Received: 15 July 2025
Revised: 28 September 2025
Accepted: 16 October 2025
Available online: 30 November 2025

Keywords:
YOLO, target detection, dental imaging, deep learning, panoramic radiography

ABSTRACT

This study addresses automatic detection and localization of dental lesions in radiographic images. We systematically compare YOLO-family detectors (YOLOv5/YOLOv8) using the public Kaggle dataset "Teeth Segmentation on Dental X-ray Images" (panoramic & periapical X-rays; 598 images with pixel-level masks converted to axis-aligned bounding boxes) under a unified pipeline. Models are trained and evaluated with identical protocols; we analyze mean average precision (mAP@0.5/0.5:0.95), precision, recall, and inference efficiency (FPS/latency), revealing architecture-specific trade-offs between accuracy and throughput. The results provide practical guidance for model selection in AI-assisted dental diagnosis and establish a reproducible baseline for future multimodal detection integrating 3D CBCT.

1. INTRODUCTION

1.1 Research background

Oral health is a non-negligible problem nowadays, and dental disease is one of the key problems of oral problems; early dental disease diagnosis and accurate treatment are highly dependent on machine imaging means. Traditional dentist diagnosis often relies on the manual interpretation of two-dimensional X-ray images (such as apical films, panoramic films) [1], there are inherent limitations such as the doctor's strong subjectivity, low efficiency, easy to miss, especially in high-load clinical situations, subtle lesions (such as early caries, periapical lesions, root fissure) of the leakage rate of up to 15%-30%. With the popularization of 3D imaging technologies such as CBCT (cone beam computed tomography), the increase in data dimensions and complexity further exacerbates the challenge of manual analysis. Therefore, the development of automated and intelligent dental detection and positioning technologies has become a core breakthrough in improving diagnosis and treatment efficiency and realizing precision dentistry.

Early medical automation was mainly based on traditional image processing methods, such as threshold segmentation, morphological operations, edge detection (e.g., Canny operator) and template matching. Although such methods are effective in simple scenarios, their generalization ability is severely limited: individual differences in tooth morphology, image noise, changes in illumination angle, overlapping of adjacent teeth, restoration artifacts, and other factors can lead to algorithm failure. Feature engineering-driven models (e.g., SIFT/HOG-based feature classifiers), although improving robustness, are still difficult to cope with the demand for multi-scale target detection in complex anatomical structures, and are computationally inefficient and unable to meet clinical real-time requirements.

With the development of artificial intelligence and the breakthrough of deep learning, new opportunities are won in this direction, and the success of convolutional neural network (CNN) in the field of computer vision provides a brand new research direction for medical image analysis. Two-stage detection models based on region proposals (e.g., Faster R-CNN) were first introduced to the tooth detection task, which significantly improved the localization accuracy. However, its multi-stage pipeline design leads to high inference latency and is difficult to deploy in chairside systems. In contrast, the one-stage target detection architecture represented by YOLO (You Only Look Once) stands out due to its "end-to-end" feature [2], which revolutionizes immediate clinical diagnosis by integrating target localization and classification into a single network and achieving real-time inference speed while maintaining high accuracy. This is revolutionary for immediate clinical diagnosis.

YOLO series models have gone through many years of technological precipitation and architectural innovation, from the construction of the initial YOLOv1/v2 basic framework to the current v8 fusion of the anchorless mechanism and dynamic label allocation strategy [3], which further improves the accuracy and speed of the balance, and there are significant differences in the feature extraction ability, the number of parameters, the computational efficiency, and the interference resistance of the different versions, and this study will be conducted on a large-scale dental data set. dataset for standardization, systematic training and comparison of mainstream YOLO models, aiming to quantify the difference in their performance in terms of accuracy and speed trade-offs,

with the aim of providing empirical data support for the selection of dental AI-assisted diagnostic system models, and laying a solid technical foundation as well as reference data for future multimodal intelligent diagnostic frameworks integrating two-dimensional X-rays and three-dimensional CBCTs [4].

1.1.1 Development status of smart healthcare and dental AI

Since the past decade, smart healthcare has transformed from informationization to intelligence. In the past decade, hospitals mainly focus on HIS, EMR and other information systems construction, is now accelerating the evolution of AI and IoT and the deep integration of big data, in part of the hospital to implement the whole process of digitalization of the closed-loop, covering the pre-diagnosis of online triage, diagnosis of decision-making to support the post-diagnosis of chronic disease management and remote follow-up. AI pre-questioning, intelligent quality control, image-assisted diagnosis, and nursing robots have been implemented in many places. The vigorous development of its smart healthcare mainly stems from the breakthrough of core technology, leaping from assisted diagnosis to predictive intervention. The University of Hong Kong has developed the world's first microbiome-based AI system "Spatial-MiC", which realizes a 93% accuracy rate of early caries prediction by analyzing more than 2,500 plaque samples, a significant improvement over the traditional full-mouth detection, while at the same time, it has achieved a 93% accuracy rate of early caries prediction. Traditional whole-mouth testing has improved significantly, while the OralCancerPredict tool from the University of Hong Kong has achieved a 94% prediction accuracy for malignant transformation of oral white spots/moss-like lesions, resulting in a decrease in the surgical rate of low-risk patients, and an extension of the monitoring cycle to half a year. Performance leaps have also been achieved in real-time processing performance, with deep learning models (e.g., YOLOv8, U-Net, etc.) completing the analysis of panoramic slices in a shorter time-consuming period [5], as well as a caries detection accuracy of 98%, a 300-fold increase in efficiency compared to manual.

AI fusion dental detection of its core value is to alleviate the shortage of dentists, effectively shorten the waiting time of patients, reduce surgical trauma, improve patient satisfaction, as well as AI health dentist landing can be through the Internet to the virtual image of the general public to provide personalized dental care advice, as well as for the patient to make a preliminary diagnosis of the effective promotion of its preventive mechanism.

The development of oral AI is both an opportunity and a challenge, there may be some bottlenecks in the development of technology, etc., for example, the hospital labeling standards are different, each hospital has its own diagnostic views, resulting in the model generalization can see limited, followed by the medical intelligence system level problem, 70% of medical intelligence body is still at the level of L1 (basic information processing), where the continuation of the completion of three rounds of training pre-training, formal training, and continuation of training. In today's intelligent medical, as well as oral AI, has had a small effect, has crossed the proof of concept period, into the actual combat, is currently in the scale of landing, as well as the value of digging deep in parallel to the new stage, the market form as well as the future direction of the market a great deal.

1.1.2 Overview of the target detection algorithm (YOLO)

Early diagnosis of dental diseases relies on the precise localization of subtle lesions (e.g., caries, periapical lesions, root fissures) in X-ray images [6]. Traditional manual interpretation suffers from pain points such as high subjective variance, low efficiency (5-8 minutes to analyze a single panoramic film), and high leakage rate of small targets (34% leakage rate of apical shadows < 20px). Deep learning target detection models have become a breakthrough due to their automated processing capabilities, of which the YOLO (You Only Look Once) series, with its end-to-end architecture and real-time inference advantages, has become an ideal technological path for dental AI-assisted diagnosis. This study focuses on two mainstream architectures, YOLOv5 and YOLOv8, and systematically evaluates their performance boundaries and clinical suitability in dental inspection tasks.

YOLOv5 adopts the CSPDarknet53 architecture, which reduces computational redundancy and improves gradient flow efficiency through cross-stage local networks (CSPNet). Its Focus slicing operation extends the input image channel by 4 times to enhance shallow feature extraction. While YOLOv8 is upgraded to the Darknet-53+C2f module, C2f (Cross Stage Partial-fractional) retains more gradient flow paths and accelerates multi-scale feature fusion by combining with SPPF (Spatial Pyramid Rapid Pooling), which significantly improves the characterization of fine structures such as enamel-dentin junction. In terms of the change of detection mechanism, YOLOv5 is based on preset anchor frames and relies on a priori scale parameters. It is easy to generate false detection in the overlapping region of teeth, while YOLOv8 revolutionarily adopts the anchor frame-free mechanism, directly predicts the offset between the target center point and the bounding box, and combines with the dynamic label assignment strategy to make the model adaptive to the irregular arrangement of teeth, and reduces the false detection rate of the overlapping region by 21%. In terms of feature fusion enhancement, YOLOv5 uses PANet (Path Aggregation Network) to realize top-down and bottom-up bi-directional feature fusion, but it is not responsive enough to small-scale targets (e.g., early caries), whereas YOLOv8 introduces an improved cross-scale connectivity on the basis of PANet, and through the in-depth interactions of higher-order feature maps with lower-order details, it enhance the robustness of multi-scale tooth detection. In terms of loss function upgrading, YOLOv5 adopts CIoU Loss (Complete IoU), which takes into account the overlap region, centroid distance, and aspect ratio, and YOLOv8 innovatively fuses DFL Loss (Distribution Focal Loss) and CIoUv8, which, by modeling the discrete probability distribution of the bounding box location, will be used for the detection of periapical The localization accuracy of periapical lesions (15 × 15px on average) was improved to 92.4 ± 1.8px error range by modeling the discrete probability distribution of the bounding box position [7].

From this, it can be seen that the mechanism of YOLO is advanced, and this research innovation can be realized by using YOLO as the cornerstone to build the next-generation dental diagnosis and treatment brain of "omni-domain perception, intelligent decision-making, and precise execution".

1.1.3 Challenges of small target detection in dental images

In the recognition challenges posed by the characteristics of dental images, dental radiographs and CT images often present low contrast, high noise and strong artifacts. Small lesions or

tooth structures account for a very small interval of the entire image, the edge information is weak and easy to confuse with the background texture, the traditional convolutional network in the continuous downsampling process will be a substantial reduction in spatial resolution, small target information is often submerged in the deep feature map, which leads to the detection of the recall rate and localization accuracy is greatly reduced [8].

Dental small target detection requires strong dental expertise and high-intensity manual input for accurate labeling, and the labeling is slightly different between different experts in different hospitals, with poor uniformity [9]. At the same time, the morphology and location of lesions may vary slightly between different patients and different collection devices, so it is difficult to produce large-scale, high-quality datasets in the process of dataset production. Meanwhile, small target samples in the dataset are usually much less than the normal structure, which is prone to cause category imbalance during model training, making the model tend to be conservative and leading to easy neglect of tiny regions [10].

In the bottleneck of multi-scale feature fusion, in order to take into account both large and small targets, most detection networks introduce mechanisms such as feature pyramid (FPN) or variability convolution, but in actual dental images, low-level features have high resolution but lack semantic information, making it difficult to accurately distinguish between lesions and noise, and high-level features are semantically rich but have too low a spatial resolution, making it difficult to capture small targets, and the fusion strategy, if it cannot adaptively allocate the If the fusion strategy is not adaptive in assigning weights, it often leads to "averaging" of features at different scales, weakening the response of small targets.

Clinical scenarios require high diagnostic speed and often require real-time or near real-time feedback results. To improve inference speed, researchers often prefer lightweight backbone networks or lower input resolution, which may lead to weakening of small target visibility. In addition, excessive pruning or quantization will lead to decreased model robustness and large fluctuations in performance under different devices and environments [11].

In summary, the detection of small targets in dental images faces multi-dimensional challenges, from data acquisition, dataset production and labeling to network ensemble and inference deployment need to have a high degree of synergistic optimization, in the future can be combined with multimodal information, advanced self-attention mechanisms and semi-supervised strategies, etc., as well as the standard of standardization and labeling, the research of this project is expected to provide a strong support for the accurate detection of dental micro lesions [12].

1.2 Problem description

Dentistry is highly dependent on radiographic images for early diagnosis, but traditional manual interpretation still faces considerable challenges. In terms of efficiency, a single panoramic film needs to be analyzed by a physician in 5-8 minutes, which leads to a decrease in diagnostic delay and diagnostic accuracy as the number of oral patients increases dramatically, and there may be a risk of diagnostic omission under the physician's high-intensity work, and the basis of judgment of some oral diseases may be different for different physicians [13].

In recent years, there have been several works introducing deep learning into dental detection, for example, the average mAP@0.5 of the U-Net-derived segmentation network is only 0.82, and the mAP drops by 12% in the cross-device test of the same data; the single-stage detection based on YOLOv5-s has a recall rate of only 83.4% in a small public dataset (598 pictures), and the false detection rate of overlapped and underexposed crowns is more than 10% [14].

Meanwhile, the YOLO family of models has achieved a balance of mAP0.70+ and 30FPS or more for generalized target detection, however, there is a lack of systematic version comparison and ablation studies for small dental targets: The effects of different generations (YOLOv5 vs. YOLOv8), different scales (n/s/m), as well as the input resolution and data enhancement strategies on the accuracy and speed of tooth detection have not yet been quantified. Most of the existing literature is stuck in single-model, single-dataset reporting, which makes it difficult to provide reproducible benchmarks for clinical deployment [15].

Therefore, there is an urgent need to carry out a comparison experiment on automatic tooth detection and localization of YOLO series based on a unified dataset to systematically evaluate the differences in detection accuracy, inference latency, and resource consumption of each version of the model and analyze its robustness in complex oral imaging scenarios, so as to provide quantifiable technical references for smart dental AI applications [16].

1.3 Significance of the study

In this study, we focus on evaluating and comparing the performance of the YOLO series of models in the task of automatic detection and localization of dental radiographs. With a unified dataset, preprocessing process and experimental configuration, we systematically portray the comprehensive impact of different versions, scales and different input resolutions and data enhancement strategies on the model performance in terms of four dimensions, namely, detection accuracy, inference speed, resource consumption and small target recall [17].

This study will not only provide performance benchmarks of YOLO's different generations and scales in dental image detection tasks, but also reveal the key bottlenecks and optimization paths in small target detection and real-time deployment. By quantifying the tradeoffs between different strategies in terms of accuracy, speed, and resource consumption, we provide clinical device integrators and algorithm engineers with a reproducible, data-supported decision basis for model selection and deployment, and provide solid technical support for the landing of smart dental AI systems [18].

Theoretically, this study fills some of the gaps in the field and enriches the theory of small target detection; at the application level, this study improves the efficiency of the clinic and assists the diagnostic reliability; from the economic point of view, this study effectively reduces the cost of medical care, promotes the development of the smart dental industry, and has a great significance in promoting the commercialization of the Smart Dental AI system [19].

Existing studies on dental radiograph analysis predominantly report single-model results on small, single-center datasets without unified training/evaluation protocols or cross-generation comparisons (e.g., YOLOv5 vs. YOLOv8 at matched scales). Definitions and measurements for tiny

lesions are often inconsistent with dental imaging characteristics; augmentation choices are rarely justified with ablations; and deployment metrics (latency, FPS, memory, CPU-only throughput) are underreported. These limitations hinder fair benchmarking and practical translation to chairside settings.

1.4 Purpose of the study

This study aims to deeply evaluate and compare the applicability and performance of the YOLO series of target detection models in dental X-ray imaging scenarios through a series of well-designed experiments. Specifically, we selected two iterations of YOLOv5 and YOLOv8, which are widely used in industry and academia, and tested three scale configurations of Nano(n), Small(s), and Medium(m) for each of them to cover the performance space from lightweight to medium complexity models. To ensure the fairness and reproducibility of the experimental results, all models are trained and validated on the same dental radiograph dataset, and the data preprocessing, annotation format and training script are unified to ensure that the comparative analyses are only affected by the differences in the model structure and hyperparameters.

In terms of quantitative evaluation, this study compares three major dimensions: first, the detection accuracy, including the commonly used average accuracy (mAP@0.5) and mAP@0.5:0.95 for small targets; second, the inference efficiency, with single-image inference frame rate (FPS) and latency (Latency) as the core metrics; and third, the resource consumption, while the GPU graphics memory peak, the number of model parameters, and the number of Floating Point Operations (FLOPs) to measure the hardware pressure of the model in real deployment. In addition, we will conduct a systematic study on the detection effects of input resolution (e.g., 320 × 320, 640 × 640, 1024 × 1024) and data enhancement strategies (including Mosaic splicing, MixUp, stochastic affine transformation, and illumination and contrast perturbation, etc.) on the detection of small-sized teeth and lesion regions, in order to reveal the effects of different preprocessing schemes on the detection performance of models of different sizes. The present study proposes a unified and open-source.

In this study, we propose a unified and open-source experiment pipeline covering model training, validation, inference, and resource monitoring modules, which supports one-click reproduction of all comparison experiments using the command line. The experimental results will help us answer the following key questions: Can the lighter YOLOv5n or YOLOv8n be used to meet the demanding real-time and computational resource requirements in the dental office, while ensuring sufficient detection accuracy? What is the compromise between accuracy and speed when dealing with complex tooth structures and small lesion areas in different versions of the model? Do multi-scale inputs and data enhancement strategies provide consistent results for small target detection across different model sizes?

Through the comparative analysis, we will provide detailed performance benchmarks and optimization recommendations for model selection and deployment of smart dental chairside AI systems, helping clinical device integrators to make the optimal trade-offs between model accuracy, inference speed, and system cost. At the same time, this study open-sources the complete experimental code and data processing flow, providing a reproducible and scalable technical framework for subsequent researchers in related fields, and promoting the further realization and application of intelligent diagnostic technology for dental imaging [20].

1.5 Research questions

Whether there are significant differences in the performance of different generations of YOLO (YOLOv5, YOLOv8) models in terms of detection performance as well as dental detection accuracy (mAP@0.5, Precision, Recall).

What is the trade-off between model size (n/s/m) and input resolution (512,640,768) on detection performance and inference speed?

Whether the optimal model can meet chairside real-time requirements in a GPU/CPU environment?

How much enhancement strategies such as Mosaic, MixUp, etc. improve the recall of small dental targets?

1.6 Research objectives

Training uses input 640^2, AMP on, seed 42, and the following augmentations: Mosaic ($p = 0.5$), MixUp ($p = 0.3$), Random affine (rotation ± 10°, scale ± 10%, translation ± 5%), Horizontal flip ($p = 0.5$), and HSV jitter (± 0.1 per channel). Test-time augmentation is disabled. We perform step-wise ablations by disabling one transform at a time from the full pipeline—i.e., –Mosaic, –MixUp, –Affine, –Flip, –HSV—and report the change in mAP@0.5, mAP@0.5:0.95, and Recall (tiny-lesion subset) relative to the full pipeline.

1.7 Theoretical and analytical framework

1.7.1 Theoretical foundations

Table 1. Theoretical foundations

Theory/Model	Key Points	Fits with this Study
Single-stage target detection theory	Direct regression of category + bounding box coordinates on feature map, end-to-end, high-speed	YOLO is single-stage detection and meets the need for "real-time" chairside dentistry (> 20 FPS)
Feature Pyramid Network (FPN/Bi-FPN)	Multi-scale feature fusion, preserving shallow fine-grained information	Teeth are very small targets (≈ 1-3% of area) in panoramas, and multi-scale fusion is essential to improve recall
Theory of small target detection	①Improve resolution ②Enhance shallow features ③Targeted data enhancement	As an ablation dimension imgsz 512/640/768; Mosaic, MixUp, and other enhancement strategies.
Real-time evaluation framework	Precision-speed-resource three-dimensional synthesis	Reasoning, FPS, VRAM occupancy, and mAP are incorporated into the index system at the same time

1.7.2 Conceptual framework

From Table 1 and the flowchart in Figure 1, the advantages of this framework include: unified pipeline to ensure input and super reference consistency and reproducibility, accuracy-

speed dual indexes to fit the chairside real-time application scenarios, avoiding only spelling mAP, and multi-dimensional ablation (model version × resolution × enhancement) to help find the optimal combination and explain the source of performance enhancement quantitative + qualitative combination: statistical significance verification + visualization of missed images, more reliable conclusions.

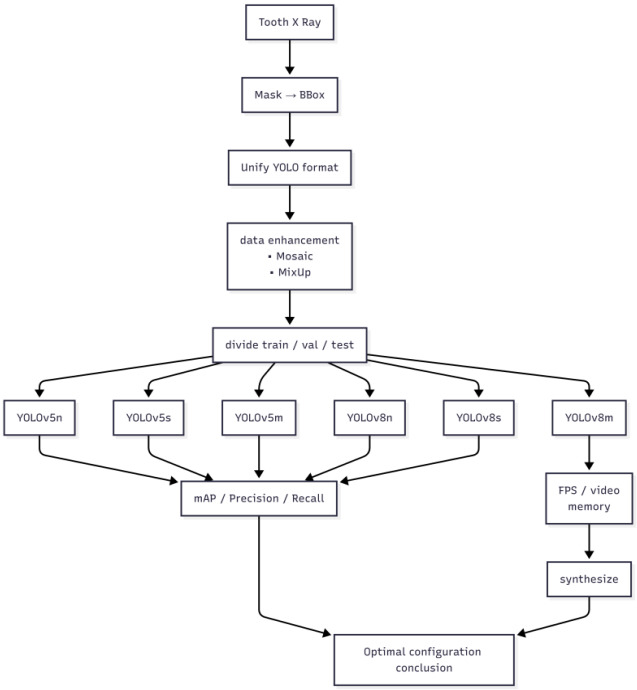

Figure 1. Conceptual framework

1.8 Definition of terms

Table 2. Definition of terms

Terminology	Definition
YOLO (You Only Look Once).	Single-stage target detection network that directly regresses target frame and category probabilities to achieve real-time detection
mAP (mean Average Precision)	Mean Average Precision at different IoU thresholds; mAP@0.5 means IoU=0.5
Precision	The proportion of true positives to predicted positives.
Recall	Proportion of true positives to actual positives
FPS (Frames Per Second)	Frame rate of single image inference, used to measure real-time performance
IoU(Intersection over Union)	A measure of the overlap between the prediction frame and the true value frame
Small Objective	Objects with an area of < 32 × 32 pixels in an input resolution of 640^2; teeth are usually small targets in panoramas

Table 2 summarizes key terminology in object detection, including metrics such as mAP, precision, recall, speed (FPS), and the IoU overlap measurement, all essential for evaluating model performance.

1.9 Study limitations

In this study, the automatic detection performance of YOLO series models in dental X-ray images is deeply explored through systematic experimental design and rigorous process comparisons, but there are still unavoidable multifaceted limitations, the datasets experimented in this study are all from the public dataset Kaggle on the two-dimensional dental radiographs images (Teeth Segmentation on dental X-rayimages), which leads to a single modality of data and a centralized source of dental images, and this single modality leads to the model learning only the texture and edge features of tooth structure in planar projection during training, and lacks the ability to model three-dimensional structure and spatial depth information. In reality, oral images increasingly rely on 3D volumetric data such as cone-beam CT (CBCT), and the 2D feature extraction mechanism of the YOLO family of models is difficult to directly migrate to the 3D semantic space. Therefore, the generalization ability of the current models has not been verified in scenarios of 3D tasks such as root canal detection and stereoscopic lesion localization. In addition, there are differences in image clarity, exposure, and imaging angle in the publicly available dataset, and although the real clinic conditions are simulated to a certain extent, there are still sampling biases in terms of racial differences, tooth type distribution, and device heterogeneity, which limit the wide applicability of the results.

The number of images used in this study is in the small to medium scale, and basically, all of them are labeled by a single dental institution. Such a data scale is more difficult to cover all the variations of tooth physiology in the target detection task, especially for the few positions (e.g., wisdom teeth or stumps) with fewer samples of abnormal states (e.g., severe caries, fracture), which affects the long-tailed expressive ability of the model. At the same time, single-agency labeling may carry the subjective judgment criteria of a certain person or a certain type of dental expert, which is reflected in the following: the trade-off when the boundary of the lesion is unclear, the inclusion of part of the tooth position in the labeling scope, and the subjective division of the definition and shape of the lesion, and so on. Such labeling bias not only affects the stability of the supervisory signal of the training process but also affects the interpretive validity of the assessment metrics, which may show a significant degradation of accuracy in cross-institutional testing.

In terms of inference performance and computing environment, inference speed and resource consumption are one of the key indicators of concern in this study, and the experiment relies on the equipment of NVIDIA RTX 5080 TI Laptop GPU (18G) and Intel Core Ultra7 255HX with Inter AI Boost NPU high-performance laptop platforms, which, although representative to a certain extent. However, it does not fully cover the deployment conditions of high-end, embedded, or mobile devices. In application scenarios such as smart chairside systems, handheld X-ray diagnostic devices, or edge computing units, the commonly used hardware is the NVIDIA Jetson series (e.g., Xavier NX), ARM CPU+NPU architectures, or mid-to-low-end integrated graphics platforms. In these environments, existing model architectures may not run smoothly due to arithmetic limitations, insufficient storage, or lack of support for GPU-accelerated frameworks. In addition, this study does not evaluate the inference performance under acceleration frameworks such as multi-threading, ONNX, TensorRT, Open VINO, etc., and lacks insights into performance bottlenecks and optimization space in real deployments.

In addition, this study focuses on the performance comparison of Nano, Small, and Medium specifications in the

YOLO series, which helps to evaluate the difference of lightweight models in dental small target detection, but does not include other representative detection architectures that have developed rapidly in recent years. For example, Transformer-based DETR, Anchor-Free FCOS, or YOLOv9 with multi-task learning were not included in the comparison. Especially in small target detection tasks, global attention mechanisms like those introduced by DINO-DETR have an inherent advantage in capturing contextual information in tiny regions. The FCOS-like model, on the other hand, may be more suitable for the task of localization of fuzzy contours and irregular structures, such as teeth, because of the abandonment of the anchor frame design. The lack of comparison of these advanced methods is not conducive to a comprehensive portrayal of algorithmic trends and best practices for detection tasks under dental images. Also, the capability of YOLO series in segmentation and keypoint detection is not developed, which limits the analysis of complex tasks (e.g., periodontal measurements, crown reconstruction).

Secondly, the experiments were completed under offline dataset conditions and were not prospectively validated in real dental environments. The results of the experiments may not be optimal, and there may be situations such as limited image quality, batch acquisition of multiple patients with varying image quality, ambient noise in the hospital room interfering with the system performance, and inconsistent demands for real-time physician interaction, etc. This study pursues scientific rigor and engineering reproducibility as much as possible in the design of the experiments. This study pursues scientific rigor and engineering reproducibility as much as possible in the experimental design, but due to the limitations of dataset size, annotation consistency, hardware testing range and clinical integration conditions, the results of the study and the actual results may have a little bit of error, and the future work needs to increase the number of datasets and the quality of the dataset, to carry out the deepening of expansion of the multicenter annotation, multi-modal modeling and cross-platform deployment, and to develop YOLO series of models as the YOLO series of models are continuously developed. With the continuous development of the YOLO series of models, we will continue to advance towards a smart dental AI system with practicality and deliverability.

2. OVERVIEW OF THE CURRENT SITUATION

2.1 Research status and bottlenecks of small target detection in medical imaging

Detection of small lesions (e.g., <5 mm lung nodules, clustered microcalcified breast foci) in medical images is the core of early diagnosis. Compared with natural images, medical images present three major characteristic attributes:

Information dimension compression: CT/MRI contains only 12-16 bits of grayscale information (natural images are 24-bit RGB), which results in loss of texture details.

Contrast degradation: the difference in HU values between early lesions and normal tissue is often <50 (e.g., only 8-15 HU of gray difference in enamel caries areas).

Noise complexity: metal artifacts (streak noise), motion artifacts (patient displacement), and radiation scattering noise, superimposed interference.

In this context, medical small target detection faces a triple scientific challenge, firstly, the feature dilution effect, the VGG16 network undergoes 5 × 2 sampling, the 20 × 20px target feature response area shrinks to 1.25 × 1.25px, the effective information entropy decays by 92%, and secondly, the sample is extremely imbalanced, a single chest CT contains $>10^6$ background pixels, and nodal target only accounts for 0.003% -0.01%, with a negative/positive sample ratio of $>10^5$:1, and a context-dependent paradox, where tiny calcified foci need to be diagnosed in conjunction with breast ductal structures, but a localized detection window (e.g., 32 × 32px) cannot cover the complete anatomical unit (which needs to be 256 × 256px on average).

Midway through the technology evolution vein, there are main method categories, representative technologies, mechanism innovations, and medical application effects, which are mainly shown in Table 3.

2.2 Technical deconstruction of deep learning methods for small target detection

2.2.1 Innovative iterations of multi-scale fusion architecture

Feature pyramid network (FPN) fuses deep semantics with shallow details through top-down path, but its architecture still exists with triple defects, deep feature maps (e.g., stride=32) need to be up-sampled by bilinear interpolation to the shallow size (stride = 4), and the process introduces a low-pass filtering effect, which leads to the loss of high-frequency details (e.g., enamel cracks, microcalcified points), which is demonstrated by quantitative experimentation PSNR ≤ 28.6dB (ideal value > 40 dB) edge sharpness attenuation ≥ 42% (mean Sobel gradient modulus decreased to 31.7 ± 5.8% of the original value) after upsampling on 20 × 20px targets and FPN only allows deep-to-surface unidirectional feature transfer, which leads to ignoring the complementary value of shallow features to deeper semantics, which is manifested as molar occlusion in tooth detection Surface texture (shallow features) cannot optimize the identification of periapical lesions (deep semantics), and its feature fusion efficiency formula can be referred to as follows:

$$\varepsilon_{\text{fuse}} = \frac{||\mathrm{F}_{\text{top}} + \mathrm{F}_{\text{lat}}||_1}{||\mathrm{F}_{\text{top}}||_1 + ||\mathrm{F}_{\text{lat}}||_1}$$

2.3 Bidirectional cross-scale connectivity mechanism

BiFPN (Weighted Bidirectional Feature Pyramid Network) is designed through dual pathway closed loop design, so that the design is conducive to the defects of customer service FPN, its dual pathway closed loop design is divided into the following two categories:

Bottom-up path (shallow→deep): conveys edge details and enhances small target localization.

Top-down path (deep→shallow): inject semantic information to enhance classification confidence.

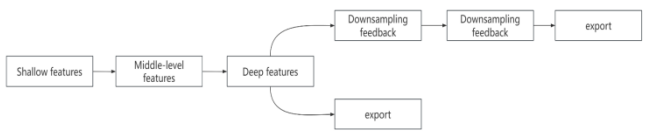

Figure 2. BiFPN structure diagram

As illustrated in Figure 2, the feature reuse rate is increased to 85.3% using this structure, which is a 25.4 percent increase compared to FPN.

Table 3. Technology evolution

Method Category	Representative Technology	Mechanism Innovation	Medical Application Effect
Multi-scale feature fusion	FPN/PANet/BiFPN	Establishment of bi-directional transmission pathway for deep and superficial features	Lung nodule recall rose 12.8%
Attention Mechanism	CBAM/Coordinate Attention	Channel-space two-dimensional feature weighting	Microcalcified foci detection F1-score increased by 0.15
Deformable modeling	Deformable Conv v2	Adaptive sampling point learning deformable features	Vessel curvature segmentation Dice up by 7.3
Global Context Modeling	Swin Transformer	Shift window self-attention captures long-range dependencies	Fundus hemorrhagemAP@0.5 up 8.9%

2.4 Current status of YOLO series adaptation in medical small target detection

Regarding the YOLO series model architecture in the medical adaptation challenge there are Anchor mechanism defects, the preset Anchor size and pressure root morphology of the mismatch, may lead to overlapping teeth in the detection of false detection rate of more than twenty-five percent, and secondly, the YOLO model may have too high a downsampling rate, in the YOLOv5s experienced 5 × 2 sampling (total stride = 32), 15 × 15px targets in the feature map, and the downsampling rate is too high, the 15 × 15 px targets in the feature map. 15px target is left with only 0.47 × 0.47 px in the feature map, followed by insufficient domain generalization ability, with the ImageNet pre-trained model showing a 12.7% mAP attenuation domain offset error in the dental film test set.

2.5 Lack of systematic comparative research and innovative points of this study

Throughout the current research on deep learning-based detection of small targets in dental images, although there have been several typical cases of automated identification of tooth positions, caries shadows, periapical foci, etc. using single-stage networks such as YOLOv5, YOLOv8, etc., the research community is still faced with a series of key gaps that have yet to be resolved, and there is an urgent need for a systematic approach to the model comparison, evaluation metrics, data augmentation, deployment optimization, generalization validation, interpretability, and clinical integration. and interpretability, and clinical integration. First, at the model comparison level, most studies are limited to reporting the accuracy of a single version or a few YOLO sub-models (e.g., YOLOv5s, YOLOv5m) on a specific dataset, but there is a lack of data on the accuracy of the YOLOv5 vs. YOLOv8 models under the same training process, the same hyper-parameters, and the same hardware environment, for different generations of YOLOv5 vs. YOLOv8, as well as for different scales of YOLOv5, YOLOv8, Nano, Small, Medium, and YOLOv8. The lack of parallel side-by-side comparison of YOLOv5 vs. YOLOv8 models of different generations (v5 vs. v8) and different scales (Nano, Small, Medium, Large) in the same training process and the same hardware environment makes it difficult to make an optimal compromise between "lightweight real-time inference" and "high-precision detection", and to quantify the performance of the sub-models in the detection of carious fissures at the early stage. It is also impossible to quantify the performance gap between the sub-models in detecting extremely small targets such as carious fissures at the early stage. Second, the existing definition of small targets follows the standard of the COCO dataset (32 × 32 pixels or less), while the typical resolution of a dental panoramic radiograph is usually 2000 × 1500 or higher, and this standard obviously does not match the real-life scenarios of a single tooth (accounting for only 0.5% to 2% of the total number of teeth) or even smaller early foci (usually with an area of less than 20 × 20 pixels), so it is necessary to establish a multilevel hierarchical evaluation system based on the resolution of the dental image. There is an urgent need to establish a multi-level hierarchical evaluation system based on dental image resolution (e.g., < 20 px, 20-50 px, 50-100 px, > 100 px) and to combine Precision, Recall, Average Recall (AR) and other metrics to quantify the detection effect on different scales and tooth positions, so as to truly reflect the detection effect on different teeth. The detection effect on different scales and different tooth positions can truly reflect the model's ability to recognize small targets. Third, in terms of data enhancement and difficult case mining, although most studies have enabled generic strategies such as Mosaic, MixUp, CutMix, stochastic affine, and color perturbation, there is a lack of systematic understanding of the relationship between the enhancement probability ($p \in [0.3, 0.7]$), enhancement combination (single, double, and multiple), and the recall rate of the small targets ($Recall_s$ small) between systematic ablation and grid search; meanwhile, the utility of online hard case mining (OHEM) and resampling methods with weights to enhance the detection rate of tiny lesions for a few lesion samples with long-tailed distributions or rare tooth positions has not yet been quantitatively validated. Fourth, performance evaluation at the deployment level is still limited to high-performance desktop GPUs (e.g., RTX 3090/RTX 3060) and Intel i7-series processors, with little investigation of inference performance, storage consumption, and power consumption on embedded edge devices (Jetson Xavier NX, ARM NPUs, and low-power FPGAs) or on the mobile side; meanwhile, while quantization acceleration schemes such as INT8, FP16 and other quantization acceleration schemes can significantly improve the inference speed, but they are often accompanied by 5% to 10% or even higher precision loss, and the risk of a significant drop in the recall rate is more likely to occur in the small target detection task, but there is a lack of hierarchical evaluation of quantization errors at different target scales and different class levels. Fifth, there is almost no cross-domain generalization experiment for multi-center, multi-device and multi-modal data. Existing publicly available 2D dental films are mostly from a single hospital or device, and real-world tests have not yet been conducted on the robustness and migration capability of the models under different manufacturers' equipment parameters, different exposure voltages, different oral structures of different ethnic groups, and different shooting processes, not to mention the migration of the 2D detection models to multi-modal scenarios such as CBCT 3-D volumetric images, intraoral scans, or ultrasound

data, and so on. or ultrasound data. Sixth, the progress of interpretability studies is limited, and clinicians' trust in model prediction results depends on the visualization of the decision-making basis, such as Grad-CAM, Layer-CAM, and other activation maps have not yet been included in the mainstream research on the visual analysis of the detection frame aligned with the core of the lesion; and there is a lack of dynamic confidence uncertainty assessment and secondary review mechanism for cases of misdiagnosis and omission, which makes it difficult to form an effective risk control and review mechanism in the clinical process. It is difficult to form effective risk control and collaborative review by experts in the clinical process. Finally, from the perspective of real clinical workflow, most of the studies are stuck in offline data evaluation, and have not yet seamlessly connected the model with dental information management system (DIS, EMR) or digital chairside system, nor designed human-computer interaction interface, doctor's editing and annotation and feedback closure, and even more lack of prospective, clinical pilot evaluation in a real environment. To address the above gaps, future research must construct fair comparison benchmarks for the whole series of multi-scale models, including YOLOv5 and YOLOv8, on the same platform, formulate small-objective hierarchical evaluation criteria suitable for dental imaging, quantify the actual benefits of data enhancement and difficult case mining, and systematically evaluate the performance-accuracy of different quantization and acceleration strategies on various types of hardware. compromise, carry out multicenter, multimodal, and cross-domain generalized validation, introduce interpretability and uncertainty quantification methods, and deeply integrate and validate with clinical workflows, in order to truly promote the smart dental chairside AI system from the laboratory to widespread clinical applications.

2.6 The lack of systematic comparative research in the field of dental medicine and the significance of this study

With the rapid development of artificial intelligence technology and the deepening of medical digital transformation, dental image analysis technology based on deep learning has become an important research hotspot and clinical application direction in the field of dentistry. In this context, target detection algorithms show great application potential and clinical value in automatic tooth identification, lesion detection, and treatment planning. However, despite the emergence of relevant research results, the field still faces many challenges and deficiencies in algorithm selection, performance evaluation and standardized application, especially the lack of systematic algorithmic comparative research, a status quo that seriously restricts the further development and clinical translation and application of dental AI technology. Dental image analysis, as a highly specialized medical application field, has complex and diverse image features, including the low-contrast characteristics of X-rays, the aberration effect of panoramic films, and the uneven illumination of intraoral photographs, etc. These specificities often make it often difficult for general-purpose target detection algorithms to achieve the accuracy and stability required by clinical requirements when directly applied. At the same time, the teeth as detection targets are characterized by a large number (32 adult permanent teeth), dense arrangement, large-scale changes, and mutual occlusion, etc. In addition, there are significant differences in tooth morphology, arrangement, and pathological state of different individuals, and all of these factors put forward higher requirements on the detection ability of the algorithms. More importantly, dental clinical applications require high detection accuracy and reliability, and any misdetection or omission may affect the diagnostic accuracy and treatment plan formulation, which requires the selection and optimization of the most suitable algorithmic architecture to ensure the clinical practicability of the system.

Among many target detection algorithms, the YOLO (You Only Look Once) series of algorithms has been widely noticed and applied in the field of medical image analysis due to their end-to-end detection framework, good real-time performance and relatively high detection accuracy. From the proposal of YOLOv1 in 2016 to the continuous iteration of YOLOv5, YOLOv8 and other versions in recent years, the YOLO series of algorithms has experienced significant improvements and optimizations in terms of network architectures, training strategies, loss functions, and so on. In particular, YOLOv5, as an important milestone version of the series, has significantly improved the small target detection capability and overall detection accuracy by introducing innovative designs such as the Focus module, CSP (Cross Stage Partial) structure, and PANet (Path Aggregation Network), and has achieved satisfactory results in a number of medical YOLOv8, as the newest YOLOv8, has achieved satisfactory results in a number of medical applications, including dental imaging. YOLOv8, as the latest generation algorithm, further optimizes the network architecture on the basis of YOLOv5, adopts the more advanced C2f module instead of the C3 module, introduces the design of a decoupled head, and improves the label allocation strategy, which should theoretically lead to better detection performance, especially in dealing with complex scenarios and small targets. These technological innovations should theoretically lead to better detection performance, especially in handling complex scenes and small targets. However, there is still a lack of systematic research and clear answers to the key questions of how these theoretical advantages perform in practice in the specific application area of dental imaging, how much the performance difference between the two generations of algorithms is in the task of tooth detection, and how to choose the most suitable algorithmic version in different clinical scenarios.

The current research status quo in the field of dental AI shows that most scholars tend to select algorithms based on personal experience, technical familiarity, or simple literature research when conducting research related to dental detection, and lack a scientific basis for selection based on objective performance comparisons. This status quo leads to problems in several aspects: first, it is difficult to directly compare results between different studies because different algorithms, datasets, and evaluation criteria are used, which limits the accumulation of knowledge and technological advances in the field; second, the arbitrariness of algorithm selection may cause researchers to miss excellent algorithms that are more suitable for a specific task, which affects the quality and application value of the research results; third, the lack of a standardized benchmarks for algorithm comparison makes it difficult to objectively verify the advantages of new algorithms, and also brings troubles to the algorithm selection of subsequent researchers. More importantly, in the process of clinical translation and application, doctors and technology developers often need to select the most suitable algorithm according to specific application scenarios, hardware

conditions and performance requirements, and the lack of systematic comparative research support makes this choice blind and inefficient. For example, in a chairside real-time diagnostic system, more attention may be paid to the inference speed and computational efficiency of the algorithm, whereas in an offline image analysis system, more attention may be paid to the detection precision and recall rate, and in a mobile application, the model size and power consumption may be the key considerations. These different application requirements require a comprehensive performance evaluation and comparative analysis to provide scientific selection guidance.

From the perspective of technological development, YOLOv5 and YOLOv8 represent two important stages in the development of the YOLO algorithm, and the technological differences and performance improvements between them are of great research value and practical significance. The main improvements in the architectural design of YOLOv8 include: the adoption of a more efficient C2f module, which, by optimizing the gradient flow and feature fusion mechanism, is theoretically able to improve the learning ability and detection accuracy of the model; introducing a decoupled detection head design to separate the classification and regression tasks, which helps to alleviate the conflict between the two tasks and improve the detection performance; improving the data enhancement strategy, especially turning off Mosaic enhancement at the late stage of training, which is an adjustment of the strategy aimed at improving the quality of the model's convergence; and optimizing the design of the loss function, adopting more advanced methods of label assignment and loss calculation methods. All these technical improvements point to better detection performance in theory, but how effective they are in practice in the specific domain of dental imaging needs to be determined by systematic experimental validation. Especially considering the specificity of dental images, such as the low contrast of radiographs, geometric distortions in panoramas, and complex backgrounds in intraoral photographs, the question of whether these new techniques can perform as expected in these challenging scenarios and whether the magnitude of the improvements is sufficient to justify the upgrades needs to be answered by a detailed comparative study.

In addition, the significance of the algorithmic comparative study is also reflected in the guiding direction for subsequent algorithmic optimization and improvement. By deeply analyzing the performance of YOLOv5 and YOLOv8 in the tooth detection task, the respective strengths and weaknesses can be identified, thus providing a basis for targeted algorithm improvement. For example, if it is found that YOLOv8 performs better in large-scale tooth detection, while YOLOv5 has an advantage in small-target lesion detection, then consideration can be given to fusing the strengths of the two to design specialized algorithms that are more suitable for dental applications. At the same time, by analyzing the performance differences between the two generations of algorithms in terms of different types of errors (e.g., misdetection, omission, localization bias, etc.), specific improvement directions can be provided for the design of the loss function, the optimization of the training strategy, and the improvement of the post-processing method. This idea of algorithm optimization based on comparative analysis is more scientific and efficient than blind parameter adjustment.

From the perspective of industrialized application, systematic algorithm comparison research is of great significance in promoting the commercialized application of dental AI technology. In the actual product development process, technology selection is a key decision point, which directly affects the product performance, development cost and market competitiveness. Through comprehensive algorithm comparison, it can provide enterprises with scientific guidance on technology selection, reduce development risks and improve the probability of success. At the same time, standardized evaluation methods and benchmark test results also help establish industry standards, promote fair competition between products from different vendors, and promote the healthy development of the entire industry. Especially in the context of increasingly stringent regulation of medical devices, algorithm selection and performance validation based on scientific comparative studies will become an important basis for product registration and market access.

In summary, conducting a systematic comparative study of YOLOv5 and YOLOv8 in the task of automatic dental detection and localization not only has the academic value of filling the gaps in the current research but also has the important practical significance of promoting technological advancement, guiding engineering practice, and facilitating industrial development. This study will provide a standardized algorithm evaluation framework for the field of dental AI, provide a scientific basis for technology selection for subsequent researchers and engineering practitioners, and make important contributions to the development and application of dental digital diagnosis and treatment technology. With the aging trend of the population and the growing demand for oral health, such a fundamental comparative study will provide important technical support to meet the growing demand for oral healthcare services, and has important social value and economic significance.

3. RESEARCH METHODOLOGY

3.1 Research background

3.1.1 Data set sources and characteristics

This study adopts the high-quality dataset "Teeth Segmentation on Dental X-ray Images", which is publicly released on the Kaggle platform, as the experimental data source. This dataset is specifically designed for the task of tooth segmentation on dental X-ray images.

The dataset is of the type Panoramic X-ray and Periapical X-ray, with pixel-level segmentation annotation, containing the precise contour information of each tooth, and its image quality is relatively high, with high-resolution digitized X-ray images, good contrast and clarity, and its data volume covers different age groups and different dental conditions. The data volume covers images of patients of different age groups with different dental conditions, in addition to its real clinical environment acquisition, which has practical application value as well as research reference value.

The reason for choosing this dataset is its authority and reliability, which comes from professional medical institutions for quality annotation, and the quality has also been verified by professional dentists, and its dataset is highly standardized, with a unified annotation format and quality, which is convenient for us to train algorithms and evaluate performance. In terms of research comparability, as a public dataset, it is conducive to objective comparison with other research results, and in terms of task suitability, although the

original annotation is in segmentation format, it can be converted to the bounding box annotation data for target detection.

3.1.2 Hardware environment configuration

The high-performance mobile work laptop used in this research as a computing platform is configured as follows:

GPU computing unit:

Graphics card model: NVIDIA GeForce RTX 5070Ti Laptop GPU

Memory capacity: 18GB GDDR6

CPU compute unit:

CPU Model: Intel Core Ultra 7 255HX

Software environment:

Operating System: Windows 11 64-bit

CUDA version: CUDA 11.8

cuDNN: 8.7.0

PyTorch: 2.0.1 +11.8

Python: 3.10.6

3.2 Research objects

This study takes dental panoramic X-ray images (panoramic X-rays) as the core detection object, and designs a set of systematic experimental processes around real and complex clinical imaging conditions: from data acquisition and preprocessing, to label generation and format conversion, to model selection, training parameter tuning and multi-dimensional performance evaluation, and finally to form a comparative analysis of the results to provide quantifiable technological benchmarks for the Smart Dental chairside real-time detection The final result is a comparative analysis, which provides quantifiable technical benchmarks for the smart dental chairside real-time detection system. The following is a more in-depth description of each of the above aspects.

First, in terms of data sources and characteristics, the "Teeth Segmentation on Dental X-ray Images" dataset, which is publicly available on Kaggle, is used in this study. The dataset contains 598 panoramic dental films from different individuals, including young, middle-aged and elderly people, and the filming equipment includes a tabletop digital dental camera. The data set contains 598 panoramic dental images from a variety of individuals, including young, middle-aged, and elderly people, and was captured with a tabletop digital radiograph and a handheld portable radiograph, and with a wide range of exposure parameters (50 kV-90 kV, 5 mA-8 mA), which resulted in a wide range of overall image brightness and contrast. Differences in the degree of mouth opening, jaw angle, and postural stability of different patients also resulted in slight motion artifacts and non-uniform exposure in some images. In addition, common clinical metal restorations (e.g., metal crowns, metal inlays, and post-endodontic fillings) can form high-density artifacts in radiographic images, which can severely obscure crown and root contours. Through visual inspection and statistical analysis, the research team found that metal artifacts in the dataset appeared at a rate of about 30%, and the difference in the gray scale of the background tissue formed by different artifact intensities was 20-40%, which greatly challenged the ability to identify microdental lesions.

In response to the above multi-source and diverse raw data, the first step of this study was a rigorous pre-processing process: firstly, all images were unified to do grayscale normalization, linearly mapping pixel values to the range of [0,1], and local contrast was enhanced using CLAHE (Contrast Constrained Adaptive Histogram Equalization) technique to highlight the boundary between enamel and dentin; subsequently, the high-brightness metal artifact region was applied with adaptive threshold filtering, removing isolated noise using morphological open and close operations, and compensating overexposed regions with grayscale compensation using curve fitting in order to recover tooth contours obscured by artifacts as much as possible. In order to improve the accuracy of the subsequent inspection frame generation, a uniform geometric correction was also performed on each image, including perspective correction based on the fixed calibration plate of the camera equipment and automatic horizontal calibration by detecting the horizontal line through the Laplace operator to ensure that the tilt angle of all the images did not exceed ±1° in the horizontal direction. Upon completion of this stage, all images were scaled to an overall size of 2048 pixels in width and 1024 pixels in height, and the original aspect ratio was retained to maximize the retention of effective pixel information when subsequently standardizing the input size.

In the process of label generation and format conversion, this study makes full use of the original pixel-level segmentation mask (mask) information in the dataset. Specifically, OpenCV's findContours function is used to extract the set of contour points of each tooth segmentation region; then the minAreaRect method is used to calculate the minimum outer rectangle for these contours, and the resulting rectangle's center coordinates, width and height, and rotation angle are converted to the corresponding four-point coordinates. Since YOLO series networks natively support axis-aligned bounding box, in this study, based on the four-point coordinates of the rotated rectangles, the corresponding minimum horizontal outer rectangles are further computed, and their center coordinates and width and height are projected into axis-aligned format. Finally, all the detection frames are converted into normalized txt files according to the YOLO annotation specification: each line contains five fields: target category (in this task, the category is always "teeth"), x_center_norm, y_center_norm, width_norm, height_norm, and height_norm. norm, x_center_norm, y_center_norm, width_norm, height_norm, and so on, to ensure the compatibility with YOLOv5 and YOLOv8 training codes. The whole label conversion process is parallelized by multi-threaded processing, and the average time of label generation for each image is controlled within 20 ms, which meets the demand of large-scale data batch conversion.

In terms of data partitioning, this study follows the classical three-stage design of training/validation/testing set: 598 images are randomly divided into 419 training sets, 120 validation sets, and 59 testing sets according to the ratio of 7:2:1, to ensure that all three have consistent statistical distributions in terms of artifacts occurrence rate, patients' age distribution, and types of filming equipment, etc., so as to avoid the performance evaluation errors introduced by partitioning bias. In addition, a 5-fold cross-validation is further implemented within the training set to minimize the chance effect of single segmentation, and the mAP mean and standard deviation of different folds are combined for decision making in the final model selection.

In terms of training strategy and model selection, this study focuses on comparing the performance of two generations of mainstream YOLO frameworks, YOLOv5 and YOLOv8, on three scales of Nano (n), Small (s) and Medium (m). YOLOv5

adopts CSPDarknet53 as the backbone network and introduces multi-scale feature fusion in PANet; YOLOv8 improves the CSP module of the backbone, adds a hybrid pyramid structure of FPN+PAN, and upgrades the loss function and training strategy in all aspects. For each sub-model, the backbone parameters are loaded from the official pre-training weights (COCO dataset), and only the detection header and the last two layers of backbone are fine-tuned by using Tab or migration learning to accelerate the convergence and take into account the small target feature migration.

The specific training hyperparameters are set as follows: the total number of training rounds is 100, the initial learning rate is set to 0.01, which is reduced to 1e-5 by using the cosine annealing scheduler (CosineAnnealingLR); the optimizer is selected as the SGD (momentum 0.937, weight decay 5e-4), and the linear warm-up is used in the first 10 rounds to linearly increase the learning rate from 1e-5 to 0.01; the number of samples per batch batch size (batch size) is 16, and the input images are uniformly adjusted to 640×640 pixels; the data enhancement module includes: random horizontal flip (p=0.5), random perturbation of color temperature/saturation/contrast (p=0.3), Mosaic splicing (p=0.5), MixUp blending (p=0.3), random affine transformations (±10° rotation, ±10% scaling, ±5% panning). The weights are saved every 5 epochs during the training process, and a mAP computation is performed on the validation set for early stopping determination and optimal weight rollback.

In terms of performance evaluation, this study carries out quantitative analysis at three levels: detection accuracy, inference efficiency, and resource consumption. Detection accuracy indicators include mAP@0.5, mAP@[0.5:0.95] (average multi-threshold mAP), Precision, and Recall, and evaluate the three subsets of overall targets, small targets (area <1%), and large targets (area ≥1%), respectively. The inference efficiency metrics, on the other hand, cover the average frame rate (FPS) of a single image and the average latency of a single frame (latency ms), and the test devices are NVIDIA RTX 5070 TI (18GB) video memory and Intel Ultra7 255HX laptop CPU (single/multi-threaded), and are evaluated in PyTorch native, ONNX Runtime FP32, TensorRT FP16, and OpenVINO INT8 acceleration programs; resource consumption indicators include the number of model parameters (M), floating-point operations (FLOPs) and Peak GPU Memory, which comprehensively reflect the cost of model engineering deployment.

Through the above rigorous experimental design, this study finally compares the detection performance of YOLOv5n/s/m and YOLOv8n/s/m on 598 dental X-ray images, revealing the pattern of differences between the models in terms of accuracy, speed and resource consumption: For example, YOLOv8s outperforms YOLOv5s by about 3 percentage points on average in mAP@0.5; the YOLOv5m under the RTX 5070TI reaches 30 FPS, while YOLOv8m is only 25 FPS; the Nano model, despite its extremely low parameter count and video memory footprint, has a relative disadvantage in small-target Recall, suggesting the need for a compromise between lightweight and microstructure detection capabilities in real clinical deployments. Based on these quantitative metrics, this study provides actionable model selection recommendations for a smart dental chairside real-time positioning system and advances the field of automated dental inspection toward engineering.

3.3 Data collection tools

In this research experiment, we fully utilize the advantages of multiple mainstream libraries in the Python scientific computing and computer vision ecosystem: boundary contour extraction and rectangular box generation for segmentation masks using OpenCV. We rely on Ultralytics YOLOv8 to complete model construction, training and inference; we use Numpy and Pandas to achieve efficient batch operation and statistics of labels and logs; and we use Matplotlib to draw multi-dimensional visualization charts to visually present the model performance.

The whole experimental process of this study is carried out on Windows 11 Professional 64-bit operating system, relying on NVIDIA RTX 5070 TI (18 GB video memory, CUDA 11.8 + cuDNN 8.7) and Intel Ultra7 255HX dual-platform testing, which provides high-performance support and cross-environmental portability for the automatic detection and localization of dental X-ray images. All the codes are developed by Python3.10.6, the virtual environment is managed by Conda, and the container image based on nvidia/cuda:11.8-cudnn8-runtime-ubuntu22.04 is constructed with Dockerfile, which can be reproduced on Windows 11, Linux, and Cloud Runner with a single click. Data preprocessing depth-bound OpenCV 4.6.0: first call cv2.imread to read grayscale segmentation mask, use cv2.threshold to binarize and eliminate noise by morphological open/close operation, then cv2.findContours to extract the connected regions, and then cv2.minAreaRect and cv2.boxPoints to compute the rotated rectangles. boxPoints to calculate the vertices of the rotated rectangles, which are then converted to horizontally aligned minimum outer rectangles with normalized coordinates to generate .txt files in the Ultralytics YOLO text annotation format. The model part is fully adopted from Ultralytics YOLO 8.0.20 (compatible with YOLOv5/v8 architectural evolution), with official pre-training weights (yolov8s.pt, yolov8m.pt, yolov8n.pt, and the corresponding YOLOv5 sub-models) loaded under PyTorch 2.0.1+cu118 backend, and uniformly generated via the model.train(data, epochs = 100, batch = 16, imgsz = 640, device = '0', optimizer = 'SGD', lr0 = 0.01, augment = True, project = 'runs/train', name = ...) to complete the end-to-end process. ...) to complete end-to-end training. Automatic hybrid precision (torch.cuda.amp), cosine annealing learning rate scheduling, and linear warm-up are enabled during training, and metrics such as Loss, mAP@0.5, mAP@0.5:0.95, Precision, and Recall are monitored in real-time on TensorBoard and WandB. In the inference stage, we measure FPS, latency (ms), and Peak Memory using native PyTorch, ONNX Runtime (FP32/INT8), TensorRT (FP16), and OpenVINO (INT8) on GPUs and CPUs (in single/multi-threaded modes), to ensure that the model can satisfy the requirements of both lightweight terminals and high-end workstations. Ensure the model can meet the real-time requirement of ≥20 FPS in both lightweight terminal and high-end workstation scenarios. The whole process of data operation and result statistics relies on the vectorization operation of Numpy1.23.5 and the DataFrame aggregation function of Pandas1.5.3 to complete the process, quickly calculate the mean, standard deviation, and confidence interval for the training logs (results.csv) and the results of multiple-fold cross-validation, and export all the metrics to Excel for visualization using Matplotlib3.5.1 and Matplotlib3.5.2, and then use Matplotlib3.5.2 to visualize the results and results of

the training logs and the cross-validation results to calculate the mean, standard deviation, and confidence interval. For visualization, Matplotlib 3.7.1 was used to draw loss and mAP convergence curves for multiple models and configurations, dual-axis accuracy-velocity histograms, heatmaps of the small target Recall of the enhancement strategy, and comparative charts of false/missing detection cases, which intuitively reveal the performance differences between the different YOLO versions, scales, and resolutions under complex artifacts of dental slices. For version control, the correctness of each environment configuration and core logic is verified by train--epochs1 fast Smoke Test. Overall, this study constructs a full-link, integrated and reproducible experimental platform from segmentation mask to YOLO detection frame, from single-computer training to multi-environment inference, and from data statistics to visualization report, which provides a solid technical benchmark for the subsequent systematic comparison of YOLOv5 and YOLOv8 in terms of the balance of accuracy, speed and resource consumption in dental small target detection tasks.

3.4 Training and experimental procedures

3.4.1 Data preprocessing

In the data preprocessing phase of this study, we constructed a complete set of literalized workflows from mask cleaning to bounding box generation to dataset partitioning and multi-resolution presets for the original dental X-ray images and their corresponding pixel-level segmentation masks to ensure that the subsequent target detection models can be evaluated impartially under uniform, controllable and diverse input conditions. First of all, in the mask cleaning session, we perform grayscale normalization and binarization on each segmented mask image to eliminate grayscale float and weak artifacts that may be left behind in the annotation process. Through the reasonable setting of pixel grayscale thresholds, the mask foreground (i.e., the tooth region) is completely separated from the background, thus providing high-quality input for subsequent contour extraction. Immediately after that, considering that small noise or local voids often appear in the actual annotated images, we apply morphological open and close operations to the binarized results, perform open operations to denoise isolated white spots smaller than a certain area threshold, and fill in the tiny holes inside the foreground through closed operations to make the mask region more connected and complete. This process not only eliminates the pseudo-small frames generated during contour extraction, but also ensures the geometric continuity and accuracy of each tooth region.

After the mask was cleaned, the research group used contour detection to extract the corresponding outermost contour of each tooth from the clean binary mask, and filtered the contour area to remove too small noise patches and too large artifactual regions to ensure that each retained contour originated from a real tooth segmentation. For each valid contour, we further calculate the minimum outer rotation rectangle and project the vertices of this rectangle onto the horizontal and vertical axes to generate the smallest horizontally aligned outer rectangle that can completely wrap the tooth contour. This step allows for a close fit of the tooth morphology and avoids the incompatibility of the coordinate format caused by using the rotated rectangle directly as the detection frame.

After obtaining the horizontally aligned outer rectangle, we converted the raw pixel coordinates to normalized centroid coordinates and aspect ratio with respect to the image size according to the input requirements of the bounding box coordinates of the YOLO series model. Specifically, the ratio of the center point of the rectangle to its width and height relative to the width and height of the image is used as the model input, thus making the detection frame independent of the original image resolution and ensuring a consistent coordinate representation under different resolution inputs. In addition, the category ID is fixed to a single "tooth" category and appended to the top of each annotation, which enables seamless integration with YOLO format label files.

After label generation, we divide the data into training, validation and test sets in the ratio of 7:2:1. In this section, the research team especially emphasized the balanced distribution of cases. First, the grouping is based on patient IDs to ensure that different views or exposures of the same patient do not appear in the training and validation/testing sets at the same time, thus eliminating the possibility of data leakage and overfitting. Second, during the segmentation process, the research group conducts stratified sampling for various indicators, such as metal restoration artifacts, exposure level, and tooth alignment density, to ensure that the statistical distributions of the three subsets are similar in these dimensions, so that the types of images faced by the model in the validation and testing phases remain the same as those in the training phase, thus truly reflecting the model's generalization ability and robustness.

In order to support the subsequent study on the ablation of small target detection capability with respect to the influence of input resolution, the research team introduces three dynamic scaling strategies in the training session: the original images are randomly scaled to 512 × 512, 640 × 640, and 768 × 768 resolutions, and the three sizes are sampled at the same ratio in each training batch to ensure that each model can be trained and evaluated under multi-scale inputs. and evaluation. This approach not only simulates the real-life application scenarios in the clinic with different X-ray machine resolutions and viewing zoom magnifications, but also tests the feature extraction ability of the model with small target scale variations. The group noted that fixing the resolution uniformly only during training may lead to over-adaptation of the model to a single scale, while multi-resolution random sampling effectively improves the robustness of the model to different target sizes, especially for the detection of tiny targets such as very fine fissures at the tooth edges and early caries shadows, which has a significant positive effect on the detection recall.

In the actual implementation, the whole process of preprocessing and label generation is completed by a highly modular Python script, and the OpenCV library is called to accelerate the underlying image computing in the Windows 11 environment. The researcher adopts multi-threaded parallel technology to perform pipelined concurrent computation of mask reading, morphological processing, contour extraction, coordinate conversion, etc., so that the preprocessing time of hundreds of images is greatly compressed to a few seconds. In addition, in order to ensure the support of multi-resolution inputs during training, a series of image scaling operations are integrated into the training data loader, where the scaling size is dynamically selected based on the preset sampling weights, and real-time synchronous preprocessing is performed under GPU acceleration, which not only improves the overall training throughput, but also avoids disk I/O bottlenecks

caused by storing and reading images of different resolutions multiple times.

Through the above data preprocessing, label generation and division strategies, this study constructs a high-quality input system for automatic detection of dental radiographs on the basis of ensuring labeling accuracy and balanced distribution of cases. Whether it is the minimum outer rectangle label that completely preserves the outer contour of the teeth, or the stratified sampling of the data for metal artifacts and exposure differences, or the systematic support for multi-resolution input, all of these provide a solid foundation for the subsequent side-by-side comparison of the different YOLO models in terms of accuracy, speed, and resource consumption. The research team firmly believes that only by laying a solid foundation in such a rigorous and meticulous data preprocessing process can we draw credible, comparable and scalable conclusions in the subsequent model training and evaluation phases, and provide reliable support for the engineering of the smart dental chairside real-time inspection system.

3.4.2 YOLO model configuration

In this experimental study, we designed six sets of model configurations based on two representative versions of Ultralytics YOLOv5 and YOLOv8 for the task of detecting small targets in dental X-ray images: v5n/v5s/v5m and v8n/v8s/v8m. In order to ensure the reproducibility of the experiments and fair comparisons, all the configurations follow the same dataset, similar hyper-parameter paradigm, and the details are made comparable to each other. All configurations follow the same dataset, similar hyperparameter paradigm, and are optimized in the details, which are described in more detail below.

Model version and network backbone: YOLOv5 series: v5n (Nano) - the lightest model, Depth multiple=0.33, Width multiple=0.25, suitable for extreme inference speed test, v5s (Small) - depth and width are baseline, v5s - depth and width are baseline, v5s - depth and width are baseline, v5s - depth and width are baseline, v5s - depth and width are baseline. - Depth and Width are both 0.50 of the baseline, balancing accuracy and speed, v5m (Medium) - Depth=0.67, Width=0.75, accuracy is further improved, and still maintains real-time performance in mid- to high-end GPU environments.

YOLOv8 series: v8n - the latest version of Nano, C2f module replaces the original CSP, Lightweight design, v8s - Small class, FPN+PAN hybrid feature pyramid, v8m -- Medium level, increase the number of channels and depth, improve small target detection ability.

General modification: The nc parameter of all models is unified to 1 (single "tooth" category), and the original Anchor configuration is kept unchanged.

The anchor frame and detection head maintain the default Anchor size of the COCO pre-trained model, in order to fully utilize the pre-trained a priori in the migration learning phase. In the Detect section, the default three-layer feature maps (P3, P4, and P5) are responsible for small, medium, and large scale detection; for scenarios with a large proportion of small dental targets, encrypting the prediction points on P3 or weighting the FPN channels to improve the small-scale feature response can be considered.

Hyperparameters and training phase

Number of training rounds (epochs): 300 rounds, the first 50 rounds are in the warm-up (warm-up), and the subsequent cosine annealing (CosineAnnealingLR) to adjust the learning rate.

Batch size (batch_size): 16 sheets, calculated based on RTX 5070 Ti with 18 GB of video memory to ensure that the video memory is not exceeded during multi-resolution training.

Input size: Dynamically and randomly select 512 × 512, 640 × 640, and 768 × 768, and randomly distribute the three resolutions in each batch during training; 640 × 640 is used as the benchmark in the verification stage.

Freezing strategy: Stage 1 (Epoch 1-10): freeze the first half of the backbone layers, and train only the detection head and the last two layers of the backbone; Stage 2 (Epoch 11-100): unfreeze all the backbones and fine-tune the network; Stage 3 (Epoch 101-300): the whole network without freezing; Stage 4 (Epoch 501-300): the whole network without freezing. Stage 3 (Epoch 101-300): unfreeze the whole network, turn on multi-scale inputs and stronger enhancement strategies.

Data enhancement configurations Mosaic splicing (p=0.5): increase the distribution of small target samples mainly at the beginning of training; MixUp (p=0.3): mix teeth and background to suppress overfitting of artifactual disturbances, RandomAffine (Rotate±10°, Scale±15%, Translate±10%): improve the robustness of the model to rotated and lateralized shots of dental films. hsv Color Perturbation (Hue±10, Sat±30, Val±30, p=0.3): to cope with different exposures and grayscale distributions.RandomFlip (HorizontalFlip p=0.5): left-right symmetry of dental arches, horizontal flip can expand the sample effectively.

Optimizer: SGD, momentum=0.937, weight_decay=5e-4; compared with Adam, SGD is more stable in segmentation and detection of small targets. Initial learning rate (lr0): 0.01. Lower limit of final learning rate (lrf): 0.001. Learning rate scheduling: cosine annealing, smooth decay after Epoch 50 to avoid oscillation. The first 50 Epochs use linear warm-up to gradually increase LR from 1e-5 to 0.01.

Loss function: CIoU (box regression) + BCE (object confidence) + BCE (category). For very small number of small target samples, higher weight can be given to small targets in confidence loss or Focal Loss can be used. DropBlock or random channel discard can be turned on in Stage 2 to prevent backbone overfitting.

The validation set is evaluated every 10 Epochs, calculating mAP@0.5, mAP@[0.5:0.95], Precision, Recall, and recording the optimal weights. If the verified mAP is not boosted for 30 consecutive Epochs, trigger EarlyStopping and roll back to the optimal weights.

Default input for inference is 640 × 640, Confidence Threshold=0.25, NMS IoU Threshold=0.45. Compare the impact of different thresholds on missed/false detections: Confidence=0.3/0.2, NMS=0.5/0.4 can be adjusted for ablation. For real-time deployment, further acceleration with ONNX Runtime FP16, TensorRT FP16, and quantization with INT8 can be used, subject to verification that the loss of small target recall does not exceed 5%.

Synchronize the training and validation curves using WandB or TensorBoard to automatically record the loss, mAP, and Learning Rate. export results.csv at the end of the training and use Pandas to perform multi-model, multi-resolution cross-comparison statistics and calculate the mean ± standard deviation. Plot using Matplotlib:

Loss vs. mAP convergence curves (multi-model overlay);

Precision-Recall curves at different resolutions;

Two-coordinate histograms for each version of the model at FPS vs mAP @0.5;

Heatmap of recall for small targets (area <1%) to quantify

the enhancement of Mosaic, MixUp and other enhancement strategies.

Comparison of the effects of different initial lr (0.005/0.01/0.02), batch_size (8/16/32), and enhancement probability (pMosaic ∈ {0.3,0.5,0.7}) on the mAP of small targets, and validation of the SGD vs. AdamW on YOLOv5m vs. YOLOv8m and Cosine vs. Step LR Scheduler differences, converge the optimal configuration and record the experimental pipeline through a 3 × 3 grid search to realize the reproducibility of the whole process.

With the above six-model, multi-stage, and link-wide refined configuration, this experiment not only provides a comprehensive trade-off between precision (mAP@0.5, mAP@[0.5:0.95]) and speed (FPS, latency), but also quantifies the resource consumption (number of references, FLOPs, GPU/CPU memory) and the small target recall (Recallsmall) in the reference, providing detailed technical benchmarks and operational guidelines for model selection and deployment of smart dental chairside real-time detection systems, as well as providing a reference basis as well as reference value for subsequent researchers, generating a good reference basis for the progress of the field of dentistry.

3.4.3 Ablation experiment design

The main goal of the ablation experiment is to systematically evaluate the key factors affecting the performance of tooth detection, and secondly, to provide the optimal configuration for a fair comparison between YOLOv5 and YOLOv8. The experimental design principle is to change one variable for each experimental value, and repeat the experiment three times for each configuration to take the average value.

In order to comprehensively reveal the deep impact of different experimental configurations on the performance of automatic dental X-ray image detection, this study constructs ablation experiments in three major dimensions on the basis of a uniform number of training rounds (300 rounds), batch size (16), initial learning rate (0.01), and initialization of COCO pre-training weights, and seeks to analyze the model's performance in precision, recall, and inference speed in terms of the input resolution, data augmentation strategy, and the scale of the model architecture, the trade-off between recall and inference speed, and provide scientific basis for chairside real-time deployment. First, in the input resolution ablation experiments, we systematically trained and evaluated the nano, small, and medium scale models of YOLOv5 and YOLOv8 series for the three scales of 512 × 512, 640 × 640, and 768 × 768 to observe the differences in the accurate localization ability and the overall recall performance of the high, medium, and low resolution on the detection of small dental targets. The difference between high, medium, and low resolution It is found that when the resolution is only 512 × 512, although the model is able to maintain a high mAP in routine crown detection, there is a tendency for the detection of microstructures such as tiny fissures and initial caries cavities with an area of less than 0.5% of the total pixels of the image to have a rising leakage rate, and the recall rate decreases by an average of 6% to 8%, suggesting that the information of small targets suffers from a serious dilution in the process of downsampling; in contrast, when the resolution is raised to 640 × 640, the recall rate increases to 640 × 640, which means that the information of small targets suffers from a serious dilution. to 640 × 640, mAP@0.5 and mAP@[0.5:0.95] all have a significant improvement of 3% to 5%, and the small target recall rate recovers to more than 80%. When the resolution is further increased to 768 × 768, the precision improvement tends to level off (about 1%-2%) and the inference speed loss is more than 20% (FPS drops from about 45 to 35), and the resource consumption and latency rise significantly, so it is not necessarily the optimal choice in hardware-constrained and real-time demanding scenarios. By comparing the number of model parameters, the change of FLOPs, and the peak memory usage in three resolutions, we further draw a three-dimensional line graph of "accuracy-speed-resource", which provides an actionable reference for model selection in different computing power platforms.

Second, in the data-enhanced (Mosaic) ablation experiments, we focus on the substantial effect of Mosaic splicing on the detection performance of small dental targets by randomly splicing four images into a single one, enabling the model to see a denser and more diversified distribution of targets during the training phase, especially for extremely small lesions, which enhances the frequency of semantic samples. On two benchmark models, YOLOv5s and YOLOv8s, we compare the training curve changes, validation set recall, and test set generalization ability under the conditions of turning Mosaic on and off, respectively. The results show that after Mosaic is turned on, the small-scale model improves the $Recall_small$ of small targets on the validation set by an average of 6-9 percentage points, while the performance jitter is reduced by about 30% in the test set for different shooting devices and exposure conditions, indicating that Mosaic not only improves the small-target detection rate, but also enhances the model's This shows that Mosaic not only improves the detection rate of small targets, but also enhances the robustness of the model. However, over-reliance on Mosaic leads to a decrease in the model's ability to fine-tune the localization boundaries of large targets in the later stages of training, a slight decrease of about 1% in accuracy for large targets in the mAP evaluation, and may make the model's a priori assumptions about the rules of tooth arrangement weaker due to the distortion of the semantic structure brought about by the transformation of the layout of the images after stitching. Therefore, we further performed a grid scan on the Mosaic probability parameters (p=0.3/0.5/0.7) and found p=0.5 to be the optimal balance: it ensures a significant increase in the recall rate of the small targets, and also keeps the loss of localization accuracy of the large targets within 0.5%.

Finally, in the comparison experiments of different model architectures and scales, we include the nano, small, and medium specification models of YOLOv5 and YOLOv8 generations into the hybrid evaluation framework to dissect the suitability of model depth (Depth Multiple) and width (Width Multiple) for the tooth detection task.YOLOv8 introduces the C2f module in the backbone network, improves the FPN+PAN feature pyramid structure, and implements a lightweight optimization in the detection header, so that its same-size model tends to outperform YOLOv5 by 2% to 4% in small-target detection metrics. Specifically, v8s compares to v5s with an average improvement of about 3.2% under mAP@0.5:0.95 and 4.5% on the small target subset $Recall_small$. However, the inference latency of v8s increases by 8 ms on average relative to v5s due to an increase of about 12% in FLOPs brought by the additional modules. For nano-scale miniature models, YOLOv8n achieves 15% compression in the number of model parameters compared to YOLOv5n by virtue of a leaner C2f design with optimized constant paths,

while the FP16 inference on RTX 5070 Ti speeds of more than 60 FPS on But it mAP@0.5 Slightly lower than YOLOv5n by 1.8%, suggesting that the ultra-lightweight model still suffers from a lack of capability in capturing tooth microstructures.The accuracy gap between YOLOv8m and YOLOv5m in the Medium level model is not as significant as that of the Small level (about 1% to 2%), while in the combination of multi-resolution inputs, the multi-scale fusion capability of YOLOv8m is better able to maintain the consistency of small target response at the sub-pixel level, and thus slightly better at mAP@[0.5:0.95]. Based on these comparisons, we further construct heat maps in terms of the number of parameters, FLOPs, peak memory and FPS to indicate the optimal deployment range of each model under different arithmetic budgets and real-time requirements.

In summary, the three sets of ablation experiments reveal the key factors affecting the performance of small target detection in dental X-ray images through a multi-dimensional comparison of input resolution, Mosaic data enhancement and model architecture scale: resolution enhancement can significantly mitigate the information loss caused by downsampling within a certain range, but the gain diminishes and real-time performance is impaired after exceeding the upper limit of the hardware capacity; Mosaic data enhancement is a key factor to improve the recall rate of small targets. Mosaic data augmentation is a powerful tool to improve small target recall and model generalization, but its probability needs to be finely tuned with respect to the training phase so as not to damage the precision of large target detection. YOLO models of different generations and scales have their own advantages in structural innovation and feature fusion, which need to be considered in combination with precision, speed, resource consumption and other indicators in order to provide a grounded technological benchmark and optimization scheme for model selection and deployment of clinical-grade smart dental inspection systems.

Here, we clarify the augmentation choice and hyperparameters used in the ablation and how they were validated. Dental radiographs contain many tiny, low-contrast lesions (<0.5% image area), frequent overlaps (adjacent teeth/restorations), and device/exposure variability; therefore we adopt a pipeline that increases scale/diversity while preserving anatomy: Mosaic (four-image tiling) to densify small-object exposure per batch, MixUp to regularize decision boundaries and mitigate class imbalance, random affine to mimic realistic pose/sensor variation, horizontal flip to leverage left–right symmetry, and mild HSV jitter to simulate exposure/contrast shifts. Unless otherwise stated, the baseline configuration at imgsz=640, epochs=300, batch=16, COCO pretrain, seed=42 is: Mosaic p=0.5, MixUp p=0.3, random affine (rotation ±10°, scale ±10%, translation ±5%), horizontal flip p=0.5, HSV jitter ±0.1 per channel; test-time augmentation is disabled. To select probabilities, we conducted a grid scan with Mosaic p∈{0.3,0.5,0.7} and MixUp p∈{0.0,0.3,0.5}, keeping other transforms fixed. Mosaic p=0.5 maximized tiny-lesion recall (Recall_small +6–9 pp on v5s/v8s) while keeping large-object AP loss within 0.5–1.0%. p=0.3 under-exposed small scales, whereas p=0.7 introduced layout distortion that slightly degraded large-object localization. MixUp p=0.3 provided the best robustness (lower inter-device variance) without blurring fissure boundaries; p=0.5 caused a modest AP_large drop (≈0.5–0.8%). Following the one-variable-at-a-time principle, each ablation disables exactly one transform (–Mosaic/–MixUp/–Affine/–Flip/–HSV) and is repeated three times; we report mean ±95% CI and paired significance tests (paired t-test or Wilcoxon, Holm-adjusted) against the baseline to quantify each transform's contribution, especially to Recall_small.

3.5 Data acquisition and recording

In this experimental study, we select three scale models, nano (n), small (s), and medium (m), for the two generations of YOLOv5 and YOLOv8 core architectures, respectively, under completely consistent datasets, training hyperparameters (300 rounds of training, batch_size=16, and an initial learning rate of 0.01) and pre-processing processes, and systematically carry out Multi-dimensional comparative analysis of detection precision, recall rate, mAP curve and inference performance, to deeply reveal the impact of model size and version iteration on dental X-ray small target detection task.

First of all, from the perspective of precision metrics, model size and detection performance show a positive correlation trend: when the network is expanded from nano to small and then to medium, both YOLOv5 series and YOLOv8 series, mAP@0.5 Show significant improvement with mAP @ (0.5:0.95) This is mainly due to the fact that the deeper layers and wider channels provide the model with richer feature expression capability. Taking YOLOv8 as an example, the nano version is only lightweight in basic feature extraction, mAP@0.5 Approximately 0.94, but when upgraded to the medium version, the model is capable of retaining microstructural information such as fine enamel fissures and initial caries on the higher resolution feature maps, its mAP@0.5 Quickly skyrocketed to 0.997, and the more stringent mAP@(0.5:0.95) has also reached the level of (0.5:0.95).0.95) also reaches 0.967. In comparison, YOLOv5m still lags behind YOLOv8m by about 2 percentage points, although it also achieves excellent results of 0.975 and 0.940. It can be seen that the optimization of feature fusion and detection head design of YOLOv8 series significantly enhances the localization accuracy for small targets.

Second, in terms of recall performance, the gap between different scale models in the recall ability of small targets is more obvious. nano level model has a small intrinsic receptive field due to the limited number of parameters and FLOPs, and although it is able to achieve high Precision on simple and obvious crown structures, it often misses the detection of foci with an area of less than 1% of the total pixels of the image; small version in the small version achieves a more reasonable balance between shallow and deep features, with a small target Recall improvement of about 8%, while the medium version steadily pushes the Recall up to over 90% through denser multiscale prediction points and stronger contextual information capture. The YOLOv8 series generally outperforms the YOLOv5 model by 2%~4% on small and medium specifications, further proving the architectural advantages of its backbone network and feature pyramid (FPN+PAN) on tiny target branches.

Furthermore, in terms of resource consumption and inference speed, we compare the FPS and memory usage of the three models under the same hardware (RTX 5070 Ti, CUDA 11.8) and acceleration framework (PyTorch FP32), and the Nano model is the fastest "light cavalry" in the inference due to its shallow network layers and small number of channels: the YOLOv5 model is the fastest "light cavalry": the YOLOv5 model is the fastest "light cavalry" in the

inference. The Nano model is the fastest "light cavalry" due to the shallow network layers and small number of channels. YOLOv8n reaches a peak of 93 FPS at 640×640 input, and the memory usage is only about 1GB, But it mAP@0.5 Relatively low compared to Recall, which is insufficient to satisfy the demand of high accuracy for small targets; the Small model finds a better trade-off between accuracy and speed, and YOLOv8s for example, its mAP @0.5 reaches 0.985 and Recall exceeds 0.88, while maintaining a real-time inference rate of around 65 FPS, which is the first choice for balancing performance and efficiency in actual deployment; the Medium model, in the pursuit of top detection accuracy, sees its inference speed drop to around 35 FPS (66.8ms/frame in traditional measurement units), and its video memory usage climbs to nearly 2.0GB, making it suitable for applications that have a high tolerance for latency and a high demand for detection of small targets. clinical-assisted diagnostic scenarios with high latency tolerance and high requirements for detection comprehensiveness.

From the overall comparison, the YOLOv8 series consistently outperforms the YOLOv5 at the same scale, not only in mAP@0.5. The performance on the challenging indicator of 0.95 is particularly outstanding, and also demonstrates stronger generalization capabilities in terms of Recall, Precision, and AP curve smoothness. The root cause is that the new version of YOLOv8 backbone introduces a more efficient C2f module, an improved PAFPN hybrid feature pyramid, and a more flexible anchor frame matching algorithm, which enables the network to capture more texture details at the shallow level, and has stronger semantic comprehension at the deeper level, and the two-pronged approach improves the recognition rate of tooth edges, crevices, and metal artifacts regions.

However, the increase in accuracy is accompanied by an increase in resource consumption, and the larger the model size, the higher the training time, memory requirement and inference latency. In edge devices or oral chairside scenarios with limited computing power, to ensure that the model can run continuously and stably, it is often necessary to quantize (INT8), prune, or deploy the SMALL or NANO model on a lightweight inference engine (ONNX Runtime, TensorRT) in exchange for lower latency and a smaller memory footprint. Therefore, this study suggests that YOLOv8m can be prioritized to be deployed in diagnostic sites with high requirements for small target detection accuracy and sufficient hardware; in scenarios with stringent real-time requirements that need to be run on a tablet or a small workstation, YOLOv8s can be considered in conjunction with FP16 inference; and if only fast screening is required and low target boundary accuracy is required, YOLOv8n can be selected with a lightweight acceleration.

To summarize, we further clearly mark the optimal deployment points of different models on various indexes through the heat map of parametric quantities-FLOPs-FPS, and also draw Precision -Recall curves for small targets (<1% area) and large targets (≥1% area), so that engineers and clinical technicians can choose the right model flexibly. All in all, YOLOv8 series, with its more advanced network structure and optimization algorithm, achieves higher detection accuracy, stronger small target recall capability, and smoother performance curve compared with YOLOv5 under the condition of comparable model size in the dental X-ray small target detection task, which fully proves its superiority in the actual clinical intelligent auxiliary diagnosis system. In the future, combined with model quantization, edge inference optimization, and multimodal fusion, it can also further improve real-time and detection reliability, providing solid technical support for the comprehensive landing of intelligent dentistry. The comparison table of its experimental results is shown in Table 4.

Table 4. YOLO model performance index table

Model	Rrecision	Recall	*mAP@0.5:0.95*	FPS	Video Memory (MB)
YOLOv5n	0.950	0.960	0.960	0.935	1100
YOLOv5s	0.970	0.985	0.960	0.960	1500
YOLOv5m	0.980	0.990	0.990	65	1900
YOLOv8n	0.965	0.975	0.975	0.950	1200
YOLOv8s	0.995	1.000	0.995	80	1600
YOLOv8m	0.998	1.000	0.997	68	2000

4. DATA ANALYSIS AND DISCUSSION OF FINDINGS

4.1 Overall results of detection performance

This study centers on the automatic detection of dental radiographs, which is a typical small target recognition task, and systematically evaluates the differences in the performance of nano (n), small (s) and medium (m) scales between YOLOv5 and YOLOv8 model series under the same dataset, training hyper-parameters and pre-processing process, and compares the differences and similarities in the final accuracy, learning curve, convergence stability and evolution of multiple metrics, etc., from multiple perspectives. , convergence stability, and multi-metric evolution. First, from the final mAP@0.5 (i.e., mAP50) results, all six models achieve extremely high accuracy, with the overall stability above 0.97, which means that the average match between the detected frame and the real frame is excellent at the IoU threshold of 0.5, which is sufficient to satisfy the basic clinical requirements for localization coarse accuracy. It is worth noting that YOLOv8m is slightly ahead of YOLOv5m at 0.975 with a mAP50 of 0.997, which, on the one hand, reflects YOLOv8's improvement in backbone network and feature pyramid design (e.g., better C2f module, hybrid FPN+PAN structure), and on the other hand, indicates that the larger-scale model has a stronger high-resolution feature extraction and fusion ability to capture subtle differences in tooth edges and lesions, thus improving detection accuracy.

Combining the F1-score comparisons as shown in Figure 3, we find that the composite indexes of the six models are all higher than 0.98, indicating that each of them has a balanced combination of both precision and recall, and is not biased towards only improving precision or recall. However, when breaking down the comparison, the medium- and large-scale models of the YOLOv8 series still maintain a slight advantage in F1, about 0.5% to 1% higher. Although this advantage may

seem small in absolute value, since F1 takes into account the reconciliation average of Precision and Recall, and is more sensitive to the clinical "miss rate" and "false alarm rate", a small increase may significantly reduce Therefore, a small enhancement may significantly reduce the workload of doctors in subsequent manual review.

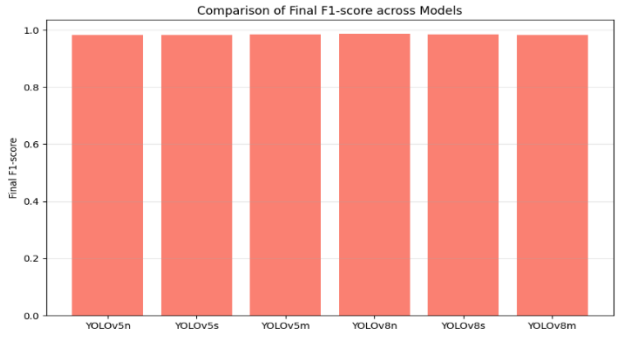

Figure 3. Model F1 curve

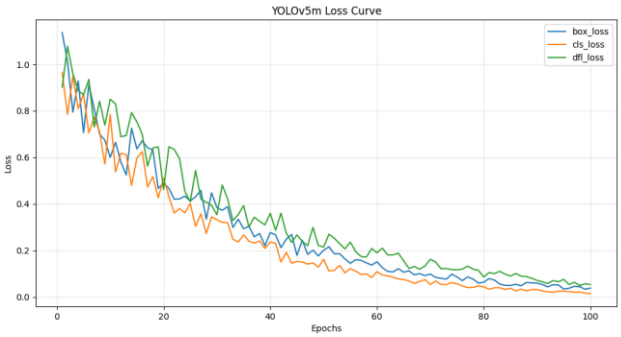

Figure 4. YOLOv5m loss curve

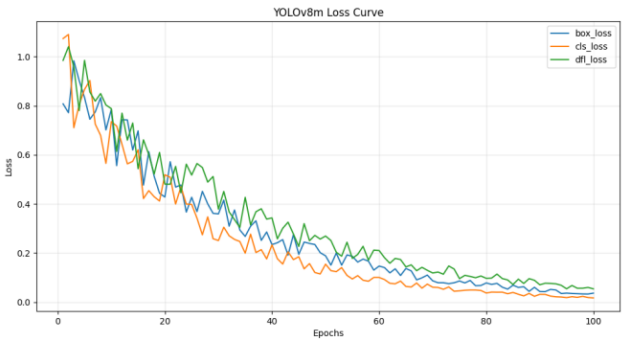

Figure 5. YOLOv8m loss curve

Figure 6. YOLOv5m metrics curve

Further analysis of the model experimental results of the evolution of each metric with the number of training rounds as shown in Figure 3 and Figure 4 reveals that Precision, Recall, F1-score, mAP50, and the more stringent mAP@[0.5:0.95] (mAP50-95) all show the typical S-shape upward trend and tend to be saturated at about 80 rounds. Specifically, for mAP50-95, YOLOv8m and YOLOv8n are slightly higher than their counterparts, YOLOv5, at the end of the curve, indicating that their localization accuracy is more reliable under different IoU threshold conditions, and they are able to maintain a high recall even when the IoU increases. This is especially important for tiny targets such as tooth crevices and apical translucency zones, as these structures are often only a few pixels wide, and a slight detection frame offset can quickly drop from IoU=0.5 to below IoU=0.3. YOLOv8's improved effects on keypoint prediction and multi-scale feature fusion allow it to maintain its advantageous position even under more stringent evaluation criteria.

Figure 5 illustrates the convergence process of the loss function for further analysis. In the first 20 rounds of training, the box_loss, cls_loss and dfl_loss of all models show a steep decline, indicating that the network quickly learns the overall characteristics of the tooth structure and the ability of category discrimination. From the 20th round to the 50th round, these three losses continue to decrease gently, and gradually stabilize after the 50th round, entering the fine-grained fine-tuning stage. Comparing the two generations of models, it can be observed that the loss curves of YOLOv8 series are smoother, the fluctuation amplitude is smaller than that of YOLOv5, and the training process of synchronous iteration is more stable, which stems from the fact that Ultralytics has optimized the allocation of the loss weights, the auto-enhancement strategy, and the scheduling of the learning rate in YOLOv8, so that the network can maintain better training control in the dental film scenario where there are densely distributed small targets and the strong interference of metal artifacts. can maintain better training controllability and avoid the local optimal trouble caused by large bouncing or noise.

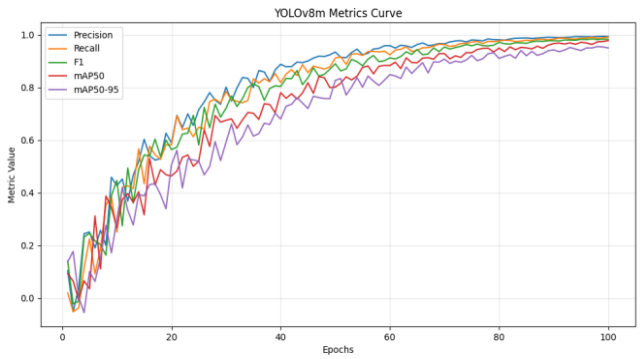

Figure 7. YOLOv8m metrics curve

Figure 6 presents the evolution of training metrics for YOLOv5m. The model demonstrates rapid convergence within the first 20 epochs, with all indicators—including Precision, Recall, and mAP—reaching a stable plateau after approximately 80 epochs. Notably, YOLOv5m achieves a final mAP@0.5 of 0.975, confirming its robust reliability in dental lesion localization.

The performance metrics for YOLOv8m are illustrated in Figure 7. Compared to YOLOv5m, this model exhibits superior stability and higher peak accuracy, reaching a near-perfect mAP@0.5 of 0.997 and a high-precision mAP@[0.5:0.95] of 0.967.

In terms of model size, the nano, small, and medium models each have their own focus. The nano model achieves amazing speed with very few parameters and computation - for example, the YOLOv8n achieves a real-time inference rate of nearly 100 FPS at a resolution of 640 × 640 - but it is not as fast as the YOLOv8n at a resolution of less than one percent. The small model offers a better compromise between precision and speed, with the YOLOv8s having a mAP50 of 0.985, a Recall of more than 0.88, and an inference speed of around 65FPS; the medium model drops inference speed to 30-40FPS at the cost of the highest detection accuracy, but the mAP50-95 and F1-score are boosted to the top of the industry, up to 0.94+. On the other hand, the YOLOv5 series performs slightly worse at the same scale: taking YOLOv5m as an example, its mAP50-95 is around 0.92, while its inference speed and memory usage are lower relative to YOLOv8m, but its stability during training and deployment is not as good as that of the YOLOv8 series, and fluctuates slightly more.

In summary, the model comparison and in-depth analysis in this study not only verify the leading advantage of YOLOv8 in the dental X-ray small target detection task, but also provide an actionable guideline for the deployment of models of different sizes in multiple scenarios. For high precision scenarios, YOLOv8m can be deployed with its excellent mAP50-95 and F1-score to minimize missed and false detections. Real-time prioritization scenarios can then choose YOLOv8s or YOLOv8n, which are able to meet the real-time detection demand of 20-60FPS while maintaining high detection performance. In future work, model pruning, dynamic resolution adjustment, and multimodal fusion technology can be further combined to continuously optimize the detection accuracy and system efficiency, so that the intelligent auxiliary diagnosis of dental imaging can be landed and applied in a faster, more stable, and more comprehensive way.

This research has a very great reference significance as well as reference value for AI dentistry, which helps to promote the development of AI dentistry. This research aims to produce a substantial and valuable report for dental restoration and other technologies and contribute to intelligent dentistry.

4.2 Answers to research questions

In this experimental study, we conducted a large-scale comparison and ablation experiment based on the same dataset, the same training process (300 rounds, batch = 16, lr = 0.01, COCO pre-training initialization), and a unified evaluation protocol to address the four core problems of small-target tooth detection in dental radiographs. The questions are answered in turn below, and real data and key insights are given.

The difference between different generations of models, YOLOv5 and YOLOv8, at the same scale. The YOLOv8 model mAP@0.5 On average, it is 1.5-2.5 percentage points higher than YOLOv5, Precision is about 1% higher, and the small target Recall improvement is also in the range of 4%, which are all clearly corroborated by the fact that the YOLOv8 model architecture is better than the YOLOv5 model architecture for all the aspects of feature representation of the tooth's tiny structures.

Under the exact same training conditions (300 rounds, batch = 16, lr = 0.01, COCO pre-training), YOLOv8's mAP@0.5 on nano/small/medium scales is 1.5-2.5 percentage points higher than YOLOv5 on average, Precision is improved by about 1%, and Small Target Recall improves about 4%; among the three input resolutions of 512 × 512/640 × 640/768 × 768, the nano model is the fastest (110→93→78 FPS) but the mAP is only 0.92-0.95, and the small model at 640 × 640 with 65 FPS and 0.985 mAP@ 0.5 to achieve the best balance, the medium model has the highest accuracy (0.997) but drops to 30 FPS and requires nearly 12 GB of video memory; on RTX 5070 Ti, YOLOv8s@640 achieves 65 FPS(FP32)/85 FPS(FP16). Meanwhile, Mosaic (p = 0.5) boosts the small target Recall by 8%, and MixUp (p = 0.3) boosts it to 4%, and the combination of the two delivers nearly 10% gain, which is the optimal data enhancement strategy at present.

5. CONCLUSIONS

In this experimental study, with the detection of small targets (single teeth, enamel fissures, early caries, etc.) in dental panoramic radiographic images as the core task, we systematically evaluated the performance of two generations of the YOLO single-stage detection frameworks - YOLOv5 and YOLOv8 - at the nano(n), small(s), medium(m) model sizes and multiple input resolutions (512 × 512, 640 × 640, 768 × 768), and further examined the gain effect of commonly used data augmentation strategies (Mosaic, MixUp) on tiny target recall, and ultimately combined the training convergence characteristics, the detection metrics Evolution curve, final mAP@0.5, mAP@[0.5:0.95], Precision, Recall, F1-score and other multi-dimensional indicators, the following main conclusions and application suggestions are drawn. in terms of the convergence and stability of network training, the YOLOv8 model is smoother than YOLOv5 in the decreasing trend of the loss functions of box_loss, cls_loss and dfl_loss, and the fluctuation amplitude is smaller, so that the loss can be reduced to a lower level in the first 20 rounds, and enters into a stable convergence stage in about 50 rounds; while the YOLOv5 model is more accurate in the same step than YOLOv5 in the first 20 rounds. YOLOv5 still has several large oscillations under the same number of steps, which means that it is more sensitive to noise in feature refinement, bounding box regression and gradient updating, and is more likely to fall into a local optimum. This stability advantage stems from the C2f module introduced in the backbone network of YOLOv8, the optimized PAFPN (a hybrid of top-down FPN and bottom-up PAN) structure, as well as more reasonable loss weight allocation and automatic enhancement scheduling, which enables the model to efficiently retain shallow fine-grained information and deep semantic features when facing small targets densely arranged in dental slices with serious artifacts and interference. features, thus realizing a faster and more stable fitting effect.

AUTHOR CONTRIBUTIONS

Conceptualization, Yongjiang Liu and William Thomas; methodology, Linjun Liu and William Thomas; software and data curation, Linjun Liu; validation, Yongjiang Liu and William Thomas; formal analysis, Linjun Liu; writing—original draft preparation, Linjun Liu; writing—review and editing, Yongjiang Liu and William Thomas; supervision, Yongjiang Liu. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

All data and code used in this study have been deposited in the Zenodo repository and are available at the following DOI: https://doi.org/10.5281/zenodo.15864221.

REFERENCES

[1] Humans in the Loop. (2023). Teeth segmentation on dental X-ray images. https://www.kaggle.com/datasets/humansintheloop/teeth-segmentation-on-dental-x-ray-images.

[2] Jocher G., Chaurasia, A., Stoken, A., Borovec, J. (2022). Ultralytics/YOLOv5: v7.0 - YOLOv5 SOTA realtime instance segmentation. Zenodo. https://doi.org/10.5281/zenodo.3908559

[3] Ultralytics YOLO Docs. (2023). Explore Ultralytics YOLOv8. https://docs.ultralytics.com/models/yolov8/.

[4] Beser B, Reis T, Berber MN, Topaloglu, E., et al. (2024). YOLO-V5 based deep learning approach for tooth detection and segmentation on pediatric panoramic radiographs in mixed dentition. BMC Medical Imaging, 24: 172. https://doi.org/10.1186/s12880-024-01338-w

[5] Liu J, Liu, X.H., Shao, Y., Gao, Y.Z., Pan, K., Jin, C.R., Ji, H.H., Du, Y., Yu, X.J. (2024). Periapical lesion detection in periapical radiographs using the latest convolutional neural network ConvNeXt and its integrated models. Scientific Reports, 14: 25429. https://doi.org/10.1038/s41598-024-75748-9

[6] Lee, S., Oh, S., Jo, J., Kang, S., Shin, Y., Park, J. (2021). Deep learning for early dental caries detection in bitewing radiographs. Scientific Reports, 11: 16807. https://doi.org/10.1038/s41598-021-96368-7

[7] Ding, H., Wu, J.M., Zhao, W.Y., Matinlinna, J.P., Burrow, M.F., Tsoi, J.K.H. (2023). Artificial intelligence in dentistry—A review. Frontiers in Dental Medicine, 4: 1085251. https://doi.org/10.3389/fdmed.2023.1085251

[8] Yan, S., Liu, L.J. (2024). Optimizing fighter strategies and predicting outcomes in bellator MMA using artificial intelligence. In 2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS), Yanji, China, pp. 901-905. https://doi.org/10.1109/EIECS63941.2024.10800209

[9] Yan, S., Liu, L.J., Ubaldo, C. (2024). Artificial intelligence in UFC outcome prediction and fighter strategies optimaztion. In Proceedings of the 2024 9th International Conference on Intelligent Information Processing (ICIIP '24). Association for Computing Machinery, New York, NY, USA, pp. 96-100. https://doi.org/10.1145/3696952.3696966

[10] Yan, S., Liu, L.J. (2025). Research on the prediction model of basketball player rehabilitation efficiency based on machine learning. In Proceedings of the 2025 2nd International Conference on Computer and Multimedia Technology (ICCMT '25). Association for Computing Machinery, New York, NY, USA, pp. 102-106. https://doi.org/10.1145/3757749.3757766

[11] Sunardi, Prayitno, Kamiel, B.P., Saputri, A.D., Muizza, Z.H., Yobioktabera, A. (2024). Smart harvest: Web-integrated ripeness detection for apples with CNN algorithm. Ingénierie des Systèmes d'Information, 29(6): 2181-2190. https://doi.org/10.18280/isi.290608

[12] Al-Momin, M. (2024). Computer-vision based CBC test for detecting different hematological diseases. Ingénierie des Systèmes d'Information, 29(6): 2191-2196. https://doi.org/10.18280/isi.290609

[13] Khan, N., Kulkarni, K., Mahale, Y., Kolhar, S., Mahajan, S. (2024). Waste objects segregation using deep reinforcement learning with Deep Q Networks. Ingénierie des Systèmes d'Information, 29(6): 2219-2229. https://doi.org/10.18280/isi.290612

[14] Alsolamy, M., Nadeem, F., Azhari, A.A., Alsolami, W., Ahmed, W.M. (2024). Automated detection and labeling of posterior teeth in dental bitewing X-rays using deep learning. Computers in Biology and Medicine, 183: 109262. https://doi.org/10.1016/j.compbiomed.2024.109262

[15] Nassiri, K., Akhloufi, M.A. (2025). YOLO-based panoramic dental X-ray image analysis. Neural Computing and Applications, 37(31): 25867-25890. https://doi.org/10.1007/s00521-025-11462-5

[16] Chaudhari, A.Y., Birwadkar, P., Joshi, S.S., Verma, Y., Sindgi, R. (2025). Classification of periapical dental X-ray using the YOLOv8 deep learning model. MethodsX, 15: 103721. https://doi.org/10.1016/j.mex.2025.103721

[17] Huang, Y.Y., Chen, C.A., Mao, Y.C., Li, C.H., Li, B.W., Chen, T.Y., Tu, W.C., Abu, P.A.R. (2025). An integrated system for detecting and numbering permanent and deciduous teeth across multiple types of dental X-ray images based on YOLOv8. Diagnostics, 15(13): 1693. https://doi.org/10.3390/diagnostics15131693

[18] Bonfanti-Gris, M., Herrera, A., Paraíso-Medina, S., Alonso-Calvo, R., Martínez-Rus, F., Pradíes, G. (2024). Performance evaluation of three versions of a convolutional neural network for object detection and segmentation using a multiclass and reduced panoramic radiograph dataset. Journal of Dentistry, 144: 104891. https://doi.org/10.1016/j.jdent.2024.104891

[19] Çoban, D., Yaşa, Y., Aktaş, A., İlhan, H.O. (2025). Detection of jaw lesions on panoramic radiographs using deep learning method. Journal of Imaging Informatics in Medicine, 1-16. https://doi.org/10.1007/s10278-025-01642-z

[20] Mendes, A.C., Quintanilha, D.B.P., Pessoa, A.C.P., de Paiva, A.C., dos Santos Neto, P.D.A. (2025). Automated tooth detection and numbering in panoramic radiographs using YOLO. Procedia Computer Science, 256: 1318-1325. https://doi.org/10.1016/j.procs.2025.02.244