Version of Record: https://www.sciencedirect.com/science/article/pii/S0010482522005881
Manuscript_36158112685d6356779aaae097ab7bc4

# Collaborative Deep Learning Model for Tooth Segmentation and Identification using Panoramic Radiographs⋆

Geetha Chandrashekar[a,1], Saeed AlQarni[a,b,1], Erin Ealba Bumann[c] and Yugyung Lee[a,*]

[a]Department of Computer Science Electrical Engineering, University of Missouri - Kansas City, Missouri, USA
[b]Department of Computing and Informatics, Saudi Electronic University - KSA
[c]Department of Oral and Craniofacial Sciences, University of Missouri – Kansas City, Missouri, USA

ARTICLE INFO

*Keywords*:
Collaborative Learning
Ensemble Learning
Summarization
Tooth Segmentation
Tooth Identification
Panoramic Radiographs

ABSTRACT

Panoramic radiographs are an integral part of effective dental treatment planning, supporting dentists in identifying impacted teeth, infections, malignancies, and other dental issues. However, screening for anomalies solely based on a dentist's assessment may result in diagnostic inconsistency, posing difficulties in developing a successful treatment plan. Recent advancements in deep learning-based segmentation and object detection algorithms have enabled the provision of predictable and practical identification to assist in the evaluation of a patient's mineralized oral health, enabling dentists to construct a more successful treatment plan. However, there has been a lack of efforts to develop collaborative models that enhance learning performance by leveraging individual models. The article describes a novel technique for enabling collaborative learning by incorporating tooth segmentation and identification models created independently from panoramic radiographs. This collaborative technique permits the aggregation of tooth segmentation and identification to produce enhanced results by recognizing and numbering existing teeth (up to 32 teeth). The experimental findings indicate that the proposed collaborative model is significantly more effective than individual learning models (e.g., 98.77% vs. 96% and 98.44% vs.91% for tooth segmentation and recognition, respectively). Additionally, our models outperform the state-of-the-art segmentation and identification research. We demonstrated the effectiveness of collaborative learning in detecting and segmenting teeth in a variety of complex situations, including healthy dentition, missing teeth, orthodontic treatment in progress, and dentition with dental implants.

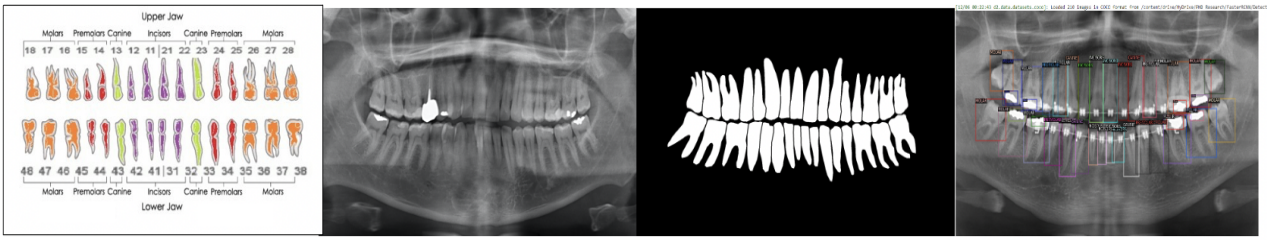
## 1. Introduction

Recent years have seen significant advancements in deep learning, which has heightened its profile in healthcare, notably dentistry. Deep learning-based image processing algorithms have made substantial progress in healthcare imaging applications such as radiographs, cone-beam computed tomography (CBCT), and magnetic resonance imaging (MRI). Deep learning-based image processing techniques have the potential to aid in accurate diagnosis, allowing dentists to identify appropriate dental treatments. For instance, orthodontists could use deep learning-based processing techniques to investigate root absorption from panoramic radiographs to inform a patient's treatment plan [1]. Furthermore, precise tooth segmentation techniques would be advantageous for determining dental age, forensic identification, and the location of impacted teeth [2].

Deep learning enables the identification and classification of features from complex and diverse medical images, resulting in a quantifiable forecasting model that aids clinicians in developing the most effective treatment plans [3]. Panoramic radiographs are used to visualize the patient's mineralized oral health in two dimensions [4]. Thus, a comprehensive dental radiograph examination is a critical component of the diagnostic technique in daily clinical practice. Tooth segmentation is a technique that allows for the separation and isolation of teeth from specific areas of the mouth based on their morphologies, numbers, and positions [5, 6]. One example of a difficulty encountered when successfully reading a panoramic radiograph is determining the precise location of teeth while monitoring these images. As a result, a comprehensive, accurate dental radiograph examination is a critical component of the diagnostic technique used in daily clinical practice. Deep learning techniques can assist with this by enabling fully automated approaches while still allowing for human interpretation. Many dentists work in single-practice settings and regularly evaluate radiographs independently.

One analysis of panoramic radiographs by dentists includes tooth numbering and detection. Occasionally, these diagnoses are inaccurate, impeding the best possible treatment planning approach. Diverse deep learning algorithms may be beneficial for resolving issues encountered during numbering and detection, such as radiographic artifacts, manual labeling, asymmetric development, and anatomical complexity [7]. As the value of dental imaging applications has increased, new paradigms for deep learning have emerged. For instance, deep learning ensembles are a novel collaboration across deep learning models that aims to improve overall accuracy by combining the results of individual models [8]. Collaboration can take place for a single task

⋆This paper is the results of the research which is funded in collaboration with National Science Foundation
*Corresponding author
gc3n3@umsystem.edu (G. Chandrashekar); saacfb@umsystem.edu (S. AlQarni); bumanne@umsystem.edu (E.E. Bumann); leeyu@umsystem.edu (Y. Lee)
https://info.umkc.edu/leeyu/ (Y. Lee)
ORCID(s):
[1]The first and second author has an equal contribution for this paper.




(a) Teeth ISO Numbering (b) Original Teeth Image (c) Teeth Segmentation Mask (d) Teeth Identification Annotation

**Fig. 1:** ISO Numbering and Dental Panoramic Radiographs with Masks and Annotations

or numerous tasks among multiple models [9, 10, 11, 12, 13], for example, via voting or average weight. Moreover, ensemble deep learning approaches have been developed to investigate the relationship between group accuracy and variation measurement [14].

Meta learning is a method that allows for the combination of predictions from multiple independent models. Deep ensembles have several advantages, including collective intelligence based on various models and inherent scalability [15]. However, determining the optimal methods for combining the predictions of multiple models is not straightforward. Developing deep ensemble models with distinct machine learning tasks or learning performances, on the other hand, presents significant challenges. It raises the question of whether multi-tasking should be incorporated into the learning process or should be learned separately [16]. By comparing the performance of various models, the Akaike information criterion for differential weights was used to identify the collaborative model [17]. Attempts to aggregate multiple-objective losses by performing a weighted linear sum of the losses for each task typically result in [18].

In this paper, we propose an ensemble model capable of performing various tasks using a variety of models and then enhancing overall performance through model collaboration. The rationale for and innovations in the proposed method are as follows: We hypothesized that collaborative inference using multiple models would be more effective than inference using a single model. This deep ensemble model was created for multiple machine learning tasks, including segmentation and recognition of teeth in panoramic radiographs. We believe that two stages can be beneficial when learning deep ensemble models for feasibility and robustness: The first stage of local learning aims to construct a model independent of other models by utilizing its data. The second stage involves obtaining inference results from the models and then collaboratively tuning them, a process known as ensemble inferencing.

The proposed collaborative learning model is a novel approach for summarizing and refining the inferencing results of the two models. The collaborative model has the following advantages: (i) It reduces reliance on individual model building because models are built independently using their own data. (ii) Collective inferencing results are summarized as outcomes. (iii) Refinement can improve both individual model and overall inferencing results.

This article's major contribution may be summarized as follows:

- Two high-performance classifiers (tooth segmentation and tooth identification) have been created and optimized using publicly available dental panoramic radiograph image datasets (see Figure 1).
- We created a collaborative learning model for tooth segmentation and identification using these two high-performance classifiers.
- We have developed a unique strategy for collaborative learning that enhances performance by utilizing the learning refinement process.
- Our newly developed two classifiers and collaborative methods were superior to previously reported techniques.

## 2. Related Work

This study focuses on dental anatomy recognition by the use of two-dimensional (2D) radiographs with distinct color codes and identifying each tooth, which aided in tooth detection. We intend to someday assist dentists in validating and explaining such interpretations to their patients. We follow the International Standards Organization's (ISO) norm for tooth numbering, which divides the teeth into four quadrants: upper right, upper left, bottom right, and lower left. Each quadrant has eight teeth, for a total of 32 in the permanent dentition, which can be classified as molar, premolar, canine, or incisor (see Fig. 1(a)).

The following research study identifies and describes the previously most effective deep learning methods for tooth detection and segmentation. Recently, clinical image segmentation tasks, such as segmentation of teeth on radiographs, have focused on tackling various perplexing issues, such as automated diagnosis and overlapping teeth. These strategies can be categorized into two categories: classical techniques that rely on prior knowledge and image highlights, and deep learning-based solutions that are powered by data.

**Table 1**
Cutting Edge Research: Deep Learning based Tooth Segmentation and Identification

| Work | Approach | Objective | Limitation |
|---|---|---|---|
| **Deep Learning Modeling for Dental Panoramic Radiographs** | | | |
| Our Work | Collaborative Learning (Mask R-CNN and Faster R-CNN) | Autonomous tooth segmentation and tooth identification | Incapable of correctly segmenting overlapping teeth and dental implants. Applied only to panoramic dental radiographs; not fully applicable to CBCT images. |
| Zhao et al. [2] | Two-staged attention segmentation network (TSAS-Net) | Autonomous tooth segmentation and identification using attention network | Incapable of precisely segmenting the foreground pixels into tooth regions. |
| Koch et al. [19] | FCN based on U-Net with Network ensembling, data augmentation and symmetry exploitation | Bootstrapping of low quality annotations for teeth segmentation | Deficiency in the design and evaluation of multi-task predictive models. |
| Jader et al. [4] | Mask R-CNN with ResNet-101 | Detection of teeth or missing teeth, constituent parts, and prosthesis | Lack of capability for segmenting mouth and teeth components, detecting missing teeth, or generating medical reports |
| Oktay et al.[20] | Mask regions using convolutional neural network (RCNN) | Recognition, segmentation, and numbering of teeth in panoramic X-ray images | Inability to recognize missing and implanted teeth; Lack of flexibility and extensive modeling required to manage multiple tasks. |
| Pinheiro et al. [21] | Mask R-CNN for dental panoramic X-ray | Tooth numbering and segmentation | Incapable of numbering overlapping and deciduous teeth or ones with implants; and absence of experts' evaluation. |
| Lee et al. [22] | Mask R-CNN for dental panoramic images | Automated teeth segmentation and identification | severe shortage of evaluation for abnormal teeth, such as missing or overlapped teeth, and a dearth of multi-task network modeling. |
| Wirtz et al. [23] | Couple Shaped Model | Segmentation and labeling of 28 individual teeth | Insufficient data, particularly for third molars, results in low segmentation accuracy and robustness. |
| Silva et al. [5] | Analyzed four neural networks (Mask R-CNN, HTC, PANet, and ResNeSt) | Segmentation and numbering of the tooth on complex dental radiographs | Insufficient data sets and model support for numbering and classifying teeth in noisy environments. |
| **Deep Learning Modeling for 3D Dental Images** | | | |
| Cui et al. [3] | Two-staged network architecture (ToothNet) | Autonomous tooth segmentation and identification | Less generalizable due to the scarcity of 3D datasets. |
| Cui et al. [24] | Learning-based segmentation approach (TSegNet) | Tooth segregation and address uncertainties generated by missing, crowding and misaligned teeth | Limited tooth segmentation results in inaccurate performance, particularly for third molars and primary (baby) teeth. |
| Lee et al. [25] | Point-based tooth localization network for CBCT | Effective heatmap regression by separating Gaussian distributions from the network. | Absence of systematic and realistic multi-tasking modeling. |
| Kakehbaraei al. [26] | Marker-controlled watershed (MCW) algorithm and local threshold approaches | Segmentation of tooth, pulp tissue and tooth enamel | Limited training data and models for enamel and pulp segmentation, as well as the performance delay associated with segmentation. |
| Zhang et al. [27] | Deep learning-based tooth segmentation model in harmonic parameter space | Autonomous and accurate segmentation. | The model's overall segmentation accuracy was compromised due to a lack of training data and lengthy image processing. |

## 2.1. Ensemble and Collaborative Model

It is critical to consider how to design ensemble models in order to maximize performance by combining multiple models [14]. Sagi et al. [13] demonstrated how ensemble models could be used to enhance the predictive performance of a single model by training and integrating multiple models. Kendall et al. demonstrated that collaborative networks outperform task-trained networks [28].

Suhail et al. [29], using a collaborative model that included the R technique for feature analysis, an n-net-based neural network, and random forests to classify teeth using

decision trees. Clinicians may benefit from this system that helps confirm expert findings because it will assist the dentist in selecting the best treatment plans, reduce human error, and improve uniformity. Furthermore, experts verified the decision tree's practicality by cross-validation of the data.

Hasan et al. [30] proposed using a multi-feature fusion model in conjunction with an ensemble classifier to determine the optimal dental impression tray from maxillary arch images. In the face of a restricted dataset, a unique multi-feature fusion model combined with an ensemble classifier would improve image labeling. Finally, the goal was to automate the dental process to assist the dentist's clinical judgment and provide a second level of analysis confirmation.

Lee et al. [31] used a variety of transfer learning approaches in conjunction with deep convolutional neural networks (CNNs) to monitor osteoporosis in dental panoramic radiographs (DPRs). Several transfer learning strategies affect deep CNN models, including the basic CNN3 and the Visual Geometry Group 16 (VGG-16). VGG-16 was more optimal since transfer learning and fine-tuning improved the overall effectiveness of the deep CNN in screening osteoporosis in DPRs.

Yaduvanshi et al. [32] explored the use of automated segmentation techniques, more specifically ensemble-based segmentation methodologies, to diagnose oral cancer in its early stages and thereby boost the survival rate via computer-aided diagnosis (CAD). While numerous ensemble models for segmentation problems exist, no combination is sufficiently dynamic to handle every dataset. For example, the model in [29] is valid only for non-surgical procedures and does not support the extraction of unusual features.

Similarly, the requirement for a larger dataset and the usage of EfficientNet-based designs is designed to facilitate future work to address the limitations of PaXNet [33]. Additionally, incorrect categorization occurs due to the absence of original data, which requires additional training images for deep learning algorithms to function correctly. Therefore, the importance of having more qualified, labeled, and validated datasets, as well as an adequate amount of datasets, to achieve outcomes by combining deep learning methodologies has been emphasized [31].

## 2.2. Tooth Segmentation and Identification for Panoramic Radiographs

Due to deep learning's efficacy, numerous tooth segmentation techniques have demonstrated promising results. Some of works on tooth segmentation or tooth identification tasks are based on U-Net [34]. Krois et al. [35] investigated the generalizability of expert systems for segmenting and identifying apical lesions on panoramic radiographs. The training and testing of U-Net-based CNNs with a root-canal fillings dataset reveals that dental practice experience in the training dataset is more essential than image features for improved results. Additionally, when segmenting panoramic radiographs, the unclear behavior of deep learning architectures in terms of generalizability is observed. It is critical to evaluate models using neutral datasets to avoid unduly optimistic outcomes due to data memory. According to [23, 26], small training datasets worsen the model's impracticality due to lower data variances. Their work is based on U-Net, whereas our collaborative model is robust enough to use a variety of individual models for tooth segmentation and tooth identification. The collaboration may be broadened to encompass a range of distinct individual models and datasets.

Mask R-CNN based works [2, 19, 4, 21, 20] are the most pertinent for segmenting teeth and accurately identifying and numbering teeth. Zhao et al. [2] created a two-staged attention segmentation network (TSASNet) to localize and classify teeth in radiographs. The first stage uses the attention model to determine the tooth's approximate location. Following that, the precise tooth borders are recognized using a fully convolutional network with an accuracy of 96.94%. It demonstrates the superiority of the fully convolutional network over previous models. Koch et al. [19] proposed an accurate tooth segmentation model for panoramic radiographs that combines fully convolutional networks (FCNs) [36]. Several strategies, such as network grouping, symmetrical data management, test-time extension, and bootstrapping of low-quality annotations, were used to improve segmentation performance. Jader et al. [4] investigated the use of Mask R-CNN (regional convolutional neural network) for segmenting individual teeth in the most difficult panoramic radiographs. Oktay et al. [20] proposed concurrently detecting, segmenting, and counting teeth in panoramic X-ray images employing Mask regions with convolutional neural network features (RCNN) and multi-class labeling each tooth type with a unique class name. Pinheiro et al. [21] built an end-to-end deep learning architecture for deciduous teeth segmentation and numbering using Mask R-CNN and PointRend.

Lee et al. [22] proposed developing a deep learning solution for automated teeth segmentation on dental panoramic images using a mask R-CNN algorithm with a custom annotated datasets. This approach applies to both interpretable diagnostic systems and forensic classification, which need comparable segmentation tasks. Wirtz et al. [23] provided a coupled shaped model for robust and accurate tooth segmentation in low-quality panoramic radiographs, so assisting dentists in their diagnostic work. The model employs a deep neural network to obtain the binary mask of the teeth to statistically identify form and space changes and therefore improve segmentation quality. Silva et al. [5] tested four neural networks on testing datasets, including Mask R-CNN, HTC (hybrid task cascade), PANet (path aggregation network), and ResNet (residual neural network), to perform tooth numbering and segmentation on difficult dental radiographs. The results indicate that while all frameworks are possible in certain situations for estimating the size, number, and placement of teeth, the accuracy of the PANet in the particular case was superior to that of the other frameworks in the competition. Furthermore, the models above performed

well when teeth were in good condition but failed when teeth were damaged or incorrectly labeled.

However, these studies have not explored collaborative models or learning via multi-task refinement. Our approach to these problems, on the other hand, is fundamentally different, as our work is focused on collaborative modeling aided by autonomous models for multi-task scenarios. For instance, rather than combining both tasks into a single model, we demonstrate the collaboration of two distinct models for tooth segmentation and tooth identification. Due to the adaptability of the proposed work, other tasks such as recognizing dental restorations or clinical situations can easily be added and their outcomes can be summarized. We discuss the benefits and shortcomings of these studies in Table 1 and explain in Section 5.3 how we surpassed them on both tasks using individual and collaborative models trained on the UFBA's panoramic radiography dataset [37]. However, we omitted some of recent tooth segmentation works, such as Wu et al. [38], Lian et al. [39], Wu et al. [40], Tian et al. [41], Tian et al. [42], from our comparative evaluation. This is because their proposed work did not report their performance measures such as accuracy, F-1, and mAP. Lei et al. [43] was not included since it was proposed for retinal fundus images.

### 2.3. Tooth Identification

Lai et al. [44] proposed a Learnable Connected Attention Network to accurately match panoramic radiographs, recognizing the practical value of human recognition based on tooth identification in forensic odontology. The design collects the interdependent information from the coordinating features retrieved from the channel attention module and the learnable connected module to forecast the precise results. Sathya and Neelaveni [45] recommended a novel three-step transfer learning method for automatically detecting human-based features in panoramic radiographs. The first stage involves determining the tooth's position and classifying it into one of the four previously outlined categories. Finally, the teeth are compared to the source images to positively identify persons in forensics once they have been allocated numbers. The suggested framework beats CNN, D-CNN, Seven-layer CNN, and ResNet-50 under the specified circumstances, with a 95% accuracy rate. Thanathornwong and Suebnukarn [46] introduced the Faster R-CNN model [47] to detect compromised teeth on the panoramic radiograph to reduce dentists' diagnostic work. The Faster R-CNN was successfully trained on a minimal amount of labeled imaging data to detect unhealthy teeth.

Chung et al. [48] developed a spatial distance regularization loss-based method for teeth localization based on point regression. The proposed network recognized each tooth autonomously using center point regression for all anatomical teeth (i.e., 32 points in the permanent dentition). The L2 regularization loss for Laplacian spatial distances improved center point detection accuracy. The final detection was accomplished using a multitasking, class-agnostic identification neural network with parallel training of center offsets. The proposed approach accurately identifies both missing and existing teeth. In terms of restrictions, the scarcity of training data is the primary problem [49]. Moriyama et al. [50] increased accuracy by examining pocket locations using radiographs and blood test data in conjunction with concurrent training. Sathya et al. [45] examined dataset expansion, the use of contemporary CNNs, and advanced augmentation approaches.

Krois et al. [35] proposed using a seven-layer deep CNN with global average and max pooling to categorize teeth into four categories: molar, premolar, canine, and incisor. Experiments demonstrated that this method outperforms three contemporary teeth classification methods, with an average accuracy of 87%. Moriyama et al. [50] developed a MapReduce-like approach for estimating the depth of periodontal pockets that includes mapping, CNN, and reduced phases. The mapping process identifies tooth numbers and photographs of pocket regions. CNN estimates pocket depth based on pocket premises, and the lowering component totals the projected depths for all identical pockets. Experimental results indicate that the suggested approach can detect acute periodontal disease autonomously.

In comparison to some previous research, which identified just four unique tooth types (molars, canines, premolars, and incisors), our collaborative model enables the accurate recognition and identification of 32 individual teeth through model improvement. The tooth identification models have been improved by the aggregation of tooth identification and tooth segmentation, followed by the improvement of collaborative learning with better accuracy. The qualitative and quantitative comparative evaluations for tooth identification have been presented in Table 2 and Section 5.3.

### 2.4. 3D Tooth Image Segmentation and Identification

Cui et al. [3] introduced a two-staged network architecture, ToothNet, in which a supervised deep learning approach is used to capture the edge map from CBCT images. After concatenating the learned map features, these learned map features are directed to the Region Proposal Network (RPN) to achieve autonomous tooth segmentation and identification. Cui et al. [24] proposed the TSegNet approach for proficient tooth segregation. They demonstrated incomplete segmentation while classifying wisdom teeth. Restricted datasets, manual labeling, insufficient masking, and segmentation dependence on the algorithm utilized are only some of the significant shortcomings of 3D segmentation models [27].

Lee et al. [25] established a framework for individual tooth segmentation based on points that do not require additional classification. However, despite the enhanced performance of point-based recognition networks on dental images, it is challenging to discriminate adjacent teeth due to their similar topologies and proximity. Kakehbaraei et al. [26] combined a marker-controlled watershed (MCW) algorithm with local threshold techniques for segmenting teeth. Primarily, the noise-free image is preprocessed by filling

**Table 2**
Deep Learning-based Tooth Identification Cutting Edge Research

| Work | Approach | Objective | Limitation |
|---|---|---|---|
| Our Work | Collaborative Learning using Faster R-CNN model [47] and Mask R-CNN model [51] | To obtain identification of 32 distinct teeth on 2D panoramic radiographs | Required to evaluate the effectiveness of 32 individual tooth detection model, particularly for aberrant detection, and increased performance through collaborative learning. |
| Lai et al. [44] | Channel attention mechanism and cosine loss | Classification with 2D panoramic radiographs | Restricted owing to the extra data from varied angles and the comprehensive model development for accurate detection. |
| Sathya et al. [45] | Three-step transfer learning method using AlexNet | Automatic feature extraction, tooth classification and numbering in panoramic radiographs | Due to the domain-specific approach and metrics based on four classes (Molar, Premolar, Canine, or Incisor), it is limited in its ability to number teeth in cases of overlapped or missing teeth. |
| Thanathornwong et al. [46] | Faster R-CNN model [47] | Panoramic radiography for the detection of compromised teeth | Lack of experts' evaluation in diagnosing teeth; Unable to utilize new augmentation approaches for improving dataset and more recent CNN architectures. |
| Chung et al. [48] | Point regression for spatial distance regularization loss | Promising high performance in identification | Advanced data augmentation techniques are required to increase the variety of tooth forms used to validate the model. |
| Li et al. [49] | Seven-layer deep CNN with global average and maximum pooling | CBCT Image Teeth categorization for molar, canine, premolar, and incisor | Limited classification for four teeth types and low performance owing to the use of only a few images from the original source. |
| Moriyama et al. [50] | MapReduce-inspired model for estimating the periodontal pocket depth | Estimation of pocket depth and aggregation of the pocket depth | Low performance due to two unrelated models for tooth recognition and depth estimation; Inability to compose the models end-to-end |

holes, maintaining intensity. Then the MCW algorithm is performed on the gradient image updated with markers to complete the segmentation process.

Zhang et al. [27] proposed a model that isomorphically transfers the 3D tooth prototype into a 2D harmonic parameter space to produce the image. The image is passed to a deep (CNN) for accurate and autonomous segmentation, and the resulting boundary mask is projected back to 3D models. The fuzzy clustering and cuts algorithm is then used to refine the results further. Wu et al. [38] created an innovative method for automatically dividing tooth forms to preserve energy and analyze orthodontic qualities. Similarly, MeshSegNet [39] was proposed for an end-to-end deep learning approach for automatic tooth identification that takes a range of raw surface features as inputs and extracts multi-scale local contextual data. Wu et al. [40] extended MeshSegNet techniques for classifying teeth and identifying landmarks in raw intraoral images and selecting a ROI on the original mesh to build a lightweight PointNet variant for regressing the corresponding landmark heatmaps.

For large numbers of missing teeth in a random arrangement, deep adversarial-driven dental inlay restoration (DAIS) may provide efficient occlusal surface ends [41]. The technique of generative adversarial network (GAN) centered image synthesis (IS) was proposed for the objective of creating images of the latent transitional space between the source and target domains [42, 43]. Tian et al. [42] showed a novel two-stage conditional GAN for replicating the surface of a dental crown.

Deep learning approaches for 3D CBCT data are attracting increasing interest. Some are based on deep learning modeling in two dimensions, while others are based on three-dimensional modeling. Due to the computational requirements of 3D modeling, it would be more practical and efficient to use 2D modeling for CBCT. We did not discuss our collaborative models for CBCT in this paper. However, because of the flexibility of adding additional models, distinct views or structures can be developed independently and dynamically integrated into ensemble inferencing and summarization via collaborative learning. This is the direction in which our future work will take.

## 3. Background and Motivation

We now discuss the fundamental methods that support the collaborative paradigm in consideration. We considered the underlying established framework ($\mathcal{M}_c$) when developing a collaborative model. First, we used the Non-Max Suppression (NMS) technique in the refinement process of the collaborative model. Second, for object detection, our teeth identification model ($\mathcal{M}_i$) was created using the Region Proposal Network from Faster R-CNN [47], and YOLO-v5

(You Only Look Once version 5) [52]. Third, we constructed a model for teeth segmentation ($\mathcal{M}_s$) by combining the instance segmentation approach Mask R-CNN [51] with the semantic segmentation technique U-Net [34].

### 3.1. Bounding Box Regression

The bounding box regression [53] was performed using four coordinates: $x$, $y$, $w$, and $h$, which identify the box's center coordinates, as well as its width and height. Scale-invariant transformations between two centers and log-scale transformations between widths and heights were computed using Eq. 1 given a predicted bounding box coordinate $p = (t_x, t_y, t_w, t_h)$ (center coordinate, width, height) and its corresponding ground truth box coordinates $g = (t_x^*, t_y^*, t_w^*, t_h^*)$.

$$\begin{aligned} t_x = \frac{x - x_a}{w_a}, t_y = \frac{y - y_a}{h_a} \\ t_w = \log(\frac{w}{w_a}), t_h = \log(\frac{h}{h_a}) \\ t_x^* = \frac{x^* - x_a}{w_a}, t_y^* = \frac{y^* - y_a}{h_a} \\ t_w^* = \log(\frac{w^*}{w_a}), t_h^* = \log(\frac{h^*}{h_a}) \end{aligned} \tag{1}$$

All the bounding box correction functions are $d_i(p)$ where $i \in \{x, y, w, h\}$. The bounding box regression can be expressed as a function of an anchor box and a nearby ground-truth box by minimizing the SSE loss using Eq. 2.

$$\mathcal{L}_r = \sum_{i \in \{x,y,w,h\}} (t_i \neg d_i(p))^2 + \lambda \parallel w \parallel^2 \tag{2}$$

The regularization term is essential in this case, and the cross-validation is to choose the optimal $\lambda$. Additionally, not all anticipated bounding boxes match to corresponding ground truth boxes. Box regression is irrelevant when there is no overlap. Thus, while training the box regression model, only predicted boxes that are within the IoU threshold ($\delta$) of a neighboring ground truth box are preserved using Eq. 3.

$$IoU = \frac{x_a \cap x^*}{x_a \cup x^*} \tag{3}$$

When the precision is $\delta = 0.7$, the overlapping is solved using an optimal non-maximum suppression (Algorithm 1), in which any bounding boxes with a value less than $\delta$ are eliminated. The algorithm illustrates the usage of non-maximum suppression in our suggested model: First, select the box with the highest score as the first step. Then, calculate its overlap with all other boxes and delete any that exceed the IoU threshold ($\delta = 0.7$). Finally, repeat step 1 until no more boxes have a lower score than the currently selected box. The remainder is maintained and utilized to generate the final forecasts.

**Algorithm 1:** Non-Max Suppression (NMS)

**Input:** Model $\mathcal{O}_c$, Threshold $\delta$
**Output:** Filteredd Bounding Boxes $B$

```
function NMS(O_c, δ)
    B ← BoundingBox(O_c) ▷ Bounding box set
    C ← Confidence(O_c)  ▷ Confidence score set
    B_f ← ∅              ▷ Filtered Bounding box set
    ▷ Delete overlapping boxes with lower scores
    for b_i ∈ B and c_i ∈ C do
        discard ← False
        for b_j ∈ B do
            if IoU(b_i, b_j) > δ then
                if score(c_j, b_j) > score(c_i, b_i) then
                    discard ← True
        end
        if ¬discard then
            B_f ← B_f ∪ b_i
    end
    return B_f
end
```

### 3.2. Object Detection Modeling

We carried out object identification modeling with YOLO-v5 [52], and Faster R-CNN [54], which are composed of convolutional layers for training the extract filters and classification layers for predicting classes and bounding boxes. YOLO-v5 predicts the relative offset of the predicted bounding box's center point from the linked cell's top left corner, whereas Faster R-CNN is based on RPN indicating the offset of the prediction box and anchor. The following sections detail the steps involved in detecting objects using YOLO-v5. It will begin by taking an image as input, reshaping it, and then extracting its features via a CNN architecture. The data is then transferred to two entirely connected layers that reshape and transform it into a predetermined grid. Once the entire image has been transformed into a grid, the data for object detection is provided. Next, it will attempt to determine whether an object is present or not in each grid. After establishing these values, the bounding box is determined using NMS (non-maximal separation) (Algorithm 1).

RPN is a fully convolutional network that anticipates object borders using object scores at each detection. Based on the RPN model based on non-maximum suppression (Algorithm 1) was designed. The positive label will be assigned to one of two types of anchors (Eq. 3): (i) the anchor/anchors with the greatest Intersection-over-Union (IoU) overlap with a ground-truth box, or (ii) any anchor with an IoU overlap more significant than $\delta = 0.7$ with any ground-truth box.

The object detection loss is calculated as the product of the log and bounding-box losses. The projected probability $p_i$ corresponds to an anchor (an object's index $i$) in a $nxn$ mini-batch. $p_i^*$ ={1,0} indicates whether the anchor is positive or negative. The vector representing the anticipated bounding box $t_i$ is compared to the related ground-truth box to establish whether the anchor is positive. The word $p_i * \mathcal{L}_b$ indicates that the regression loss is active only when the

anchors are positive ($p_i^* = 1$) and is deactivated otherwise ($p_i^* = 0$).

$$\begin{aligned}\mathcal{L}_o &= \mathcal{L}(\{p_i\}\{t_i\}) \\ &= \frac{1}{N_c}\sum \mathcal{L}_c(p_i, p_i^*) + \lambda \frac{1}{N_b}\sum p_i^* \mathcal{L}_b(t_i, t_i^*)\end{aligned} \tag{4}$$

For the regression loss, the loss function $S_{\mathcal{L}_1}$ defined in [54] is used, i.e., $\mathcal{L}_b(t_i, t_i^*) = S_{\mathcal{L}_1}(t_i - t_i^*)$ as defined in Eq. 5.

$$S_{\mathcal{L}_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{5}$$

Thus, the object detection model $\{p_i\}$ and $\{t_i\}$ make predictions based on the classification $c$ and region detection $b$ layers' compositions, respectively.

### 3.3. Image Segmentation Modeling

We employed image semantic segmentation and instance segmentation approaches to segment teeth. U-Net [34] is a well-known model for segmenting biological images semantically. It has shown exceptional performance on several natural and medical image segmentation tasks. The U-Net generates a lower-dimensional representation of the images using a CNN network, which is then upsampled to create the final segmentation map. The weight map is a segmentation of the ground truth that has been predefined. The ground truth segmentation is subjected to morphological image processing to establish the fine borders that separate cells. A weight map is generated, with a significant weight assigned to these brief cell division borders. By incorporating this weight map into the calculation of cross-entropy loss, the U-Net is severely penalized for failing to establish these specific cell borders or for doing so in an inefficient manner. The loss function of the U-Net is defined to compare the predicted mask to the ground truth mask, hence optimizing the model's parameters for the upcoming training sample.

For the instance segmentation, Mask R-CNN [51] was created to forecast the class label used to pick the output mask. For each RoI, the mask branch produces a $Km^2$-dimensional output encoding $K$ binary masks of resolution $m * m$, one for each of the $K$ classes. Assume that $\mathcal{L}_c$ is the classification loss, $\mathcal{L}_b$ is the bounding-box loss, and $L_m$ is the RoI associated with the kth mask's ground-truth class. On each sampled RoI, the multi-task loss was estimated as defined in Eq. 6.

$$\begin{aligned}\mathcal{L}_i &= \mathcal{L}_c(p, u) + \mathcal{L}_b(t^u, v) + \mathcal{L}_m \\ &= -log p_u + \sum_{i \in (x,y,w,h)} smooth_i(t_i^u - v_i) \\ &- \frac{1}{m^2}\sum_{1 \le i,j \le m} [y_{ij} log \hat{y}_{ij}^k + (1 - y_{ij} log(1 - \hat{y}_{ij}^k)]\end{aligned} \tag{6}$$

$\mathcal{L}_c$ is a multinomial cross-entropy loss that can be calculated using softmax on a per-pixel basis. $\mathcal{L}_b$ is calculated using IoU (Eq. 3), whereas $smooth_{\mathcal{L}_1}$ is the smooth $\mathcal{L}_1$ loss (Eq. 5). $\mathcal{L}_m$ is defined as the average binary cross-entropy loss calculated using a sigmoid on a per-pixel basis. Thus, the network can construct masks for each class without regard for a class competition.

## 4. Design of Collaborative Learning Model

We proposed a collaborative learning approach to supplement existing object recognition and instance segmentation algorithms (see Figure 2). Collaborative learning can be defined as integrating two deep learning models to obtain better results. First, the segmentation and identification models are trained and developed individually. The optimizer is Stochastic Gradient Descent (SGD) with momentum that attempts to accelerate gradient vectors in the right direction, resulting in a more rapid convergence. The configuration of the deep neural network models used for tooth segmentation and identification training is shown in Table 3. Then, the outputs of the two models are combined to create a panoramic radiograph image that incorporates detection from both models. When findings are related, refinement is utilized to further refine them.

### 4.1. Step 1. Modeling and Inferencing

We developed two distinct models for tooth segmentation and tooth identification. Also, we performed inference utilizing these models, which are forwarded to the consequence phase for aggregation of the inferencing results.

**Model 1: Tooth Segmentation Modeling.** To segment teeth, we used two segmentation models, $\mathcal{M}_{s1}$: Mask R-CNN [51] and $\mathcal{M}_{s2}$: U-Net [34] with the panoramic radiograph dataset. While both $\mathcal{M}_{s1}$ and $\mathcal{M}_{s1}$ are CNN-based techniques for recognizing and segmenting teeth, Mask R-CNN ($\mathcal{M}_{s1}$) is instance segmentation that recognizes 32 individual teeth using dental masks, and U-Net ($\mathcal{M}_{s2}$) is semantic segmentation that does not distinguish between distinct individual teeth. As demonstrated in a panoramic radiograph in Figure 1(b), the normal number of teeth is 32 in the permanent dentition, and the teeth mask displayed in Figure 1(c) is annotated for the 32 teeth segmentation. Mask R-CNN uses ResNet-101 as the backbone for feature extraction. After extracting features using ResNet-101, FPN (Feature pyramid Network) with anchors are created with identified ROIs (Region of interests). After the ROIs are aligned, classification and localization are applied by regressing the bounding boxes. Finally, each object is detected and segmented by the complete convolution network indicated using bounding boxes. The U-Net model creates a binary mask comprised of 1s and 0s from an input image, a grayscale radiographic image of teeth (including borders between teeth). Until the instance segmentation (Mask R-CNN), all teeth are labeled as "tooth" without distinction of individual teeth. The tooth segmentation model was created using the panoramic radiograph dataset [49] and Facebook Research's Detectron2 Library [55] for python 3.7.

**Model 2: Tooth Identification Modeling.** Our modeling of tooth identification is based on two models, $\mathcal{M}_{i1}$ and $\mathcal{M}_{i2}$ using the most recent object detection algorithms, the

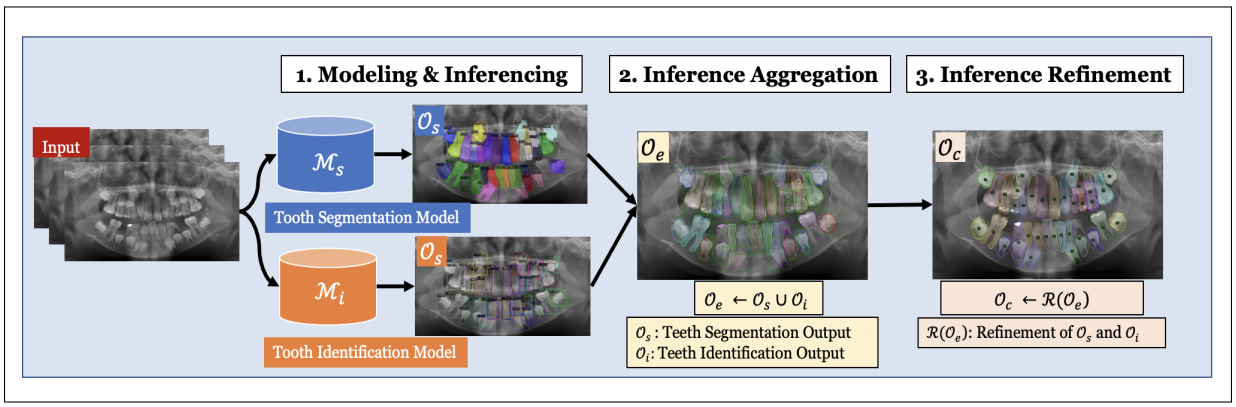



**Fig. 2:** Collaborative Deep Learning Architecture

**Table 3**
Configuration for Individual Models

| Parameter/Task | Segmentation $\mathcal{M}_{s1}$ | Identification $\mathcal{M}_{i1}$ |
|---|---|---|
| Network | Mask R-CNN | Faster R-CNN |
| Pretrained Model | ResNet-101 | |
| Stepsize | 50000 | |
| Learning Rate | 0.001 | |
| Batch Size | 2 | |
| Max Iterations | 300 | |
| Threshold IoU | 0.7 | |
| ROI Heads/image | 64 | |
| Optimizer | SGD with Momentum | |

Faster R-CNN framework [47] and YOLO-v5 [52]. These models aim to construct a tooth identification model capable of sorting teeth into four categories (molar, premolar, canine, incisor). The $\mathcal{M}_{i1}$ model based on the Faster R-CNN framework [47] is also one of the most acceptable frameworks for object detection due to its region classification and RPN architecture based on anchor boxes and non-maximum suppression (Algorithm 1). The $\mathcal{M}_{i2}$ model based on YOLO-v5 [52] recognizes objects in real-time with high precision by incorporating anchors into the detection process and employing a pre-trained version generated from the COCO dataset. First, the image is divided into cell boxes since Anchor boxes are essential for a high detection rate. Second, if the bounding box's center is in a cell, the bounding box is predicted by each cell box.

For the $\mathcal{M}_{i1}$ modeling, the bulk of human tooth structures resemble one another; the model detected multiple overlapping bounding boxes for each object during initialization. When the precision is 0.7, an optimal non-maximum suppression algorithm is used to resolve the overlapping, eliminating any bounding boxes with a value less than 0.7. The balance of the data is retained and used to make final forecasts. The technique for achieving optimal non-maximum suppression in the $\mathcal{M}_{i1}$ model is illustrated in Algorithm 1. The $\mathcal{M}_{i2}$ model (YOLO-v5 network [52]) is composed of a single-stage, whereas the $\mathcal{M}_{i1}$ model (Faster R-CNN network) is composed of two stages. $\mathcal{M}_{i1}$ is an FPN-based network, whereas $\mathcal{M}_{i2}$ is based on SPP (Spatial pyramid pooling) and PANet for multi-channel feature fusion with mosaic training and self adversary training.

Due to a scarcity of available radiographs for training, the architecture is utilized to practice multi-class detection (molar, premolar, canine, incisor). First, 100 radiographs were annotated, and the augmentation technique was used to produce additional images during the pre-processing portion of the identification model. Next, the tooth identification model was used in conjunction with the collaborative approach is depicted in Figure 2. When compared to the $\mathcal{M}_{i1}$ (Faster R-CNN) model, the $\mathcal{M}_{i2}$ (YOLO-v5) model performed well for the four different tooth kinds (molar, premolar, canine, incisor). The detailed results are reported in Section 5.3.

## 4.2. Step 2: Inferencing Aggregation

The collaborative model's second phase involves aggregating the predictions of various models to summarize the detection results (see Figure 2). This stage of the collaborative model is referred to as the *ensemble model* since it does not require any refinement and is solely focused on synthesizing observations from two independent models: tooth segmentation and tooth identification. At this stage, the accuracy is calculated as the weighted average of the accuracy of two distinct models. The initial weights are equal, and the final weights can be calculated by evaluating the contributions of each model after the refinement. The summary of integrated inferences will be saved in a standard format (MS coco) for usage in the collaborative inference process.

## 4.3. Step 3: Inferencing Refinement

Collaborative Inference (CI) is the final stage of the collaborative learning model; it utilizes an ensemble of teeth segmentation and identification algorithms to refine the integrated inference from the previous phase. CI is a highly effective method for creating collaboration between two classifiers since it enables the refinement of independent models produced using a mapping schema.

Two separate classifiers, tooth segmentation and tooth identification are refined in conjunction via mapping. The

mapping strategy for the CI is designed to boost detection during inference by employing the combined inferencing summary created by these two models. For instance, the identification technique is inefficient when teeth are missing or overlapping. The feedback from the tooth segmentation model improves the accuracy of teeth identification due to the model's refinement. Similarly, collaboration with the identification model boosted the segmentation outcomes. For example, because the identification model can determine the center of each tooth, any false positives that extend beyond the confines of the tooth segmentation model are easily repaired. This process may occur continuously or sequentially, depending on the inference platform's design.

### 4.4. Collaborative Learning Algorithm

The collaborative learning model that is necessary for the end-to-end process is described in the following steps by Algorithm 2. First, before developing individual classifiers, we perform the preprocessing. Second, we independently train the tooth segmentation model ($\mathcal{M}_s$) and the tooth identification model ($\mathcal{M}_i$) using Mask R-CNN and Faster R-CNN, respectively. Third, we perform inference to determine the segmentation of 32 classes using the segmentation model ($\mathcal{M}_s$). Fourth, we determine the four tooth types with the testing data ($\mathcal{D}_s$) using the identification model ($\mathcal{M}_i$). Fifth, we obtain the ensemble model output ($\mathcal{O}_e$) by combining these two outputs. Sixth, we fine-tune the collaborative model's composite output ($\mathcal{O}_c$) through the refinement operation $\mathcal{R}(\mathcal{O}_e)$ (Algorithm 3): (i) Filter overlapping boxes by applying non-maximum suppression to the identifying bounding boxes. (ii) Converge and map the segmentation and identification outputs to detect bounding boxes that do not contain a target object. (iii) Identifying 32 distinct individual teeth by locating the midpoint of each enclosing box and applying the ISO standard for tooth numbering. (iv) Eliminate segmentation findings that are beyond boundaries. Finally, produce a summary of the refining results, including the collaborative model output and an accuracy report.

The loss function for the collaborative learning model $\mathcal{L}_c$ was computed using Eq. 7 to reduce model selection bias and uncertainty by refining integrated predictions. First, $\mathcal{L}_e$ was calculated using Eq. 7 by combining the loss functions of the two models, such as segmentation loss $\mathcal{L}_s$ for tooth segmentation model $\mathcal{M}_s$ and identification loss $\mathcal{L}_i$ for tooth identification model $\mathcal{M}_i$.

Second, following the computation of $\mathcal{L}_e$, the collaborative learning model's loss function $\mathcal{L}_c$ is computed to obtain the average prediction by combining $\mathcal{L}_e$ and the refinement function $\mathcal{R}$ across the two models, resulting in an increase in predicted accuracy.

$$\begin{aligned} \mathcal{L}_e &= \mathcal{L}_s(\mathcal{M}_s) + \mathcal{L}_i(\mathcal{M}_i) \\ \mathcal{L}_r &= \mathcal{R}(\mathcal{M}_s, \mathcal{M}_i) \\ \mathcal{L}_c &= \mathcal{L}_e + \mathcal{L}_r \end{aligned} \tag{7}$$

**Algorithm 2:** Collaborative Learning Model

**Data:** Training $\mathcal{D}_r$, Validation $\mathcal{D}_v$, Testing $\mathcal{D}_s$
**Result:** Summary $\mathcal{O}_c$
preprocessing($\mathcal{S}$, $\mathcal{D}_r$, $\mathcal{D}_v$)
preprocessing($\mathcal{I}$, $\mathcal{D}_r$, $\mathcal{D}_v$)
$\mathcal{M}_s \leftarrow$ training($\mathcal{S}$, $\mathcal{D}_r$, $\mathcal{D}_v$) ▷ *Training segmentation model.*
$\mathcal{M}_i \leftarrow$ training($\mathcal{I}$, $\mathcal{D}_r$, $\mathcal{D}_v$) ▷ *Training identification model.*
**repeat**
  $\mathcal{O}_s \leftarrow$ Prediction($\mathcal{S}$,$\mathcal{M}_s$, $\mathcal{D}_s$) ▷ *Inference segmentation model.*
  $\mathcal{O}_i \leftarrow$ Prediction($\mathcal{I}$, $\mathcal{M}_i$, $\mathcal{D}_s$) ▷ *Inference identification model.*
  $\mathcal{O}_e \leftarrow \mathcal{O}_s \cup \mathcal{O}_i$ ▷ *Ensemble of segmentation and identification models.*
  $\mathcal{L}_e = \mathcal{L}_s(\mathcal{M}_s) + \mathcal{L}_i(\mathcal{M}_i)$
  $\mathcal{O}_c \leftarrow \mathcal{R}(\mathcal{O}_e)$ ▷ *Refinement of $\mathcal{O}_e$ through the collaboration of $\mathcal{O}_s$ and $\mathcal{O}_i$.*
  $\mathcal{L}_c = \mathcal{L}_e + \mathcal{L}_r(\mathcal{M}_s, \mathcal{M}_i)$
**until** no more refinement is possible
$\mathcal{O}_c \leftarrow \mathcal{O}_c \cup \mathcal{L}_c$ ▷ *Summary of collaboration*
**return** $\mathcal{O}_c$

**Algorithm 3:** Refinement

**function** Refinement($O_s$, $\mathcal{O}_i$)
  ▷ *Perform non maximum suppression in $\mathcal{O}_i$.*
  $\mathcal{O}_c \leftarrow$ NMS($\mathcal{O}_i$)
  ▷ *Filter bounding boxes in $\mathcal{O}_c$ with no targeted object.*
  $\mathcal{O}_c \leftarrow$ Filter($\mathcal{O}_s \cup \mathcal{O}_i$)
  ▷ *Locate the midpoint of each enclosing box B for ISO standard numbering.*
  **for** ($\forall$ $B \in \mathcal{O}_c$) **do**
    $\mathcal{M}_b \leftarrow (\frac{max(x)-min(x)}{2}, \frac{max(y)-min(y)}{2})$
    $\mathcal{O}_c \leftarrow$ ISO($\mathcal{M}_b^2$)
  ▷ *Remove out-of-bounds segmentation.*
  **for** ($\forall$ $b \in \mathcal{O}_c$) **do**
    $\mathcal{B}_b \leftarrow (\max(b_x), min(b_x), max(b_y).min(b_y))$
    $\mathcal{O}_c \leftarrow$ BoundFilter($\mathcal{B}_b$)
  **return** $\mathcal{O}_c$

## 5. Experimental Results

### 5.1. Dataset and Evaluation Measures

Our key dataset for tooth segmentation and tooth identification is the UFBA-UESC dental dataset [37], 1500 panoramic radiographs were used to train the tooth segmentation model, which was then extended for identification and labeling. The dataset is described in detail in Table 4. The technique of deep neural network training is characterized as achieving the convergence criterion and continuing until the optimal learning results are obtained. The criterion for model selection is based on a loss function

that minimizes the error associated with preset labels as determined by empirical and cross-validated residual sums of squares.

Individual models (tooth segmentation and tooth identification) as well as collaborative models are evaluated. We employed a variety of metrics in this review, including accuracy, precision, recall, F1 score, and mAP (mean average precision). TP denotes the true position, TN denotes the true negative, FP denotes the false positive, and FN denotes the false negative. Precision assesses the proportion of positive class predictions (TP+FP) that are genuinely positive class predictions (TP). Recall quantifies the number of positive class predictions (TP) made from the dataset's positive examples (TP+FN). F-Measure generates a single score that accounts for both precision and recall concerns in a single number. The mean average precision (mAP) is calculated as the average of the precision score for each query, where $Q$ is the total number of inquiries.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{11}$$

$$\begin{aligned} AveP &= \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{\mathrm{D}} \\ mAP &= \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \end{aligned} \tag{12}$$

where $D$ denotes the total number of relevant dental images and $rel$(k) denotes an indicator function equal to one if the item at rank $k$ is a relevant dental image and zero otherwise.

## 5.2. Results

Individual models for tooth segmentation models ($\mathcal{M}_{s1}$ and $\mathcal{M}_{s2}$) and tooth identification models ($\mathcal{M}_{i1}$ and $\mathcal{M}_{i2}$) were developed separately. The dataset and performance for tooth segmentation, tooth identification, and collaborative model construction using panoramic radiographs are shown in Table 4. First, the two tooth segmentation models ($\mathcal{M}_{s1}$ and $\mathcal{M}_{s2}$) were constructed using about 193 annotated training images, 83 validation images with approximately 300 epochs, and were evaluated using approximately 1224 testing images. Second, the tooth identification models ($\mathcal{M}_{i1}$ and $\mathcal{M}_{i2}$) were developed using 750 training images, 150 validation images, and 100 testing images. The identification models were trained on 100 annotated panoramic radiographs from the UFBA and 650 images created with data augmentation techniques (flip, saturation, and contrast) of Detectron2 [55]. Roboflow [56] was utilized to generate the annotated images. Multiple images were generated for training, validation, and testing utilizing augmentation.

The learning and loss curves for four independently trained models are shown in Figure 3 (Faster R-CNN and YOLO-v5 for tooth identification; Mask R-CNN and U-Net for tooth segmentation). Examining models' learning and loss curves during training enables us to demonstrate our work's convergence criterion and optimization strategy, which are based on a loss function that minimizes the error associated with predefined labels.

Third, the collaborative model was inferred by merging the outputs of two of the best individual models: tooth segmentation $\mathcal{M}_{s1}$ and tooth identification $\mathcal{M}_{i1}$. Through collaborative learning, multiple bounding boxes were recognized around the final detected images as part of the identification model. The bounding boxes with poor accuracy were removed using non-maximum suppression. We create an ensemble model to segment, recognize, and number 32 individual teeth using those two models $\mathcal{M}_{s1}$ and $\mathcal{M}_{i1}$. 150 images were utilized to evaluate the collaborative model from the UFBA [37] dataset by combining the output from two distinct models and performing refining on the outcome.

Table 5 shows the testing datasets that were used for evaluation of segmentation, identification, collaborative models. This table presents the UFBA's ten detailed categories of the dataset based on the number of images used to test teeth segmentation, tooth identification, and collaborative models. It is worth noting that the testing images used to evaluate three distinct models can overlap. Figure 4 and Figure 5 illustrate an example of the testing results for the tooth identification models ($\mathcal{M}_{i1}$ and $\mathcal{M}_{i2}$), the tooth segmentation models ($\mathcal{M}_{s1}$ and $\mathcal{M}_{s1}$), and collaborative model $\mathcal{M}_c$ for each of the ten distinct categories of UFBA panoramic radiographs [37].

We validated our findings using images that were not included in our collection. As illustrated in Figures 6-9, we obtained the findings from the two models. After applying U-Net to the training data, we obtain a 90% accuracy on the testing data. We also tested the model on actual images, and it performed admirably. After detecting and labeling teeth, an accuracy of approximately 98% was obtained. The accuracy of the multi-class label detection using Detectron2 [55] was approximately 85%.

## 5.3. Comparison with State-of-the Art Research

In Tables 6 and 7, we compared the accuracy of teeth segmentation, recognition, and collaborative learning models to the state-of-the-art research in terms of accuracy, F-Score, and mean average precision (mAP). These results demonstrated the enhancement of our proposed works compared to existing works via a comparative evaluation. Because most state-of-the-art studies in tooth segmentation and identification do not provide their model, source code, or

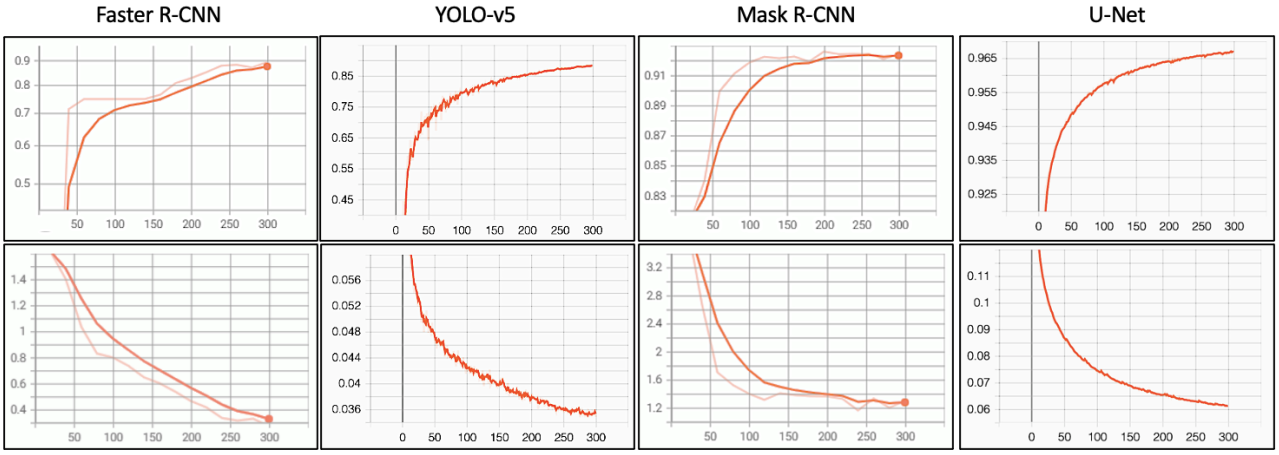



**Fig. 3:** Training Performance for Tooth Identification Models (Faster R-CNN and YOLO-v5) and Tooth Segmentation Models (Mask R-CNN and U-Net) (Top Row: Learning Curve and Bottom Row: Loss Curve)

**Table 4**
Dataset and Performance for Model Building with Panoramic Radiographs
$\mathcal{D}_r$: Training Data, $\mathcal{D}_a$: Augmented Data, $\mathcal{D}_v$: Validation Data, $\mathcal{D}_s$: Testing Data (Collaborative Learning's testing data are randomly selected from the Tooth Segmentation's testing data. For the identification model, augmented data ($\mathcal{D}_a$) generated by Detectron2 [55] were used.)

| Model | Dataset | $\mathcal{D}_r$ ($\mathcal{D}_a$) | $\mathcal{D}_v$ ($\mathcal{D}_a$) | $\mathcal{D}_s$ | Epoch# | Training Time | Testing Time |
|---|---|---|---|---|---|---|---|
| Tooth Segmentation ($\mathcal{M}_s$) | UFBA [37] | 193 | 83 | 1224 | 300 | 30 min | 120 min |
| Tooth Identification ($\mathcal{M}_i$) | | 100 (650) | 45 (105) | 100 | 300 | 30 min | 10 min |
| Collaborative Learning ($\mathcal{M}_c$) | | - | - | 150 | - | - | 30 min |

data, we were unable to replicate their results and compare them to our own in the same context. Additionally, as previously stated, our work is not directly comparable to several prior studies on tooth segmentation and identification [3, 24, 26, 27]. This is because they concentrated their efforts on deep learning on CBCT or 3D dental images.

Specifically, Table 6 compares our approach to the cutting edge research in deep learning-based tooth segmentation. For the tooth segmentation task, we selected $\mathcal{M}_{s1}$: Mask R-CNN over $\mathcal{M}_{s2}$: U-Net because to its superior instance segmentation (32 distinct classes) compared to U-Net's semantic segmentation (a single class). Through collaborative learning with the tooth segmentation and tooth identification models, the performance (accuracy, F-1, and mAP) was improved from 96%, 98%, 95% ($\mathcal{M}_{s1}$) to 98.77%, 98.83%, 97.30% ($\mathcal{M}_c$). As previously stated, we excluded several recent teeth segmentation studies [38, 39, 40, 41, 42] from our comparative evaluation in Table 6, due to their absence of performance metrics such as accuracy, F-1, and

**Table 5**
Dataset for Segmentation, Identification, Collaborative Models
$\mathcal{M}_s$: Tooth Segmentation Model, $\mathcal{M}_i$: Tooth Identification Model, $\mathcal{M}_c$: Collaborative Model;
$\mathcal{D}_r$: Training Data, $\mathcal{D}_a$: Augmented Data, $\mathcal{D}_v$: Validation Data, $\mathcal{D}_s$: Testing Data

| ID | Category | Total | Segmentation ($\mathcal{M}_s$) | | | Identification ($\mathcal{M}_i$) | | | $\mathcal{M}_c$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | $\mathcal{D}_r$ | $\mathcal{D}_v$ | $\mathcal{D}_s$ | $\mathcal{D}_r(\mathcal{D}_a)$ | $\mathcal{D}_v(\mathcal{D}_a)$ | $\mathcal{D}_s$ | $\mathcal{D}_s$ |
| C1 | 32 teeth+Dental Appliance+Restoration | 73 | 0 | 0 | 73 | 6 (54) | 2 (10) | 6 | 25 |
| C2 | 32 teeth-Dental Appliance+Restoration | 220 | 116 | 44 | 60 | 5 (49) | 3 (13) | 10 | 15 |
| C3 | 32 teeth+Dental Appliance | 45 | 31 | 12 | 2 | 22 (163) | 12 (29) | 5 | 10 |
| C4 | 32 teeth-Dental Appliance | 140 | 46 | 27 | 67 | 8 (66) | 5 (11) | 40 | 50 |
| C5 | < 32 teeth +Dental Implant | 120 | 0 | 0 | 120 | 14 (62) | 6 (4) | 4 | 10 |
| C6 | > 32 teeth | 170 | 0 | 0 | 170 | 6 (60) | 4 (10) | 5 | 5 |
| C7 | < 32 teeth+Missing+Dental Appliance | 115 | 0 | 0 | 115 | 11 (72) | 2 (9) | 5 | 10 |
| C8 | < 32 teeth+Missing-Dental Appliance | 457 | 0 | 0 | 457 | 12 (33) | 4 (5) | 10 | 5 |
| C9 | < 32 teeth+Missing+Dental Appliance | 45 | 0 | 0 | 45 | 10 (47) | 3 (8) | 10 | 5 |
| C10 | < 32 teeth+Missing-Dental Appliance | 115 | 0 | 0 | 115 | 6 (44) | 4 (6) | 5 | 15 |
| Total | | 1500 | 193 | 83 | 1224 | 100(650) | 45(105) | 100 | 150 |

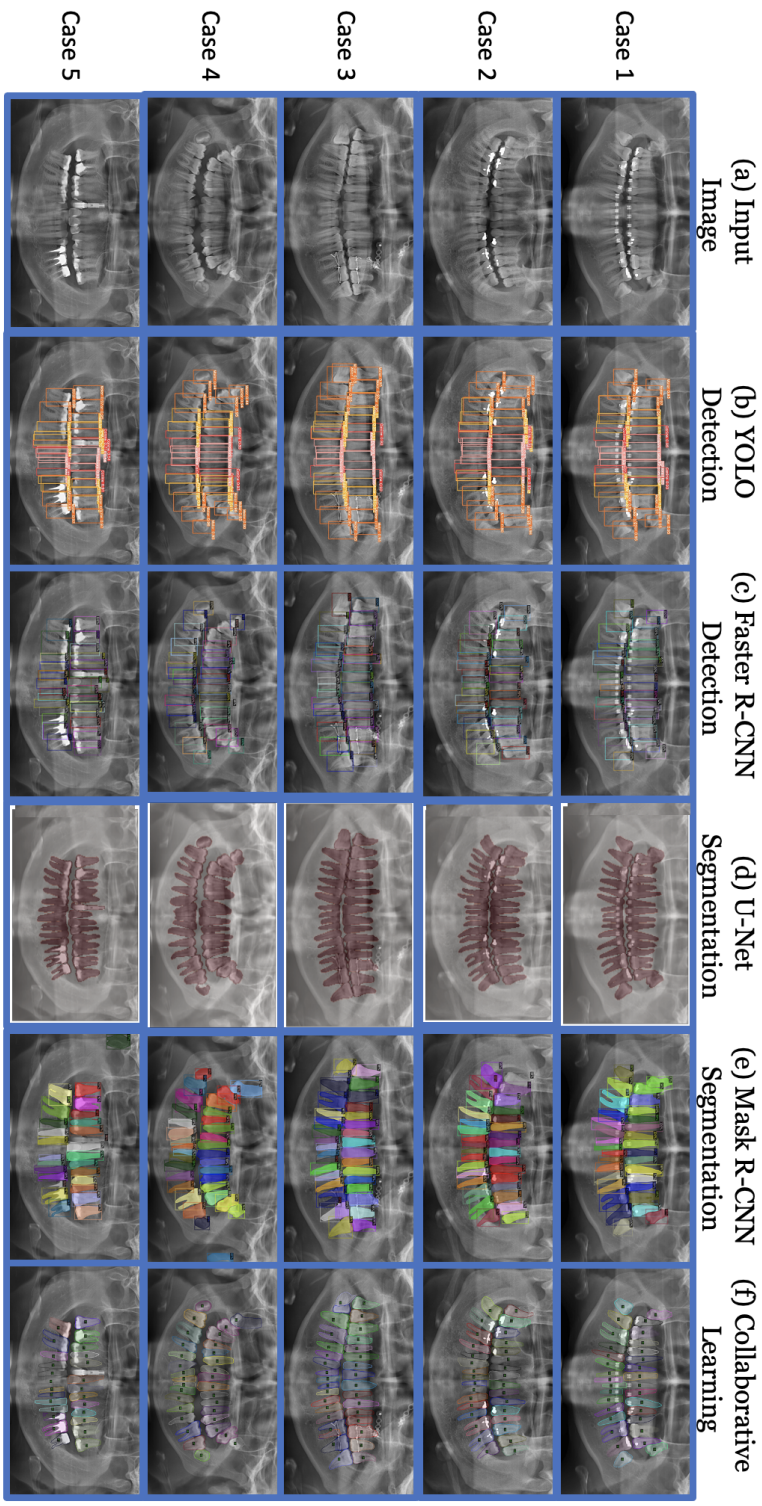



**Fig. 4:** Comparative Evaluation for Categories 1 - 5: (a) Input (b) ($\mathcal{M}_{i2}$): YOLO-v5 Detection (c) ($\mathcal{M}_{i1}$): Faster R-CNN Detection (d) ($\mathcal{M}_{s2}$): U-Net Segmentation (e) ($\mathcal{M}_{s1}$): Mask R-CNN Segmentation (f) ($\mathcal{M}_{c}$): Collaborative Learning

mAP. Additionally, we omitted Lei et al. [43] since their study was presented for retinal fundus images.

The comparative evaluation with the state-of-the-art research in deep learning-based tooth identification is shown in Table 7. Collaborative learning improves teeth identification models by expanding the task: Individual models ($\mathcal{M}_{i1}$ and $\mathcal{M}_{i2}$) classify teeth into four unique categories (molar, canine, premolar, and incisor), whereas the collaborative model classifies teeth into 32 distinct categories. $\mathcal{M}_{i2}$ is a more precise model than $\mathcal{M}_{i1}$ in terms of accuracy, F-1,

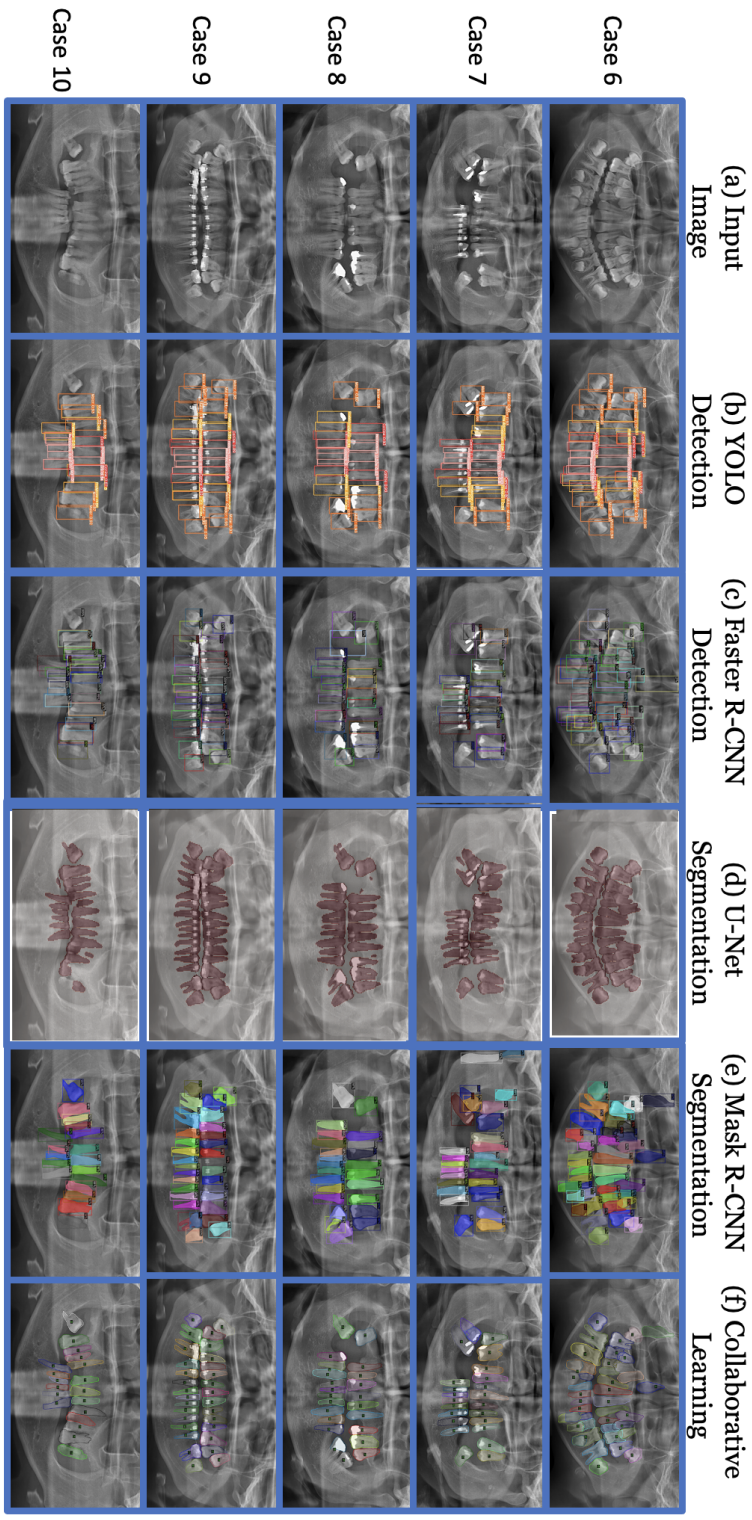



**Fig. 5:** Comparative Evaluation for Categories 6 - 10: (a) Input (b) ($\mathcal{M}_{i2}$): YOLO-v5 Detection (c) ($\mathcal{M}_{i1}$): Faster R-CNN Detection (d) ($\mathcal{M}_{s2}$): U-Net Segmentation (e) ($\mathcal{M}_{s1}$): Mask R-CNN Segmentation (f) ($\mathcal{M}_{c}$): Collaborative Learning

and mAP (99.5%, 99.85%, 99.5% vs. 91%, 90%, 91%). The collaborative model $\mathcal{M}_c$ attained an accuracy of 98.44%, 98.75% for F-1, and 97.78% for mAP, all of which are comparable to state-of-the-art accuracy. Additionally, because to the expanded task (32 classes vs. 4 classes), the collaborative model $\mathcal{M}_c$ outperforms $\mathcal{M}_{s1}$. In general, the proposed model, collaboration model $\mathcal{M}_c$, outperforms existing deep learning models for a variety of dental tasks, including tooth segmentation $\mathcal{M}_{s1}$ and tooth recognition $\mathcal{M}_{i1}$.

**Table 6**
Cutting Edge Research: Testing Accuracy for Tooth Segmentation and Identification Models with Dental Panoramic Radiographs, CBCT Datasets ($\mathcal{D}_r$: Training, $\mathcal{D}_v$: Validation, $\mathcal{D}_s$: Testing)

| **Deep Learning Modeling for Dental Panoramic Radiographs** | | | | | |
|---|---|---|---|---|---|
| Research | Dataset | Data Split (Images) | Accuracy | F1-Score | mAP |
| $\mathcal{M}_c$: Collaborative (Ours) | Panoramic Radiographs: UFBA [37] | $\mathcal{D}_s$: 150 | **98.77%** | **98.83%** | **97.30%** |
| $\mathcal{M}_{s1}$: Mask R-CNN (Ours) | | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 193, 83, 1224 | 96% | 98% | 95% |
| $\mathcal{M}_{s2}$: U-Net (Ours) | | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 193, 83, 1224 | 96.97% | 93.63% | 92.08% |
| Zhao et al [2] | | $\mathcal{D}_r$, $\mathcal{D}_s$: 1200, 150 | 96.94% | NA | NA |
| Koch et al. [19] | | $\mathcal{D}_r$, $\mathcal{D}_s$: 80%, 20% | 94.76% | NA | NA |
| Jader et al. [4] | | $\mathcal{D}_r$, $\mathcal{D}_s$: 193, 1224 | 98% | 88% | NA |
| Oktay et al. [20] | | $\mathcal{D}_r$, $\mathcal{D}_s$: 200, 278 | 98.11% | 93% | 82% |
| Pinheiro et al. [21] | | 6-fold CV: 450 | NA | NA | 77.3% |
| Lee et al. [22] | Panoramic Radiographs | $\mathcal{D}_r$, $\mathcal{D}_s$: 40, 10 | NA | 87.5% | NA |
| Wirtz et al [23] | Panoramic Radiographs | $\mathcal{D}_r$, $\mathcal{D}_s$: 10, 14 | 81.8% | 80.3% | NA |
| Silva et al. [5] | Jader Dataset [4] | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 324, 108, 778 | 96.7% | 91.6% | NA |
| **Deep Learning Modeling for 3D Dental Images** | | | | | |
| Research | Dataset | Data Split (Images) | Accuracy | F1-Score | mAP |
| Cui et al [3] | CBCT Scans | $\mathcal{D}_r$, $\mathcal{D}_s$: 12, 8 Subjects | 99.55% | NA | NA |
| Cui et al [24] | 3D Dental Models | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 1500, 100, 400 | NA | 94.2% | NA |
| Lee et al. [25] | CBCT Scans | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 80, 20, 20 | NA | NA | 90.91% |
| Kakehbaraei et al [26] | CBCT Scans | 30 Subjects | **99.93%** | NA | NA |
| Zhang et al. [27] | 3D Dental Models | $\mathcal{D}_r$, $\mathcal{D}_s$: 100, 20 | 98.87% | NA | NA |

**Table 7**
Cutting Edge Research: Testing Accuracy for Tooth Identification for Dental Panoramic Radiographs, CBCT, and Oral Photographs Datasets ($\mathcal{D}_r$: Training, $\mathcal{D}_s$: Testing, $\mathcal{D}_v$: Validation)

| Research | Dataset | Class# | Data Split (Images#) | Accuracy | F1 | mAP |
|---|---|---|---|---|---|---|
| $\mathcal{M}_c$: Collaborative (Ours) | Panoramic UFBA [37] Radiographs | **32** | $\mathcal{D}_s$: 150 | **98.44%** | **98.75%** | **97.78%** |
| $\mathcal{M}_{i1}$: Faster R-CNN (Ours) | | 4 | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 750, 150, 100 | 91% | 90% | 91% |
| $\mathcal{M}_{i2}$: YOLO-v5 (Ours) | | 4 | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 750, 150, 100 | 99.5% | 99.85% | 99.5% |
| Lai et al.[44] | Panoramic Radiographs | NA | $\mathcal{D}_r$, $\mathcal{D}_s$: 22,262, 1,168 | 87.21% | NA | NA |
| Sathya et al.[45] | Panoramic Radiographs | (s1)2 (s2)4+4 (s3)6+6 | $\mathcal{D}_r$, $\mathcal{D}_s$: (s1)120042, 200 (s2)120042, 800 (s3)191449, 1600 | Precision: (s1)100% (s2)95.24% (s3)90.5% | | |
| Thanathornwong et al.[46] | Panoramic Radiographs | 1 | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 70, 10, 20 subjects | NA | 81% | NA |
| Chung et al. [48] | Panoramic Radiographs | 32 | $\mathcal{D}_r$, $\mathcal{D}_v$, $\mathcal{D}_s$: 574, 162, 82 | NA | 98.43% | 91% |
| Li et al.[49] | CBCT Dataset | 4 | $\mathcal{D}_r$, $\mathcal{D}_s$: 200, 200 | 87% | NA | NA |
| Moriyama et al.[50] | Oral Photographs | 2, 3, 15 | $\mathcal{D}_r$, $\mathcal{D}_s$: 2100, 525 | Accuracy: Screening: 76.5%, Severity: 73.1%, Depth: 47% | | |

## 5.4. Case Study

Five case examples are presented to illustrate the performance of the tooth segmentation model ($\mathcal{M}_{s1}$) and the tooth identification model ($\mathcal{M}_{i1}$) and collaborative model ($\mathcal{M}_c$). These five cases include those involving restored teeth, missing teeth, and dental implants.

**Example 1:** As illustrated in Figure 6, each of the two individual models, as well as the ensemble (pre-refinement)

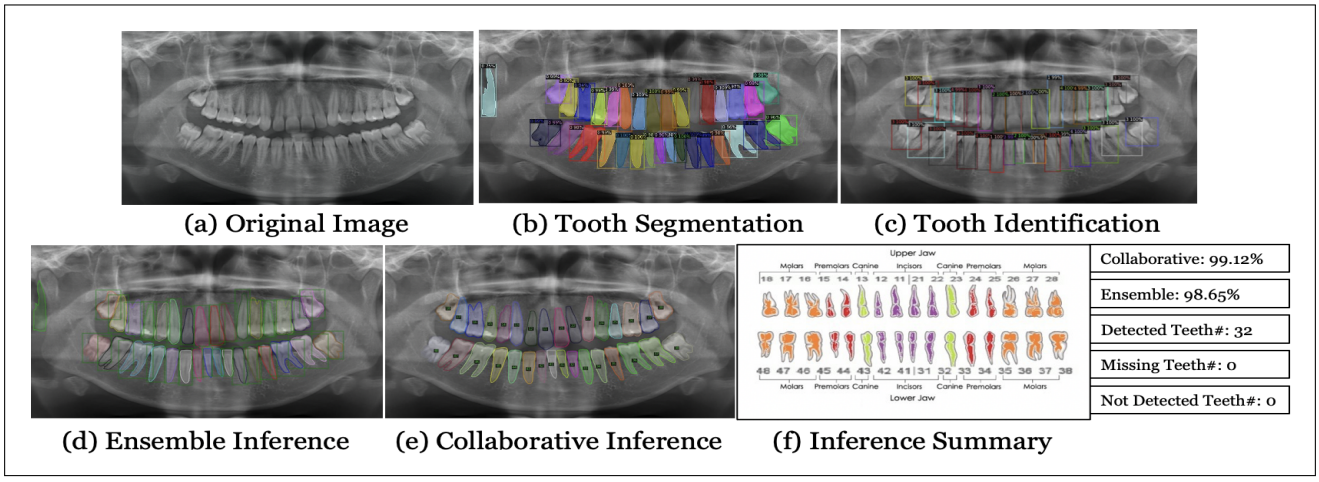



**Fig. 6:** Case 1: Healthy dentition example: This is a set of 32 healthy teeth with no dental treatment that was successfully segmented and identified using both segmentation and identification models. All 32 teeth were discovered. The ensemble and collaborative models perform the best in terms of tooth detection and identification, with a mean average precision (mAP) scores of 98.65% and 99.12%, respectively.

and collaborative (post-refinement) model, were effectively implemented. The output of the segmentation and identification models is integrated, and the ensemble model and summary of the results are displayed. In addition, individual model detection findings were incorporated into the ensemble model. The ensemble model was subjected to a post-refinement process to improve the accuracy of the collaboration model. First, segmentation accuracy was 98.75%, as all 32 permanent teeth were accurately segmented and recognized. Second, the identification model allocated each tooth a unique number; however, one tooth was mistakenly labeled with two separate numbers. It correctly identified the number and type of all teeth, including molars, premolars, canines, and incisors, and then achieved an accuracy of 98.50%. Third, when the conclusions from these two models were combined, the ensemble model had an accuracy of 98.65%. Finally, after post-processing, the image and using ISO numbering standards, the collaborative model's accuracy was 99.12%.

**Example 2:** As illustrated in Figure 7, the segmentation model accurately identified all 32 permanent teeth, with a 99.20% mAP score. In addition, the identification model accurately recognized the tooth numbers and types, including molars, premolars, canines, and incisors; however, a few teeth were duplicated for the same type. For example, a premolar in the mandible, also known as the lower jaw, was recognized as a molar using the Identification model. One of the incisors in the maxilla, also known as the upper jaw, was likewise misidentified as canine but retained a 90.7% mAP score. The ensemble model included the results of the segmentation and identification models, whereas the collaborative model improved accuracy through a collaborative refining process. As a result, the combined accuracy of these two models on the ensemble model was 94.97%. However, after post-processing the ensemble model findings to compensate for incorrectly identified tooth numbers using ISO numbering standards, the collaboration model's accuracy was 97.77%. Figure 7 illustrates the respective output images, as well as a summary of the number of teeth and missing teeth.

**Example 3:** As illustrated in Figure 8, we analyzed a panoramic radiographs image with orthodontic braces. Although the braces appear radiopaque on the panoramic radiograph, the segmentation model with an mAP score of 98.63% accurately segmented and recognized 32 teeth. While most teeth were accurately recognized with their given tooth numbers, a handful was mistakenly identified. The identification model correctly classified an incisor as both an incisor and a canine. In the maxilla, a canine was categorized as both a premolar and a canine, and a premolar as both a premolar and a canine in the mandible, with an accuracy (mAP) of 88.83%. These two model detection findings were incorporated into the ensemble model with an accuracy (mAP) of 93.73%, which was then refined further. The accuracy of the collaborative model was increased through a refining process. By comparison, post-refinement of the ensemble model findings improved the collaborative model's accuracy (mAP) to 98.83% by appropriately numbering the incisor, canine in the maxilla, and premolar on the mandible that the identification model had incorrectly recognized.

**Example 4:** We examined a panoramic radiological image of a tooth set with dental implants and multiple missing teeth, as shown in Figure 9. The segmentation model correctly recognized the 27 mandible teeth (mAP 95.20%). Although the identification model correctly identified all teeth with their associated numbers and types (premolars, canines, and incisors), dental implants were incorrectly identified as teeth with an accuracy of 94.33% (mAP). The combined accuracy of these two models was 94.76% on the ensemble model, but following the post-refinement of the ensemble model's findings, the cooperation model's accuracy was 99.33%.

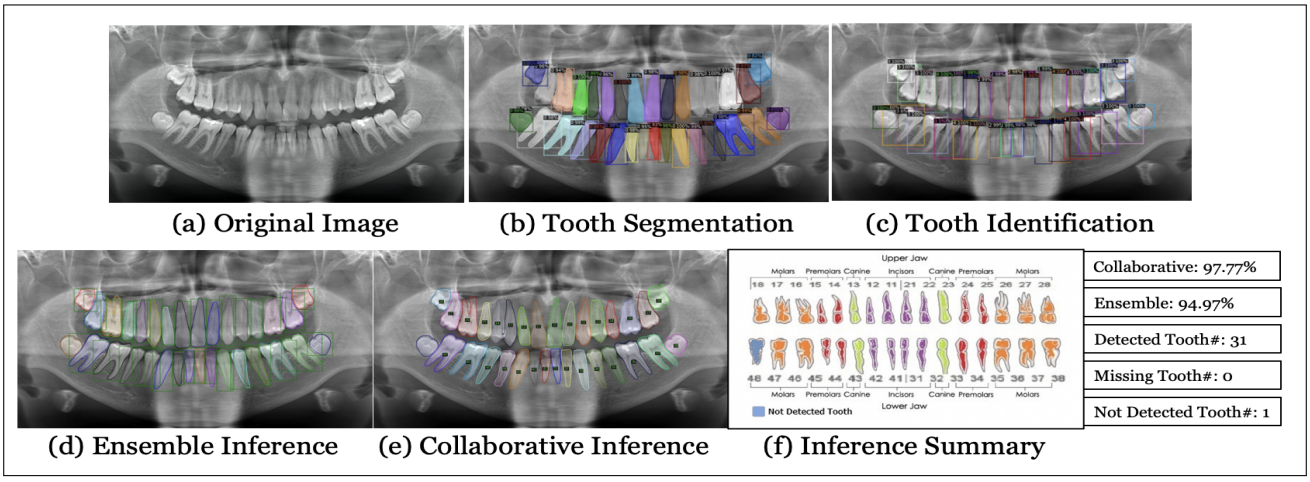



**Fig. 7:** Case 2: Teeth Not Detected: This example has 31 permanent teeth with 3 treated teeth successfully detected by both segmentation and identification models, but one tooth of the 31 was not detected. The ensemble and collaborative models detected and identified the teeth with the mean average precision (mAP) of 94.97% and 97.77%, respectively.

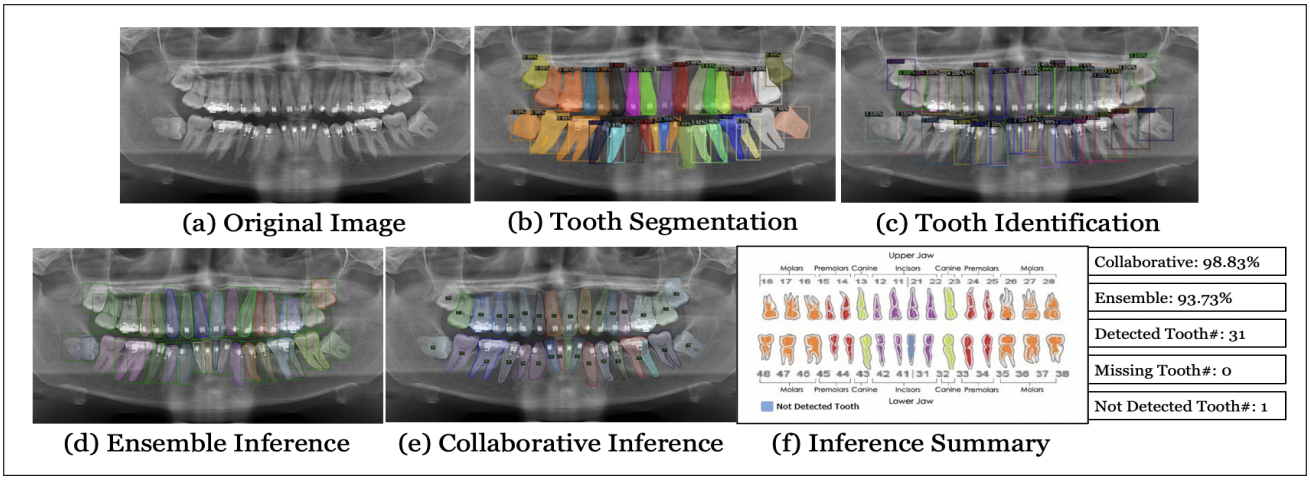



**Fig. 8:** Case 3: Active orthodontic treatment example: An active orthodontic case can regularly be seen in panoramic radiographs, since mid-treatment panoramic radiographs are important in orthodontic care. Both segmentation and identification models correctly detected the 31 permanent teeth and three treated teeth, however one tooth was not detected. Even with braces, the ensemble and collaborative models had a mean average precision (mAP) of 93.73% and 98.83%, respectively, in detecting and identifying teeth.

**Example 5:** As illustrated in Figure 10, a patient had many missing teeth, yet all 8 teeth were successfully segmented with an accuracy (mAP) of 85.33%. The identification model accurately recognized all teeth, with an accuracy (mAP) of 90.88%, except for a few teeth with several tooth numbers for the same tooth. On the maxilla, the identification model correctly classified a premolar as a canine, a canine as both a canine and a premolar, and an incisor as both an incisor and a canine. A premolar is appropriately identified as a canine on the mandible. The ensemble model's accuracy (mAP) was 88.10%. Still, after post-refinement, the collaboration model's accuracy (mAP) climbed to 98.74% by numbering the premolar, canine, and incisor on the maxilla and the premolar on the mandible.

Table 8 summarizes the testing accuracy for the case studies discussed. The accuracy of each model was determined by comparing the observed and predicted results. The ensemble model's testing accuracy was calculated by averaging the two individual models. After merging the outputs, ensemble refinement was done to each to obtain the combined output. Finally, the collaborative approach's testing accuracy is calculated as the average of the two refined models.

## 6. Discussion

Our contribution is to construct a collaborative model by developing two distinct task models: tooth recognition and teeth segmentation. Collaborative learning provides the following advantages: (1) The tooth identification task's performance was increased from 4 to 32 different types of teeth

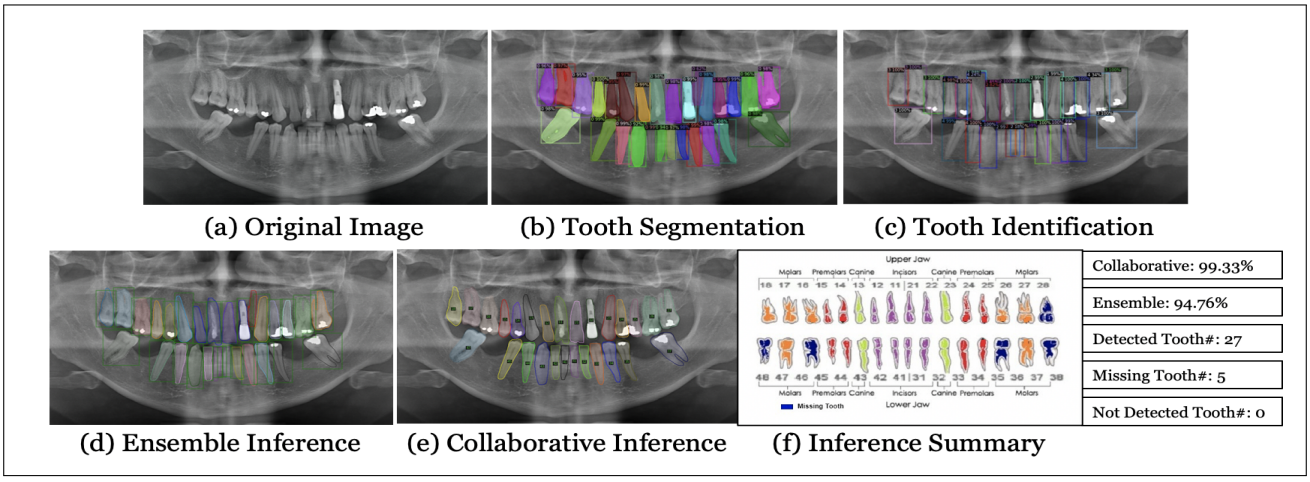



**Fig. 9:** Case 4: Natural dentition with dental implants and some missing teeth example: Dental implants are becoming a regular part of dental care. This example shows a patient with 27 permanent teeth and five missing teeth were effectively detected using both segmentation and identification models, although teeth were not discovered. Interestingly, the collaborative model enhanced the ensemble model's mAP score from 94.76% to 99.33%

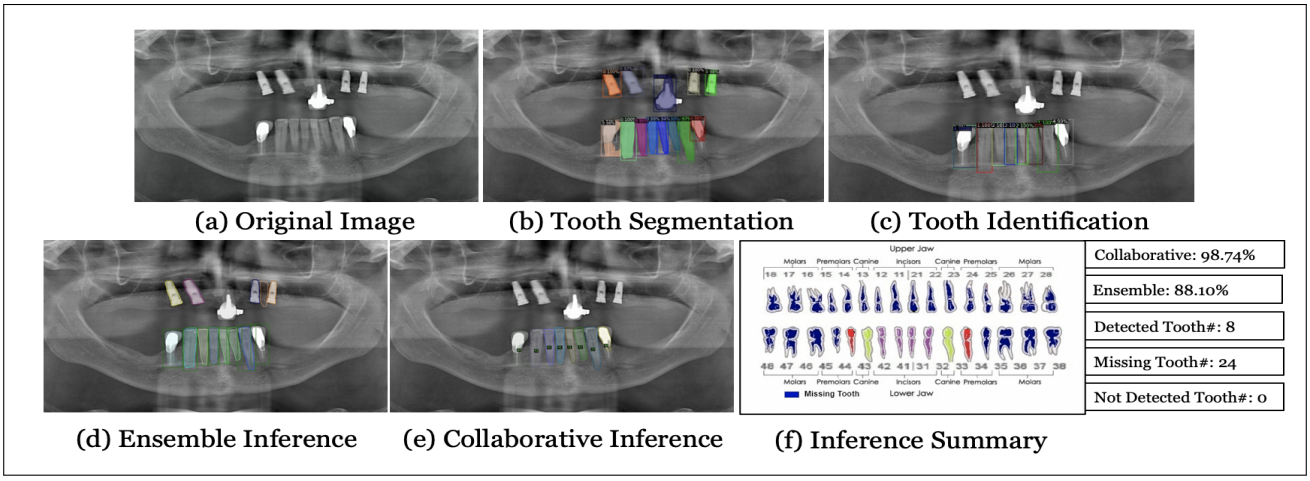



**Fig. 10:** Case 5: Patient with implants and multiple missing teeth example: This case is extremely complicated, involving only 8 permanent teeth and twenty-four missing teeth that were effectively discovered using both segmentation and identification models. Nonetheless, even in this difficult scenario, when compared to the ensemble model, the collaborative model's performance in recognizing and labeling teeth improved considerably, with mean average precision (mAP) increasing from 88.10% to 98.74%.

through collaboration with a tooth segmentation model; (2) The tooth segmentation model's performance was improved through a refinement process for false positive cases to make it more consistent with the tooth identification output; and (3) The two models were integrated into a collaborative model using the inferencing ensemble approach. While the models are not physically connected, their interaction provides insight into the outcomes of the inference of these two models.

First, regarding tooth identification, either YOLO-v5 or Faster R-CNN model can be employed to identify teeth. Faster R-CNN and YOLO-v5 are adequate to achieve a 91% and 99.5% accuracy rate when detecting four distinct tooth kinds. YOLO-v5 ($\mathcal{M}_{i2}$) outperforms Faster R-CNN ($\mathcal{M}_{i1}$) in detecting four unique tooth kinds. However, regardless of whatever one is picked, there is no discernible change in the overall performance of the collaborative model. This is because only tooth numbering is derived from the tooth identification model. Along with Mask R-CNN [51], Faster R-CNN [47] was constructed utilizing the Detectron2 framework [55]. Thus, we built the collaborative model with Mask R-CNN ($\mathcal{M}_{s1}$) and Faster R-CNN ($\mathcal{M}_{i1}$) while achieving an accuracy of 98.44% for $\mathcal{M}_c$'s tooth segmentation and identification tasks.

Second, we have made significant improvements to two tasks, such as tooth segmentation $\mathcal{M}_s$ and tooth identification $\mathcal{M}_i$, by upgrading the models through a thorough training and validation method. Regarding tooth segmentation, while the Mask R-CNN model ($\mathcal{M}_{s1}$) excels in segmenting 32 distinct types of teeth with a 96% accuracy, the U-Net model ($\mathcal{M}_{s2}$) is limited to semantic segmentation of the single class "tooth" with a 97.05% accuracy. Mask R-CNN takes $\mathcal{M}_{s1}$ substantially more processing power than U-Net

**Table 8**
Case Study Results: Testing Accuracy

| | | **Individual Models** | | | | **Ensemble $\mathcal{M}_e$** | | **Collaborative $\mathcal{M}_c$** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Tooth Segmentation $\mathcal{M}_{s1}$** | | **Tooth Identification $\mathcal{M}_{i1}$** | | **Before Refinement** | | **After Refinement** | |
| Case# | Teeth# | mAP | F-1 | mAP | F-1 | mAP | F-1 | mAP | F-1 |
| **Case 1** | 32 | 98.75% | 98.15% | 98.50% | 98.93% | 98.65% | 98.54% | 99.12% | 99.45% |
| **Case 2** | 31 | 99.20% | 98.33% | 90.75% | 89.25% | 94.97% | 93.79% | 97.77% | 98.33% |
| **Case 3** | 31 | 98.63% | 98.20% | 88.83% | 90.15% | 93.73% | 94.17% | 98.83% | 99.11% |
| **Case 4** | 27 | 95.20% | 97.15% | 94.33% | 95.88% | 94.76% | 96.51% | 99.33% | 99.51% |
| **Case 5** | 8 | 85.33% | 87.48% | 90.88% | 93.25% | 88.10% | 90.365% | 98.74% | 98.91% |

$\mathcal{M}_{s2}$ so that it is better appropriate for our application due to its instance-based segmentation.

Third, collaborative learning is improved for multi-task learning through the integration of two separate models. In fact, selecting the appropriate weight for each task is not simple, and the problem becomes even more complicated when dealing with complex models performing multiple tasks. Our collaborative modeling methodology is unique in compared to prior strategies for aggregate modeling. Rather than combining multiple models, we combined the inference findings from these two distinct models. Additionally, because each model is specialized for a specific task, post-processing considerably improves the outcome by fine-tuning the inferencing outputs from several models. Multitask learning thus outperforms two separate tooth segmentation and identification models.

The collaboration model improved their tooth identification accuracy from 91% ($\mathcal{M}_{i1}$: four types of molar, premolar, canine, incisor) to 98.77% ($\mathcal{M}_c$: 32 types of teeth) and the tooth segmentation accuracy from 96% ($\mathcal{M}_{s1}$) to 98.44% ($\mathcal{M}_c$), respectively. Furthermore, this technological advancement is made possible by the combination of these two models, which enables greater adaptability to dental applications of varied magnitudes.

We conducted a thorough evaluation of the collaborative model's effectiveness in enhancing overall performance through collaboration with cutting-edge works. Our evaluation revealed that our study outperforms existing studies in the tasks such as tooth segmentation and identification. We were, however, unable to reproduce and compare their cutting edge research findings in the same setting. This was due to a lack of models, source code, and data. As a consequence, we compared the collaborative model $\mathcal{M}_c$ to the underlying deep learning networks by comparing Mask R-CNN ($\mathcal{M}_{s1}$) vs. U-Net ($\mathcal{M}_{s2}$) and Faster R-CNN ($\mathcal{M}_{i1}$) vs. YOLO-v5 ($\mathcal{M}_{i2}$). After conducting a thorough evaluation, we were able to justify our methodology for the collaborative learning $\mathcal{M}_c$.

Our current study has limitations as well. For example, the proposed framework may not suit unusual dental conditions, such as those containing maxillary and mandibular advancements and setbacks with oral surgery appliances, maxillary nance appliances, mandibular lower lingual holding arches, pathologies, and others. Furthermore, our models may not be the most robust when dealing with low-quality data, such as low resolution, partial information, small pixel size images, or different formats, such as 3D images. Nevertheless, by addressing these inadequacies, the proposed models can benefit dentists by allowing them to evaluate a tooth's suitability for treatment and identifying many possible dental restorations. Additionally, this research might be expanded to include the identification of teeth based on their structure and a 3D data set and analysis to detect and cure growth irregularities in children while their teeth are still developing.

## 7. Conclusion

We have proposed and demonstrated the efficacy of a novel method for collaborative learning in this study. The proposed collaborative learning approach combines inference results from two sequentially created tooth segmentation and identification learning models to generate a summary of the combined findings from inferencing the individual models. Significant improvement is achieved through post-processing and fine-tuning of the two models. Collaborative learning $\mathcal{M}_c$ outcomes significantly outperformed those of individual learning, e.g., 98.77% vs. 96% and 98.44% vs. 91% for tooth segmentation $\mathcal{M}_s$ and tooth identification $\mathcal{M}_i$, respectively. Additionally, comparable or superior learning outcomes are obtained compared to state-of-the-art accuracy in tooth segmentation and tooth identification. Finally, we examined five case studies to demonstrate the proposed model's robustness: healthy dentition, missing teeth, orthodontic treatment in progress, natural dentition with dental implants and missing teeth, and patients with implants and multiple missing teeth.

## References

[1] W. W. Chee, N. Mordohai, Tooth-to-implant connection: a systematic review of the literature and a case report utilizing a new connection design, Clinical implant dentistry and related research 12 (2) (2010) 122–133.

[2] Y. Zhao, P. Li, C. Gao, Y. Liu, Q. Chen, F. Yang, D. Meng, Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network, Knowledge-Based Systems 206 (2020) 106338.

[3] Z. Cui, C. Li, W. Wang, Toothnet: Automatic tooth instance segmentation and identification from cone beam ct images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6368–6377.

[4] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, L. Oliveira, Deep instance segmentation of teeth in panoramic x-ray images, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2018, pp. 400–407.

[5] B. Silva, L. Pinheiro, L. Oliveira, M. Pithon, A study on tooth segmentation and numbering using end-to-end deep neural networks, in: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2020, pp. 164–171.

[6] A. Haghanifar, M. M. Majdabadi, S.-B. Ko, Automated teeth extraction from dental panoramic x-ray images using genetic algorithm, in: 2020 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2020, pp. 1–5.

[7] S. Lee, S. Woo, J. Yu, J. Seo, J. Lee, C. Lee, Automated cnn-based tooth segmentation in cone-beam ct for dental implant planning, IEEE Access 8 (2020) 50507–50518.

[8] Z.-H. Zhou, Ensemble methods: foundations and algorithms, Chapman and Hall/CRC, 2019.

[9] H. Li, J. Y.-H. Ng, P. Natsev, Ensemblenet: End-to-end optimization of multi-headed models, arXiv preprint arXiv:1905.09979 (2019).

[10] B. Brazowski, E. Schneidman, Collective learning by ensembles of altruistic diversifying neural networks, arXiv preprint arXiv:2006.11671 (2020).

[11] S. Fort, H. Hu, B. Lakshminarayanan, Deep ensembles: A loss landscape perspective, arXiv preprint arXiv:1912.02757 (2019).

[12] C. Shui, A. S. Mozafari, J. Marek, I. Hedhli, C. Gagné, Diversity regularization in deep ensembles, arXiv preprint arXiv:1802.07881 (2018).

[13] O. Sagi, L. Rokach, Ensemble learning: A survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (4) (2018) e1249.

[14] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine learning 51 (2) (2003) 181–207.

[15] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, arXiv preprint arXiv:1612.01474 (2016).

[16] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, S. Savarese, Which tasks should be learned together in multi-task learning?, in: International Conference on Machine Learning, PMLR, 2020, pp. 9120–9132.

[17] D. R. Anderson, K. P. Burnham, Avoiding pitfalls when using information-theoretic methods, The Journal of wildlife management (2002) 912–918.

[18] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).

[19] T. L. Koch, M. Perslev, C. Igel, S. S. Brandt, Accurate segmentation of dental panoramic radiographs with u-nets, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 15–19.

[20] A. B. Oktay, A. Gurses, Detection, segmentation, and numbering of teeth in dental panoramic images with mask regions with convolutional neural network features, in: State of the Art in Neural Networks and their Applications, Elsevier, 2021, pp. 73–90.

[21] L. Pinheiro, B. Silva, B. Sobrinho, F. Lima, P. Cury, L. Oliveira, Numbering permanent and deciduous teeth via deep instance segmentation in panoramic x-rays, in: 17th International Symposium on Medical Information Processing and Analysis, Vol. 12088, SPIE, 2021, pp. 95–104.

[22] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, I. Kim, Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs, Oral surgery, oral medicine, oral pathology and oral radiology 129 (6) (2020) 635–642.

[23] A. Wirtz, S. G. Mirashi, S. Wesarg, Automatic teeth segmentation in panoramic x-ray images using a coupled shape model in combination with a neural network, in: International conference on medical image computing and computer-assisted intervention, Springer, 2018, pp. 712–719.

[24] Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, W. Wang, Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model, Medical Image Analysis 69 (2021) 101949.

[25] J. Lee, M. Chung, M. Lee, Y.-G. Shin, Tooth instance segmentation from cone-beam ct images through point-based detection and gaussian disentanglement, arXiv preprint arXiv:2102.01315 (2021).

[26] S. Kakehbaraei, H. Seyedarabi, A. T. Zenouz, Dental segmentation in cone-beam computed tomography images using watershed and morphology operators, Journal of medical signals and sensors 8 (2) (2018) 119.

[27] J. Zhang, C. Li, Q. Song, L. Gao, Y.-K. Lai, Automatic 3d tooth segmentation using convolutional neural networks in harmonic parameter space, Graphical Models 109 (2020) 101071.

[28] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7482–7491.

[29] Y. Suhail, M. Upadhyay, A. Chhibber, et al., Machine learning for the diagnosis of orthodontic extractions: a computational analysis using ensemble learning, Bioengineering 7 (2) (2020) 55.

[30] M. A. Hasan, N. A. Abdullah, M. M. Rahman, M. Y. I. B. Idris, O. F. Tawfiq, Dental impression tray selection from maxillary arch images using multi-feature fusion and ensemble classifier, IEEE Access 9 (2021) 30573–30586.

[31] K.-S. Lee, S.-K. Jung, J.-J. Ryu, S.-W. Shin, J. Choi, Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs, Journal of clinical medicine 9 (2) (2020) 392.

[32] V. Yaduvanshi, R. Murugan, T. Goel, An automatic classification methods in oral cancer detection, in: Health Informatics: A Computational Perspective in Healthcare, Springer, 2021, pp. 133–158.

[33] A. Haghanifar, M. M. Majdabadi, S.-B. Ko, Paxnet: Dental caries detection in panoramic x-ray using ensemble transfer learning and capsule classifier, arXiv preprint arXiv:2012.13666 (2020).

[34] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[35] J. Krois, A. G. Cantu, A. Chaurasia, R. Patil, P. K. Chaudhari, R. Gaudin, S. Gehrung, F. Schwendicke, Generalizability of deep learning models for dental image analysis, Scientific reports 11 (1) (2021) 1–7.

[36] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[37] G. Silva, L. Oliveira, M. Pithon, Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives, Expert Systems with Applications 107 (2018) 15–31.

[38] C.-H. Wu, W.-H. Tsai, Y.-H. Chen, J.-K. Liu, Y.-N. Sun, Model-based orthodontic assessments for dental panoramic radiographs, IEEE Journal of biomedical and health informatics 22 (2) (2017) 545–551.

[39] C. Lian, L. Wang, T.-H. Wu, F. Wang, P.-T. Yap, C.-C. Ko, D. Shen, Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners, IEEE transactions on medical imaging 39 (7) (2020) 2440–2450.

[40] T.-H. Wu, C. Lian, S. Lee, M. Pastewait, C. Piers, J. Liu, F. Wang, L. Wang, C. Jackson, W.-L. Chao, et al., Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans, arXiv preprint arXiv:2109.11941 (2021).

[41] S. Tian, M. Wang, F. Yuan, N. Dai, Y. Sun, W. Xie, J. Qin, Efficient computer-aided design of dental inlay restoration: A deep adversarial framework, IEEE Transactions on Medical Imaging 40 (9) (2021) 2415–2427.

[42] S. Tian, M. Wang, N. Dai, H. Ma, L. Li, L. Fiorenza, Y. Sun, Y. Li, Dcpr-gan: Dental crown prosthesis restoration using two-stage generative adversarial networks, IEEE Journal of Biomedical and Health Informatics (2021).

[43] H. Lei, W. Liu, H. Xie, B. Zhao, G. Yue, B. Lei, Unsupervised domain adaptation based image synthesis and feature alignment for joint optic disc and cup segmentation, IEEE Journal of Biomedical and Health Informatics (2021).

[44] Y. Lai, F. Fan, Q. Wu, W. Ke, P. Liao, Z. Deng, H. Chen, Y. Zhang, Lcanet: Learnable connected attention network for human identification using dental images, IEEE Transactions on Medical Imaging 40 (3) (2020) 905–915.

[45] B. Sathya, R. Neelaveni, Transfer learning based automatic human identification using dental traits-an aid to forensic odontology, Journal of Forensic and Legal Medicine 76 (2020) 102066.

[46] B. Thanathornwong, S. Suebnukarn, Automatic detection of periodontal compromised teeth in digital panoramic radiographs using faster regional convolutional neural networks, Imaging Science in Dentistry 50 (2) (2020) 169.

[47] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015) 91–99.

[48] M. Chung, J. Lee, S. Park, M. Lee, C. E. Lee, J. Lee, Y.-G. Shin, Individual tooth detection and identification from dental panoramic x-ray images via point-wise localization and distance regularization, Artificial Intelligence in Medicine 111 (2021) 101996.

[49] Z. Li, S.-H. Wang, R.-R. Fan, G. Cao, Y.-D. Zhang, T. Guo, Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling, International Journal of Imaging Systems and Technology 29 (4) (2019) 577–583.

[50] Y. Moriyama, C. Lee, S. Date, Y. Kashiwagi, Y. Narukawa, K. Nozaki, S. Murakami, A mapreduce-like deep learning model for the depth estimation of periodontal pockets., in: HEALTHINF, 2019, pp. 388–395.

[51] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[52] G. Jocher, K. Nishimura, T. Mineeva, R. Vilariño, Yolov5 (Mar. 19 2022).
URL https://github.com/ultralytics/yolov5

[53] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[54] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[55] Y. Wu, A. Kirillov, F. Massa, W. Lo, R. Girshick, Detectron2 [www document], URL https://github. com/facebookresearch/detectron2 (accessed 3.3. 21) (2019).

[56] B. Dwyer, J. Nelson, Roboflow (version 1.0) (Dec. 20 2021).
URL https://roboflow.com