Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Improving cervical maturation degree classification accuracy using a multi-stage deep learning approach

**Parisa Motie**
Medical Image and Signal Processing Research Center, Medical University of Isfahan
**Hossein Mohammad-Rahimi**
Topic Group Dental Diagnostics and Digital Dentistry, ITU/WHO Focus Group AI on Health
**Sahel Hassanzadeh-Samani**
Topic Group Dental Diagnostics and Digital Dentistry, ITU/WHO Focus Group AI on Health
**Negar Razzaghi**
Craniofacial Research Center, Tehran Islamic Azad University of Medical Sciences
**Mohammad Behnaz**
Department of Orthodontics, School of Dentistry, Shahid Beheshti University of Medical Sciences
**Shahriar Shahab**
Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Shahed University
**Saeed-Reza Motamadian**
drmotamedian@gmail.com

Dentofacial Deformities Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Science

**Article**

**Keywords:** Growth evaluation, Cervical vertebrae maturation, Classification, Artificial intelligence, deep learning

**Posted Date:** March 7th, 2024

**DOI:** https://doi.org/10.21203/rs.3.rs-3969639/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

Classifying the cervical vertebral maturation (CVM) degree is helpful in determining the peak period of growth and predicting the growth rate and pattern. The current study proposed a multistage framework for automated CVM classification.The dataset consisted of 2325 lateral cephalograms. Two orthodontists independently classified the images into six classes. One object detection (Faster RCNN) and two classification models (ResNet 101) were designed using the Python programming language and PyTorch library. The First classification model classified images into two main groups (i.e., C1-C3 and C4-C6) based on the C4 vertebrae shape. The second one classified each group into its subcategories. Each classification model was trained and tested using a 10-fold cross-validation strategy. The general framework reached an accuracy of 82.96%. The object detection of ROI extraction reached the mAP50 and mAP75 of 100%. The first classifier model had an accuracy of 99.10% on the hold out test set. The classifier of C1-C3 images had higher accuracy than the C4-C6 classification model (86.49% versus 82.80%) The accuracy of this fully automated framework was promising. Considering the gradual changes in cervical vertebrae morphology the use of visualized data by gradient-weighted class activation maps (Grad-CAM) is suggested to improve the model's performance.


# Introduction

Artificial intelligence (AI) refers to the practice of simulating human behavior in machines. Rapid advancements in medical imaging technologies are generating a large amount of visual data. AI healthcare applications have risen exponentially as AI-based tools provide powerful hands in collecting and evaluating this massive amount of data. Several research have documented the successful use of AI models in detistry[1].

Biological age assessment according to dental and skeletal maturation and psychosomatic development provides more accurate information than chronological age in determining the peak period of growth and predicting the growth rate and pattern[2]. The hand and wrist bones were initially used to evaluate the skeletal age[3]. However, Lamparski et al.[4] further suggested that cervical vertebral maturation (CVM) assessment eliminates the additional patient radiation. Traditionally, classification has been performed visually, which is subjective, expert-dependent, time-consuming, and often prone to inter-observer variability[5,6]. In addition, since vertebrae maturation is a continuous process, differentiating the stages' borderline is difficult for human observers[7].

The recent advancements in the use of machine learning techniques have shown promising results in the automated classification of cervical maturation[6,8]. However, the task is still challenging due to the complex and variable morphology of the cervical vertebrae. Classical machine learning-based models were designed based on manual feature extraction, and manually labeled datasets, as the input features[9–11]. Further models could learn the features automatically; however, the cropping and segmentation of the regions of interest (ROIs) had to be performed manually during the preprocessing.

Although these methods had a relatively high level of accuracy[7], they necessitated significant human involvement and expert knowledge, and were extremely time-consuming. Another approach is the employment of a full end-to-end framework. However, the only study that developed an end-to-end approach for CVM six-class classification using a three-step model had a relatively low accuracy (62.5%)[6]. In another study[12] the U-Net and Alex-Net architectures were designed for automated segmentation of cervical vertebrae and to classify the CVM stages into three categories of pre-pubertal (CVM1 and CVM2), Pubertal (CVM3 and CVM4), and post-pubertal (CVM5 and CVM6), respectively. Despite the general framework's high accuracy (91%, 80%, and 87%, respectively), the original six stages of maturation identified by Baccetti[13] were not classified in their study.

To address these challenges, we proposed a multi-stage deep learning framework for the classification of cervical maturation degrees. To the best of our knowledge, this is the first study to design a multi-stage deep learning framework for CVM classification that takes into account the shape of C4 and separates the first three classes from the later ones.

## Materials and Methods

# Study design

The current study was reported following the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [14]. This is the improved version of our pilot study[8]. A three-stage deep-learning based model was designed for end-to-end classification of the stages of CVM degree using the lateral cephalograms. The first model was a Faster-RCNN object detection model to automatically retrieve the regions of interest (ROIs). The second and third models were ResNet50s for two-stage classification of ROIs, one of which classified the images based on the shape of C4 into C1-C3 and C4-C6. Another classification model was trained to classify each group to its subcategories (Fig. 1). This study was approved by the Ethics Committee of the Shahid Beheshti University of Medical Sciences (IR.SBMU.DRC.REC.1400.007). All experiments were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants. Study participant names (and other personally identifiable information) were removed from all images.

# Data

The dataset used to develop the models, consisted of a total of 2325 lateral cephalograms from three different sources (Table 1). C2, C3 and C4 cervical vertebrae were completely visible in all included images. Non-standard (i.e., with improper patients' head position or signs of patients' movement) and low-quality images (i.e., blurry or noisy images), were excluded.

Table 1
the sources of the dataset used for developing the models

| Class | Source | | | |
|---|---|---|---|---|
| | University | Private dental clinic | IEEE International Symposium | Total |
| C1 | 111 | 82 | 19 | 212 |
| C2 | 45 | 56 | 20 | 121 |
| C3 | 27 | 35 | 12 | 74 |
| C4 | 210 | 371 | 152 | 733 |
| C5 | 149 | 713 | 167 | 1029 |
| C6 | 35 | 99 | 22 | 156 |
| Total | 577 | 1356 | 392 | 2325 |

A Promax Dimax 3 Digital Pan/Ceph device (Planmeca, Helsinki, Finland) was used for image acquisition for first two dataset. All mages were anonymized and converted to jpg format with the size of 1088 × 896.

# Ground truth

The cervical vertebrae were classified into six stages from C1-C6 following the guide proposed by McNamara and Franchi [15]. To avoid any possible source of bias, the classifications were performed on pre-processed images. Two orthodontists with six and ten years of experience classified the images, independently. The orthodontists were calibrated through in-person meetings after classifying 50 random images. Controversies were resolved by consensus. If no consensus was reached, the image was removed from the training and test datasets. They also manually annotated the ROIs containing C2, C3 and C4 using the LableMe software (http://labelme.csail.mit.edu) by bounding boxes. The posterior border of C5 vertebrae and the Throat soft tissue border were the posterior and anterior borders of ROIs. The interface of the C2 spinous process and Dens body was considered the superior border, while the most anterior point on the upper mesial corner of C5 vertebrae was considered the inferior border of ROIs. The final bounding boxes enclosed all parts of C2, C3, and C4 (Fig. 2). Each orthodontist confirmed the drawn ROIs by the other one.

# Data Partitions

For the object detection model, 571 images were randomly chosen and divided into the training (421 images), test (75 images), and validation (75 images) datasets. The classification models were trained and tested on a total of 2325 images. Initially, 10% data separated for hold-out test using stratified sampling based on data source and class. Each classification model was trained and tested using a 10-fold cross validation strategy.

# Models

One object detector model and two classifiers were trained in this study.

## • Object detection

The Faster RCNN model was designed to automatically extract the ROIs. the C2-C4 region was detected from the other parts of the images using bounding boxes so that the following classifier models can focus on the proposed ROI. The model was developed using the Python programming language through detecteron library, version 2. an adaptive moment estimation (ADAM) optimizer was implemented to update the network weights. Sum of the classification loss over 2 classes (if there was object or not) and the regression loss of bounding boxes were used as loss function while training the model.

## • Classification

Since, ResNet101 showed the best performance in our pilot study, it was used for two stage classification. The first stage classifier, classified the images based on the lower half part of the images (C4 vertebrae shape) into two main groups (i.e., C1-C3 and C4-C6). At second stage classifications, the growth stage of each image was detected in each group. We also trained a single-stage ResNet101 as a baseline for comparison [8].

The Python programming language and PyTorch library were used to develop the models. Extracted ROIs from the object detection model were used as input for these classifiers. Zero padding was used to convert ROIs to squares. Then images resized to 224 * 224 pixels. ADAM optimizer was implemented. binary cross-entropy was used as the loss function.

# Training

The training was conducted on a Tesla K80 graphic processor unit (Nvidia Corporation, Santa Clara, CA, USA) using the Google Collaboratory platform. All the models were trained for 50 epochs. On the basis of the best model weights, early stopping was used to prevent overfitting. All hyperparameters of the implemented approaches were set in their original implementation.

# Evaluation

The evaluation technique relied on recognizing object-level errors. This was performed by matching target detected objects with anticipated objects given by participants and then computing true positives and false positives. The object detection model was evaluated with mean absolute precision (mAP), at a minimum intersection-over-union (IOU) threshold 0.5 and 0.75 to identify correctly detected objects. Any other predicted object detection that was less than the threshold was regarded as an error.

$$\text{mAP} = \frac{1}{n}\sum_{1}^{n}\frac{|TP|}{|TP|+|FP|}$$

Where TP, TN, FP, FN, were the number of true positive, true-negative, false-positive, and false-negative samples, respectively. A confusion matrix was developed using all TP, FP, TN, and FN.

The general model's performance evaluated using the following criteria:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\#All\ Samples}$$

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

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

$$\text{F1–score} = = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

# Results

The object detection of ROI extraction reached the mAP50 and mAP75 of 100%. The first classifier model reached an accuracy of 99.10 on the hold out test set. The classifier of C1-C3 images had higher accuracy than the C4-C6 classification model (86.49 versus 82.80) (Table 2).

Table 2
total accuracy of the classification models

| Models | 10-fold cross validation | Hold out test set |
|---|---|---|
| **Model II (2-Class classification model)** | 98.66 ± 1.53 | 99.10 |
| **Model III-a (3-Class classification of CS 1, 2 and 3)** | 82.70 ± 3.59 | 86.49 |
| **Model III-b (3-Class classification of CS 4, 5 and 6)** | 84.68 ± 3.57 | 82.80 |

The study compared the performance of a baseline one-stage classification model with the proposed framework for cervical maturation degree classification. The baseline model achieved an accuracy of 72.20 on the test set. The precision values for the different classes ranged from 0.428 to 0.85, while the recall values ranged from 0.36 to 1.00. The F1-scores for the different classes ranged from 0.444 to 0.85, with an average F1-score of 0.647. The results show that the baseline model performed moderately in

terms of accuracy and precision, but with lower performance in recall and F1-scores for some classes (Table 3). Figure 3 shows the confusion matrices of the one-stage classification model.

Table 3
Performance measurements of baseline one-stage classification model

| Class | Measurement criteria | | | |
|---|---|---|---|---|
| | Precision | Recall | F1-score | Total accuracy |
| **C1** | 0.85 | 0.85 | 0.85 | 72.20 |
| **C2** | 0.80 | 0.36 | 0.50 | |
| **C3** | 0.43 | 1.00 | 0.60 | |
| **C4** | 0.67 | 0.89 | 0.76 | |
| **C5** | 0.84 | 0.64 | 0.72 | |
| **C6** | 0.50 | 0.40 | 0.44 | |
| **Average** | 0.68 | 0.69 | 0.64 | |

The proposed framework achieved a higher accuracy of 82.96 on the test set, outperforming the baseline model. The precision values for the different classes ranged from 0.6 to 0.95, while the recall values ranged from 0.6 to 0.91. The F1-scores for the different classes ranged from 0.6 to 0.92, with an average F1-score of 0.790. The proposed framework showed an improvement in all metrics, especially in precision for CS1, which improved from 0.85 to 0.95. The results suggest that the proposed framework can achieve higher accuracy and precision in cervical maturation degree classification than the baseline model (Table 4). The confusion matrices of the general framework is presented in Fig. 4.

Table 4
Performance measurements of the general framework (two-stage object detection-classification model)

| Class | Measurement criteria | | | |
|---|---|---|---|---|
| | Precision | Recall | F1-score | Total accuracy |
| **C1** | 0.94 | 0.90 | 0.92 | 82.96 |
| **C2** | 0.83 | 0.90 | 0.86 | |
| **C3** | 0.67 | 0.67 | 0.67 | |
| **C4** | 0.84 | 0.83 | 0.84 | |
| **C5** | 0.83 | 0.84 | 0.84 | |
| **C6** | 0.60 | 0.60 | 0.60 | |
| **Average** | 0.79 | 0.79 | 0.79 | |

# Discussion

Timing is an important issue that can determine whether or not growth modification treatments are successful[12]. Calculating growth spurts based on the skeletal maturity evaluation is a valid method that is independent of individual variances[16]. In the current study, a new two-stage classification model using the C4 vertebrae shape in the classification process was developed and reached a relatively high accuracy for an end-to-end CVM classification.

It was assumed that by dividing the functions between different models and reducing the features which need to be learn, the model's complexity and thus the training time of the model on larger amounts of data would be less[7]. In comparison to our previous pilot study, higher number of lateral cephalometric radiographs were used in the current study, and the problem of lower data number in some classes was resolved by the use of few-shot technique. In addition, end-to-end framework was designed rather than the use of cropped images for classification. Another strength point of the current study was implementing a public dataset (ISBI2015 dataset) so that state-of-the-art comparison of designed model with previous in-paper models would be possible. It was tried to bypass the lack of generalizability problem by the use of multi-center data source. A previously end-to-end three-step model was developed received resized images as input, cropped the regions of interest, segmented the cervical vertebrae and produced binary images comprising the background (0) and foreground (1). These binary images were fed into the classification model for the final CVM staging. The three-step model reached an accuracy of 62.5% which was a higher accuracy than the one-step classifier (53%) and the two-step ROI detector-classifier (59%) models[6]. In the current work, we have shown that the development of a two-stage classifier combined with an ROI detector model improved classification accuracy by 82.96%. The designed model also outperformed the best model of our previous study regarding accuracy (61.62%)[8].

One noteworthy finding of the present study was that most incorrectly classified stages differed by one stage from the actual stages (i.e., the images of $C_n$ were incorrectly categorizing into a single preceding ($C_{n-1}$) or next stage ($C_{n+1}$)) (Fig. 3). In accordance with this finding, Kim et al.,[6] estimated that, when the one-stage errors were considered correct, the models' accuracy increased about 30%. This is due to the fact that in many cases, cervical vertebrae maturity encompasses a period between the previous and next stage of development. this might explain the reason why more accurate classification results were obtained by Radwan et al.,[12] by merging each two subsequent stages into a single class. Similarly, our previously developed three-class classification model, which categorized the CVM stages into three classes of pre-pubertal (C1, C2), growth spurt (C3, C4), and post-pubertal (C5, C6), achieved a higher accuracy than our six-class CVM diagnosis model (82.83% V.S. 61.62%)[8].

Regarding class-based accuracy, the highest F1-score value was in C1 (92%), while C6 (60%), C3 (67%) had the lowest F1-score values, respectively. The F1-score criterion assesses model accuracy by combining the model's precision and recall scores. It is especially useful when there is an uneven distribution of false positives and false negatives. In accordance with the results of the current study, Radwan et al.,[12] found the lowest F1-score (57%) in C3, and C4 stages. Also, Zhou et al.,[11] proved that C3 had the lowest F1-score (31%). Seo et al., reported that in comparison to other stages, the C3 had the lowest AUC value[7]. According to another study, a deep learning-based architecture (SciKit-Learn and Keras with tensorflow backend) showed the lowest accuracy when classifying C3, and C4 [17]. Kim et al., stated that the one stage error highly occurred in C3. In addition high levels of two stage and three stage errors led to incorrect classification of C3 and C4 [6]. It seems that there are two main reasons for lower accuracy results in C3 classification; the first is that pubertal stages have shorter durations, making them harder to encounter in clinical practice. Therefore, a smaller portion of the training dataset consists of this stage[11]. Insufficient sample size could also be the cause of low F1-score of C6 classification in our study. As a second explanation, there could be difficulties deciding whether C3 and C4 represent squares, rectangles horizontally and vertically, or trapezoids[12]. Since, rather than C6, C3 and C4 are considered critical time-points for orthodontic interventions and present optimal growth potential for treatment of class II malocclusion[13,18], in terms of clinical performance, our model outperformed earlier models by having a higher F1-score value, indicating that it was both more sensitive and precise.

This should be noted that CVM staging, like all stage-based radiographic growth indicators, has limitations to its diagnostic reliability due to the drawbacks of variability and unpredictability of stages' durations as well as individual-based variations[19–21]. Therefore, it would better to be used as an auxillary diagnostic tool not the main one. An inherent limitation of the CVM classification studies was the difficulty in ground truth annotation. During the evolution of the cervical vertebrae, there is a continuous rather than a gradual change in the morphology[12]. discretizing the continuous CVM level, makes the CVM staging a challenging task especially duo to low reliability and reproducibility[22]. A previous study reported intra-rater agreement ranged from 77-87.3% and inter-rater agreement was lower than 50% (mean : 42.82%)[23]. To overcome this obstacle, we removed images from the dataset that did not receive

agreement from the observers on the precise stage. The proposed framework of the current study was based on a three-stage object detection-classification model, which may increase the computational requirements and processing time. Considering continuous growth-related changes in cervical vertebrae morphology, in order to improve the accuracy of future end-to-end models, we suggest the use of visualized data by gradient-weighted class activation maps (Grad-CAM) instead of linear measurements and reference points[7]. In addition, the models' generalizability would be enhanced by utilizing data augmentation methods such as applying directional filters and grayscale jitteringfor data and patch augmentation[2].

# Conclusion

The suggested multi-stage deep learning architecture was more efficient and reliable than single stage classification models for cervical maturation degree staging. This end-to-end design provides an automated technique that assists clinical practitioners in more accurate timing of growth modification orthodontic treatments.

# Declarations

- **Authors' contributions:** P. M. contributed with data labeling process, writing the manuscript and submitting the article, H. MR helped with writing the method section and developing the algorithm, S. HS and N. R contributed with data labeling process, M. B and SH. SH helped with data collection and SR.M. provided the research idea, and guided the research procedure.
- **Availability of data and materials:** The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.
- **Competing interests:** The authors declare that they have no competing interests.
- **Funding:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

# References

1. Motie, P. *et al.* in *Emerging Technologies in Oral and Maxillofacial Surgery* 287–328 (Springer, 2023).
2. Atici, S. *et al.* Classification of the Cervical Vertebrae Maturation (CVM) stages Using the Tripod Network. *arXiv preprint arXiv:2211.08505* (2022).
3. Flores-Mir, C. *et al.* Correlation of skeletal maturation stages determined by cervical vertebrae and hand-wrist evaluations. The Angle Orthodontist 76, 1–5 (2006).
4. LAMPALSKI, D. Skeletal age assessment utilizing cervical vertebrae. *Master of Science Thesis, University of Pittsburgh* (1972).
5. Gabriel, D. B. *et al.* Cervical vertebrae maturation method: poor reproducibility. American Journal of Orthodontics and Dentofacial Orthopedics 136, 478. e471-478. e477 (2009).

6. Kim, E.-G. *et al.* Estimating Cervical Vertebral Maturation with a Lateral Cephalogram Using the Convolutional Neural Network. Journal of Clinical Medicine 10 (2021).
7. Seo, H., Hwang, J., Jeong, T. & Shin, J. Comparison of Deep Learning Models for Cervical Vertebral Maturation Stage Classification on Lateral Cephalometric Radiographs. Journal of Clinical Medicine 10, 3591 (2021).
8. Mohammad-Rahimi, H. *et al.* Deep learning for the classification of cervical maturation degree and pubertal growth spurts: A pilot study. Korean journal of orthodontics 52, 112–122 (2022).
9. Amasya, H., Cesur, E., Yıldırım, D. & Orhan, K. Validation of cervical vertebral maturation stages: Artificial intelligence vs human observer visual analysis. American Journal of Orthodontics and Dentofacial Orthopedics 158, e173-e179 (2020).
10. Kök, H., Izgi, M. S. & Acılar, A. M. Evaluation of the artificial neural network and Naive Bayes models trained with vertebra ratios for growth and development determination. Turkish Journal of Orthodontics 34, 2 (2021).
11. Zhou, J. *et al.* Development of an artificial intelligence system for the automatic evaluation of cervical vertebral maturation status. Diagnostics 11, 2200 (2021).
12. Radwan, M. T., Sin, Ç., Akkaya, N. & Vahdettin, L. Artificial intelligence-based algorithm for cervical vertebrae maturation stage assessment. Orthodontics & Craniofacial Research (2022).
13. Baccetti, T., Franchi, L. & McNamara Jr, J. A. in Seminars in Orthodontics. 119–129 (Elsevier).
14. Mongan, J., Moy, L. & Kahn Jr, C. E. Vol. 2 e200029 (Radiological Society of North America, 2020).
15. McNamara Jr, J. A. & Franchi, L. The cervical vertebral maturation method: A user's guide. The Angle Orthodontist 88, 133–143 (2018).
16. Bittencourt, M. V. *et al.* Accuracy of dental development for estimating the pubertal growth spurt in comparison to skeletal development: a systematic review and meta-analysis. Dentomaxillofacial Radiology 47, 20170362 (2018).
17. Makaremi, M., Lacaule, C. & Mohammad-Djafari, A. Determination of the cervical vertebra maturation degree from lateral radiography. *Multidisciplinary Digital Publishing Institute Proceedings* 33, 30 (2020).
18. Graber, T. M. & Vanarsdall, R. L. *Orthodontics: current principles and techniques*. (Mosby, 1994).
19. Perinetti, G., Bianchet, A., Franchi, L. & Contardo, L. Cervical vertebral maturation: An objective and transparent code staging system applied to a 6-year longitudinal investigation. American Journal of Orthodontics and Dentofacial Orthopedics 151, 898–906 (2017).
20. Perinetti, G., Contardo, L. & Primozic, J. Diagnostic accuracy of the cervical vertebral maturation method. European Journal of Orthodontics 40, 453–454 (2018).
21. Perinetti, G., Primozic, J., Sharma, B., Cioffi, I. & Contardo, L. Cervical vertebral maturation method and mandibular growth peak: a longitudinal study of diagnostic reliability. European Journal of Orthodontics 40, 666–672 (2018).

22. Nestman, T. S. *et al.* Cervical vertebrae maturation method morphologic criteria: poor reproducibility. American Journal of Orthodontics and Dentofacial Orthopedics 140, 182–188 (2011).
23. Schoretsaniti, L., Mitsea, A., Karayianni, K. & Sifakakis, I. Cervical vertebral maturation method: reproducibility and efficiency of chronological age estimation. Applied Sciences 11, 3160 (2021).

# Figures

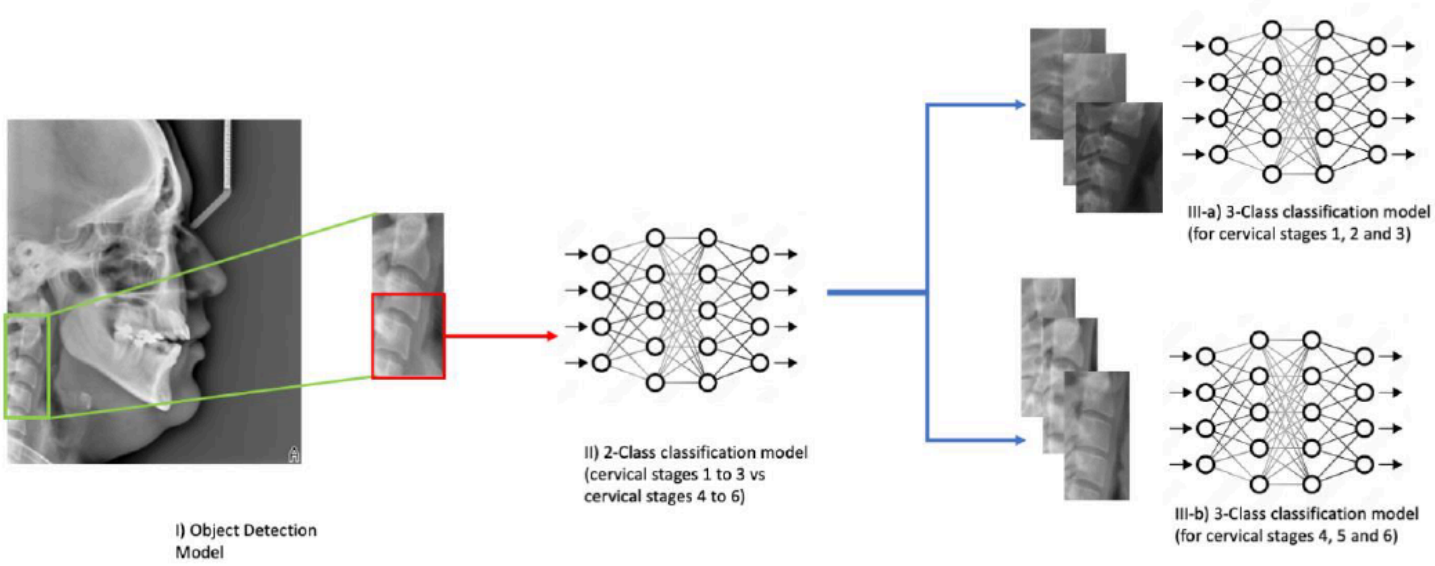
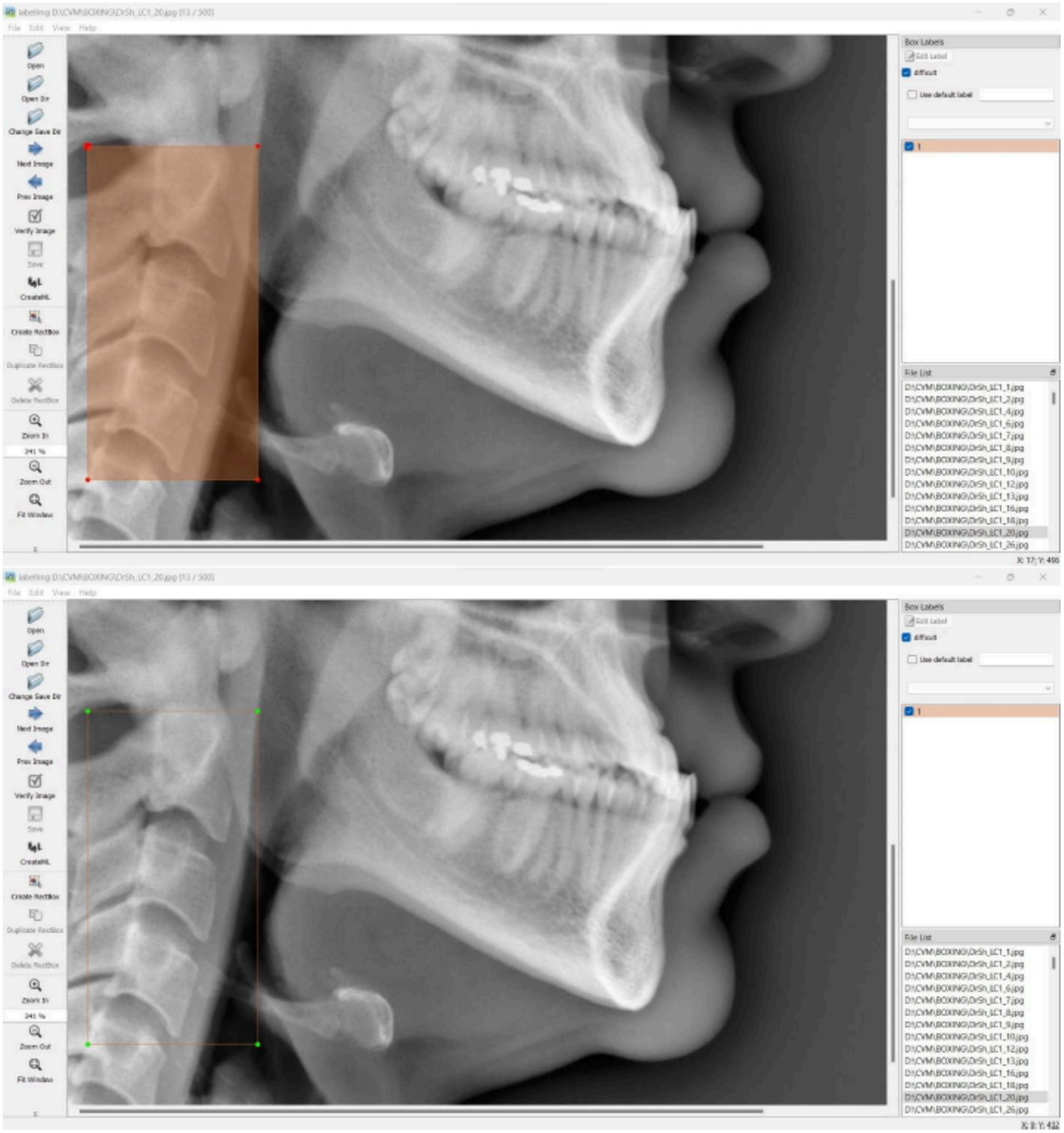


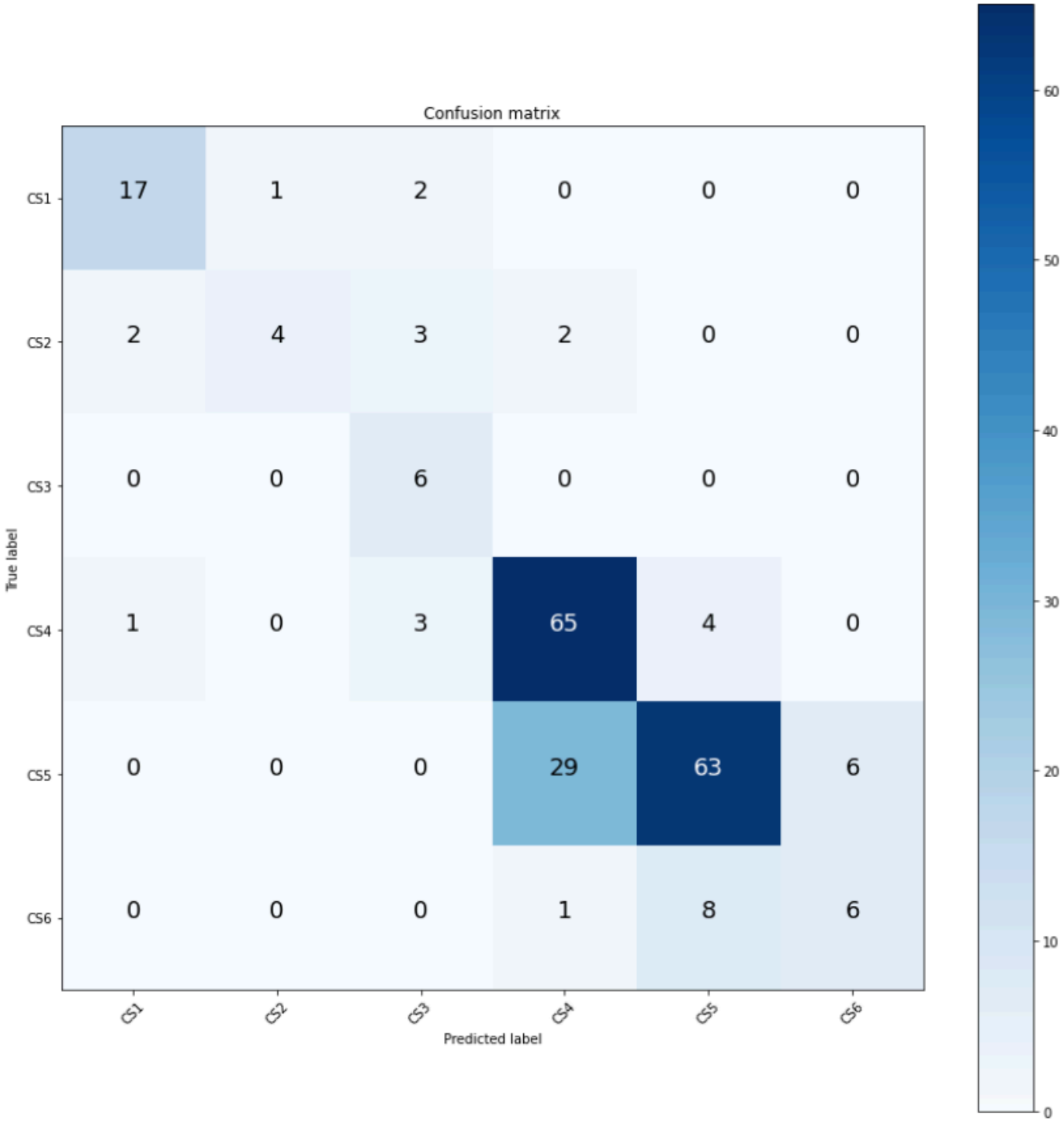

**Figure 1**

The total model structure, I: a Faster-RCNN object detection model, II, III: ResNet50s for two-stage classification of ROIs



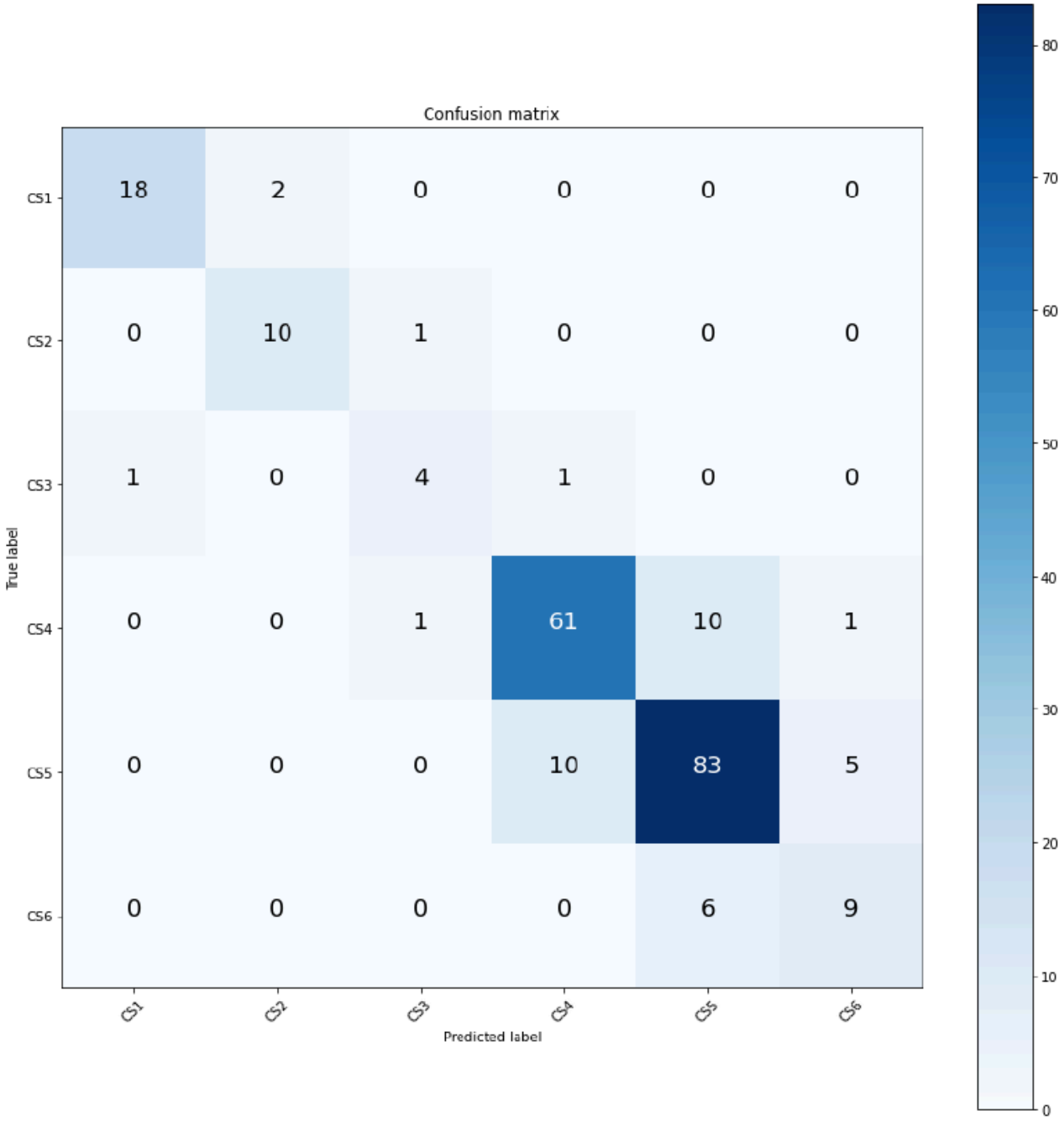

## Figure 2

The bounding boxes over the ROIs




## Figure 3

The confusion matrices of the one-stage classification model




## Figure 4

The confusion matrices of the general framework