
INSTANCE SEGMENTATION AND TEETH CLASSIFICATION IN PANORAMIC X-RAYS

A PREPRINT

 **Devichand Budagam**

Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
devichand579@gmail.com

 **Ayush Kumar**

Department of Humanities and Social Sciences
Indian Institute of Technology Kharagpur
Kharagpur, India
ayushkumar8804143@gmail.com

 **Sayan Ghosh**

Department of Chemical Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
gh.sayan203kgp@gmail.com

 **Anuj Shrivastav**

Department of Architecture and Regional Planning
Indian Institute of Technology Kharagpur
Kharagpur, India
anujshrivastav594@gmail.com

 **Azamat Zhanatuly Imanbayev**

School of IT and Engineering
Kazakh-British Technical University
Almaty, Kazakhstan
a.imanbaev@kbtu.kz

 **Iskander Rafailovich Akhmetov**

School of IT and Engineering
Kazakh-British Technical University
Almaty, Kazakhstan
i.akhmetov@kbtuedu.onmicrosoft.com

 **Dmitrii Kaplun**

Artificial Intelligence Research Institute
China University of Mining and Technology
Xuzhou, China

Department of Automation and Control Processes
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
dikaplun@etu.ru

 **Sergey Antonov**

Department of Automation and Control Processes
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
saantonov@etu.ru

 **Artem Rychenkov**

Department of Information Systems
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
artem.rychenkov18@gmail.com

 **Gleb Cyganov**

Faculty of Digital Transformation
ITMO University
St. Petersburg, Russia
gleb.geo@mail.ru

 **Aleksandr Sinitca**

Centre for Digital Telecommunication Technologies
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
amsinitca@etu.ru

ABSTRACT

Teeth segmentation and recognition are critical in various dental applications and dental diagnosis. Automatic and accurate segmentation approaches have been made possible by integrating deep

learning models. Although teeth segmentation has been studied in the past, only some techniques were able to effectively classify and segment teeth simultaneously. This article offers a pipeline of two deep learning models, U-Net and YOLOv8, which results in BB-UNet, a new architecture for the classification and segmentation of teeth on panoramic X-rays that is efficient and reliable. We have improved the quality and reliability of teeth segmentation by utilising the YOLOv8 and U-Net capabilities. The proposed networks have been evaluated using the mean average precision (mAP) and dice coefficient for YOLOv8 and BB-UNet, respectively. We have achieved a 3% increase in mAP score for teeth classification compared to existing methods, and a 10-15% increase in dice coefficient for teeth segmentation compared to U-Net across different categories of teeth. A new Dental dataset was created based on UFBA-UESC dataset with Bounding-Box and Polygon annotations of 425 dental panoramic X-rays. The findings of this research pave the way for a wider adoption of object detection models in the field of dental diagnosis.

Keywords Teeth Segmentation · Teeth Classification · Panoramic X-Rays · YOLOv8 · U-Net · BB-UNet

1 Introduction

The increasing demand for dental care and dentists is driven by population growth, increasing life expectancy, and a heightened focus on oral health. Advanced technologies are necessary to improve diagnostic quality, treatment planning, and patient care. In this evolving dental field, deep learning architectures have emerged as a promising solution [1, 2, 3, 4] despite certification difficulties.

The FDI World Dental Federation notation [5] shown in Fig.1 is a dental notation system used to identify and label teeth uniquely. It is a standard method used worldwide to communicate tooth information in a consistent and universally understood way. The FDI notation system assigns a two-digit number to each tooth in the mouth and is widely used by dental professionals.

Teeth segmentation and classification, which involves accurately delineating individual teeth, consists of classifying each image pixel into an object of interest and numbering the teeth from dental images, which is crucial for various dental applications. The traditional manual method of segmenting and annotating teeth can be quite labor-intensive and time-consuming; unclear images can make it difficult for a specialist to evaluate them. To this end, automated teeth segmentation using deep learning techniques has become an effective alternative due to its success in various complex computer vision tasks. However, due to the high variability, low contrast, and high amplitude noise in panoramic X-rays, automated teeth segmentation and classification pose many challenges in recognizing and locating each tooth and classifying it precisely. They can even mislead in the formulation of the diagnosis.

So far, most of the research on teeth segmentation is based on several unsupervised pixel-wise segmentation approaches [6] mainly studied using intraoral images (periapical X-rays or bitewing X-rays). The article by G. Jader et al. [6] experimented with small data sets for most of the methods discussed and evaluated them with very few changes to the dataset, using a dataset of 1500 panoramic X-ray images called the UFBA-UESC dental image dataset. They concluded that these segmentation methods failed to effectively isolate the teeth from neighboring teeth in the mouth. Very few authors performed experiments on panoramic X-ray images, threshold-based [7, 8, 9], region-based [10], cluster-based [11], boundary-based [12, 13, 14] and one-class segmentation [15, 16, 17]. In the paper on deep instance segmentation of teeth on panoramic radiographs [18] the authors investigated teeth segmentation on the UFBA-UESC dental images dataset using a Mask R-CNN [19] for instance segmentation task, but all teeth were classified into a single category ignoring independent tooth recognition. Silva et al.[20] proposed different deep learning architectures PANet, HTC, ResNeSt, and Mask R-CNN, in [21], and they used Mask R-CNN with different segmentation heads pointRend [22] and FCN [23] and obtained excellent results.

This paper proposes a different method to perform teeth classification and instance segmentation by exploiting spatial prior information on U-Net [24] using a one-stage object detection model, YOLO [25] rather than performing the task with a two-stage objection detection model such as Mask R-CNN as in previously mentioned methods. U-Net architecture can accurately segment the image and locate tooth positions, enabling us to efficiently segment the teeth. Integration of prior knowledge, such as the location or shape of the objects, can improve the performance of CNNs. Integration of prior knowledge in CNNs has provided good results in the paper Zotti et al. [26]. Therefore, we impose the teeth location information with the help of a boundary box prior to the skip connections of U-Net [27].

Table 1: Description of the UFBA-UESC dataset

| Category | 32 Teeth | Restoration | Dental appliance | Images | Used Images |
|----------|----------|--------------------------------------|------------------|--------|-------------|
| 1 | ✓ | | ✓ | 73 | 24 |
| 2 | ✓ | ✓ | | 220 | 72 |
| 3 | ✓ | | ✓ | 45 | 15 |
| 4 | ✓ | | | 140 | 32 |
| 5 | | Images containing dental implant | | 120 | 37 |
| 6 | | Images containing more than 32 teeth | | 170 | 30 |
| 7 | | ✓ | ✓ | 115 | 33 |
| 8 | | ✓ | | 457 | 140 |
| 9 | | | ✓ | 45 | 7 |
| 10 | | | | 115 | 35 |
| Total | | | | 1500 | 425 |

2 Materials and Method

2.1 Dataset Construction

In dental image analysis, where prediction accuracy is paramount, our preparatory work focuses on carefully selecting a representative dataset and processing it for subsequent model training. Our foundation was the UFBA-UESC Dental Images Dataset introduced by G. Jader et al.[6], an extensive collection of anonymized panoramic X-ray dental images of high variability. It comprises 1500 images of dimensions $512 \times 512 \times 3$, classified into 10 distinct categories. These categories represented different types of dental cases, with standard 32 teeth, with dental appliances, and with dental restorations. The dataset also uses images where the number of teeth is less than 32 if they are extracted and more if the jaw has an abnormal mutation. This diversity mirrored the real-world variations in dental scans due to factors such as dental anomalies or missing teeth. The structure of the open data set UFBA-UESC is presented in table 1.

Due to the lack of image labels necessary for training models in the presented dataset, we initially labelled 425 images from the set. We collected two sets of basic annotations, for instance segmentation and object detection respectively. We selected these 425 images randomly from all categories and maintained same distribution of images across all categories as in the original dataset. For manual labelling, we chose the semi-automated annotation tool Roboflow

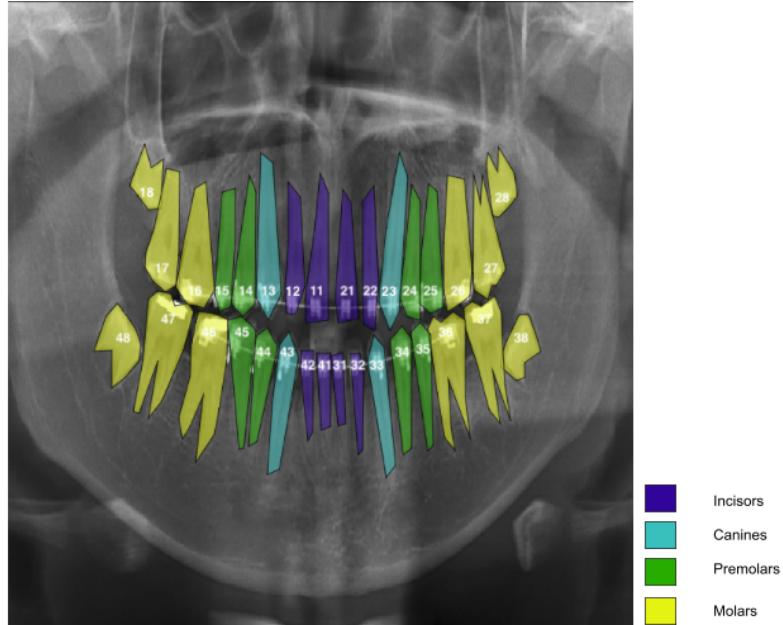


Figure 1: FDI Teeth Numbering Notation

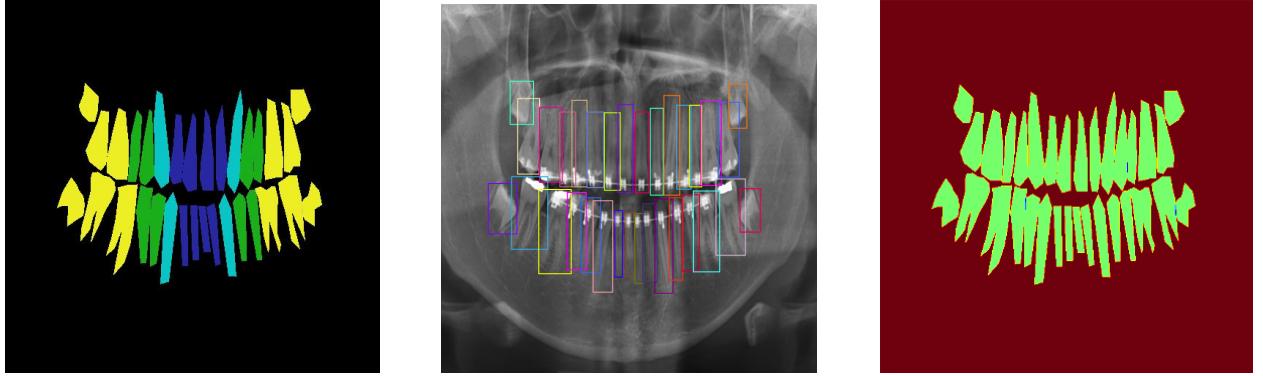


Figure 2: (a) Annotated polygon mask. (b) Annotated bounding Box scan. (c) Binary mask of polygon-based annotation

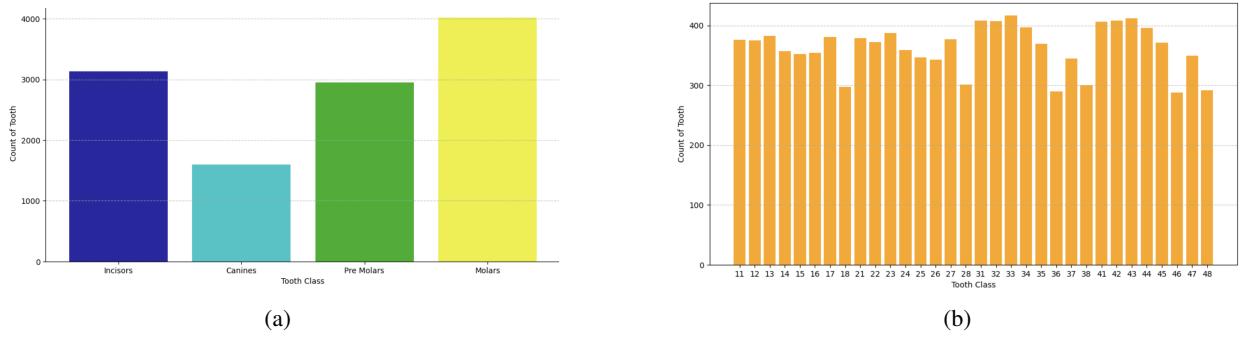


Figure 3: (a) Class distribution among 32 teeth (b) Count of each tooth

[28] for bounding box annotations needed for the objection detection task. We also used another annotation tool, Apeer [29], to help us create individual segmentation masks for each of the 32 teeth in the images in the dataset. These binary masks provided additional information by focusing on the fine contours and boundaries of the teeth. We converted the resulting segmented polygons into binary maps of size $512 \times 512 \times 32$. This comprehensive approach to annotation was pivotal in ensuring our model's success in dental image analysis. An example of annotations is shown in Fig. 2.

From Fig. 3, we can see that the dataset is not biased towards any particular class of teeth and maintains variability across all classes of teeth and all categories of panoramic X-ray images, which makes the dataset suitable for training teeth detection or segmentation models. It is possible to use this data set extensively for teeth segmentation and detection, as it is one of the largest publicly available datasets.

2.2 Model Architecture

2.2.1 YOLO Architecture

At the moment, many different models that solve the object detection problem have been developed. However, the two most commonly used approaches for object detection tasks include single-stage and two-stage neural network architectures. One-stage approaches have an advantage over two-stage approaches in that the former produces outcomes considerably more quickly and efficiently. YOLO is one of the most popular single-stage models to solve the problem of object detection tasks on images. Recently, many different versions of the YOLO model have been released by various developers. Such models include YOLO, YOLOv4, YOLOv5, YOLOv7, YOLOv8, and others.

YOLOv8 is the latest version of the YOLO model. The architecture of this model is shown in Fig.4 and includes new enhancements that provide superior detection accuracy while maintaining high speed and efficiency. The YOLOv8 backbone network architecture is based on the previous version of YOLOv5. The critical difference between this model and the old version is that the authors of YOLOv8 abandoned predefined anchors, simplifying the model's architecture and eliminating the need for additional tuning of hyperparameters. The YOLOv8 model also includes several essential modifications.

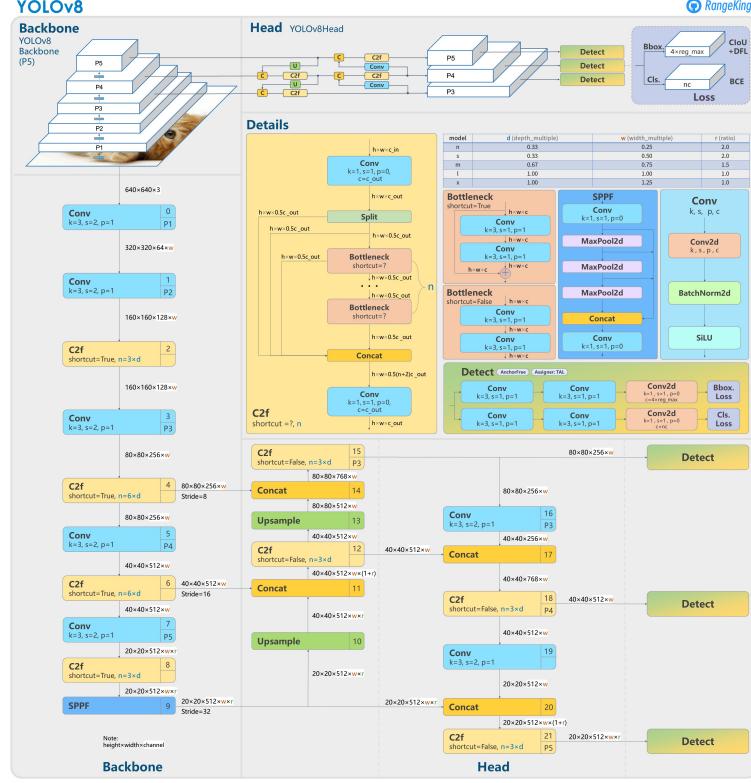


Figure 4: YOLOv8 Architecture [30]

- **CSPDarknet53 Feature Extractor:** YOLOv8 uses CSPDarknet53, a variant of the Darknet architecture, as the feature extractor. This component comprises convolutional layers, batch normalization, and SiLU activation functions. The notable difference is that YOLOv8 replaces the original 6x6 convolutional layer with a 3x3 convolutional layer to improve feature extraction.
- **Module C2f:** YOLOv8 introduces the C2f module to combine high-level features with contextual information efficiently. This is achieved by concatenating bottleneck block outputs, consisting of two 3x3 convolutions with residual connections. This architectural change aims to improve the presentation of features. In the previous version of YOLOv5, instead of C2f, the C3 block, which contains one more convolution layer than C2f module, was used. Considering that C2f is used eight times throughout the entire architecture, this change is a significant advantage in choosing a new architecture.
- **Detection head:** YOLOv8 uses an anchor-free detection strategy, eliminating the need for predefined anchor fields and directly predicting the centres of objects. A significant improvement to the model detection head is the use of a different activation function. Objectness estimation in the output layer uses a sigmoid activation function representing the probability of an object being present in the bounding box. YOLOv8 uses a softmax function for class probabilities that indicates the probability of an object belonging to each class. To optimize the model, YOLOv8 uses CIoU (complete intersection through a union) and DFL (dynamic focal loss) loss functions for bounding box regression and binary cross-entropy for classification. These loss functions effectively improve object detection, especially for small objects.

Thanks to all these enhancements, the model performed better than earlier object detection models, which makes it appropriate for classifying teeth which are difficult to classify because of numerous teeth and minute details found in panoramic X-rays.

2.2.2 BB-UNet Architecture

Our methodology is built upon the U-Net framework, a highly regarded model for the semantic segmentation of medical images. U-Net is selected for its efficiency in delivering robust results, even with limited training data. The architecture includes encoder layers, which extract contextual information from images, and decoder layers, focused

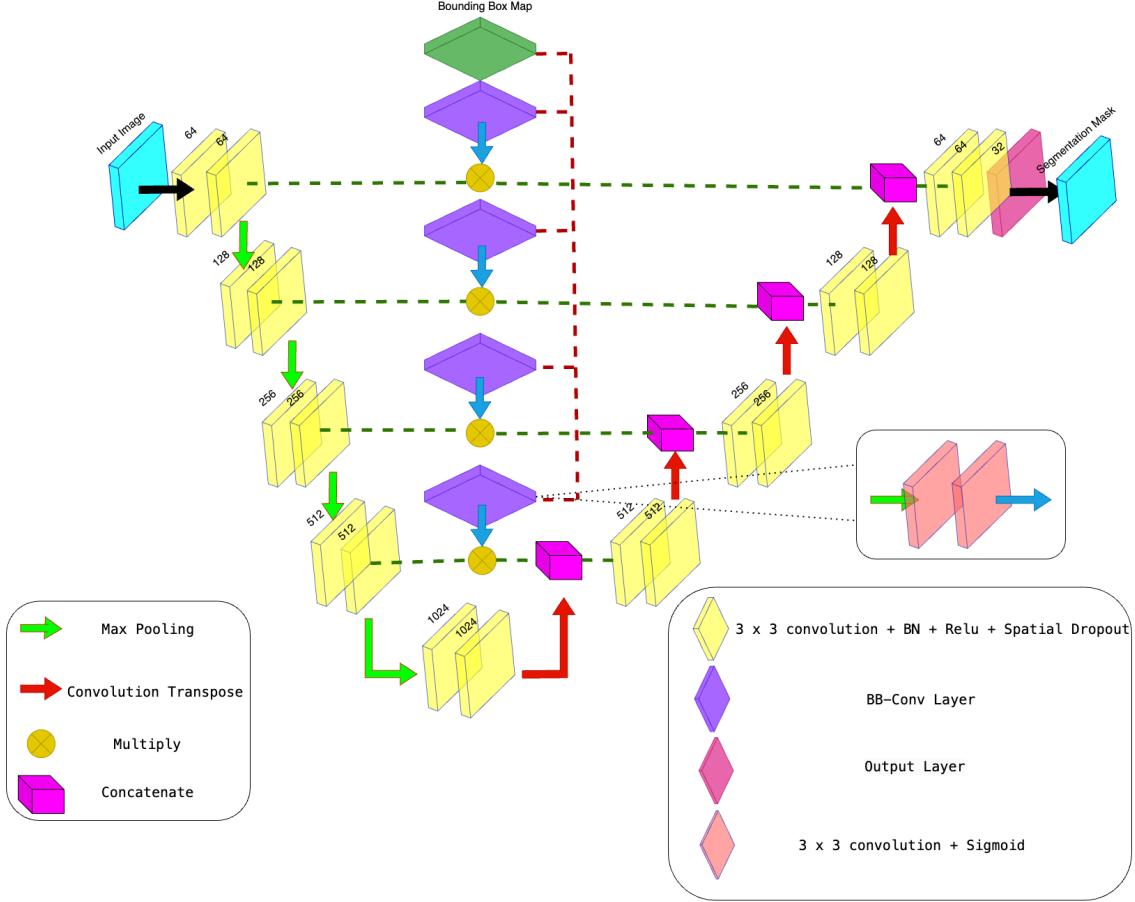


Figure 5: BB-UNet Architecture

on identifying patterns and reverting image maps to their original dimensions. We use skip connections to merge both contextual and positional information effectively. Despite U-Net's proficiency in integrating global features, it struggles to incorporate specific location data. To overcome this, we have integrated location-based constraints into the learning process, specifically using binary maps of bounding boxes, as per methods introduced in [27], previously in articles [26] and [31] have implemented constrained neural networks for medical image analysis for achieving better performance. This integration enhances the model's ability to consider location information and improves performance.

Incorporating prior knowledge via bounding boxes is achieved through the use of BB-Convolution layers added to the skip connections. A BB-Convolution layer consists of 2D max-pooling, followed by dual convolution layers, and culminates with a sigmoid activation layer. Different teeth are identified using bounding boxes, fed in through a 32 multi-channel binary map into the BB-Convolution layer. This process produces a feature map that assists the network in more accurately pinpointing the location of each tooth. These BB-Convolution layers are incorporated into each skip connection, as demonstrated in Fig.5. Our model is structured into four phases, each containing encoder and decoder elements. The encoder layers gather contextual information and features within the images. In contrast, the decoder layers focus on pattern identification and restoring the image maps to their initial size. Within the encoder phases, two convolution blocks are followed by max-pooling layers (with a stride of 2) to decrease feature resolution. The decoder phase includes a transposed convolution block followed by two convolution blocks. At every stage, the output from the BB-Convolution layer is element-wise multiplied with the encoder's output features before being concatenated with the output from the decoder's transposed convolution layers.

The final layer of our architecture consists of a 1×1 convolution, ending with a softmax activation that generates a pixel-by-pixel probability map.

We have decided to use a regularized variant of Dice loss to optimize the parameters of the BB-UNet during the model's training. Dice loss is a widely used metric in medical segmentation and computer vision tasks for calculating the similarity between two images and is defined as:

$$Loss = \frac{1}{N} \sum_{n=0}^N \frac{2 \sum_{i=1}^M P_n(i) \hat{P}_n(i)}{\sum_{i=1}^M P_n(i)^2 + \sum_{i=1}^M \hat{P}_n(i)^2} + \lambda \frac{1}{N} \sum_{n=0}^N \sum_{i=0}^M (P_n(i) - \hat{P}_n(i))^2 \quad (1)$$

Where N is the number of class labels, and M is the number of pixels in each channel of the image. $P_n(i)$ and $\hat{P}_n(i)$ are the pixel values in the predicted map and ground truth label respectively. Here λ , being a regularisation constant. The latter component of the loss function plays a significant role in preventing overfitting along with the spatial dropout layers introduced in the BB-UNet architecture.

2.3 Evaluation Metrics

We have used several metrics to evaluate the quality of the proposed method. We calculate accuracy, recall, precision, and mAP:

- *Accuracy* calculates the ratio of correctly predicted instances to the total number of instances.
- *Recall* measures the ability of the model to capture and correctly identify all relevant instances of a particular class.
- *Precision* measures the accuracy of the positive predictions made by the model. It indicates the proportion of true positives among all instances predicted as positive.
- *mAP* calculates the precision-recall area's average under the curve for multiple classes at different confidence thresholds, providing a comprehensive evaluation of model performance.
- *AP50* calculates the precision-recall area's average under the curve for multiple classes at a confidence thresholds of 0.5.

These metrics can be formulated in terms of the confusion matrix. If we have N classes and a confusion matrix M , then Equations 2, 3, 4 and 5 can be used.

$$Accuracy = \frac{\sum_{i=1}^N M_{ii}}{\sum_{i=1}^N \sum_{j=1}^N M_{ij}} \quad (2)$$

$$Precision_i = \frac{M_{ii}}{\sum_{j=1}^N M_{ji}} \quad (3)$$

$$Recall_i = \frac{M_{ii}}{\sum_{j=1}^N M_{ij}} \quad (4)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n (Recall_j - Recall_{j-1}) Precision_j \quad (5)$$

Where M_{ij} is the corresponding element of a confusion matrix, M and n is the number of thresholds. mAP is used as the primary metric for teeth classification. In contrast, these metrics cannot provide a robust assessment of instance segmentation of teeth; the dice coefficient is used as the primary metric for instance segmentation of teeth. The dice coefficient is a measure of the similarity between two sets, and it is used to quantify the agreement between the predicted segmentation masks and the ground truth masks. The dice coefficient is defined as:

$$\text{Dice Coefficient} = \frac{1}{N} \sum_{n=0}^N \frac{2 \sum_{i=1}^M P_n(i) \hat{P}_n(i)}{\sum_{i=1}^M P_n(i)^2 + \sum_{i=1}^M \hat{P}_n(i)^2} \quad (6)$$

where N is the number of class labels and M is number of pixels in each channel of the image. $P_n(i)$ and $\hat{P}_n(i)$ are the pixel values in predicted map and the ground truth label, respectively.

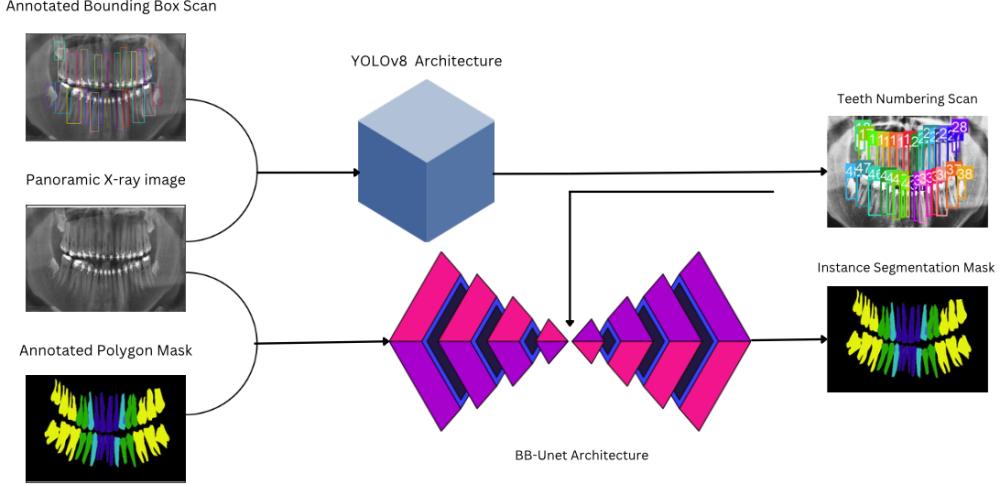


Figure 6: Pipeline of the method

2.4 Model Pipeline and Training

We propose a pipeline comprising a two-step process of YOLOv8 and BB-UNet architectures as shown in Fig.6. The first step involves training the YOLOv8 model to classify the teeth, locate the teeth, and extract the bounding boxes of each located tooth. The latter step is training the BB-UNet, for instance segmentation of teeth.

For training YOLOv8, the images were resized to 640×640 , and histogram equalization was applied to improve the contrast of the images. From the originally built dataset, 894 training images and 128 validation images were created using two augmentation approaches: random cropping with a range of 0% to 20% and brightness adjustment with a range of 0% to 10%. Throughout the training, we processed the dataset in batches, each containing ten images. To avoid overfitting, we applied dropout with a rate of 0.6 as a regularization technique. Additionally, to ensure efficient learning during the training, SGD optimizer with a learning rate of 0.005 was used. Fig.7 shows the variation in training losses, that is, box loss, class loss, DFL loss and mAP score on validation dataset varies over a span of 30 epochs for YOLOv8 training. For training BB-UNet, the bounding box information from YOLOv8 was used to

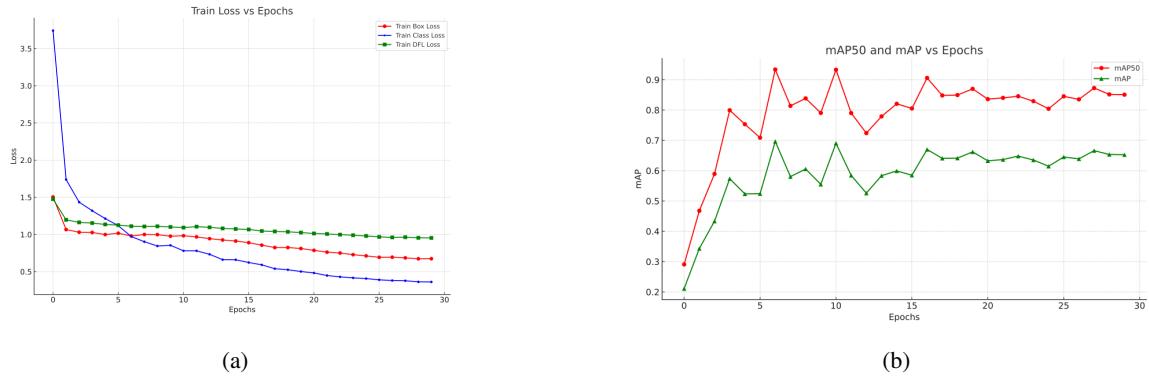


Figure 7: (a) Training loss curves for YOLOv8 (b) AP50 and mAP curves for YOLOv8

generate $512 \times 512 \times 32$ bounding box binary maps which served as inputs to BB-Convolution layers. CLAHE [32](Contrast Limited Adaptive Histogram Equalisation) with a contrast limit of 0.02 was applied to enhance the details of the image. Horizontal and vertical flipping were used as data augmentation techniques, finally generating training data of 340 images and test data of 85 images. The model was trained with a batch size of 2 tensors of size $512 \times 512 \times 35$, where 32 channels correspond to binary maps of bounding boxes and the remaining three channels correspond to the original image which served as inputs to the BB-UNet. Fig.8 shows the variation in training loss, validation loss, and dice coefficient on validation dataset over 60 epochs for BB-UNet training.

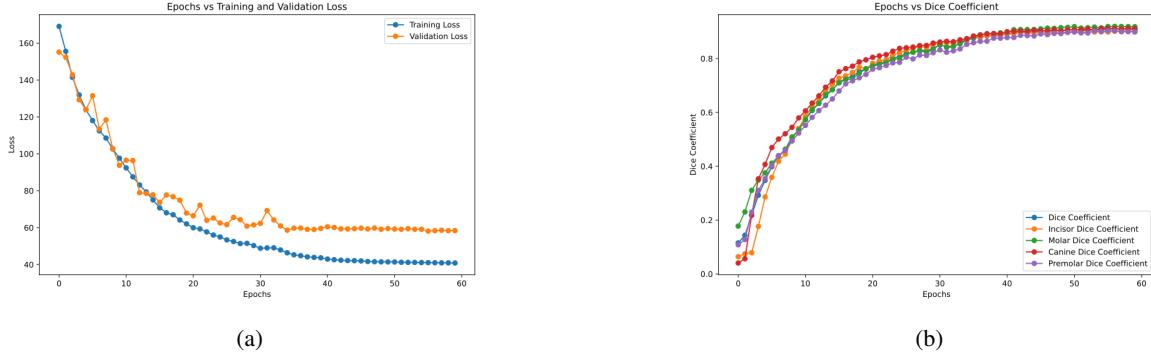


Figure 8: (a) Training and Validation loss curves for BB-UNet (b) Dice coefficient curves for BB-UNet

A dropout rate of 0.12 and regularisation weight, λ of 0.1, were used to prevent overfitting, a momentum of 0.99, and a learning rate of 0.0003 with Adam optimizer were used for the training process. The learning rate was halved if validation loss did not improve over a period of 5 epochs. A 4-fold cross-validation method is used to train the model on the training data, and the average dice coefficient over the 4-fold cross-validation is 0.82 ± 0.006 . A Nvidia A100 GPU with a 12-core CPU and 120GB RAM was used for all training and evaluation processes.

3 Experiments and Results

3.1 Quantitative Analysis

The YOLOv8 model was evaluated on the dataset containing 425 images and 11591 instances of teeth and achieved a precision of 94.3, a recall of 92.3, and a mAP of 72.9 and a mean average precision at 50 (AP50) of 94. 6% in all classes of teeth. The results indicated a striking balance between precision and recall and achieved an excellent mAP. These results suggest the effectiveness of the model in accurately localising and recognising teeth in the dataset. YOLOv8 performed better than Mask R-CNN, having a mAP of 70.5 [20] and other supervised and unsupervised methods[6] and provided more accurate results. This demonstrates that a single-stage object detection algorithm can outperform a two-stage object detection method like Mask R-CNN. Due to the intricate intersection of the box region with other teeth in the image, the teeth belonging to the incisors have the lowest mAP score of all the tooth classes, approximately 60.00. YOLOv8 was able to localize the molar teeth precisely with a mAP of nearly 80.00 even though they have quite complex shapes, which shows the effectiveness of YOLOv8.

The confusion matrix in Fig.9 for YOLOv8 predictions for a confidence threshold of 0.5 and IoU threshold of 0.5, where the last row shows the number of instances of tooth being unclassified indicates that YOLOv8 was able to accurately recognize the instances of every tooth with very few misclassifications but was unable to recognize some instances of every tooth above the confidence threshold; this indicates lack of knowledge of instances of the tooth in some images of the dataset.

For evaluation of BB-UNet and U-Net, two test datasets were prepared from the original dataset; **Test Dataset 1**: contains 85 images belonging to all categories, **Test Dataset 2**: contains 72 images belonging to all categories except category-5 and category-6; i.e., images containing dental implants and images having more than 32 teeth. Fig.10a shows that, when evaluated on Test Dataset 1 for all tooth kinds, BB-UNet performed better than U-Net and consistently achieved a dice coefficient of nearly 0.84 for all kinds of teeth, with U-Net having an overall dice coefficient of 0.70 and BB-UNet with an overall dice coefficient of 0.84. Because molars and premolars typically have complex geometries with more than two roots, U-Net had trouble precisely locating them. However, BB-UNet was able to do so because prior knowledge was incorporated into the model architecture. For all tooth kinds, the dice coefficient has increased by 10% to 15% when compared to U-Net, indicating the significance of prior knowledge in the instance segmentation of teeth. The primary limitation of BB-UNet was its incapacity to identify teeth in images with dental implants and images with more than 32 teeth because of the intricacy of the images and the presence of other dental instruments degraded the quality of prior knowledge and BB-UNet's performance. Fig.10b compares BB-UNet's performance on Test Dataset 1 and Test Dataset 2, having an overall dice coefficient of 0.84 and 0.86, respectively. This indicates that BB-UNet's performance was poor in locating teeth in category-5 and category-6 images, resulting in scattered segmentation masks.

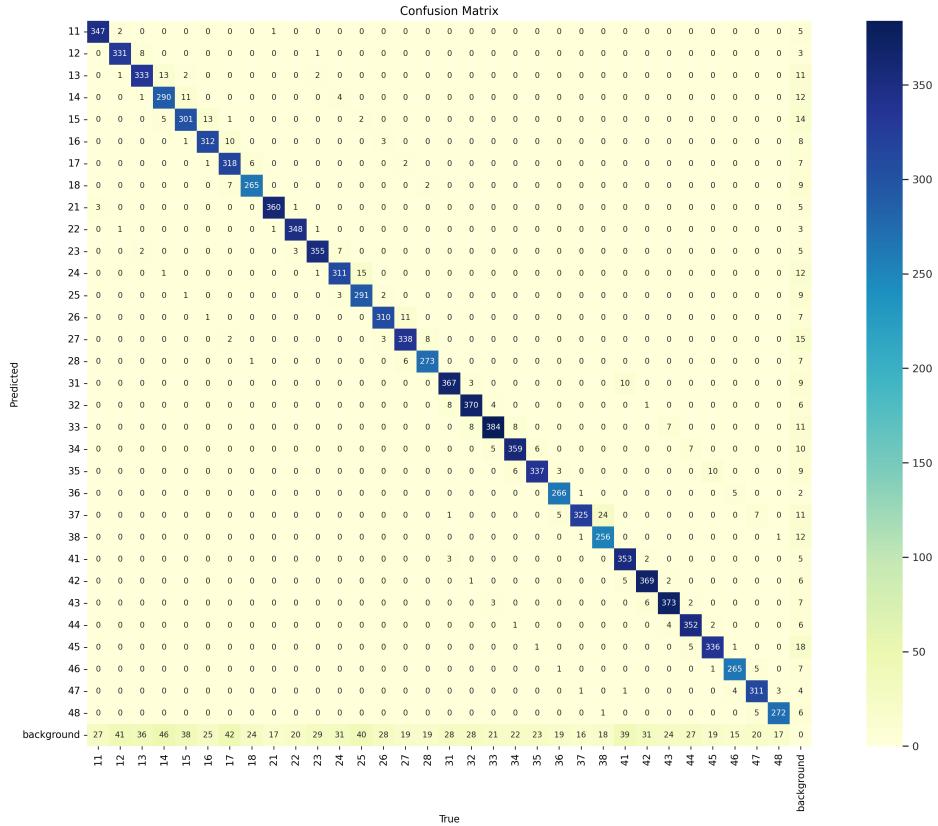


Figure 9: Confusion matrix for YOLOv8 predictions

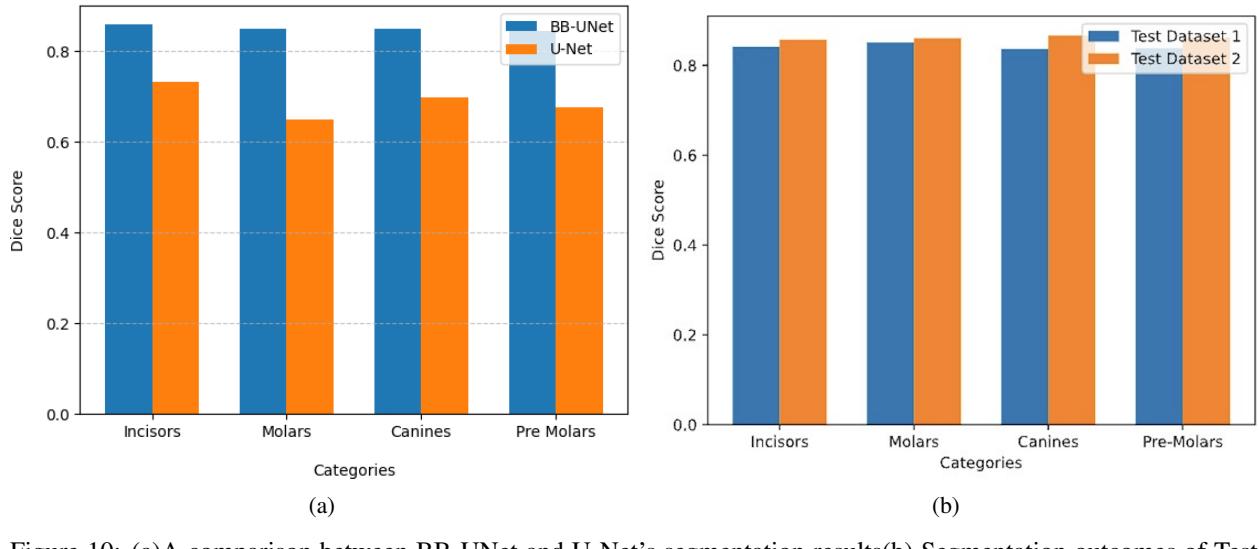


Figure 10: (a) A comparison between BB-UNet and U-Net's segmentation results (b) Segmentation outcomes of Test dataset 1 and Test dataset 2 are compared

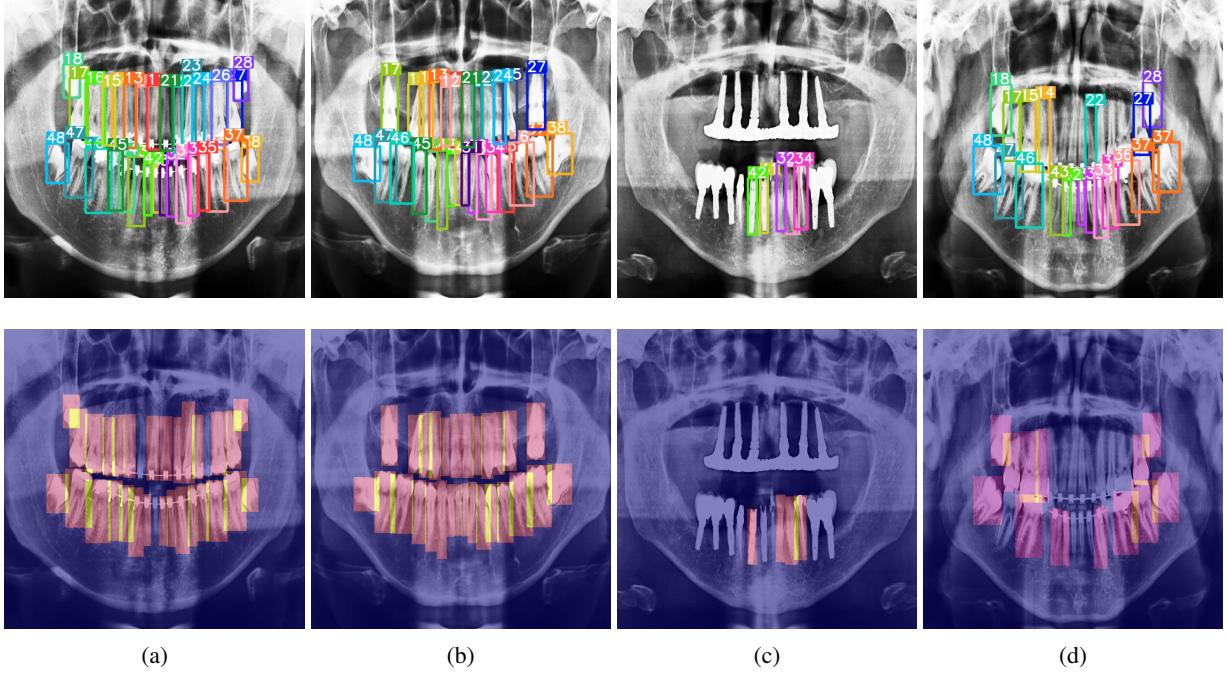


Figure 11: Classification results and corresponding heatmaps of YOLOv8

3.2 Qualitative Analysis

The YOLOv8 model's performance analysis demonstrates accurate predictions for images lacking fillings and dental implants. However, challenges arise when encountering additional teeth, fillings, or implants. Notably, the scarcity of annotated data encompassing teeth with fillings or implants within the dataset contributes to this issue. Within Figure 11a, Figure 11b, the model presents precise predictions; conversely, Figure 11c, Figure 11d exhibits missing tooth predictions, notably in instances of fillings, implants, or degraded teeth, posing difficulty for accurate detection.

Comparatively, the BB-UNet segmentation method notably outperforms the standard U-Net segmentation regarding quality. Figure 12 highlights the superior pixel-level classification of teeth achieved by BB-UNet, especially in complex tooth structures like molars and premolars, where U-Net encounters challenges. Nonetheless, in scenarios where bounding box predictions are absent, as illustrated in Figure 13, BB-UNet's performance aligns with U-Net due to the absence of prior knowledge.

The disparity in performance between models underscores the impact of dataset diversity on model proficiency, particularly in scenarios involving intricate tooth conditions like fillings, implants, or varying tooth degradation levels. Moreover, the dependence of BB-UNet's performance on bounding box predictions emphasizes the importance of comprehensive data annotation for robust segmentation outcomes.

4 Discussion

Table 2 compares the obtained results of the YOLOv8 architecture with those of other popular methods on classification of dental X-rays. YOLOv8 outperformed Mask R-CNN, HTC, and ResNetSt proposed by Silva et al.[20] but has a lower AP50 score than these model architectures. Mask R-CNN + FCN and Mask R-CNN + pointRend proposed by Pinheiro et al.[21] have better performance than YOLOv8 concerning mAP score but have less AP50 score compared to YOLOv8. PANet is the only model outperforming YOLOv8 in both mAP and AP50 scores. Most of these mentioned methods have utilized transfer learning for initializing the weights of the model architecture and have been evaluated on a smaller test dataset that does not contain complex panoramic X-rays belonging to category-5 and category-6, compared to YOLOv8 which was evaluated on an extensive dataset, including many complexities and was able to achieve superior performance for classification of teeth.

We also provide an evaluation of the BB-UNet + YOLOv8 relative to other popular methods in Table 3. BB-UNet + YOLOv8 significantly outperforms U-Net in terms of dice coefficient. Mask R-CNN and U-Net + Mask R-CNN

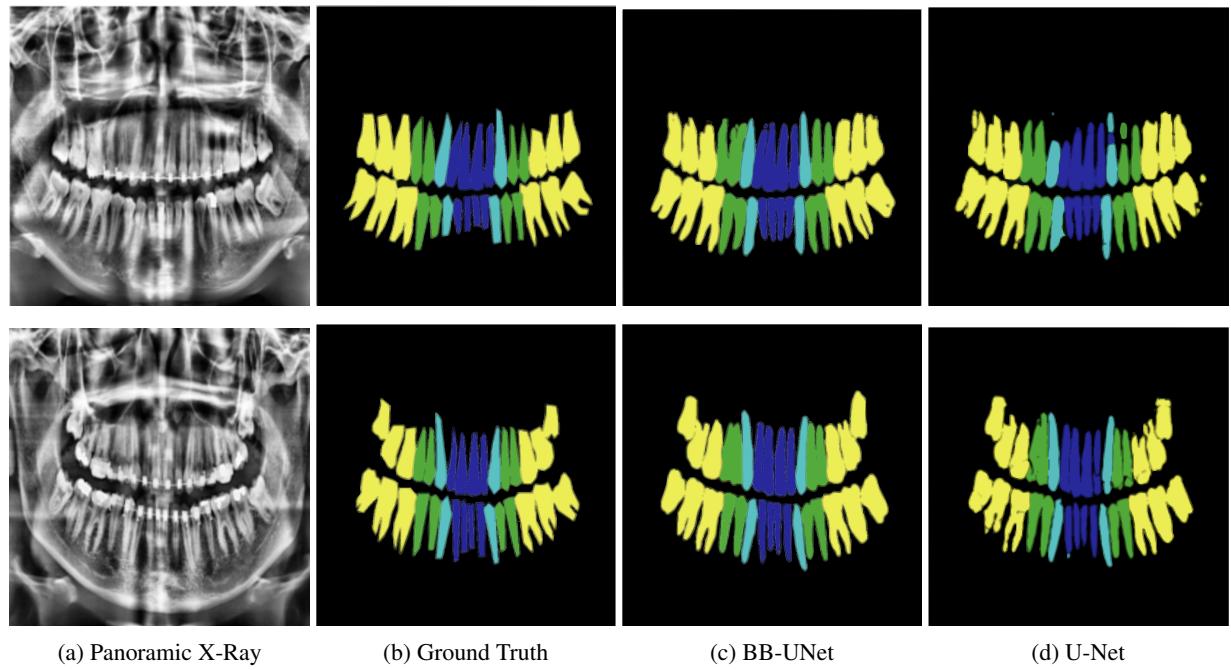


Figure 12: Superior Segmentation results of BB-UNet over U-Net

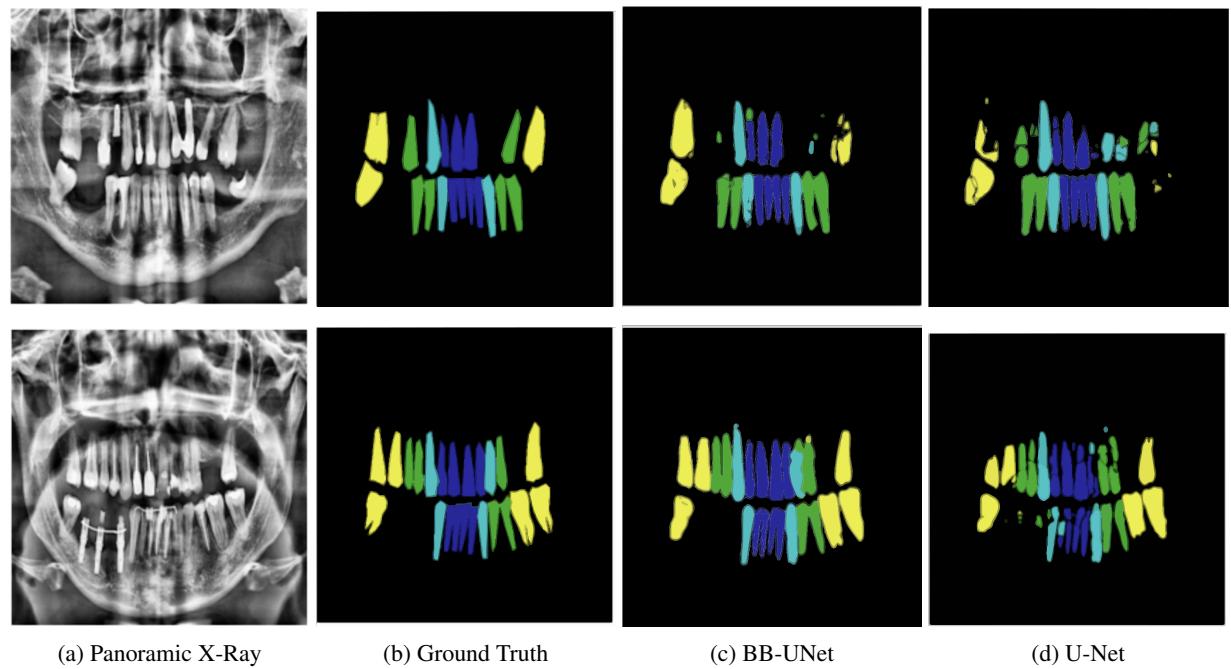


Figure 13: Poor segmentation results of BB-UNet due to missing prior knowledge

| Model Architecture | mAP | AP50 |
|----------------------------|------|------|
| Mask R-CNN[20] | 70.5 | 97.2 |
| Mask R-CNN + FCN[21] | 74.1 | 92.8 |
| Mask R-CNN + pointRend[21] | 75.3 | 94.4 |
| PANet[20] | 74.0 | 99.7 |
| HTC[20] | 71.1 | 97.3 |
| ResNeSt[20] | 72.1 | 96.8 |
| YOLOv8 | 72.9 | 94.6 |

Table 2: Analysis of Teeth Classification Results Using Different Models

| Model Architecture | Dice Coefficient(%) | | | |
|------------------------------------|---------------------|---------|-----------|--------|
| | Incisors | Canines | Premolars | Molars |
| U-Net | 73.29 | 69.92 | 67.62 | 64.98 |
| Mask R-CNN [33] | 89.56 | 89.45 | 88.70 | 87.55 |
| U-Net + Mask R-CNN [33] | 91.55 | 91.00 | 90.00 | 88.58 |
| BB-UNet + YOLOv8 (Test Dataset 1) | 85.81 | 84.91 | 84.89 | 84.40 |
| BB-UNet + YOLOv8 (Test Dataset 2) | 85.71 | 86.64 | 86.22 | 86.03 |

Table 3: Analysis of Teeth Instance Segmentation Results Using Different Models

proposed by R. Nader et al.[33] have a better dice coefficient than BB-UNet + YOLOv8. However, these architectures have utilized transfer learning to initialize weights for Mask R-CNN architecture. Our model achieved comparable performance with other model architectures even with a more challenging dataset. The performance of the BB-UNet majorly depends on the prior knowledge provided by the YOLOv8 model and can be improved by including more images with dental implants and images having more than 32 teeth in the dataset. Bharati et al.[34] demonstrated that YOLO infers more quickly than Mask R-CNN, which makes it appropriate for dentistry applications, BB-UNet + YOLOv8 was able to obtain comparable performance with other model designs at a higher speed of inference.

5 Conclusion

Our study revealed promising advancements in segmentation performance by incorporating the YOLOv8 output onto the U-Net model’s skip connection. However, despite these strides, several areas for enhancement were identified. The YOLOv8 model encountered challenges in accurately predicting numerous labels belonging to the images of categories 5 and 6 within the dataset adversely affecting overall model performance. Addressing these imbalances by expanding the dataset size rather than augmentation strategies could bolster performance. Our conclusion suggests that enlarging the dataset and refining the object detection model hold the potential for achieving superior segmentation and classification outcomes in future research work.

6 Dataset Availability

The dataset and the source code for the pipelines are available at https://drive.digiratory.ru/d/s/wEiroe9cokkVRFv1JW07S5rhyhA2NVZu/1WzaszgSZ1t--n8kdVdugbjhTUGyuqA_-NLGvkOD8Ao or on request from the authors.

References

- [1] Titus KL Schleyer, Thankam P Thyvalikakath, Heiko Spallek, Miguel H Torres-Urquidy, Pedro Hernandez, and Jeannie Yuhaniak. Clinical computing in general dentistry. *Journal of the American Medical Informatics Association*, 13(3):344–352, 2006.
- [2] Lubaina T Arsiwala-Scheppach, Akhilanand Chaurasia, Anne Müller, Joachim Krois, and Falk Schwendicke. Machine learning in dentistry: a scoping review. *Journal of Clinical Medicine*, 12(3):937, 2023.
- [3] Nuha Junaid, Niha Khan, Naseer Ahmed, Maria Shakoor Abbasi, Gotam Das, Afsheen Maqsood, Abdul Razzaq Ahmed, Anand Marya, Mohammad Khursheed Alam, and Artak Heboyan. Development, application, and performance of artificial intelligence in cephalometric landmark identification and diagnosis: a systematic review. In *Healthcare*, volume 10, page 2454. MDPI, 2022.

- [4] Eun-Hye Kim, Seunghoon Kim, Hyun-Joo Kim, Hyoung-oh Jeong, Jaewoong Lee, Jinho Jang, Ji-Young Joo, Yerang Shin, Jihoon Kang, Ae Kyung Park, et al. Prediction of chronic periodontitis severity using machine learning models based on salivary bacterial copy number. *Frontiers in Cellular and Infection Microbiology*, 10:571515, 2020.
- [5] FDI World Dental Federation. Fdi notation. <http://www.fdiworlddental.org/>. Accessed: 2023-09-12.
- [6] Gil Jader, Luciano Oliveira, and Matheus Melo Pithon. Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives. *ArXiv*, abs/1802.03086, 2018.
- [7] Aqsa Ajaz and D. Kathirvelu. Dental biometrics: Computer aided human identification system using the dental panoramic radiographs. In *2013 International Conference on Communication and Signal Processing*, pages 717–721, 2013.
- [8] Muhamad Rizal Mohamed razali, Nazatul Sabariah Ahmad, Zulkifly Mohd Zaki, and Waidah Ismail. Region of adaptive threshold segmentation between mean, median and otsu threshold for dental age assessment. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 353–356, 2014.
- [9] Rarasmaya Indraswari, Agus Zainal Arifin, Dini Adni Navastara, and Naser Jawas. Teeth segmentation on dental panoramic radiographs using decimation-free directional filter bank thresholding and multistage adaptive thresholding. In *2015 International Conference on Information & Communication Technology and Systems (ICTS)*, pages 49–54, 2015.
- [10] Guilherme Monteiro Tosoni, F. Walker, Nedra Joyner, John V. Tsimikas, and Alan G Lurie. Recursive hierarchical segmentation analysis (rhseg) of bone mineral changes on digital panoramic images. *Oral Surgery Oral Medicine Oral Pathology Oral Radiology and Endodontology*, 101, 2006.
- [11] Mutasem Khalil Sari Alsmadi. A hybrid fuzzy c-means and neutrosophic for jaw lesions segmentation. *Ain Shams Engineering Journal*, 2016.
- [12] Muhamad Rizal Mohamed Razali, Nazatul Sabariah Ahmad, Rozita Hassan, Zulkifly Mohd Zaki, and Waidah Ismail. Sobel and canny edges segmentations for the dental age assessment. In *2014 International Conference on Computer Assisted System in Health*, pages 62–66, 2014.
- [13] Md Mosaddik Hasan, Waidah Ismail, Rozita Hassan, and Atsuo Yoshitaka. Automatic segmentation of jaw from panoramic dental x-ray images using gvf snakes. In *2016 World Automation Congress (WAC)*, pages 1–6, 2016.
- [14] Lucie Grajciarová, Magdalena Kasparová, Soroush Kakawand, Aleš Procházka, and Tatjana Dostalová. Study of edge detection task in dental panoramic x-ray images. *Dento maxillo facial radiology*, 42, 05 2013.
- [15] Thorbjørn Louring Koch, Mathias Perslev, Christian Igel, and Sami Sebastian Brandt. Accurate segmentation of dental panoramic radiographs with u-nets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 15–19, 2019.
- [16] Yue Zhao, Pengcheng Li, Chenqiang Gao, Yang Liu, Qiaoyi Chen, Feng Yang, and Deyu Meng. Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network. *Knowl. Based Syst.*, 206:106338, 2020.
- [17] Qiaoyi Chen, Yue Zhao, Yang Liu, Yongqing Sun, Chongshi Yang, Pengcheng Li, Lingming Zhang, and Chenqiang Gao. Mslpnet: multi-scale location perception network for dental panoramic x-ray image segmentation. *Neural Computing and Applications*, 33:10277 – 10291, 2021.
- [18] Gil Jader, Jefferson Fontineli, Marco Ruiz, Kalyf Abdalla, Matheus Pithon, and Luciano Oliveira. Deep instance segmentation of teeth in panoramic x-ray images. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 400–407, 2018.
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [20] Bernardo Silva, Laís Pinheiro, Luciano Oliveira, and Matheus 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)*, pages 164–171, 2020.
- [21] Laís. Pinheiro, Bernardo Silva, Brenda Sobrinho, Fernanda Lima, Patrícia Cury, and Luciano Oliveira. Numbering permanent and deciduous teeth via deep instance segmentation in panoramic x-rays. In Letícia Rittner, Eduardo Romero Castro, Natasha Lepore, Jorge Brieva, Marius G. Linguraru, and Adam Walker, editors, *17th International Symposium on Medical Information Processing and Analysis*, volume 12088 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 120880C, December 2021.
- [22] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. *CoRR*, abs/1912.08193, 2019.

- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pages 234–241. Springer, 2015.
- [25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [26] Clement Zotti, Zhiming Luo, Alain Lalande, and Pierre-Marc Jodoin. Convolutional neural network with shape prior applied to cardiac mri segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1119–1128, 2019.
- [27] Rosana El Jurdì, Caroline Petitjean, Paul Honeine, and Fahed Abdallah. Bb-unet: U-net with bounding box prior. *IEEE Journal of Selected Topics in Signal Processing*, 14:1189–1198, 2020.
- [28] B. Dwyer, J. Nelson, and J. Solawetz. Roboflow (version 1.0) [software]. <https://roboflow.com/>. Accessed: 2023-20-05.
- [29] David Dang, Mu Le, Thomas Irmer, Oguzhan Angay, Bernhard Fichtl, and Bernhard Schwarz. (2021).apeer: an interactive cloud platform for microscopists to easily deploy deep learning. <https://www.apeer.com/home/>. Accessed: 2023-12-06.
- [30] Ultralytics. Rangeking yolov8 architecture. <https://github.com/ultralytics/ultralytics/issues/189>. Accessed: 2023-10-14.
- [31] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Matthias Heinrich, Wenjia Bai, Jose Caballero, Stuart A. Cook, Antonio de Marvao, Timothy Dawes, Declan P. O'Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. *IEEE Transactions on Medical Imaging*, 37(2):384–395, 2018.
- [32] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.
- [33] Rafic Nader, Andrey Smorodin, Natalia De La Fournière, Yves Amouriq, and Florent Autrusseau. Automatic teeth segmentation on panoramic x-rays using deep neural networks. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4299–4305, 2022.
- [34] Puja Bharati and Ankita Pramanik. Deep learning techniques—r-cnn to mask r-cnn: A survey. *Computational Intelligence in Pattern Recognition*, 2019.