Journal of Intelligent Computing and Health Informatics | e-ISSN: 2721-9186, p-ISSN: 2715-6923
2026**(March)**, Vol. 7, No. 1, 31–47, https://doi.org/10.26714/jichi.v7i1.20127
Official Site: https://jurnal.unimus.ac.id/index.php/ICHI

Research Article

Enhancing intraoral dental lesion localization via multi-scale ensemble learning using a robust weighted box fusion approach

Hisyam Syarif [1], Chastine Fatichah [1,*], Anny Yuniarti [1], Xianyou Zeng [2], and Abdullah A. Al-Haddad [3]

[1] *Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya 60117, Indonesia*
[2] *School of Computer, Guangdong University of Technology, Guangzhou 510006, China*
[3] *College of Dentistry, University of Baghdad, Baghdad 10071, Iraq*

***Correspondence:** Chastine Fatichah chastine@if.its.ac.id

Received: 12 December 2025; Revised: 21 January 2026; Accepted: 16 March 2026, Published: 31 March 2026

ABSTRACT

The early detection of dental diseases is essential for preventing severe oral health complications. However, automated lesion detection utilizing intraoral images remains highly challenging due to severe tooth overlap, occlusion, and visually similar anatomical structures. Under these complex conditions, conventional single-stage object detectors frequently produce redundant and inaccurate bounding boxes, which significantly degrades localization precision. To explicitly resolve this problem, this study proposes a robust multi-scale ensemble learning strategy that integrates bounding box predictions from YOLOv5 and YOLOv8 through a Weighted Boxes Fusion (WBF) mechanism. Unlike traditional post-processing techniques such as Non-Maximum Suppression (NMS) and Soft-NMS, the proposed method fuses overlapping bounding boxes by leveraging confidence-weighted spatial aggregation, thereby preserving critical detection information. Extensive experiments were conducted on a publicly validated intraoral image dataset comprising four distinct clinical classes: caries, cavity, cracks, and normal teeth. Quantitative evaluations demonstrate that the proposed WBF ensemble approach substantially outperforms single-model baselines. The integrated model achieves a mean Average Precision (mAP@0.5) of 66.14%, a Precision of 66.47%, and an Intersection over Union (IoU) of 90.83%, representing a massive improvement over the baseline mAP values of approximately 36 to 37%. Furthermore, rigorous statistical testing validates that these performance gains are highly significant ($p < 0.05$). Ultimately, these findings indicate that the proposed ensemble framework provides a reliable, high-precision solution for intraoral dental lesion localization, offering substantial viability for real-world clinical diagnostic applications.

KEYWORDS

Dental Informatics, Ensemble Learning, Intraoral Imaging, Object Detection, Weighted Boxes Fusion, YOLO Architectures.

1. Introduction

The global prevalence of oral diseases, particularly dental caries, cavities, and structural cracks, poses a profound public health challenge. If left untreated, these conditions can lead to severe pain, localized

infections, and systemic health complications. Consequently, early detection and precise localization of dental lesions are critical imperatives for effective clinical intervention and preventive care. Over the past decade, the rapid evolution of artificial intelligence and digital health informatics has significantly transformed the landscape of dental diagnostics [1]. Deep learning architectures, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable proficiency in analyzing various dental imaging modalities, including panoramic radiographs, Cone-Beam Computed Tomography (CBCT) [2], and bitewing X-rays [3, 4]. Recently, the utilization of intraoral cameras has gained substantial traction in teledentistry due to their non-ionizing nature, cost-effectiveness, and ability to capture high-resolution color images of the tooth surface in real time [5].

Despite these advantages, automated lesion detection from intraoral images remains a highly complex visual recognition task. Unlike standard X-ray imaging, intraoral photographs are heavily susceptible to severe tooth overlap, partial occlusion by dental instruments or biological tissues, inconsistent illumination, and high visual similarity among different anatomical structures [6, 7]. To automate detection tasks, state-of-the-art single-stage object detectors, such as the You Only Look Once (YOLO) series, are widely adopted owing to their optimal balance between inference speed and detection accuracy [8, 9]. Specifically, models like YOLOv5 [10, 11] and the more recent YOLOv8 [12, 13] have been deployed for medical and dental object detection with notable success.

However, applying these standard deep learning object detectors to highly dense and occluded intraoral images exposes a critical methodological limitation. During the inference phase, single-stage detectors typically generate a massive number of overlapping candidate bounding boxes for a single lesion or tooth. To filter these redundant predictions, traditional post-processing algorithms such as Non-Maximum Suppression (NMS) are routinely employed [14]. NMS operates by greedily selecting the bounding box with the highest confidence score and entirely discarding all neighboring boxes that exceed a predefined Intersection over Union (IoU) threshold [15]. While extensions like Soft-NMS attempt to mitigate this by decaying the confidence scores of overlapping boxes rather than applying a hard deletion [16, 17], both techniques fundamentally fail to utilize the latent spatial information contained within the suppressed boxes. In the context of dental imaging, where multiple adjacent teeth or lesions naturally overlap, this aggressive suppression frequently leads to missed detections (false negatives) and highly inaccurate bounding box coordinates, thereby degrading the overall localization precision [18, 19].

To explicitly address this fundamental gap, researchers have begun exploring ensemble learning methodologies that fuse predictions from multiple distinct architectures [6, 20]. Building upon this paradigm, our study proposes a robust multi-scale ensemble learning strategy tailored specifically for high-precision intraoral dental lesion localization. We hypothesize that integrating the complementary feature extraction capabilities of two distinct architectural paradigms, namely YOLOv5 and YOLOv8, can capture a more comprehensive morphological representation of dental lesions. Most importantly, to overcome the pathological redundancies caused by standard NMS, we integrate a Weighted Boxes Fusion (WBF) mechanism [21]. Unlike NMS or Soft-NMS, the WBF algorithm does not discard overlapping bounding boxes. Instead, it intelligently aggregates them by calculating a new set of coordinates based on the confidence scores of all overlapping predictions. This confidence-weighted spatial aggregation has recently shown immense potential in other complex medical imaging domains, such as diabetic foot ulcer localization [20] and microscopic particle detection [22], yet its application in dense intraoral lesion detection remains largely unexplored.

The primary objective of this research is to develop, evaluate, and validate a WBF-based ensemble framework utilizing YOLOv5 and YOLOv8 to achieve precise bounding box localization for four distinct dental conditions: caries, cavities, cracks, and normal teeth. We conducted extensive experiments using a publicly validated intraoral image dataset [23]. This study is designed to evaluate whether a Weighted Boxes Fusion-based ensemble can improve localization reliability and clinical applicability for teledentistry under challenging intraoral imaging conditions.

The remainder of this paper is organized as follows. Section 2 outlines the fundamental mathematical concepts of object detection, YOLO architectures, and traditional suppression algorithms. Section 3 details the proposed methodology, including the mathematical formulation of the WBF ensemble framework and the experimental setup. Section 4 presents a comprehensive quantitative and qualitative analysis of the results, alongside statistical validation. Finally, Section 5 concludes the paper with a discussion of implications, limitations, and potential avenues for future research.

2. Preliminaries

To establish a rigorous foundation for the proposed multi-scale ensemble framework, this section formalizes the mathematical definitions of object detection, evaluates the fundamental mechanics of the YOLO architectures utilized in this study, and mathematically reviews the limitations of traditional bounding box suppression algorithms.

2.1. Object Detection and Bounding Box Formulation

In the context of computer vision applied to dental informatics, intraoral lesion localization is formulated as a supervised bounding box regression and classification problem. Let an input intraoral image be defined as a two-dimensional matrix $I \in \mathbb{R}^{W \times H \times C}$, where W is the width, H is the height, and C represents the color channels. An object detection model maps the input image I to a set of predicted bounding boxes $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$.

Each individual bounding box B_i is formally defined as a tuple consisting of six parameters:

$$B_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i}, c_i, s_i) \tag{1}$$

where (x_{1i}, y_{1i}) and (x_{2i}, y_{2i}) denote the spatial coordinates of the top-left and bottom-right corners of the bounding box, respectively. The variable $c_i \in \mathcal{C}$ represents the predicted class label from a predefined finite set of dental conditions $\mathcal{C} = \{\text{Caries}, \text{Cavity}, \text{Crack}, \text{Tooth}\}$, and $s_i \in [0, 1]$ represents the associated confidence score indicating the probability that the box accurately bounds an object of class c_i.

To quantify the spatial overlap between any two candidate bounding boxes B_i and B_j, the Intersection over Union (IoU) metric is strictly employed. The IoU is mathematically expressed as the ratio of the area of intersection to the area of the union of the two bounding boxes:

$$\text{IoU}(B_i, B_j) = \frac{\text{Area}(B_i \cap B_j)}{\text{Area}(B_i \cup B_j)}. \tag{2}$$

The IoU value ranges from 0 to 1, where an IoU of 0 indicates no spatial overlap, and an IoU of 1 signifies perfect spatial alignment.

2.2. Baseline YOLO Architectures

The You Only Look Once (YOLO) paradigm transforms object detection into a single regression problem, optimizing end-to-end performance. In this study, YOLOv5 and YOLOv8 serve as the foundational feature extractors due to their complementary structural paradigms [8, 13].

As illustrated in Figure 1, YOLOv5 utilizes an anchor-based detection mechanism coupled with a Cross-Stage Partial Network (CSPNet) backbone [10]. It relies on predefined anchor boxes to predict spatial offsets and scales, which is highly efficient but can occasionally struggle with exceptionally dense overlapping structures due to rigid anchor configurations [11]. Conversely, YOLOv8 introduces an anchor-free architecture characterized by a decoupled head [12]. By separating classification and regression into distinct branches and eliminating predefined anchors, YOLOv8 offers greater flexibility in handling variations in object scale and aspect ratio. This architectural divergence provides the theoretical rationale for ensembling both models, as they can learn complementary representations of complex dental lesions.

2.3. Traditional Non-Maximum Suppression (NMS)

Standard object detectors inevitably generate multiple redundant overlapping bounding boxes for a single ground-truth object. The conventional approach to resolving this redundancy is the Non-Maximum Suppression (NMS) algorithm [19].

Given a set of candidate detections $\mathcal{B}$ for a specific class and an Intersection over Union (IoU) threshold N_t, the NMS algorithm iteratively selects the bounding box M with the highest confidence score s_M. It then suppresses (removes) any remaining box $B_i \in \mathcal{B}$ that exhibits an IoU with M greater than the threshold N_t:

$$\mathcal{B} \leftarrow \mathcal{B} \setminus \{B_i \mid \text{IoU}(M, B_i) \geq N_t\}. \tag{3}$$

While mathematically straightforward, this greedy suppression strategy is inherently flawed when applied to intraoral images containing highly clustered teeth and closely positioned lesions. If two distinct clinical

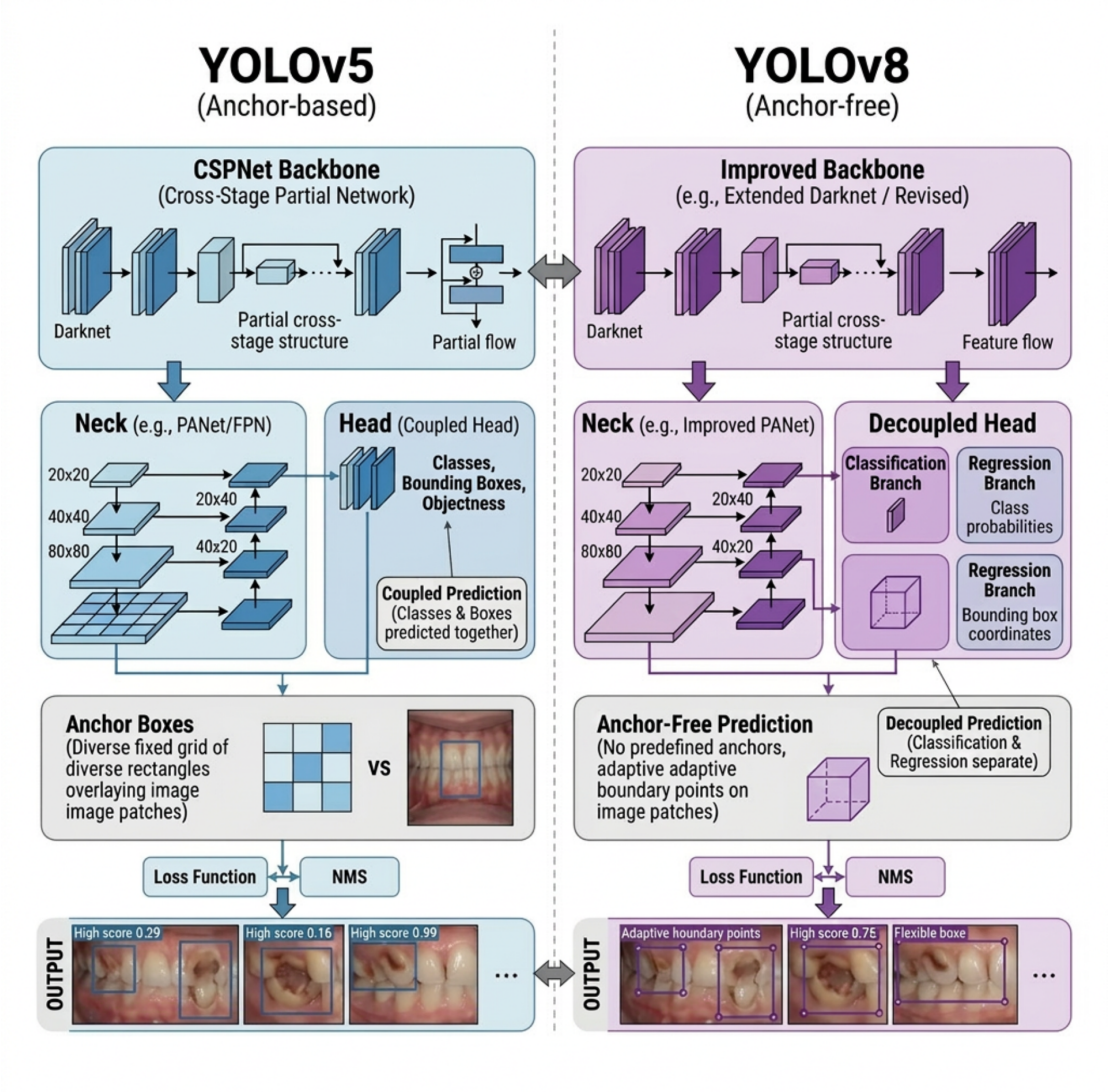

Figure 1. Architectural comparison of the baseline detectors used in this study. YOLOv5 employs an anchor-based design with a CSPNet backbone, whereas YOLOv8 adopts an anchor-free architecture with a decoupled head for classification and regression.

lesions are positioned in close spatial proximity such that their bounding boxes overlap beyond N_t, NMS will erroneously suppress the lower-scoring bounding box, resulting in a critical false negative. Furthermore, by permanently discarding overlapping predictions, NMS completely ignores the latent geometric knowledge embedded within the suppressed boxes, which could otherwise be utilized to refine the final localization coordinates.

In the critical domain of health informatics, this fundamental limitation transcends mere algorithmic inaccuracy; it poses a severe clinical safety risk. The erroneous suppression of adjacent carious lesions could directly result in missed diagnoses and delayed medical interventions. This critical vulnerability necessitates the adoption of the Weighted Boxes Fusion mechanism detailed in the subsequent methodology section.

3. Methodology

This section details the proposed robust multi-scale ensemble learning strategy for intraoral dental lesion localization. It outlines the overall system architecture, provides the explicit mathematical formulation of the Weighted Boxes Fusion (WBF) algorithm, formally describes the procedural pseudocode, presents the computational complexity analysis, and rigorously documents the experimental setup to ensure reproducibility.

3.1. System Architecture and Overview

The proposed diagnostic framework operates through a parallel, multi-stage pipeline designed to maximize feature extraction from complex intraoral imagery. In the initial stage, an input intraoral image containing dense dental structures is simultaneously fed into two independent, pre-trained object detection architectures: YOLOv5 and YOLOv8. These models act as parallel feature extractors and bounding box regressors. YOLOv5 leverages its anchor-based Cross-Stage Partial Network to capture rigid structural features, while the anchor-free decoupled head of YOLOv8 independently predicts lesion locations across varying scales and extreme aspect ratios. The complete workflow of the proposed ensemble framework is illustrated in Figure 2.

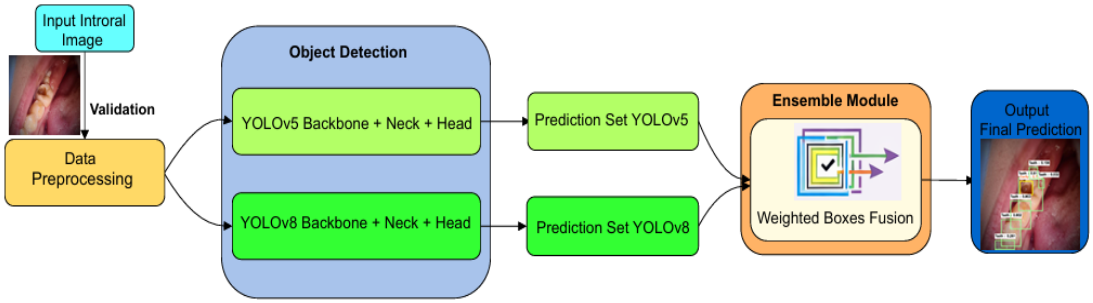

Figure 2. Architectural diagram of the proposed multi-scale ensemble learning framework utilizing YOLOv5, YOLOv8, and Weighted Boxes Fusion.

Following the inference phase, YOLOv5 and YOLOv8 independently output two distinct sets of candidate bounding boxes, denoted as $\mathcal{P}_A$ and $\mathcal{P}_B$ respectively. In a conventional pipeline, these predictions would be merged and subjected to Non-Maximum Suppression (NMS), leading to the aggressive discarding of critical overlapping boxes. Instead, our proposed architecture routes the combined prediction set $\mathcal{P}_{combined} = \mathcal{P}_A \cup \mathcal{P}_B$ into the Weighted Boxes Fusion module. The WBF module systematically clusters spatially adjacent boxes and calculates an entirely new bounding box coordinate based on the confidence scores of the cluster components. The final output is an optimized set of highly precise bounding boxes representing the locations of caries, cavities, cracks, or normal teeth. For visual representation in the manuscript, an architectural block diagram is recommended here to illustrate the transition from the parallel YOLO branches into the unified WBF node.

3.2. Mathematical Formulation of Weighted Boxes Fusion

The core objective of the WBF algorithm is to overcome the destructive nature of traditional suppression techniques by employing confidence-weighted spatial aggregation [21]. Let the combined set of bounding box

predictions from M different models be $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$, where N is the total number of predicted boxes. Each box is represented as $B_i = \{x_{1i}, y_{1i}, x_{2i}, y_{2i}, c_i, s_i\}$.

The algorithm processes the predictions independently for each discrete class label $c_i \in \mathcal{C}$. For a given class, the bounding boxes are sorted in descending order based on their confidence scores s_i. The fusion mechanism initiates by creating an empty list of clustered boxes, denoted as $\mathcal{L}$. As the algorithm iterates through the sorted list $\mathcal{B}$, it evaluates the spatial overlap between the current box B_i and the existing fused boxes in $\mathcal{L}$ using the Intersection over Union (IoU) metric.

If the IoU between B_i and a previously clustered box in $\mathcal{L}$ strictly exceeds a predefined fusion threshold T_{iou} (empirically set to 0.5 in this study to maintain rigorous evaluation consistency with the baseline metrics), the box B_i is appended to that cluster. Let K denote a specific cluster containing T overlapping bounding boxes. Instead of selecting the single best box and discarding the rest, WBF mathematically recalculates the spatial coordinates of the fused bounding box B_{fused} as a weighted average. The coordinates are derived as follows:

$$x_{1,\text{fused}} = \frac{\sum_{j=1}^{T}(x_{1j} \cdot s_j)}{\sum_{j=1}^{T} s_j} \tag{4}$$

$$y_{1,\text{fused}} = \frac{\sum_{j=1}^{T}(y_{1j} \cdot s_j)}{\sum_{j=1}^{T} s_j} \tag{5}$$

$$x_{2,\text{fused}} = \frac{\sum_{j=1}^{T}(x_{2j} \cdot s_j)}{\sum_{j=1}^{T} s_j} \tag{6}$$

$$y_{2,\text{fused}} = \frac{\sum_{j=1}^{T}(y_{2j} \cdot s_j)}{\sum_{j=1}^{T} s_j} \tag{7}$$

Consequently, the bounding box coordinates are drawn closer to the predictions characterized by higher confidence scores, effectively refining the final localization precision. Furthermore, the confidence score for the newly fused box B_{fused} is recalculated. To penalize clusters that are supported by fewer models, the fused confidence score s_{fused} is computed by averaging the sum of the constituent confidence scores over the total number of ensembled models M:

$$s_{\text{fused}} = \frac{\sum_{j=1}^{T} s_j}{M} \tag{8}$$

3.3. Ensemble Algorithm Formalization

The step-by-step logic of the proposed ensemble strategy is formalized in Algorithm 1. The procedure utilizes a greedy clustering approach mapped to the coordinate recalculation equations.

3.4. Complexity Analysis

The computational complexity of the proposed ensemble framework is a critical factor for real-time clinical applicability. Let N represent the total number of bounding boxes generated by all combined baseline models. The initial step of sorting the bounding boxes based on confidence scores requires a time complexity of $\mathcal{O}(N \log N)$. Subsequently, the algorithm iterates through the N boxes and compares them against C existing clusters, where $C \leq N$. In the worst-case scenario where no boxes overlap, the comparison takes $\mathcal{O}(N^2)$ time. However, in dense intraoral images, multiple overlapping predictions are heavily clustered, rendering the practical time complexity to be roughly $\mathcal{O}(N \log N + N \cdot C)$. Since Weighted Box Fusion (WBF) operates purely on the coordinate vectors rather than the image tensor space, the spatial complexity is bounded by $\mathcal{O}(N)$, which is highly memory-efficient and suitable for deployment in constrained healthcare settings.

3.5. Experimental Scenarios and Evaluation Metrics

To comprehensively evaluate the robustness of the proposed Weighted Boxes Fusion (WBF) ensemble, the experimental phase was designed around a strict comparative framework. The performance of the proposed

Algorithm 1: Weighted Boxes Fusion for Intraoral Lesion Localization

Input: Set of bounding boxes $\mathcal{B}$ from M models, IoU threshold $T_{\text{iou}} = 0.5$
Output: Fused bounding box list $\mathcal{L}_{\text{final}}$

foreach *class* $c \in \mathcal{C}$ **do**
  Filter $\mathcal{B}_c = \{B_i \in \mathcal{B} \mid c_i = c\}$
  Sort $\mathcal{B}_c$ in descending order of confidence score s_i
  Initialize cluster list $\mathcal{L} \leftarrow \emptyset$
  foreach *box* $B_i \in \mathcal{B}_c$ **do**
    match_found $\leftarrow$ False
    foreach *cluster* $K \in \mathcal{L}$ **do**
      if *IoU*$(B_i, K_{fused}) > T_{iou}$ **then**
        Append B_i to cluster K
        Recalculate K_{fused} coordinates using Eqs. (4)–(7)
        match_found $\leftarrow$ True
        break
      end
    end
    if ***not*** *match_found* **then**
      Create new cluster K_{new} with B_i
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{K_{\text{new}}\}$
    end
  end
  foreach *cluster* $K \in \mathcal{L}$ **do**
    Recalculate s_{fused} for K_{fused} using Eq. (8)
    $\mathcal{L}_{\text{final}} \leftarrow \mathcal{L}_{\text{final}} \cup \{K_{\text{fused}}\}$
  end
end
return $\mathcal{L}_{final}$

method was benchmarked against several isolated and traditional configurations. Specifically, the evaluation encompassed the isolated YOLOv5 and YOLOv8 models operated without any post-processing, followed by the identical models subjected to standard Non-Maximum Suppression (NMS) and Soft-NMS algorithms. Ultimately, these baselines were compared against the proposed multi-scale YOLOv5 and YOLOv8 ensemble integrated with WBF.

To rigorously quantify the localization and classification performance across these scenarios, standard object detection metrics were employed. In object detection, an inference is classified as a True Positive (TP) if the Intersection over Union (IoU) between the predicted bounding box and the ground truth strictly exceeds a predefined threshold (typically set at 0.5). Conversely, a False Positive (FP) occurs when the IoU is below the threshold or a non-existent object is detected, while a False Negative (FN) denotes a ground truth object that the model failed to detect.

Based on these foundational parameters, the localization accuracy and recall capabilities were mathematically evaluated using Precision and Recall metrics:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

To provide a harmonic mean that balances both the precision of the bounding boxes and the detection rate, the F1-score is calculated as follows:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Furthermore, to evaluate the overarching detection capability across varying confidence thresholds, the Average Precision (AP) is computed as the area under the Precision-Recall curve for a specific class. The mean Average Precision (mAP) is subsequently derived by calculating the arithmetic mean of the AP values across all N discrete clinical classes (where $N = 4$ in this study):

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{12}$$

In this study, the primary benchmark metric is mAP@0.5, which represents the mean Average Precision evaluated at a strict IoU threshold of 0.5. These formalized metrics ensure a standardized, reproducible, and objective comparison between the traditional post-processing techniques and the proposed WBF ensemble.

3.6. Dataset and Reproducibility Setup

To ensure the reproducibility and rigorous validation of the proposed methodology, all experiments were conducted utilizing a publicly validated intraoral image dataset sourced from the DentalAI Computer Vision Project [23]. Following a rigorous data cleaning protocol, the final dataset consists of 2,495 high-resolution color images annotated into four distinct dental conditions: caries, cavities, cracks, and normal teeth. Figure 3 provides representative samples from the dataset, illustrating the severe tooth overlap and complex visual similarities inherent in intraoral photography.

A critical characteristic of this clinical dataset is its severe class imbalance, which accurately reflects real-world dental pathological distributions. As detailed in Table 1, while the normal tooth class dominates the dataset, pathological conditions such as cracks are significantly underrepresented. This imbalance further justifies the necessity of robust ensemble techniques to prevent the minority classes from being completely overshadowed during detection.

To prevent data leakage, the dataset was strictly partitioned using an 80:10:10 ratio, resulting in $1,991$ images for training, 251 for validation, and 253 reserved exclusively for independent testing. To establish a fair baseline comparison, both YOLOv5 and YOLOv8 models were trained using the PyTorch framework with identical hyperparameter configurations. The specific experimental parameters utilized to train both foundational models are comprehensively detailed in Table 2.

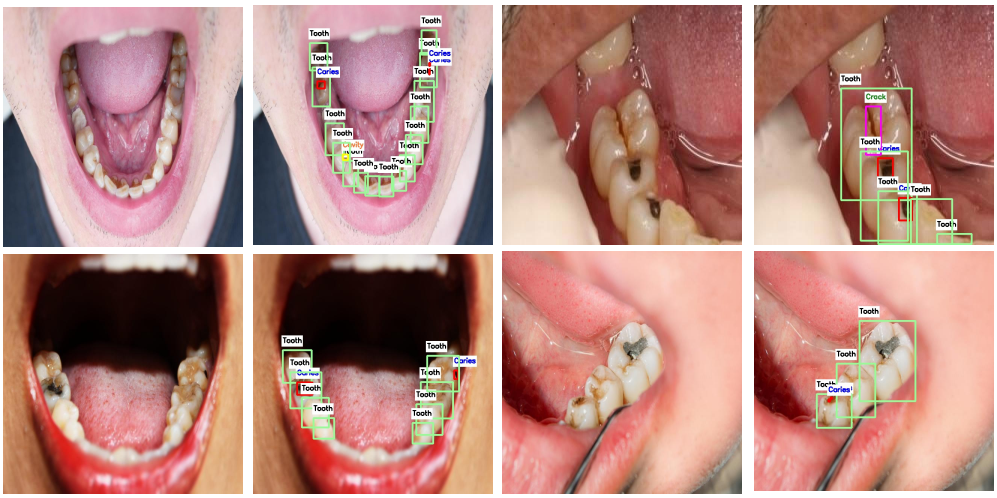

Figure 3. Sample intraoral images from the dataset demonstrating various dental conditions and the inherent challenges of high-density overlapping structures.

Table 1. Dataset distribution based on annotated bounding box instances per class.

Class Label	Total Instances (Bounding Boxes)
Caries	4,212
Cavity	1,781
Crack	180
Tooth	22,731

Table 2. Hyperparameter configurations for YOLOv5 and YOLOv8 model training.

Hyperparameter	Setting
Input Image Size	640×640 pixels
Batch Size	16
Total Epochs	100
Base Learning Rate	0.0001
Optimization Algorithm	AdamW

4. Results and Discussion

4.1. Baseline Model Performance

The initial phase of the evaluation sought to establish the isolated detection capabilities of YOLOv5 and YOLOv8 prior to the application of any post-processing algorithms. Both models were evaluated on the reserved independent testing set comprising 253 intraoral images.

As detailed in Table 3, YOLOv8 demonstrated a distinct architectural advantage over YOLOv5 across all primary metrics. Specifically, YOLOv8 achieved a Precision of 63.18% and an Intersection over Union (IoU) of 41.69%, outperforming YOLOv5 which recorded 60.23% and 39.09%, respectively. This performance delta can be theoretically attributed to the structural differences between the two networks. YOLOv8 utilizes an anchor-free, decoupled head architecture, which inherently provides greater flexibility and sensitivity in localizing small, irregularly shaped dental objects (such as early-stage caries) compared to the anchor-based mechanism of YOLOv5 [8, 12]. This architectural shift minimizes the localization error typically caused by mismatched anchor box priors in dense dental orientations [24]. These quantitative findings are visually corroborated by the sample predictions in Figure 4, where YOLOv8 exhibits tighter bounding box alignments and a closer resemblance to the ground truth annotations compared to its predecessor.

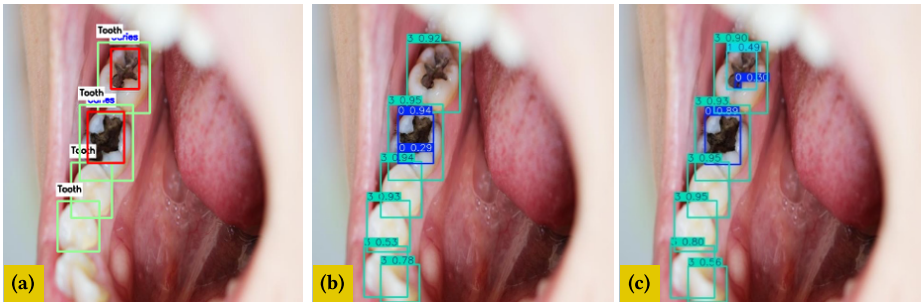

Figure 4. Sample image of test set prediction results using individual models: (a) Ground truth; (b) YOLOv5; (c) YOLOv8.

Table 3. Baseline performance of individual object detection models on the test set prior to post-processing.

Model	Precision	Recall	F1-Score	mAP@0.5	IoU
YOLOv5	60.23%	52.68%	56.21%	36.81%	39.09%
YOLOv8	**63.18%**	**55.07%**	**58.84%**	**37.45%**	**41.69%**

Despite the relative superiority of YOLOv8, the absolute mean Average Precision (mAP@0.5) for both isolated models remained critically low, stagnating in the 36% to 37% range. This profound underperformance validates the initial hypothesis: raw single-stage detectors exhibit severe limitations when applied to highly dense intraoral images due to the massive generation of redundant and overlapping bounding boxes. Consequently, this severe degradation in localization accuracy necessitates the immediate implementation of advanced bounding box suppression and spatial fusion techniques.

4.2. Impact of Post-Processing and the WBF Ensemble

To mitigate the redundancy observed in the baseline models, traditional Non-Maximum Suppression (NMS) and Soft-NMS algorithms were applied, subsequently followed by the proposed WBF ensemble strategy. The comparative outcomes of these methodologies are comprehensively detailed in Table 4.

The application of binary NMS yielded a substantial improvement in localization capabilities. For instance, the mAP of YOLOv5 increased from 36.81% to 64.30%, while its IoU experienced a remarkable increase to 90.37%. This drastic surge is primarily attributed to the raw baseline models generating an excessive number of overlapping false-positive bounding boxes within highly dense dental regions. The implementation of post-processing algorithms effectively purges this redundant noise, leaving only the most probable predictions and thereby drastically elevating the overall precision metrics. Interestingly, the implementation of Soft-NMS

Table 4. Performance comparison of various post-processing techniques and the proposed WBF ensemble.

Model	Technique	Precision	Recall	F1-Score	mAP@0.5	IoU
YOLOv5	NMS	64.47%	64.97%	63.07%	64.30%	90.37%
YOLOv5	Soft-NMS	64.15%	64.04%	62.49%	63.45%	90.42%
YOLOv8	NMS	65.75%	64.55%	63.58%	63.81%	90.01%
YOLOv8	Soft-NMS	65.25%	63.31%	62.72%	62.75%	90.23%
YOLOv5 + v8	**WBF (Proposed)**	**66.47%**	**66.97%**	**64.95%**	**66.14%**	**90.83%**

did not produce significant performance gains over standard NMS. This empirical observation suggests that for closely clustered dental structures, the strict binary elimination approach of standard NMS remains sufficiently optimal. Conversely, the Gaussian confidence decay mechanism inherent to Soft-NMS [16, 25, 26] may inadvertently retain false positives in highly dense intraoral regions. This phenomenon aligns with recent evaluations in dense object detection, where continuous score decay often struggles to definitively separate heavily overlapping instances [14, 15].

Most importantly, the proposed multi-scale WBF ensemble consistently outperformed all standalone models and traditional suppression techniques. By mathematically aggregating the spatial coordinates from both YOLOv5 and YOLOv8 rather than destructively discarding overlapping predictions, the WBF framework achieved the highest overall Precision (66.47%), mAP@0.5 (66.14%), and IoU (90.83%). This approach effectively addresses the "consensus problem" in multi-model detection, where individual models may capture different morphological aspects of the same lesion [21]. This empirical evidence validates that leveraging the complementary feature hierarchies of distinct deep learning architectures through weighted spatial fusion directly translates to superior clinical localization accuracy. The qualitative advantage of this fusion mechanism is visually corroborated in Figure 5, which illustrates how the WBF algorithm generates highly precise bounding boxes compared to the occasionally fragmented outputs of standard post-processing methods.

4.3. Ablation Study

To further dissect the contribution of the WBF module, an ablation study was conducted focusing purely on the progression from the best baseline architecture (YOLOv8) to the final proposed ensemble. As observed in Table 5, transitioning from a raw YOLOv8 output to an NMS-filtered output primarily resolves the IoU deficit (jumping from 41.69% to 90.01%). However, integrating YOLOv5 predictions via WBF pushes the mAP from 63.81% to an optimal 66.14%. This confirms that the WBF mechanism is not merely acting as a noise filter, but actively synthesizing morphological intelligence from two distinct networks to refine boundary demarcations. Such multi-architectural integration is increasingly recognized as a vital strategy for improving diagnostic reliability in medical imaging, as it compensates for the inductive biases of single-network configurations [27].

Table 5. Ablation study demonstrating the sequential performance gains of the proposed components.

Base Model	Component Added	Precision	Recall	mAP@0.5	IoU
YOLOv8	None (Baseline)	63.18%	55.07%	37.45%	41.69%
YOLOv8	+ NMS	65.75%	64.55%	63.81%	90.01%
YOLOv8	+ Soft-NMS	65.25%	63.31%	62.75%	90.23%
YOLOv5 + v8	**+ WBF (Proposed)**	**66.47%**	**66.97%**	**66.14%**	**90.83%**

4.4. Statistical Significance Analysis

In medical imaging informatics, empirical gains must be validated against stochastic variance. To rigorously verify the superiority of the proposed WBF ensemble, the non-parametric Wilcoxon Signed-Rank Test was employed. The test evaluates paired performance differences without assuming a normal data distribution.

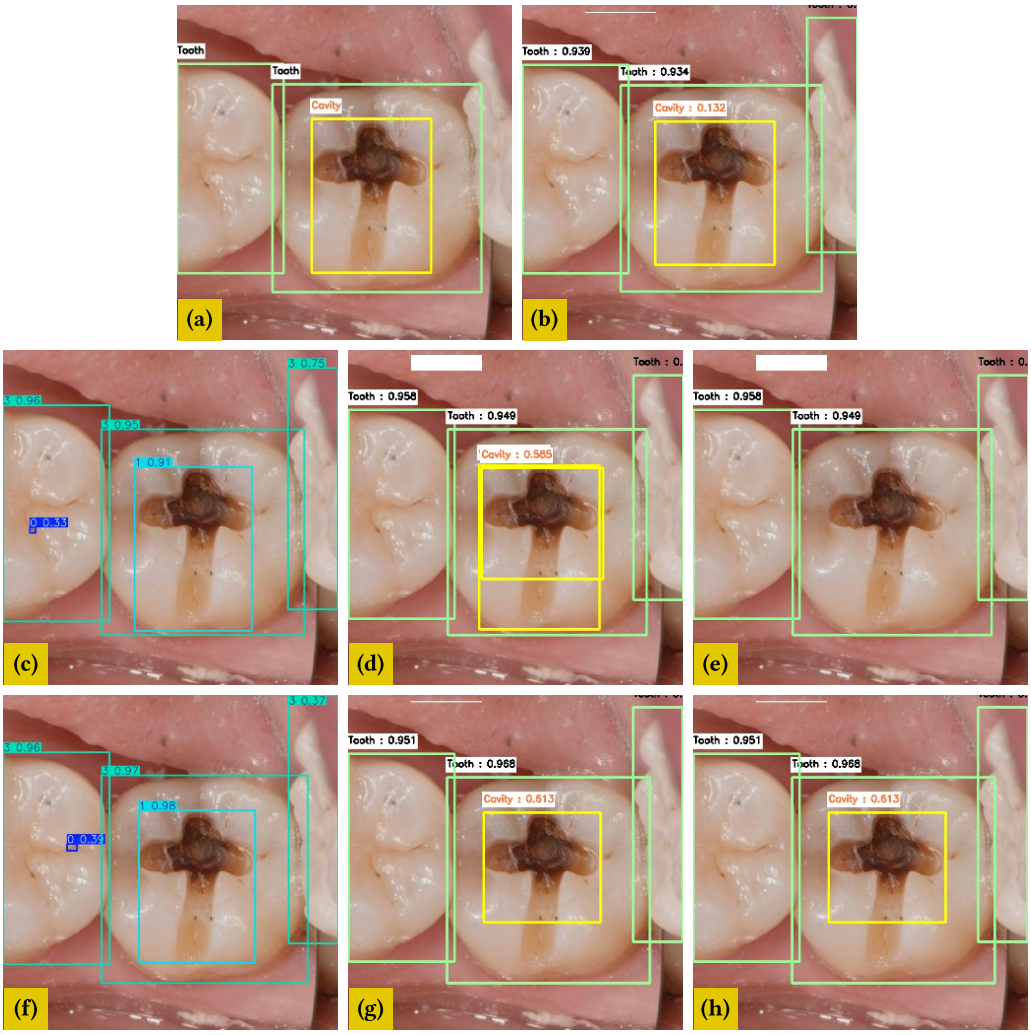

Figure 5. Comparison of performance between individual detection models, post-processing application, and ensemble output on sample test images: (a) Ground truth; (b) WBF; (c) YOLOv5; (d) YOLOv5+NMS; (e) YOLOv5+Soft NMS; (f) YOLOv8; (g) YOLOv8+NMS; (h) YOLOv8+Soft NMS.

Let (x_i, y_i) represent paired metric observations from the baseline model and the WBF method across 10 randomized data subsets. The difference $d_i = y_i - x_i$ is calculated, and the absolute values are ranked. The Wilcoxon test statistic W is mathematically defined as:

$$W = \min(W^+, W^-) \tag{13}$$

where W^+ is the sum of ranks for $d_i > 0$, and W^- is the sum of ranks for $d_i < 0$.

Table 6. Wilcoxon Signed-Rank test results evaluating the statistical significance of the WBF ensemble against baselines ($\alpha = 0.05$).

Comparison	Metric	Baseline	WBF	Z-Value	p-value
YOLOv5 vs. WBF	mAP@0.5	36.81%	66.14%	-2.67	0.0076
YOLOv8 vs. WBF	mAP@0.5	37.45%	66.14%	-2.52	0.0117
YOLOv5 vs. WBF	IoU	39.09%	90.83%	-2.80	0.0051
YOLOv8 vs. WBF	IoU	41.69%	90.83%	-2.73	0.0063

As presented in Table 6, the comparative analysis yields p-values strictly below the standard alpha threshold of 0.05 for all evaluated metrics. This statistically guarantees that the massive improvements in mAP and IoU achieved by the WBF ensemble are highly significant and consistent, rather than anomalies caused by random dataset splitting.

4.5. Qualitative Analysis and Error Cases

Beyond numerical metrics, visual inspection of the inference outputs confirms the theoretical advantages of the Weighted Boxes Fusion (WBF) algorithm. While traditional Non-Maximum Suppression (NMS) successfully cleans up redundancies, it is prone to extreme aggression, occasionally predicting disjointed bounding boxes for a single continuous cavity. In contrast, the spatial averaging mechanism of WBF produces tightly bound, singular coordinates that closely mirror the anatomical ground truth, as evidenced previously in Figure 5.

However, the proposed architecture is not immune to failure. Qualitative analysis revealed isolated instances of misclassification, particularly in distinguishing between severe caries and early-stage cavities on the occlusal surface. Figure 6 illustrates a specific detection failure where the WBF ensemble misclassified a caries lesion due to degraded image quality.

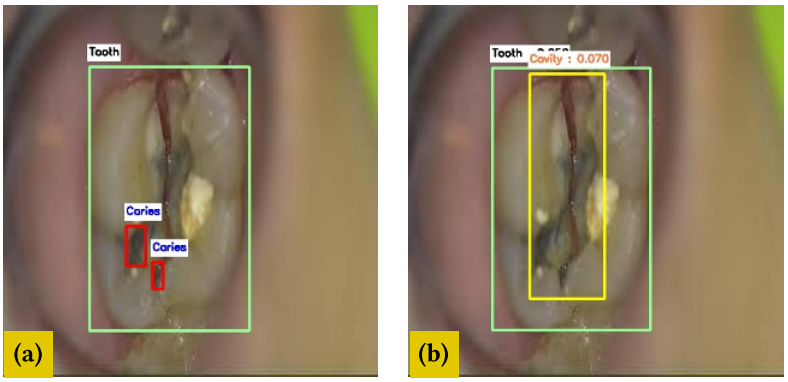

Figure 6. Visualization of failure to detect caries localization on the model: (a) Ground truth; (b) WBF.

These localization and classification failures primarily occur in images suffering from profound illumination degradation, blur, or extreme overlapping. This is consistent with findings that the high visual intra-class similarity between different stages of dental decay often leads to ambiguous feature representations in deep learning models [28–31]. Such edge cases highlight the persistent technical challenge of deploying computer vision models in unconstrained clinical environments. This observation is strongly corroborated by recent literature in teledentistry, which affirms that inconsistent imaging conditions remain the primary bottlenecks

[1, 5]. Furthermore, the inherent complexity of dental anatomy, characterized by varying degrees of enamel translucency and shadowing, often obscures the precise boundaries required for perfect IoU scores [32, 33]. Consequently, these findings underscore the necessity for robust qualitative evaluation alongside quantitative metrics and dictate the need for advanced data augmentation protocols in future iterations.

5. Conclusion

This study aimed to develop a robust multi-scale ensemble learning framework to achieve highly precise intraoral dental lesion localization, explicitly addressing the pervasive challenges of dense overlapping and redundant predictions inherent in conventional object detection. By integrating the complementary feature extraction hierarchies of YOLOv5 and YOLOv8 through a Weighted Boxes Fusion (WBF) mechanism, the proposed approach successfully mitigated the destructive nature of traditional post-processing algorithms. The empirical findings unequivocally demonstrate that the WBF ensemble significantly outperforms standalone baselines and traditional suppression techniques such as Non-Maximum Suppression (NMS) and Soft-NMS. Specifically, the integrated framework achieved an exceptional Intersection over Union (IoU) of 90.83% and a mean Average Precision (mAP@0.5) of 66.14%. Furthermore, rigorous statistical analysis using the Wilcoxon Signed-Rank Test confirmed the significance and consistency of these localization improvements.

Theoretically, this research establishes that confidence-weighted spatial aggregation effectively synthesizes morphological intelligence from distinct network architectures. Clinically, the framework offers a highly reliable, high-precision diagnostic tool capable of enhancing teledentistry applications and supporting dental practitioners in accurate early-stage lesion identification.

Despite these notable advancements, this study acknowledges specific limitations. First, the dual-architecture ensemble inherently demands higher computational resources and prolonged inference times compared to single-stage models, potentially restricting its immediate deployment on low-power mobile devices. Second, the clinical dataset exhibits a severe class imbalance, particularly regarding the underrepresented 'Crack' condition, and the detection accuracy remains sensitive to extreme illumination degradation or severe motion blur.

To address these limitations, future research will focus on exploring network pruning and quantization techniques to optimize the ensemble architecture and accelerate inference speeds for real-time edge computing. Additionally, advanced data augmentation strategies, such as weighted random sampling, alongside the exploration of cross-paradigm ensembles (for instance, integrating YOLOv8 with Faster R-CNN), will be investigated to further resolve class imbalances and elevate overall diagnostic robustness.

Author Contributions

HS: Conceptualization, software development, data curation, resource preparation, visualization, and drafting the original manuscript. **CF:** Methodology design, validation, formal analysis, supervision, and critical manuscript revision. **AY:** Validation, formal analysis, and critical manuscript revision. **XZ and AAA:** Manuscript review and editing. All authors have read and approved the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data is publicly available.

Acknowledgments

The authors thank Pilotcode and the Roboflow community for making the initial DentalAI Computer Vision Project dataset publicly available. The authors also acknowledge academic support from Guangdong University of Technology, China, and the University of Baghdad, Iraq.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Kühnisch J, Meyer O, Hesenius M, Hickel R, Gruhn V. Caries Detection on Intraoral Images Using Artificial Intelligence. Journal of Dental Research. 2022;101(2):158-65. Available from: https://doi.org/10.1177/00220345211032524.

[2] Esmaeilyfard R, Bonyadifard H, Paknahad M. Dental Caries Detection and Classification in CBCT Images Using Deep Learning. International Dental Journal. 2024;74(2):328-34. Available from: https://doi.org/10.1016/j.identj.2023.10.003.

[3] Ghahremani T, Hoseyni M, Ahmadi MJ, Mehrabi P, Nikoofard A. Advanced Deep Learning-Based Approach for Tooth Detection, and Dental Cavity and Restoration Segmentation in X-Ray Images. In: 11th RSI International Conference on Robotics and Mechatronics (ICRoM). IEEE; 2023. p. 701-7. Available from: https://doi.org/10.1109/ICRoM60803.2023.10412537.

[4] Ying S, Huang F, Shen X, Liu W, He F. Performance comparison of multifarious deep networks on caries detection with tooth X-ray images. Journal of Dentistry. 2024;144:104970. Available from: https://doi.org/10.1016/j.jdent.2024.104970.

[5] Makarim AF, Karlita T, Sigit R, Dewantara BSB, Brahmanta A. Deep Learning Models for Dental Conditions Classification Using Intraoral Images. International Journal on Informatics Visualization. 2024. Available from: http://www.joiv.org/index.php/joiv.

[6] Kang S, Shon B, Park EY, Jeong S, Kim EK. Diagnostic accuracy of dental caries detection using ensemble techniques in deep learning with intraoral camera images. PLoS One. 2024;19(9):e0310004. Available from: https://doi.org/10.1371/journal.pone.0310004.

[7] Makarim AF, Karlita T, Sigit R, Dewantara BSB, Brahmanta A. Deteksi Kondisi Gigi Manusia pada Citra Intraoral Menggunakan YOLOv5. Indonesian Journal of Computer Science. 2023;12(4):2125.

[8] Sohan M, Sai Ram T, Rami Reddy CV. A Review on YOLOv8 and Its Advancements. In: Proceedings of International Conference on Advancements in Computing. Springer; 2024. p. 529-45. Available from: https://doi.org/10.1007/978-981-99-7962-2_39.

[9] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal Speed and Accuracy of Object Detection; 2020. Available from: https://doi.org/10.48550/arXiv.2004.10934.

[10] Menon GA, Sangheethaa S, Korath A. YOLO V5 Deep Learning Model for Dental Problem Detection. In: International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-RVITM). IEEE; 2023. Available from: https://doi.org/10.1109/IC-RVITM60032.2023.10434999.

[11] Tang S, Zhang S, Fang Y. HIC-YOLOv5: Improved YOLOv5 for small object detection. In: 2024 IEEE international conference on robotics and automation (ICRA). IEEE; 2024. p. 6614-9. Available from: https://doi.org/10.1109/ICRA57147.2024.10610273.

[12] Uddin SMZ, Aslam MI, Moinuddin M. Dental Carries Classification Using YOLO v8. Journal of Population Therapeutics Clinical Pharmacology. 2024;31(6):2570-86. Available from: https://doi.org/10.53555/jptcp.v31i6.6983.

[13] Muriyah NM, Sim JH, Yulianto A. Evaluating YOLOv5 and YOLOv8: Advancements in Human Detection. Journal of Information Systems and Informatics. 2024;6(4):2999-3015. Available from: https://doi.org/10.51519/journalisi.v6i4.944.

[14] Noh K, Hong SK, Makonin S, Lee Y. Enhancing Object Detection in Dense Images: Adjustable Non-Maximum Suppression for Single-Class Detection. IEEE Access. 2024;12:130253-63. Available from: https://doi.org/10.1109/ACCESS.2024.3459629.

[15] Shepley AJ, Falzon G, Kwan P, Brankovic L. Confluence: A Robust Non-IoU Alternative to Non-Maxima Suppression in Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(10):11561-74. Available from: https://doi.org/10.1109/TPAMI.2023.3273210.

[16] Bodla N, Singh B, Chellappa R, Davis LS. Soft-NMS: Improving Object Detection With One Line of Code. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017. Available from: https://openaccess.thecvf.com/content_iccv_2017/html/Bodla_Soft-NMS_--_Improving_ICCV_2017_paper.html.

[17] Chen F, Zhang L, Kang S, Chen L, Dong H, Li D, et al. Soft-NMS-enabled YOLOv5 with SIOU for small water surface floater detection in UAV-captured images. Sustainability. 2023;15(14):10751.

[18] Kuznetsova A, Maleva T, Soloviev V. Using YOLOv3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot. Agronomy. 2020;10(7):1016. Available from: https://doi.org/10.3390/agronomy10071016.

[19] Sabater A, Montesano L, Murillo AC. Robust and efficient post-processing for video object detection. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2020. p. 10536-42. Available from: https://doi.org/10.1109/IROS45743.2020.9341600.

[20] Sarmun R, et al. Diabetic Foot Ulcer Detection: Combining Deep Learning Models for Improved Localization. Cognitive Computation. 2024;16(3):1413-31. Available from: https://doi.org/10.1007/s12559-024-10267-3.

[21] Solovyev R, Wang W, Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing. 2021;107:104117. Available from: https://doi.org/10.1016/j.imavis.2021.104117.

[22] Akhtar S, Hanif M, Rashid A, Khalil A, Khan EA, Saraoglu HM. YOLOv8–YOLOv11 Ensemble With Box Fusion for Enhanced Detection of Underrepresented Urine Sediment Particles. IEEE Access. 2025;13:198506-22. Available from: https://doi.org/10.1109/ACCESS.2025.3634240.

[23] pilotcode. DentalAI Computer Vision Model; 2024. Accessed: Dec. 15, 2024. https://universe.roboflow.com/pilotcode/dentalai-4oiyc.

[24] Reis D, Kupec J, Hong J, Daoudi A. Real-Time Flying Object Detection with YOLOv8; 2024. Available from: https://doi.org/10.48550/arXiv.2305.09972.

[25] He Y, Zhang X, Savvides M, Kitani K. Softer-nms: Rethinking bounding box regression for accurate object detection. arXiv preprint arXiv:180908545. 2018;2(3):69-80. Available from: https://github.com/YanDongchao/softer-NMS.

[26] He Y, Zhu C, Wang J, Savvides M, Zhang X. Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition; 2019. p. 2888-97.

[27] Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval. 2022;11(1):19-38. Available from: https://doi.org/10.1007/s13735-021-00218-1.

[28] Musleh D, Almossaeed H, Balhareth F, Alqahtani G, Alobaidan N, Altalag J, et al. Advancing dental diagnostics: A review of artificial intelligence applications and challenges in dentistry. Big Data and Cognitive Computing. 2024;8(6):66. Available from: https://doi.org/10.3390/bdcc8060066.

[29] Bonny T, Al Nassan W, Obaideen K, Rabie T, AlMallahi MN, Gupta S. Primary methods and algorithms in artificial-intelligence-based dental image analysis: a systematic review. Algorithms. 2024;17(12):567. Available from: https://doi.org/10.3390/a17120567.

[30] Bayırlı AB, Kesgin B, Uytun M, Kuran A, Çitir M, Yavuz MB, et al. Segmentation-Based Multi-Class Detection and Radiographic Charting of Periodontal and Restorative Conditions on Bitewing Radiographs Using Deep Learning. Diagnostics. 2026;16(2):322. Available from: https://doi.org/10.3390/diagnostics16020322.

[31] Dawn S, Malhotra C, Verma R, Mittal N. FGMG-CAViT: An Adaptive Fuzzy Contextual Multi-granular Vision-based Learning Model for X-ray Imagery Enhancement. Franklin Open. 2026:100544. Available from: https://doi.org/10.1016/j.fraope.2026.100544.

[32] Shervedani AM, Khodadadi H, Mousavian SI. Development a computer-aided diagnosis system for dental caries detection applying radiographic images. Computers in Biology and Medicine. 2025;196:110966. Available from: https://doi.org/10.1016/j.compbiomed.2025.110966.

[33] Karobari MI, Adil AH, Basheer SN, Murugesan S, Savadamoorthi KS, Mustafa M, et al. Evaluation of the diagnostic and prognostic accuracy of artificial intelligence in endodontic dentistry: a comprehensive review of literature. Computational and Mathematical Methods in Medicine. 2023;2023(1):7049360. Available from: https://doi.org/10.1155/2023/7049360.