

Orthodontic diagnosis with Artificial Intelligence

*2022 Orthodontic Faculty Development
Fellowships (OFDFA)*

Dr. Shivam Mehta

smehta@tamu.edu
O: 773-231-3050

FollowUp Form

Award Information

In an attempt to make things a little easier for the reviewer who will read this report, please consider these two questions before this is sent for review:

- Is this an example of your very best work, in that it provides sufficient explanation and justification, and is something otherwise worthy of publication? (We do publish the Final Report on our website, so this does need to be complete and polished.)*
- Does this Final Report provide the level of detail, etc. that you would expect, if you were the reviewer?*

Title of Project*

Orthodontic diagnosis with Artificial Intelligence

Award Type

Orthodontic Faculty Development Fellowship Award (OFDFA)

Period of AAOF Support

July 1, 2022 through June 30, 2023

Institution

Texas A&M University Health Science Center

Names of principal advisor(s) / mentor(s), co-investigator(s) and consultant(s)

1) Mentor: Dr. Dawei Liu; 2) Research Mentor: Dr. Madhur Upadhyay; 3) Collaborator: Yasir Suhail

Amount of Funding

\$20,000.00

Abstract

(add specific directions for each type here)

Overall Objectives: Malocclusions as a common oral health problem and are the third most prevalent oral problem after dental caries and periodontal diseases. Orthodontic treatment is perhaps the only viable solution for such problems. The patients undergoing orthodontic treatment undergo either extraction or non-

extraction of teeth for effective resolution of the malocclusion. This is a critical decision because extractions are irreversible. A wrong decision can lead to undesirable results like: suboptimal esthetics, improper bite, functional issues in mastication & speech and in the worst-case scenario; unfinished treatment. The treatment planning decision is based on the analysis of radiographs such as cephalometric radiographs and identifying the different features on the photographic images, and radiographic images dataset. Therefore, an artificial intelligence (AI) based system would be useful as an alternative set of eyes for the clinician in performing interpretation of lateral cephalograms and other feature variables from the dataset for diagnosis and treatment planning. However, literature lacks evidence regarding the lacks evidence regarding the automated diagnosis of cephalometric images with a deep learning system. Additionally, deep learning systems like Region Convolved Neural Network (RCNN) on orthodontic diagnosis are yet to be studied on heterogenous datasets.

Hypothesis: The central hypothesis of this proposal is that the evaluating heterogenous datasets with artificial intelligence deep learning based algorithms will improve the accuracy of landmark identification and diagnosis prediction that can be applied to clinical situations.

Specific Aims:

1. Create the first fully automated framework for cephalogram analysis using one of the rapidly developing deep learning techniques—Region based convolutional neural networks (RCNN).
2. Evaluate accuracy and reliability of the proposed algorithm in comparison with trained experts with different levels of experience.
3. Compare the efficiency of the algorithm in locating the desired landmarks in cephalometric images and photographic images derived from different centers across the world.

Clinical Implications & Significance: It is critical for clinicians to perform cephalometric identification, and diagnosis of orthodontic patients efficiently in an orthodontic practice. Based on the cephalometric interpretation and diagnosis of the patient, the treatment rendered for most orthodontic patients would be to do either non-extraction or extraction of teeth for effective resolution of the malocclusion. If extraction is chosen as a treatment option, a second task is to identify which teeth to be extracted. For students and inexperienced practitioners in orthodontics, orthodontic diagnosis and treatment planning can be very challenging. It takes a relatively long time for orthodontists to accumulate experience let alone students. Orthodontists with less experience often require consultation with experts Thus, AI deep learning based algorithm can help the clinicians to decrease the inconsistency in identification of landmarks and patient diagnosis. The most important criteria for application of an AI based algorithm is the variety and veracity of the labelled data. Because of the lack of literature on the AI based algorithms on heterogenous datasets, the clinical application of the AI algorithms has been challenging. The advantage of using heterogenous datasets and a large sample size is that it will be adaptable to different clinical situations and can be used by clinicians to achieve satisfactory orthodontic outcomes. The knowledge about cephalometric landmark identification and orthodontic diagnosis with deep learning AI model using RCNN will help the clinicians in understanding and decreasing the chances of misdiagnosis in their orthodontic practice. It will also provide the framework for future evaluation of AI to identify more complex volumetric datasets such as three-dimensional radiographs.

Respond to the following questions:

Detailed results and inferences:*

If the work has been published, please attach a pdf of manuscript below by clicking "Upload a file".

OR

Use the text box below to describe in detail the results of your study. The intent is to share the knowledge you have generated with the AAOF and orthodontic community specifically and other who may benefit from your

study. Table, Figures, Statistical Analysis, and interpretation of results should also be attached by clicking "Upload a file".

Final Report.pdf

Due to word limit, Please find attached.

Were the original, specific aims of the proposal realized?*

Specific aim 1 – Create the first fully automated framework for cephalogram analysis using one of the rapidly developing deep learning techniques—Region based convolutional neural networks (RCNN).

- In this aim, a collection of heterogenous datasets of 14,000 cephalograms (10,000 training data and 4000 test data) is to be analyzed. So far, we have created a fully automated framework for cephalogram analysis using Region based convolutional neural network (RCNN)

Specific aim 2 – Evaluate accuracy and reliability of the proposed algorithm in landmark identification comparison with trained experts with different levels of experience.

-This aim is achieved with the results listed above in the explanation of specific aim 2

Specific aim 3 – Compare the efficiency of the algorithm in locating the desired landmarks in cephalometric radiographs derived from different centers across the world.

- This aim is achieved with the results listed above. Additional cephalograms are being analyzed currently and the algorithm is being refined on more datasets.

Specific aim 4 – Create an automated diagnosis and treatment planning algorithm using RCNN

- In this objective, 2600 datasets with each dataset consisting of 8 photographic images, lateral cephalogram, and panoramic radiograph (2000 training data and 600 test data are to be evaluated .

Currently, 1000 such datasets have been evaluated.

- This aim is achieved partially. Additional datasets are being evaluated and the algorithm is being trained on the datasets. Additional training requires more funding support and thankfully we have received another grant from Texas SB30 funds to take this research further.

Were the results published?*

No

Have the results of this proposal been presented?*

Yes

To what extent have you used, or how do you intend to use, AAOF funding to further your career?*

The AAOF has played a pivotal role in not only this research project but also my academic career. The current project is a distillation of my previous work and is the first step in the many required to make diagnosis & treatment planning standardized, accurate & completely automated with Artificial Intelligence. It has taken a team of dedicated engineers, clinicians & software/code specialists to outline the basic pathway for this project. And the support from AAOF has been a critical source of funding for such credible orthodontic research. AAOF has help me develop my academic career and I intend to continue credible research along the lines of artificial intelligence, rapid palatal expansion, and orthodontic tooth movement and also help contribute by giving back as well as by enrolling to conduct service activities to advance the mission of AAOF.

Accounting: Were there any leftover funds?

\$384.14

Not Published

Are there plans to publish? If not, why not?*

Yes, the manuscript are under preparation and will be published soon.

Presented

Please list titles, author or co-authors of these presentation/s, year and locations:*

(i) Developing an AI Algorithm for Predicting Severity and Location Impacted Maxillary teeth. Poster Presented at AAO by An Jin et al.

(ii) Invitation for book chapter – Attached in the illustration/addendum

(iii) Mehta S, Upadhyay M, Suhail Y. Artificial Intelligence for radiographic image analysis. Seminars in Orthod. 2021;27(2):109-120 (Invited Article).

(iv) Acceptance of proposal grant for Northeastern Society of Orthodontics – adding to this research and taking it further

(v) Manuscript under preparation: Assessing agreement among orthodontists on cephalometric and clinical parameters

(vi) Received another grant from Texas SB30 funds for further continuation of the research study: 150,000 Shivam Mehta : Principal Investigator

Was AAOF support acknowledged?

If so, please describe:

Yes AAOF support was acknowledged in the presented poster. Only some results of the proposal are presented - other are currently in progress for manuscript.

Internal Review

Reviewer Comments

Reviewer Status*

File Attachment Summary

Applicant File Uploads

- Final Report.pdf

Final Report AAOF OFDFA – Shivam Mehta
Texas A&M University College of Dentistry

1. Specific Aims

List original specific aims

Specific Aim 1: Create the first fully automated framework for cephalogram analysis using one of the rapidly developing deep learning techniques—Region based convolutional neural networks (RCNN).

Specific Aim 2: Evaluate accuracy and reliability of the proposed algorithm in landmark identification comparison with trained experts with different levels of experience. Thesis

Specific Aim 3: Compare the efficiency of the algorithm in locating the desired landmarks in cephalometric radiographs derived from different centers across the world.

Specific Aim 4: Create an automated diagnosis and treatment planning algorithm using RCNN

2. Studies and Results

Summarize which studies have already been conducted and results achieved, particularly in reference to specific aims (investigated hypothesis(es) and corresponding findings).

Specific Aim 1: Create the first fully automated framework for cephalogram analysis using one of the rapidly developing deep learning techniques—Region based convolutional neural networks (RCNN).

Work done, studies conducted:

1. Infrastructure Setup:

1A. Sample Collection: Lateral cephalograms were collected from 6 different centers located in the United States (Connecticut, New York, Ohio) and Jordan. In total, more than 6000 radiographs were collected. All patient data and location of collection were deidentified.

1B. Cloud Services: The computational power necessary to run the model exceeds the abilities of most personal computers. Additional graphic processing unit's (GPU's) were purchased from Amazon web services (AWS) to accommodate this demand. Digital ocean was used to provide the hosting service for the COCO annotator application, storing the image data and the data of the landmark points.

1C: Interface for Gathering Data: COCO Annotator was used as a web-based annotation platform for the human experts, different expert-level participants, and AI to complete landmarks annotation of all cephalograms. Pixel data was extracted from each of these annotations off COCO Annotator to determine variations between experts, human participants, and the AI, to determine landmark accuracy.

2. Developing the Algorithm

2A:Consecutive Radiographs: Consecutive radiographs were taken from each center were lateral cephalograms were obtain. Specifically, all consecutive images within a set range were included when obtaining radiographs so that no images were selectively eliminated. By including images of varying quality, robustness was added to the model.

2B: Inclusion and exclusion criteria: Inclusion criteria included images of patients with fixed orthodontic appliances, implants, and/or surgical bone plates taken before, after, or during orthodontic treatment. Exclusion criteria were images of poor enough quality that landmark identification was not possible.

2C: Moving them to a AWS in folders/buckets: Images were grouped into 8 training sets consisting of 319 - 539 radiographs each.

2D: Identifying landmarks: 27 commonly used landmarks were chosen for annotation. Landmarks were chosen based on their relevance for clinical application and academic research. Figure 1 shows the landmarks included in a sample lateral cephalometric radiograph. Table 2 describes each landmark in further detail.

2E: Engineering steps:

2Ea. Literature Review: A thorough research of the previous methods of automatic localization of landmarks was completed. Techniques, algorithms, opinion papers & software were all reviewed. This helped us evaluate the best infrastructure for this project in terms of: algorithm selection, optimization / customization of the algorithm, tool selection for annotating the landmarks by experts (generating labelled data) and in documenting the whole process for future research, modifications and publication.

2Eb. Model development and training. The AI model used for the task was a convolutional neural network regression model and was used to predict the coordinates of the selected cephalometric landmarks points. The model consists of two parts: feature extractor and regression head. The feature extract was an efficientnet_b7 model. The model was trained through a data centric approach in multiple stages. In each stage, more data was added in the form of human annotated training sets, and error specific data augmentation was also performed. Data augmentation included random rotation, flipping, random translation in any direction, changing brightness, and gamma correction (Figure 2).

The machine learning model is 20% code (model architecture) and 80% data. This study focused on improving model output through a data centric approach, rather than the tradition model centric approach. The model was initially developed using a 2,000-image annotated data set. Then the model was trained in multiple stages, and in each stage more data was added and error specific data augmentation as performed. Data augmentation included random rotation, flipping, random translation in any direction, changing brightness, and gamma correction. This error specific multiple augmentation technique made the model very robust. The output of each round of training was analyzed in comparison the experts' evaluation and algorithm was optimized accordingly. Overall 8 training rounds were completed and around 6000 radiographs were used for training.

2Ef: Code cleanup and final repo creation. The leftover data structures and other unwanted materials were removed from the memory and the filesystem. This made the source code itself easier to understand, maintain, and modify.

2F: Error Calculation: The aspirational goal is to create an algorithm that has an error rate similar to human experts. Thus a gold standard for the error between the human experts needs to be established for each landmark, which the error in the algorithm can be compared to. The

two experts with the greatest training (MU, SM) each annotated a data set of 330 radiographs. The error between experts for each point was set as the aspirational goal for the error of the model, in order to be considered successful.

This gold standard was calculated as a percentage. The error used in previous papers are accuracy scores at various radius's, as measured in terms of mm. There is an inherent problem in that method because the error is dependent on the resolution of the image. As the resolution increases, error increases. In order to mitigate this, we used a new error calculation that is independent of resolution, using a percentage calculation instead of a mm dimension. The percentage error was calculated based difference between different predicted coordinates of a landmark on the image, comparing either two expert annotators, or at the model to an annotator. Put simply, in an image with a resolution of 100 pixels x 100 pixels and the coordinate of the ground truth pixel is (50,60), a 1% error would indicate that the predicted point would lie within $(50 \pm 1, 60 \pm 1)$. In other words, a 1% error would indicate that the predicted point would have an x coordinate between 49 and 51 and a y coordinate between 59 and 61. To calculate the gold standard between two experts, the percentage was calculated between the predicted landmark locations by the two annotators, looking separately at the same radiograph. To calculate the percentage error for the model, the percentage was calculated between the coordinate determined by the annotator and the coordinate predicted by the model. Since this method of calculating error was novel, it was not possible to use it when comparing the accuracy of the model to previous models. Therefore, error was calculated through two additional methods. The first was through a direct mm value. The point-to-point error was calculated as the absolute pixel distance value between the ground truth position determined by the annotator and the predicted location by the model. This pixel difference was then converted to mm. Because all images had different resolutions, the average resolution of all uploaded cephalograms was taken and then used to create an averaged conversion rate between pixel and mm to be used on all images. This gave a single numerical error for each point on each radiograph. To compare the accuracy of the model to previous models, a successful detection rate (SDR) was also calculated. The percentage of times the landmark was successfully detected within 1 mm, 1.5 mm, 2 mm, 2.5 mm, 3 mm and 4 mm was calculated.

3. Developing the User Interface

Beta-testing website has been made where users can interact with the algorithm by uploading any cephalogram and the interface will use the CNN developed by us and automatically locate the landmarks and provide the interpretation.

4. Analyze the performance of the AI Algorithm.

4A: Analyzing Directionality of Error: To determine if errors followed any specific pattern, the deviation of the model-predicted point from the human-determined coordinate was recorded as a scatterplot. Scatterplots were created for each landmark, and each plotted point in the final 319 image testing set was represented as a point on the scatterplot, showing the deviation from normal along the x, y axis.

4B: Analyzing Performance of Algorithm Relative to Experts. Accuracy was compared to experts at each stage of training, to determine how the model improved with each successive iteration. Accuracy was determined through the same percentage calculation outlined to determine the gold standard, looking at the percentage of total pixels that the model differed from the experts by. Error was calculated through this manner for a total of nine training sets.

Results:

Creation of a New Gold Standard for Human Performance

To create a new gold standard for performance, two experts annotated the same 300 image set and the difference between their performance was measured to determine the highest level of accuracy that could be expected from human annotators. Accuracy was measured in percentage, as described in the methods, so that this gold standard could be extrapolated to future studies and data sets (Figure 3, Table 2). The average error between experts was found to be 0.73%, meaning the lowest expected inter-observer variability in landmark identification that can be expected of highly training clinicians is 0.73% of a radiographs dimensions. The landmarks with the highest interobserver variability were O with 1.27%, PgS with 1.22%, and G with 1.19%. The landmarks with the lowest inter-observer variability were Uc with 0.27%, Lc with 0.35%, and S with 0.38%.

Stagewise Error

Error, measured in percentage of image area, decreased with each training stage. Error started at 2.02% in stage 1 and decreased to 0.77% by stage 9 (Figure 4, Table 3). Similarly sized training sets were used for each stage of training, however improvement was notably more substantial between earlier training stages than later training stages. The improvement between stages can be seen in Figure 5, which shows how different versions of the model performed on the same image.

Directionality of Error

The directionality of error for all landmarks in the final testing set of 319 radiographs was assessed and is represented by scatterplots in Figure 6, which shows the difference between the location of the landmark predicted by the model and the location of the landmark determined by the annotator in both direction and millimeter distance. Several landmarks exhibited pronounced directionality in their errors. Landmarks G and PgS showed a primarily vertical error, while primarily horizontal error was seen for ANS, L6D, L6M, PNS, Me, U6D, and U6M. A strong diagonal vector was seen for landmark Go.

Performance of Model

Performance of Model was assessed in multiple ways. The first was in percentage of error, as was developed for this study. Overall, the model showed an average error of 0.77%. The average millimeter error for the 27 landmarks annotated was also determined (Table 2, Figure 7). The landmarks showing they highest level of error G, Go, and P with 1.21 mm, 1.04 mm and 1.04 mm respectively. The landmarks showing the lowest error of error were Uc, Sn and Cb, with 0.54 mm, 0.55 mm and 0.55 mm respectively.

The model was also evaluated to determine it's successful detection rate, as this has been the standard method to evaluate performance by previous investigators in the field (Table 4). Overall, the model had a 74.9% SDR in the 1 mm range, a 90.5% SDR in the 1.5 mm range, and a 96% SDR in the 2 mm range. Only 1.8% of landmarks had a SDR of 2.5 mm and over. Landmarks showing the greatest number of SDR <1 mm were Cb, Ls and Uc, while landmarks showing the lowest number of SDR <1 mm were G, Go and Co. Landmarks showing the highest performance in the 2 mm range were SEM, U6M, Sn, Pn, Ls, and Li, all of which had a SDR <2 mm in over 99% of tested radiographs. Radiographs showing with the least number of SDR <2 mm were G, Go, and P, with 82.2%, 90.2% and 90.5% respectively.

Specific Aim 2: Evaluate accuracy and reliability of the proposed algorithm in landmark

identification comparison with trained experts with different levels of experience.

Work done; studies done:

2A: Recruitment of Experts: Two individuals (Shivam Mehta (PI), Madhur Upadhyay (Research Mentor) served as the human experts for determining the gold standard of human error, and for calibrating other orthodontists to assist with algorithm training. Together, they have a combined experience of over 20 years of using lateral cephalograms in clinical diagnosis and treatment planning and considerable experience with research work involving cephalometric parameters. These experts (MU, SM) then instructed additional investigators (KC, RS, RB, MC) until they were calibrated to have the same degree of error in locating cephalometric landmarks as the experts themselves. These additional investigators participated in annotation of the 3,129-image training set for the algorithm.

2B: Interface for Gathering Data: COCO Annotator was used as a web-based annotation platform for the human experts, different expert-level participants, and AI to complete landmarks annotation of all cephalograms. Pixel data was extracted from each of these annotations off COCO Annotator to determine variations between experts, human participants, and the AI, to determine landmark accuracy.

Method – different experts

Results:

Analyzing Performance of Algorithm Relative to Experts with Varying-Levels of

Training: In order to gauge where the algorithm performance ranked relative to clinicians, its error was measured relative to 4 levels of orthodontic clinicians. These levels were first year orthodontic residents, second year orthodontic residents, third year orthodontic residents and orthodontic providers practicing in a post-graduate clinical setting. Three calibrated orthodontic annotators from the study determined the true location of each landmark. Five different individuals were included in each group and each traced six radiographs, for a total of thirty radiographs per each group. The model and each group annotated the same thirty radiographs so accuracy could be compared.

Performance of the Model, Compared to Experts

Performance of the model was compared to the gold standard set by the expert annotators (Table 2, Figure 8). While the inter-experts variability was 0.73%, the model showed an error of 0.77% when compared to the expert annotators. The range of error in mm for the model was 0.55 - 1.21 mm, with the model producing the greatest error with G and the least with Uc and Ls. The range for variability between experts was 0.35 - 1.21 mm, with the experts showing the greatest variability with O and the least with Lc. Overall, the experts showed greater variability for 13 landmarks; A, B, Co, Go, Li, Lr, Ls, O, PNS, PgS, Pn, Sn and Ur. The model showed greater error for 14 Landmarks; ANS, Cb, G, L6D, L6M, Lc, Me, N, P, S, U6D, and U6M.

Comparison of the Model to Human Clinician Performance

Error was measured in percentage of total image dimensions, as outlined in the methods (Table 6). Of the 4 groups, the model substantially outperformed all 4 groups with an error of 0.83%. The top performing group was practicing clinicians with an error of 1.9%. The third highest performing group was second year residents with an error of 2.2%, followed by first year residents with an error of 2.64%, and the lowest performing group was the third year residents

with an error of 2.66%. All 4 human groups were within 1 standard deviation of the mean, which was 2.05. However, the AI model performance was 2 standard deviations from the mean.

Specific Aim 3: Compare the efficiency of the algorithm in locating the desired landmarks in cephalometric radiographs derived from different centers across the world.

Work done; studies done: Cephalometric radiographs were collected from different centers across the world in order to increase the generalizability of the algorithm.

3A. Infrastructure setup: Different annotator tools and GPU space requirements were evaluated to figure out a confirm CoCo Annotator was the best fit when choosing the platform. Overall, 3,448 images were then uploaded to the Coco Annotator Platform.

3B. Data Preprocessing: This was divided into four stages:

3Ba1: Data cleaning. This involved filling in for ‘missing’ data and removal of ‘noisy’ data (binning and regression).

3Ba2: Data integration. Because data was collected from multiple sources (locations and different machines), the data had to be consolidated and unified.

3Ba3: Data reduction. The data was condensed by employing various filters and component analysis.

3Ba4: Data transformation. This was the final step or preprocessing to make the data appropriate for data modeling. It involved: smoothing, aggregation, discretization and feature construction.

Results:

Performance of the Model, Compared to YoloV3

When evaluated on 80 cephalometric landmarks, YOLOv3 had an average SDR of 80.4% for <2mm, 87.4% for <2.5 mm, 92% for <3 mm, and 96.2% for <4mm (25). When evaluated for 27 cephalometric landmarks, the current model had an average SDR of 96.4% for <2mm, 97.1% for <2.5 mm, 97.4% for <3 mm and 98% for <4 mm (Figure 9).

Performance of the Model, Compared to the Model Developed by Kim

Our model outperformed the Kim model on all landmarks except for sella, maxillary incisor crown and mandibular incisor crown (Figure 10, Table 5). Their model looked at 20 landmarks overall and had a 56.6% SDR within 1 mm and a 83.6% SDR within 2 mm. Our model looked at 27 landmarks overall and had a 74.9% SDR within 1 mm and a 96% SDR within 2 mm. To make a more specific comparison, SDR’s were evaluated for only landmarks that were mutually assessed. This included 15 landmarks that are listed in table. Within this smaller subset, their model had a 60% SDR within 1 mm and a 80% SDR within 2 mm, while the current model showed 74.6% SDR within 1 mm and a 96.4% SDR within 2 mm. In terms of individual landmark performance, the current model outperformed the Kim model in all landmarks except Sella, the maxillary incisor crown tip, and the mandibular incisor crown tip. Their SDR on the top three landmarks outperformed the SDR of any landmarks for the current

model, with the current models highest performing mutual landmark being maxillary incisor crown tip at 90.2%.

However, the current model demonstrated more consistent performance with no landmarks performing below 50%, while the Kim model had 5 landmarks performing below 50%; mandibular incisor root, ANS, porion, PNS, maxillary molar distal, and B point.

Specific Aim 4: Create an automated diagnosis and treatment planning algorithm using RCNN

Work done and current status: In this objective, 2600 datasets with each dataset consisting of 8 photographic images, lateral cephalogram, and panoramic radiograph (2000 training data and 600 test data are to be evaluated for specific feature variables with each dataset consisting of radiographic images such as lateral cephalogram, panoramic radiograph, and eight photographic images. Currently, 1000 such datasets have been evaluated by multiple orthodontists with different levels of experience. The AI based algorithm is being trained to evaluate the data obtained by the evaluated datasets. Further work is being done in this aim for the development of the AI algorithm and adding more datasets to be evaluated for the feature variables.

Label studio prototype is created to organize, crop, and analyze the photographic and radiographic images. The following link describes the link for the same.

<http://138.197.152.103:8080/user/login/>

1. Progress Report

Explain how research is proceeding relative to original timetable and contingency plans (any changes from original plan). Accordingly, what are the future plans to carry the project to its end? Supply specific statements on human or animal subjects.

a. Progress to date

The progress to date has been made in all of the specific aims. The following explain the progress in each of the specific aim.

Specific aim 1 – Create the first fully automated framework for cephalogram analysis using one of the rapidly developing deep learning techniques—Region based convolutional neural networks (RCNN).

- In this aim, a collection of heterogenous datasets of 14,000 cephalograms (10,000 training data and 4000 test data) is to be analyzed. So far, we have created a fully automated framework for cephalogram analysis using Region based convolutional neural network (RCNN). The link for the fully automated framework for cephalogram analysis is the following.
- (Link: <http://157.230.187.183:5000/#/auth> Username: Test;password: trial).
-

Specific aim 2 – Evaluate accuracy and reliability of the proposed algorithm in landmark identification comparison with trained experts with different levels of experience.

-This aim is achieved with the results listed above in the explanation of specific aim 2

Specific aim 3 – Compare the efficiency of the algorithm in locating the desired landmarks in cephalometric radiographs derived from different centers across the world.

- This aim is achieved with the results listed above. Additional cephalograms are being analyzed currently and the algorithm is being refined on more datasets.

Specific aim 4 – Create an automated diagnosis and treatment planning algorithm using RCNN

- In this objective, 2600 datasets with each dataset consisting of 8 photographic images, lateral cephalogram, and panoramic radiograph (2000 training data and 600 test data are to be evaluated . Currently, 1000 such datasets have been evaluated.
- This aim is achieved partially. Additional datasets are being evaluated and the algorithm is being trained on the datasets.

b. Plans for the requested remaining months of support

Specific Aim 1 – Out of the 10,000 training data, over 7000 cephalograms have been analyzed. The remaining datasets are being evaluated and the algorithm is being trained on the datasets.

Specific aim 2 – Achieved

Specific aim 3 - Additional cephalograms are being analyzed currently for this aim.

Specific aim 4 – Out of the 2600 datasets, over 1000 datasets are analyzed. The remaining datasets are being evaluated and the algorithm is being trained on the datasets.

c. Subjects (detailed description of sample, including information gender and age)

All samples (cephalometric radiographs) and datasets were deidentified. Over 7000 cephalograms have been used from different centers in the study including University of Connecticut, private practices, and orthodontic clinics. The datasets for radiographic images and photographic images were obtained from the University of Connecticut.

d. Publications/presentations

- (i) Developing an AI Algorithm for Predicting Severity and Location Impacted Maxillary teeth. Poster Presented at AAO by An Jin et al.
- (ii) Invitation for book chapter – Attached in the illustration/addendum
- (iii) Mehta S, Upadhyay M, Suhail Y. Artificial Intelligence for radiographic image analysis. *Seminars in Orthod.* 2021;27(2):109-120 (Invited Article).
- (iv) Acceptance of proposal grant for Northeastern Society of Orthodontics – adding to this research and taking it further

- (v) Manuscript under preparation: Assessing agreement among orthodontists on cephalometric and clinical parameters
 - (vi) Received another grant from Texas SB30 funds for further continuation of the research study: 150,000 Shivam Mehta : Principal Investigator
- e. Listing of investigators, nature of involvement in research, and time allotted since beginning of research
- Shivam Mehta – 30% time allotted since beginning of research – Principal investigator
- Madhur Upadhyay – 30 % time allotted since beginning of research – Research Mentor
- Gaurav Sinha – 20% time allotted since beginning of research – Collaborator – Develop the Artificial Intelligence algorithm.
- Other investigators – 20% Katherine S Chapman and other experts involved in identification of landmarks and analyzing datasets.
- f. Percentage funding from AAOF and other sources; \$19,615 used from 20,000 from AAOF used
Other sources used – University of Connecticut Health Center startup funding (K), Texas SB30 funds

2. Illustrations, addendum

Figures and Tables

Figure 1: A sample cephalometric x-ray showing the true position of the cephalometric landmarks.

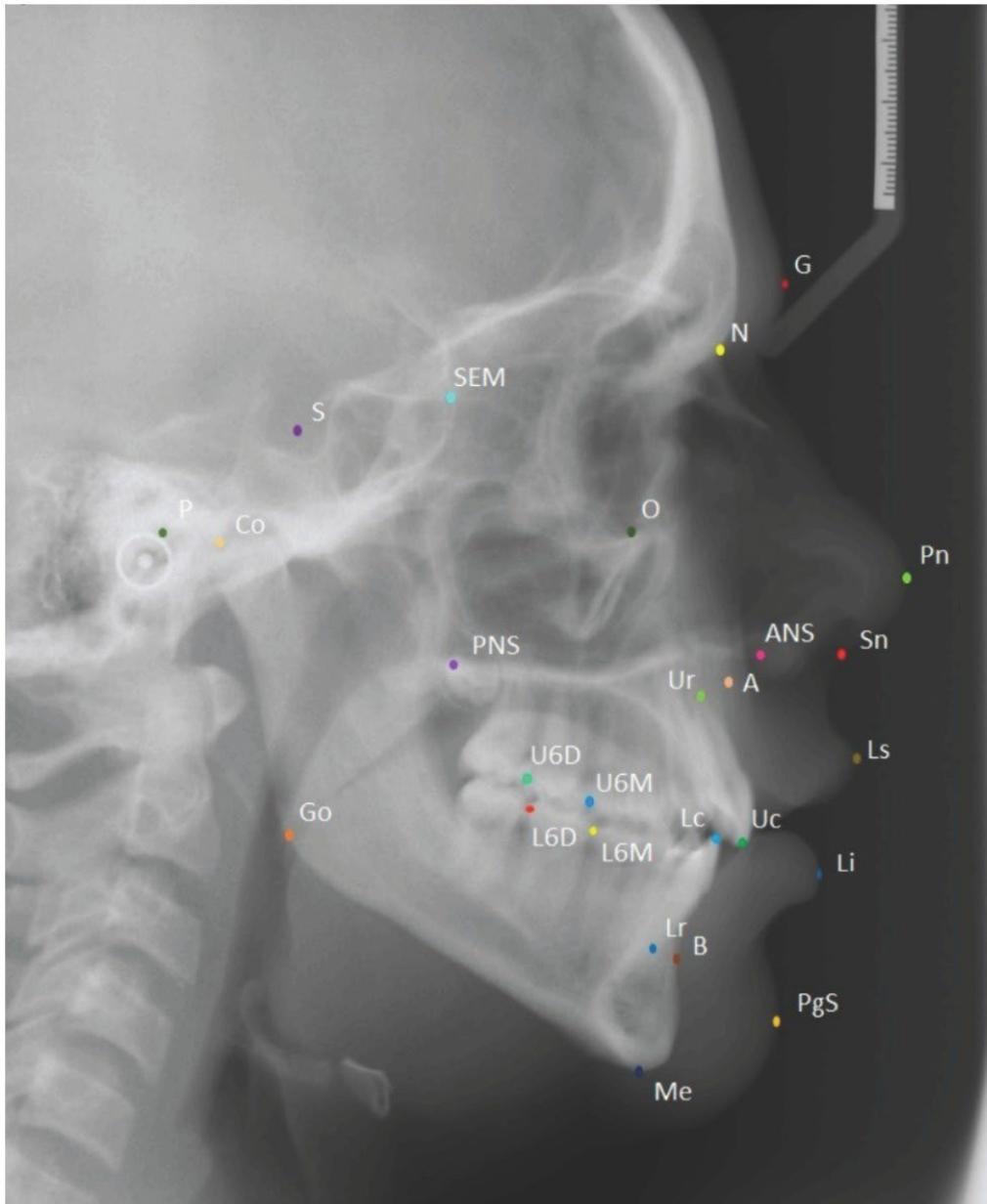


Figure 2: Sample images showing the results of data transformations.

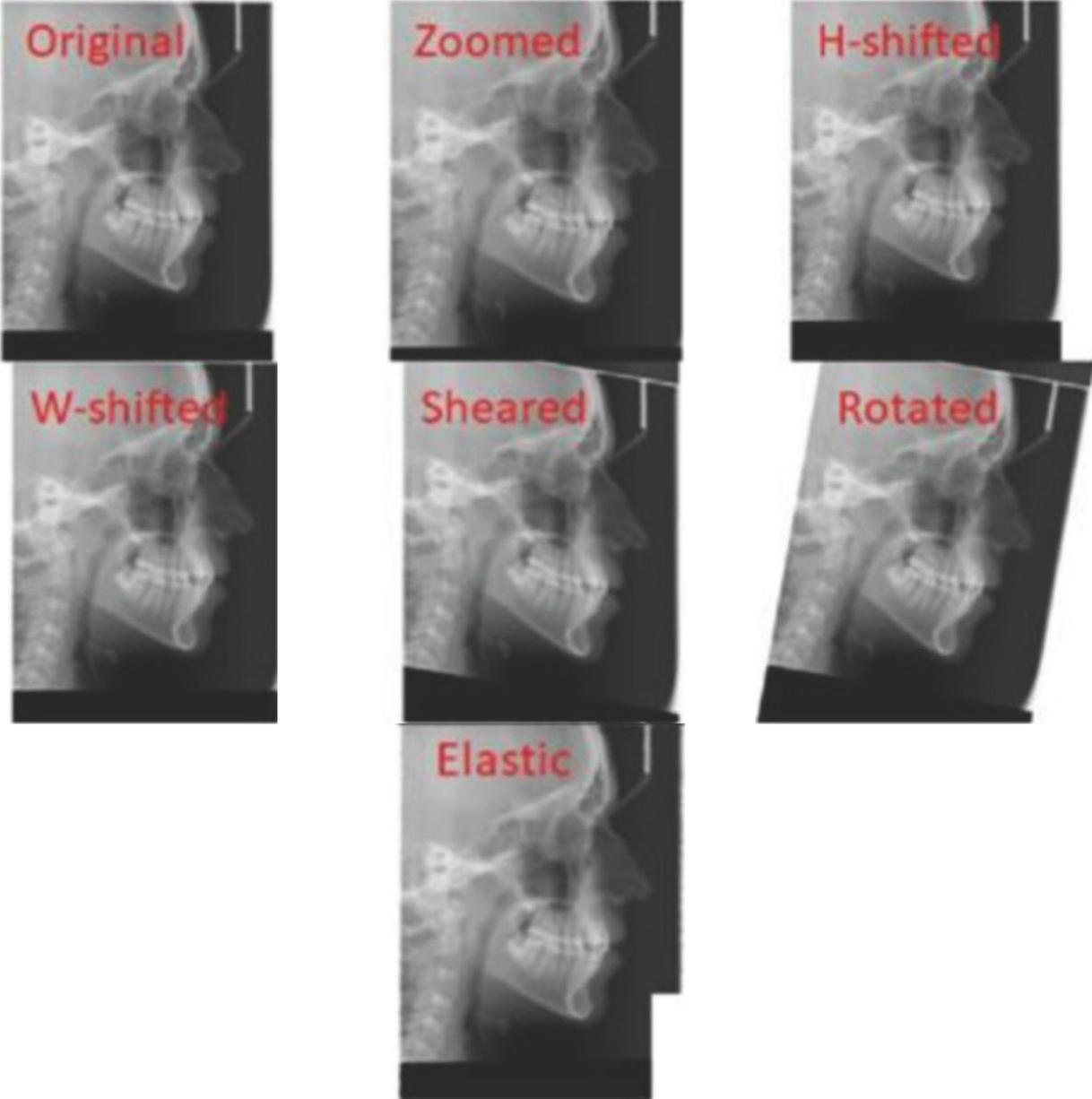


Figure 3: Development of the new gold standard, the inter-expert variability, as measure in millimeters

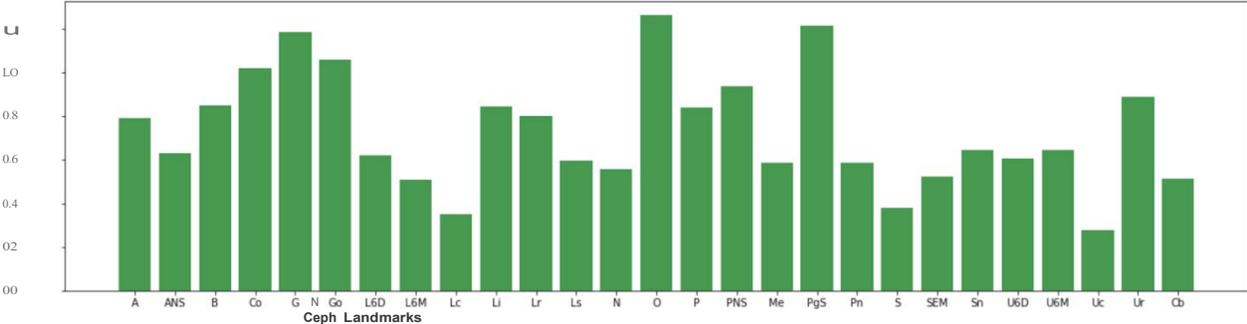


Figure 4: Improvement of model with each iteration.

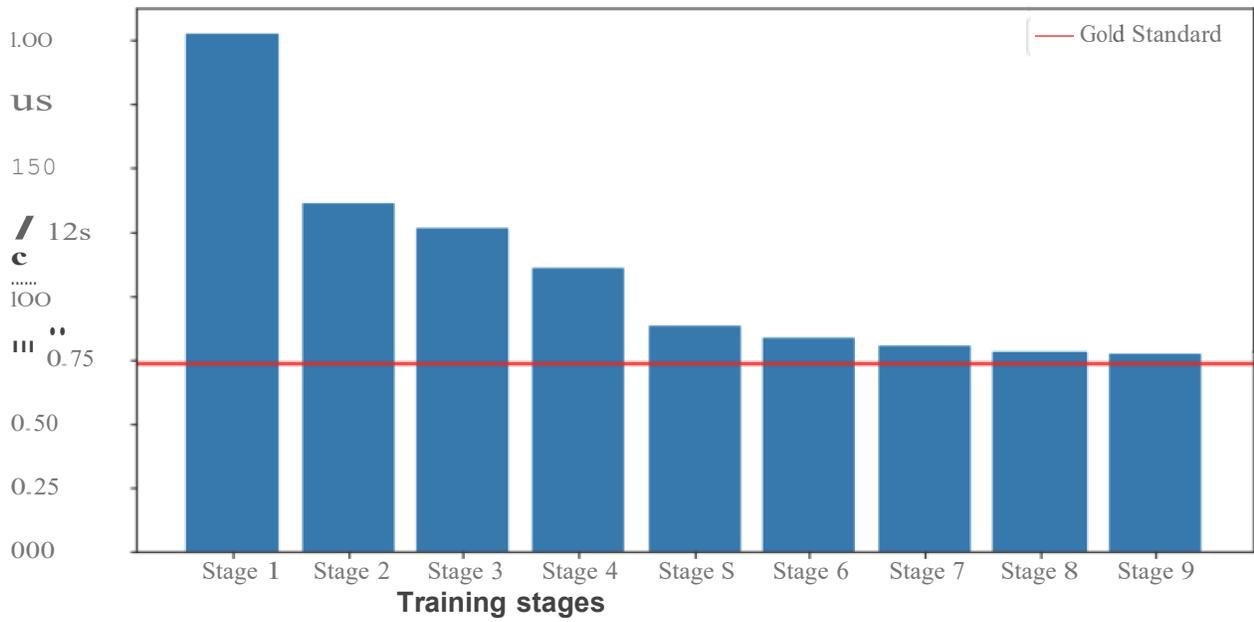
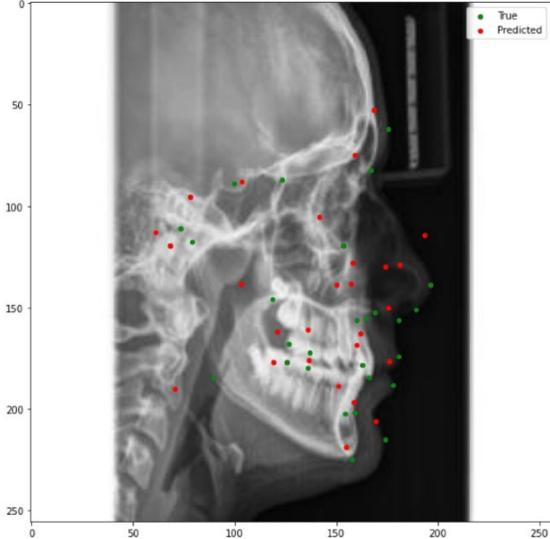
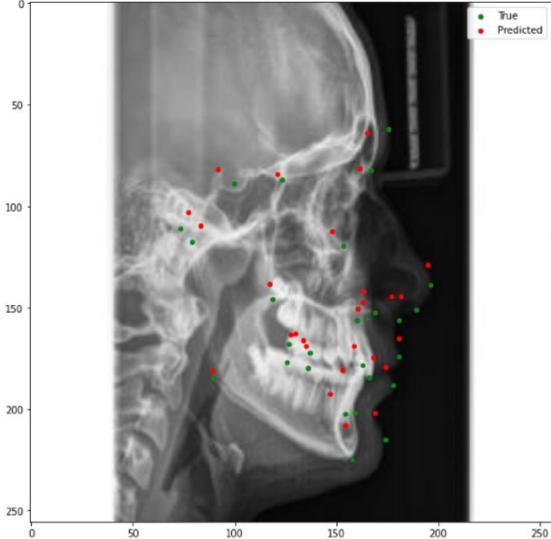


Figure 5: Iterative improvement in landmark prediction

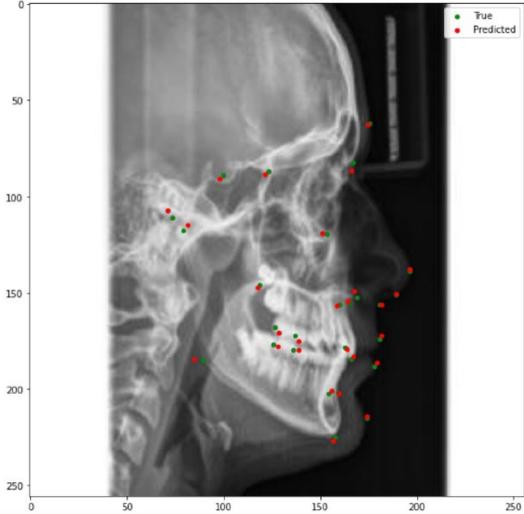
INITIAL TRAINING



MID-TRAINING 1



MID-TRAINING 2



FINAL MODEL

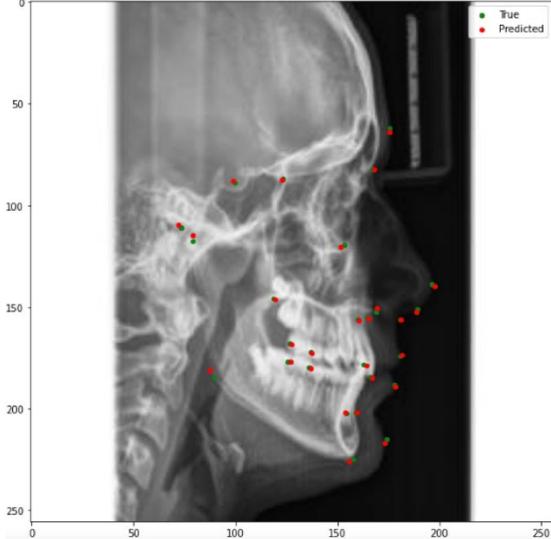


Figure 6: Scatterplots showing directionality of errors for each landmark.

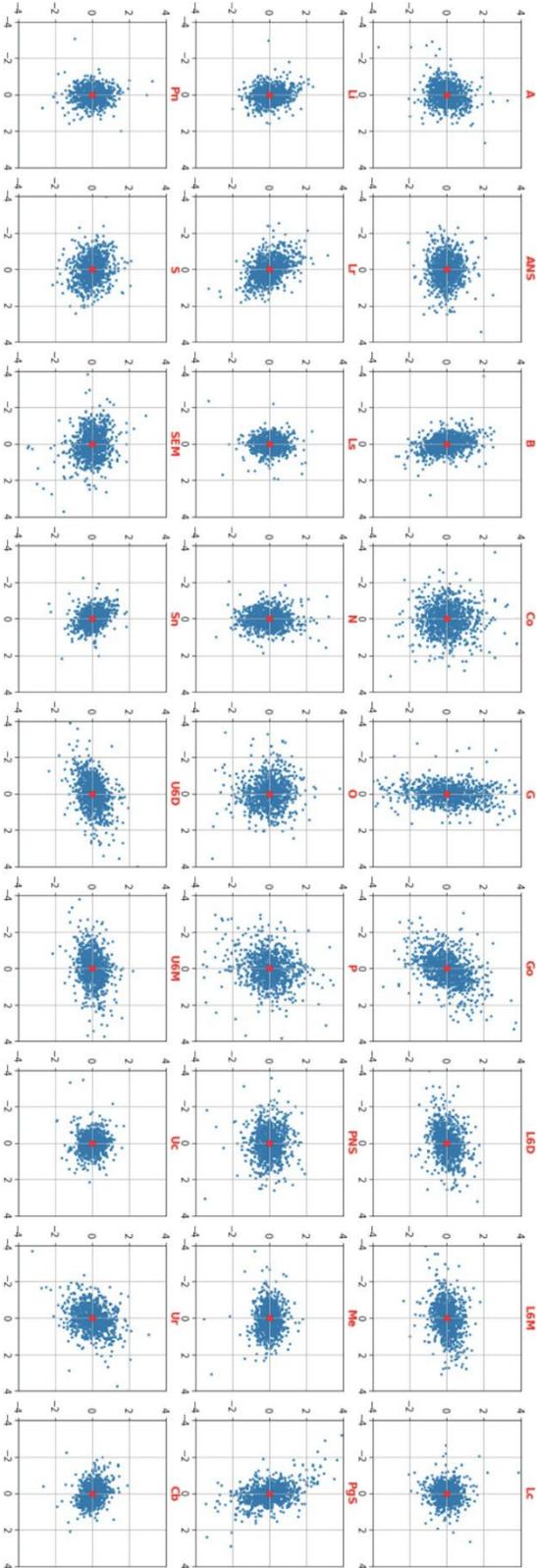


Figure 7: Accuracy of model in millimeter.

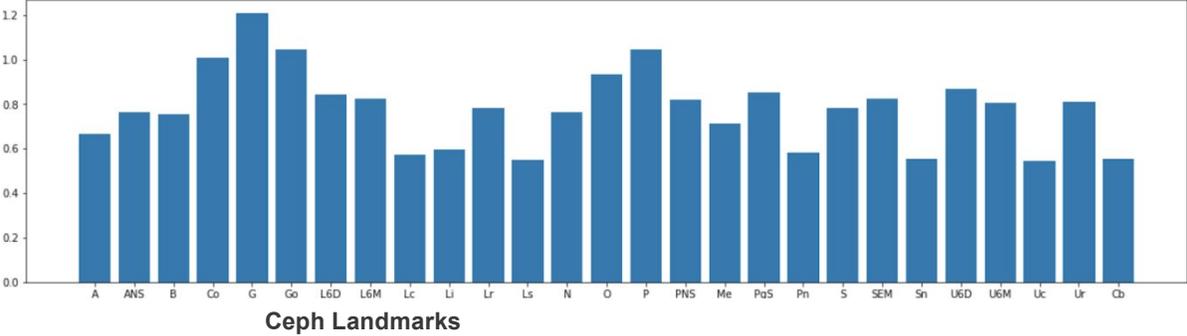


Figure 8: Comparison of model performance to inter-expert variability

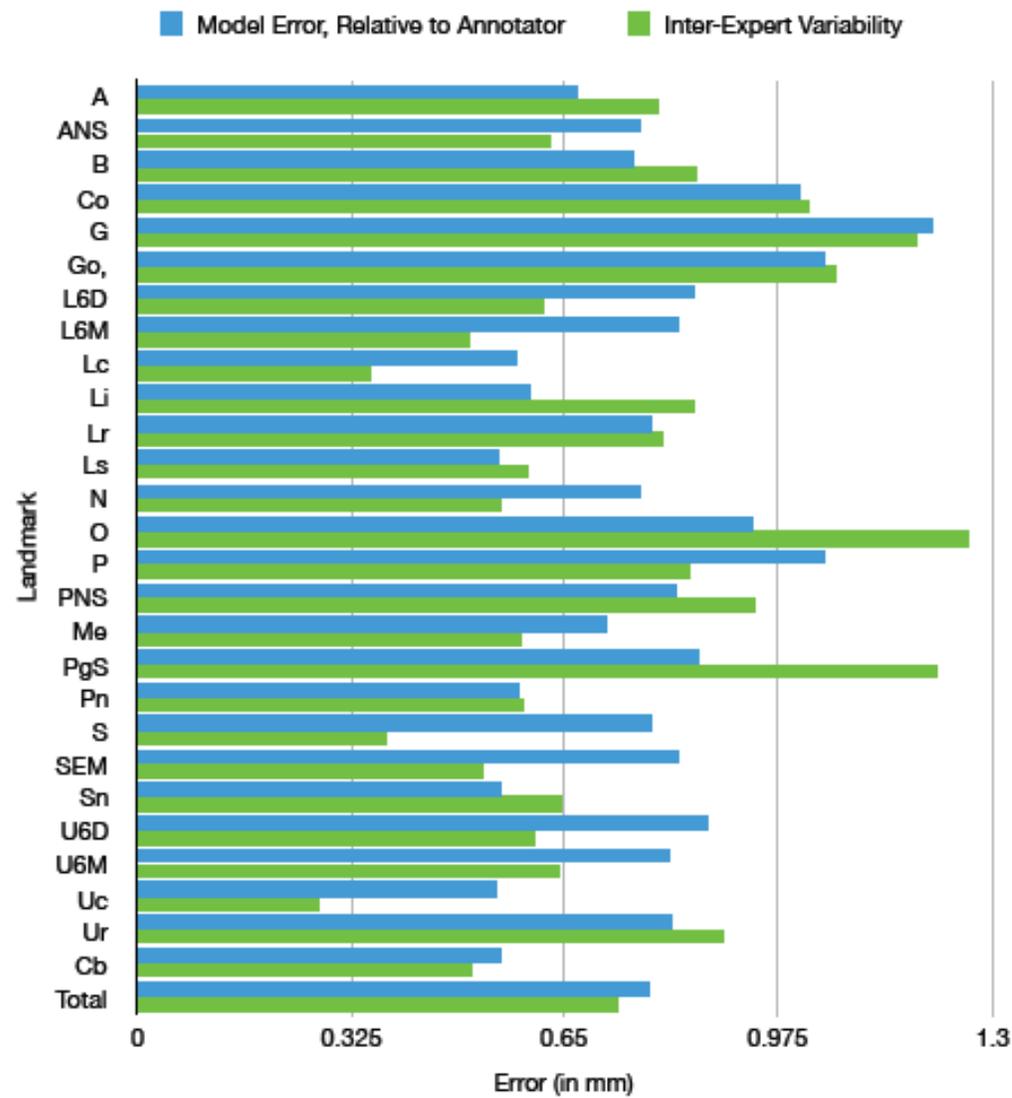
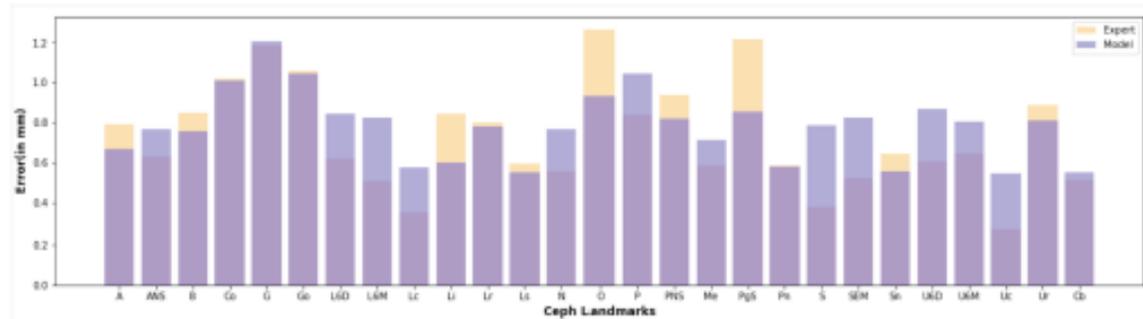


Figure 9: Accuracy compared to YOLOv3, measured in various SDR's.

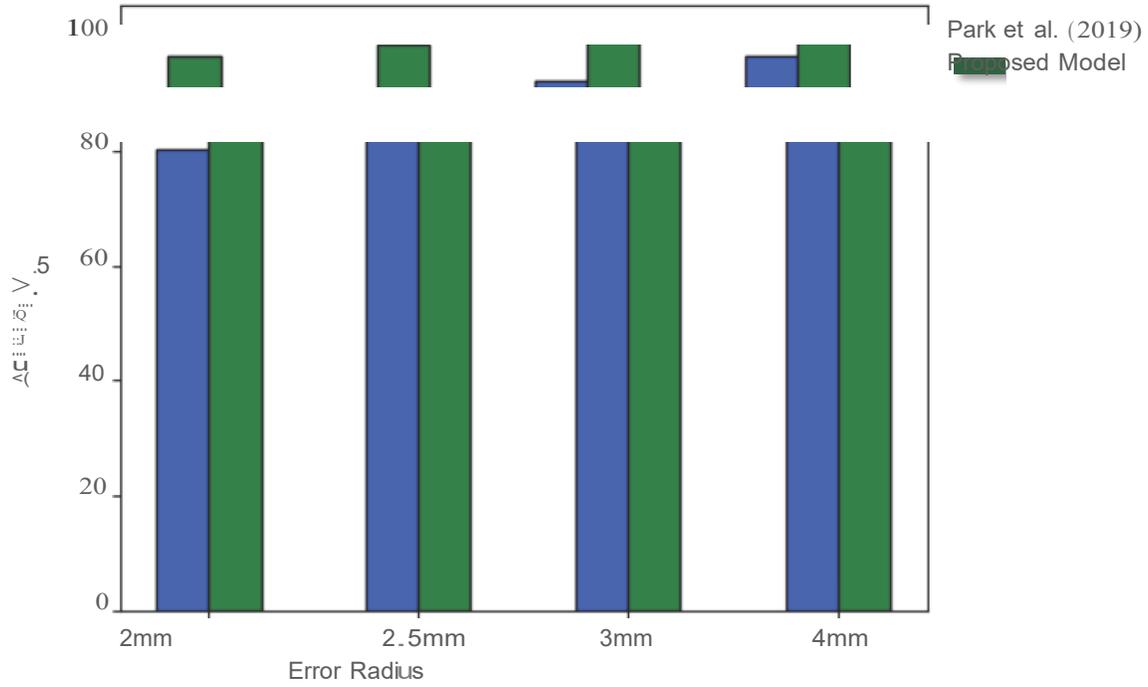


Figure 10: Comparing accuracy of Kim 2021 model to current model

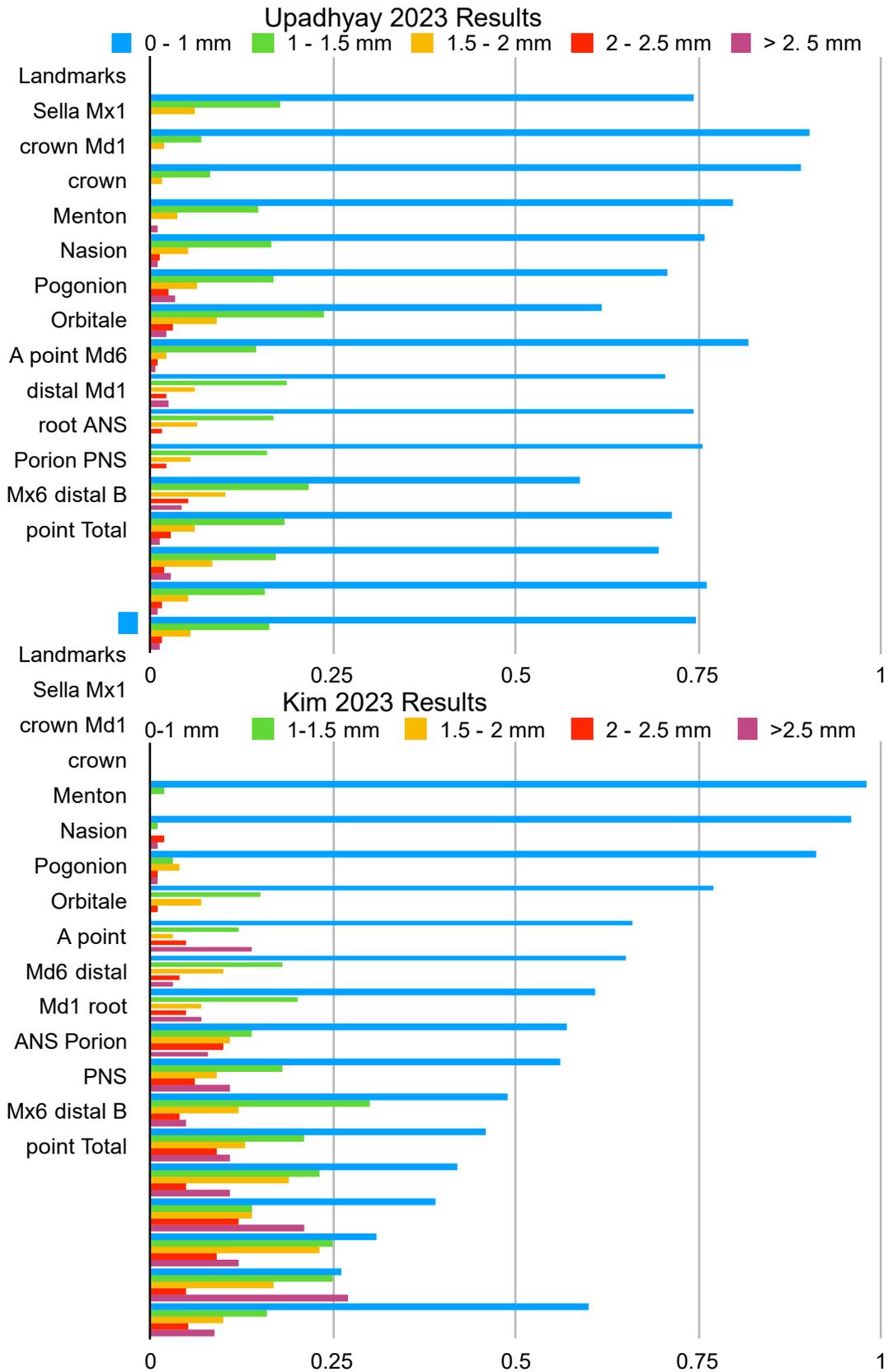


Table 1: Clinical Definitions of all included landmarks.

Landmark	Description
S (Sella)	The geometric center point of the pituitary fossa.
P(Porion)	The most superior point on the external auditory meatus.
Go (Gonion) N	A point on the curvature of the angle of the mandible.
(Nasion)	The most anterior point on the frontonasal suture.
O (Orbitale)	The lowest point on the inferior rim of the orbit.
A (Point A) B	The deepest point on the bony concavity between ANS and supradentale.
(Point b) Me	The deepest point on the bony concavity between Pogonion and infradentale.
(Menton) G	The most inferior point on the hard tissue chin.
(Glabella)	The most anterior point on the forehead.
Sn (Subnasale)	The junction of the nose and upper lip.
Ls (Labrale superius) Li	The most prominent point on the upper lip.
(Labrale inferius)	The most prominent point on the lower lip.
PgS (Soft tissue pogonion)	The most prominent point on the soft tissue chin.
Ur Uc	The root tip of the upper incisor.
Lc Lr	The crown tip of the upper incisor.
ANS PNS	The crown tip of the lower incisor.
U6M	The root tip of the lower incisor.
U6D	(Anterior nasal spine) Anterior tip of the nasal spine.
L6M L6D	(Posterior nasal spine) Posterior tip of the nasal spine.
SEM (Sphenoethmoidal point)	Upper first molar mesial tip: most prominent point on the mesial cusp.
Pn (Pronasale) Co	Upper first molar distal tip: most prominent point on the distal cusp.
(Condylion)	Lower first molar mesial tip: most prominent point on the mesial cusp.
	Lower first molar distal tip: most prominent point on the distal cusp.
	Intersection of the greater wing of sphenoid and the cranial floor.
	The most anterior point on the nose.
	The most superior and posterior point on the condylar head.

Table 2: Error of proposed model for testing set and inter-expert Variability, in millimeters

Table 2 Model Error in Millimeter

Points	Model Error, Relative to Annotator	Inter-Expert Variability
A	0.6671	0.7898
ANS	0.7657	0.6292
B	0.7556	0.8485
Co	1.0087	1.0201
G	1.2085	1.1877
Go	1.0447	1.0623
L6D	0.8456	0.6177
L6M	0.8233	0.507
Lc	0.5748	0.3521
Li	0.5986	0.8472
Lr	0.7824	0.7989
Ls	0.5494	0.5942
N	0.7665	0.5554
O	0.9341	1.2655
P	1.0445	0.84
PNS	0.8217	0.9387
Me	0.7139	0.5853
PgS	0.8555	1.2153
Pn	0.581	0.5869
S	0.7848	0.3783
SEM	0.8244	0.524
Sn	0.5545	0.6457
U6D	0.8689	0.6071
U6M	0.8076	0.6435
Uc	0.547	0.2754
Ur	0.8113	0.8899
Cb	0.5531	0.5112
Total	0.7812	0.7303

Table 3: Improvement in model accuracy with each training stage. Error is measured as percentage of total image dimensions.

Training Stage	Error(in %)
Stage 1	2.02
Stage 2	1.36
Stage 3	1.26
Stage 4	1.11
Stage 5	0.88
Stage 6	0.83
Stage 7	0.8
Stage 8	0.78
Stage 9	0.77

Table 4: Successful detection rate of each landmark by model

Landmarks	0-1 mm (Excellent)	1-1.5 mm (Good)	1.5-2 mm (Fair)	0-2 mm	(2-2.5)mm(Acceptable)	(>2.5)mm (Unacceptable)
A	817	144	22	983	9	8
ANS	756	161	56	973	21	6
B	762	157	53	972	17	11
Co	580	243	110	933	38	29
G	529	171	122	822	89	89
Go	576	229	97	902	56	42
L6D	704	186	62	952	23	25
L6M	731	162	58	951	24	25
Lc	890	83	16	989	6	5
Li	874	98	21	993	5	2
Lr	744	168	65	977	15	8
Ls	907	68	17	992	6	2
N	757	167	53	977	13	10
O	617	238	91	946	32	22
P	588	215	102	905	51	44
PNS	713	184	61	958	29	13
Me	798	147	36	981	8	11
PgS	707	170	63	940	26	34
Pn	885	93	12	990	6	4
S	744	178	61	983	13	4
SEM	705	203	55	963	21	16
Sn	895	82	16	993	6	1
U6D	695	172	86	953	20	27
U6M	733	159	58	950	28	22
Uc	902	70	19	991	4	5
Ur	717	192	61	970	19	11
Cb	913	63	15	991	6	3
Total Percentage	74.96%	90.53%	5.739%	96.04%	2.19%	1.77%

Table 5: Comparing accuracy of Kim 2021 model to current model on mutual landmarks

Comparing Accuracy of Kim 2021 Model to Upadhyay 2023 Model

Landmarks	0-1 mm (Excellent)		1-1.5 mm (Good)		1.5-2 mm (Fair)		2-2.5 mm (Acceptable)		>2.5 mm (Unacceptable)	
	Kim 2021	Upadhyay 2023	Kim 2021	Upadhyay 2023	Kim 2021	Upadhyay 2023	Kim 2021	Upadhyay 2023	Kim 2021	Upadhyay 2023
Sella	98%	74.4%	2%	17.8%	0%	6.1%	0%	0.13%	0%	0.04%
Mx1 crown	96%	90.2%	1%	7%	0%	1.9%	2%	0.04%	1%	0.05%
Md1 crown	91%	89%	3%	8.3%	4%	1.6%	1%	0.06%	1%	0.05%
Menton	77%	79.8%	15%	14.7%	7%	3.6%	1%	0.08%	0%	1.1%
Nasion	66%	75.7%	12%	16.7%	3%	5.3%	5%	1.3%	14%	1%
Pogonion	65%	70.7%	18%	17%	10%	6.3%	4%	2.6%	3%	3.4%
Orbitale	61%	61.7%	20%	23.8%	7%	9.1%	5%	3.2%	7%	2.2%
A point	57%	81.7%	14%	14.4%	11%	2.2%	10%	0.9%	8%	0.8%
Md6 distal	56%	70.4%	18%	18.6%	9%	6.2%	6%	2.3%	11%	2.5%
Md1 root	49%	74.4%	30%	16.8%	12%	6.5%	4%	1.5%	5%	0.08%
ANS	46%	75.6%	21%	16.1%	13%	5.6%	9%	2.1%	11%	0.06%
Porion	42%	58.8%	23%	21.5%	19%	10.2%	5%	5.1%	11%	4.4%
PNS	39%	71.3%	14%	18.4%	14%	6.1%	12%	2.9%	21%	1.3%
Mx6 distal	31%	69.5%	25%	17.2%	23%	8.6%	9%	2.0%	12%	2.7%
B point	26%	76.2%	25%	15.7%	17%	5.3%	5%	1.7%	27%	1.1%
Total	60%	74.6%	16.1%	16.2%	9.9%	5.6%	5.2%	1.7%	8.8%	1.4%

Table 6: Model performance, compared to clinicians with varying levels of orthodontic experience

Model Performance Relative to Orthodontic Experience

Rank	Group	Error (in o/o)
1	AI model	0.83
2	Practicing Clinicians	1.9
3	Second Year Residents	2.2
4	First Year Residents	2.64
5	Third Year Residents	2.66
	Standard Dev	Average
	0.7506	2.046



shivam mehta <mehta.shivam21@gmail.com>

Test mail: Invitation to contribute a chapter on AI

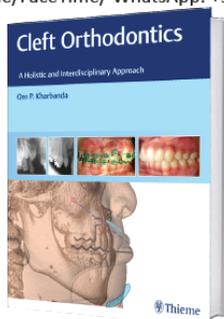
Prof. Om P Kharbanda <dr.opk15@gmail.com>
To: mehtashivam21@gmail.com

Mon, Apr 17, 2023 at 10:05 PM

Dear Shivam, Here is the content file of edition 3.
Now we're working on Ed 4.
Kind regards

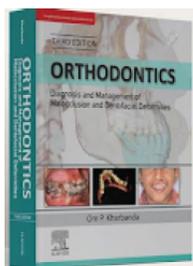
Prof. OP Kharbanda

MDS (Lucknow), M MEd (Dundee), M Orth RCS (Edinburgh),
FDS RCS (England), FDS RCS (Edinburgh) Hon, FAMS
Pro Vice-Chancellor Health Sciences
Ramaiah University of Applied Sciences (RUAS)
University House, [New BEL Road, MSR Nagar](#)
[Bengaluru, 560054, India](#)
www.ramaiah-india.org
Professor Emeritus: National Academy of Medical Sciences (NAMS)
<https://www.nams-india.in/>
Honorary Adjunct Professor La Trobe University Australia
Academic Fellow Universiti Sans Malaysia
Phone/FaceTime/ WhatsApp: +91-9899062144



www.thieme.com

2 attachments



Elsevier Ortho OPK 2.png
189K

 **ContentsandContributors.pdf**
1721K