BMC Oral Health

RESEARCH

Open Access

Impacted lower third molar classification and difficulty index assessment: comparisons among dental students, general practitioners and deep learning model assistance

Paniti Achararit[1], Chawan Manaspon[2], Chavin Jongwannasiri[1], Promphakkon Kulthanaamondhita[3,4], Chumpot Itthichaisri[3,4], Soranun Chantarangsu[5], Thanaphum Osathanon[6], Ekarat Phattarataratip[5*] and Kraisorn Sappayatosok[3,4*]

Abstract

Background Assessing the difficulty of impacted lower third molar (ILTM) surgical extraction is crucial for predicting postoperative complications and estimating procedure duration. The aim of this study was to evaluate the effectiveness of a convolutional neural network (CNN) in determining the angulation, position, classification and difficulty index (DI) of ILTM. Additionally, we compared these parameters and the time required for interpretation among deep learning (DL) models, sixth-year dental students (DSs), and general dental practitioners (GPs) with and without CNN assistance.

Materials and Methods The dataset included cropped panoramic radiographs of 1200 ILTMs. The parameters examined were ILTM angulation, class, and position. The radiographs were randomly split into test datasets, while the remaining images were utilized for training and validation. Data augmentation techniques were applied. Another set of radiographs was used to compare the accuracy between human experts and the top-performing CNN. This dataset was also given to DSs and GPs. The participants were instructed to classify the parameters of the ILTMs both with and without the aid of the best-performing CNN model. The results, as well as the Pederson DI and time taken for both groups with and without CNN assistance, were statistically analyzed.

Results All the selected CNN models successfully classified ILTM angulation, class, and position. Within the DS and GP groups, the accuracy and kappa scores were significantly greater when CNN assistance was used. Among the groups, performance tests without CNN assistance revealed no significant differences in any category. However, compared with DSs, GPs took significantly less time for the class and total time, a trend that persisted when CNN assistance was used. With the CNN, the GPs achieved significantly higher accuracy and kappa scores for class classification than the DSs did ($p = 0.035$ and 0.010). Conversely, the DS group, with the CNN, exhibited higher accuracy and kappa scores for position classification than did the GP group ($p < 0.001$).

*Correspondence:
Ekarat Phattarataratip
ekarat.p@chula.ac.th
Kraisorn Sappayatosok
kraisorn.s@rsu.ac.th
Full list of author information is available at the end of the article

Conclusion The CNN can achieve accuracies ranging from 87 to 96% for ILTM classification. With the assistance of the CNN, both DSs and GPs exhibited significantly higher accuracy in ILTM classification. Additionally, compared with DSs with and without CNN assistance, GPs took significantly less time to inspect the class and overall.

Keywords Impacted tooth, Artificial intelligence, Deep learning, Convolutional neural network, Pederson difficulty index

Impacted teeth represent one of the most common pathologies in the oral cavity. They can lead to various problems, such as caries or periodontitis of adjacent teeth, pericoronitis, and even the development of other serious pathologies, such as odontogenic cysts or tumors [1]. Impacted lower third molar (ILTM) is the most commonly found impacted tooth [2]. The treatment of choice for ILTM is extraction or surgical removal, which is a common procedure for general dentists. However, the difficult position of ILTM, particularly with respect to its proximity to the inferior alveolar nerve or when it is deeply embedded in the mandibular ramus and requires extensive bone removal, can sometimes lead to postoperative complications such as anesthesia-related complications, significant pain, and swelling. Therefore, evaluating the difficulty of ILTM surgical removal is very important for predicting postoperative complications, appointment time, and procedure duration. The Pederson index [3] is the most commonly used difficulty index (DI) for impacted third molars. This index judges the difficulty on the basis of radiographs, considering the impacted tooth's angulation (Winter classification) [4], depth and ramus relationship (Pell and Gregory classification) [5].

Deep learning (DL), a form of machine learning, is playing an increasingly significant role in the fields of medicine and dentistry. There is a growing utilization of DL to assist in disease diagnosis, encompassing both radiological imaging and histopathological diagnosis [6, 7]. In the field of dentistry, numerous studies have shown that DL can play a crucial role in assisting in the diagnosis of various dental conditions, including periodontal disease, periapical inflammation, and even different types of lesions, such as lichen planus [8] and oral squamous cell carcinoma [7]. A study by Yang et al. [7] showed that with the assistance of DL, the accuracy and speed of oral squamous cell carcinoma diagnosis from histopathology images can be improved. While some studies have explored the application of DL in assisting with the angulation, class and position of impacted teeth [9, 10], to date, no studies have analyzed the DI derived from DL and compared it to that of human assessment. Therefore, the objective of this study was to report on the ability of DL to assess the angulation, position, and classification of ILTM. Additionally, we aimed to compare the accuracy of these parameters, DI and the time required for interpretation among DL models, sixth-year dental students (DSs), and general dental practitioners (GPs) currently performing impacted surgical removal.

Materials and methods

This study was conducted in accordance with the guidelines of.

the World Medical Association Helsinki Declaration for biomedical research involving human subjects and was approved by the Institutional Review Board of Rangsit University (COA DPE. No. RSUERB2022–064). All the data were analyzed anonymously.

Part A: Convolutional Neural Networks (CNNs) and classification of impacted tooth

Data Preparation

In this study, the dataset used for training and evaluation purposes was created by utilizing panoramic radiograph images from the College of Dental Medicine, Rangsit University, from 2019 to 2023. The dataset comprised 1200 cropped photographs of ILTMs from 994 patients (mean age of 26.8 years, standard deviation of 9.23, age range of 18–55 years, and 509 males and 485 females). These images were obtained via standardized imaging protocols that were rigorously adhered to throughout the acquisition process across the dataset. Only patients with ILTM and intact second molars were included in the study.

Patients with extensive carious lesions affecting both the ILTM and second molar, severe periodontal disease distal to the second molar, or any other bony defects that could affect parameter interpretation were excluded.

Panoramic radiographs of the patients were obtained via an X-Mind Trium (Acteon, Bangkok, Thailand) according to the 78 kV, 7 mA user manual.

The region around the mandibular third molar was manually cropped into a square shape ranging from 300 to 400 pixels, ensuring the inclusion of adjacent structures relevant for impacted tooth classification parameters. These structures include the mandibular ramus and the distal area of the mandibular second molar, as specified by the Winter classification and Pell Gregory classification. The parameters assessed within

this region include the angulation of ILTM according to the Winter classification, as well as the class and position of the ILTM according to the Pell and Gregory classification.

For angulation classification according to the Winter classification, ILTMs were classified by comparison with the long axis of the second molar. The data were divided into four of the most common classes: mesioangular, distoangular, horizontal, and vertical impaction. The description for each angulation is as follows:

> Mesioangular: ILTM is tilted toward the second molar in the mesial direction (from 11° to 79°).
> Distoangular: The long axis of ILTM is angled distally and posteriorly away from the second molar (from −11° to −79°).
> Horizontal: The long axis of ILTM is horizontal (from 80° to 100°).
> Vertical: The long axis of ILTM is parallel to the long axis of the second molar (from 10° to −10°).

Regarding ILTM class and position classification according to the Pell and Gregory classification, the ILTM class was classified on the basis of the positional relationship between the occlusal plane and the second molar, whereas the classification of the ILTM position was related to the occlusal and anterior margins of the mandibular ramus.

In Class I, there is sufficient space available between the anterior border of the ascending ramus and the distal aspect of the second molar for the eruption of ILTM.

In Class II, the space available between the anterior border of the ramus and the distal aspect of the second molar is less than the mesio-distal width of the crown of ILTM.

In Class III, the ILTM is completely embedded in the bone of the anterior border of the ascending ramus because of the absolute lack of space.

In position A, the occlusal plane of ILTM is at the same level as the occlusal plane of the second molar.

In position B, the occlusal plane of ILTM is between the occlusal plane and the cervical margin of the second molar.

In position C, the ILTM is below the cervical margin of the second molar.

The dataset, which included a diverse range of cases, was created by the consensus of two dentists serving as the gold standard, one being a board-certified oral and maxillofacial surgeon (CI) and one being a board-certified oral and maxillofacial diagnostician (KS). Both observers underwent calibration. Every image was meticulously checked to ensure accuracy, relevance, and the absence of biases or artifacts that might affect the performance of the classification models.

For model training, the distribution of images across the angulation, class, and position categories was carefully balanced to ensure a representative and comprehensive dataset.

A transfer learning approach using TensorFlow's Keras applications was implemented. This approach involves the utilization of several pretrained models, including RegNetY032 [11], DenseNet201 [12], InceptionResNetV2 [13], ResNetRS101 [14], InceptionV3 [15], and Xception [16]. These models, which were originally trained on the extensive ImageNet dataset [17], have the ability to recognize a wide range of features and patterns, making them an ideal starting point for classification tasks. We adapted these models to our specific needs in ILTM classification by replacing some of their final layers.

The dataset was divided into training, validation, and test groups, with each set containing images categorized on the basis of the angulation, position, and class of the ILTM. The training process involves fitting the models to the training data and validating their performance with the validation dataset. The models were subsequently evaluated on the test dataset to identify the best-performing model. The best model was used in the second part of the study. The detailed workflow of the study, including data segmentation and the subsequent training, validation, and testing processes, is comprehensively depicted in Fig. 1. The distributions of the data for the training, validation, and testing datasets are shown in Table 1.

Data preprocessing

The dataset used for this study consists of ILTM images, which undergo a series of preprocessing steps via the OpenCV library to increase their quality and detail. The preprocessing pipeline in this study can be divided into five steps, as illustrated in Fig. 2.

Although there are many ways to convert a color image to a grayscale image, in this study, contrast limited adaptive histogram equalization (CLAHE) was used to convert three color pixel images to grayscale pixels. The threshold for contrast limiting was set to 2.0, and the size of the contextual regions was set to 8×8 pixels. After CLAHE, global histogram equalization (GHE) was applied to adjust the contrast across the entire image, enhancing the overall visibility of features. The purpose of combining CLAHE and GHE was to produce an image with both enhanced local contrasts and a generally more balanced contrast across the entire image, which could also increase the overall visibility of features in an image.

Next, unsharp masking was used to sharpen the images, emphasizing edges and fine details. In this

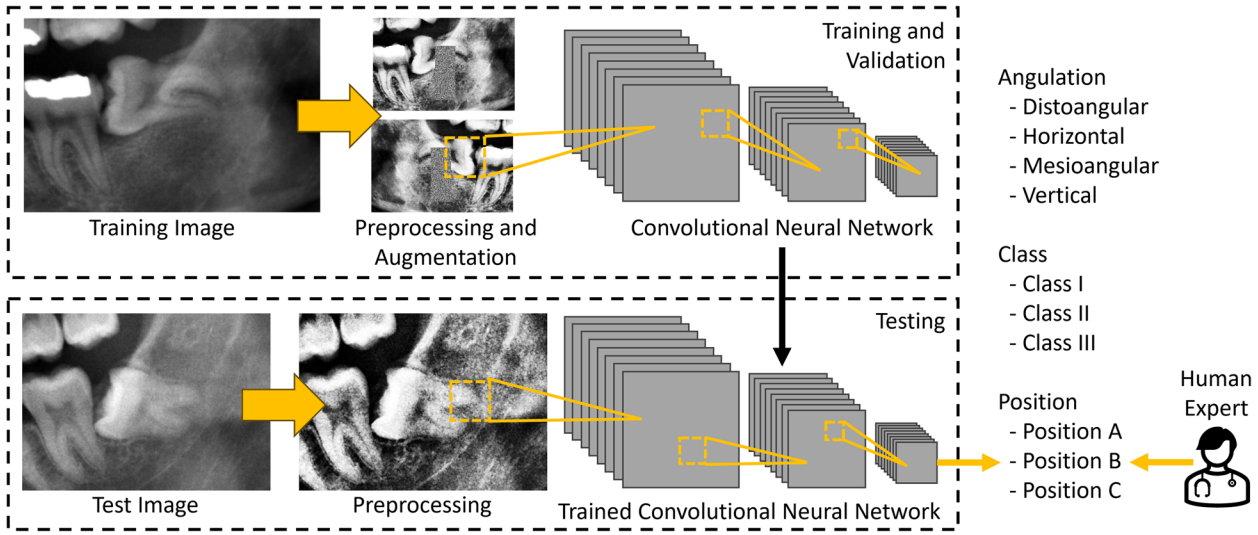

Fig. 1 Overview of the workflow for ILTM assessment via DL

Table 1 Summary of the dataset for each category in the training, validation, and test datasets

Class	Training	Validation	Test
Angulation			
Distoangular	50	10	11
Horizontal	293	10	14
Mesioangular	518	10	14
Vertical	318	10	11
Class			
I	180	10	20
II	200	10	20
III	27	10	10
Position			
A	655	10	18
B	518	10	18
C	27	10	14

step, a blurred image was created by applying Gaussian blur, with the Gaussian kernel size set to zero, to allow OpenCV to automatically choose the size of the Gaussian kernel on the basis of sigmaX, which was set to 2. Afterward, the images were combined with their blurred versions to create sharpened images by subtracting a fraction of the blurred image. Weights of 1.5 and −0.5 were used for the original and blurred images, respectively.

The morphological opening technique was subsequently used to increase image quality. This technique uses a structuring element with an elliptical shape, measuring 3×3 pixels. The elliptical form was specifically chosen for its ability to selectively target and eliminate smaller, irrelevant features from the image. This ensured that the larger, pertinent structures remained unaffected. The final step involved converting the processed images back to RGB format to ensure compatibility with the input requirements of our CNN models. Each of these steps played a vital role in preparing the images for accurate and efficient analysis in the study.

Model architecture

In this study, a transfer learning strategy was used, utilizing a variety of sophisticated pretrained models available through TensorFlow's Keras library. Those models included RegNetY032 [11], DenseNet201 [12], InceptionResNetV2 [13], ResNetRS101 [14], InceptionV3 [15], and Xception [16]. Each of these models was carefully selected for its proven effectiveness in various image recognition tasks and its ability to serve as a robust foundational architecture for specific dental classification challenges. The details of each model, as evaluated on the ImageNet dataset, are summarized in Table 2.

The key to harnessing the power of these pretrained models was in our approach to modifying them to suit our unique requirements. We began by retaining the original convolutional base of each model, which had been trained on the extensive ImageNet dataset, enabling them to recognize a wide array of general image features. We subsequently replaced the classifier of the predesigned model with our customized classification layers, as illustrated in Fig. 3.

The first layer added atop each model was a global average pooling layer. This was followed by a fully connected

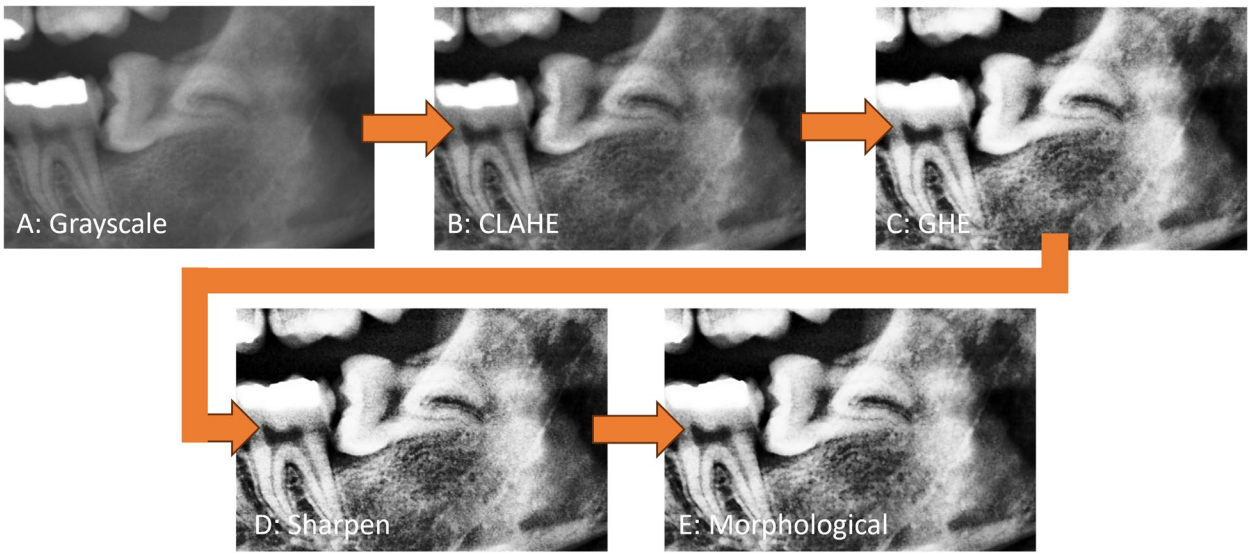

Fig. 2 Comprehensive image preprocessing steps: (**a**) conversion to grayscale, (**b**) contrast enhancement with adaptive histogram equalization, (**c**) tonal balance adjustment with global histogram equalization, (**d**) detail clarity through edge sharpening, and (**e**) noise reduction and artifact removal via morphological opening

Table 2 Comparative analysis of pretrained CNN models used for transfer learning

Model	Parameters	Test accuracy (%)
RegNetY032	19.4 M	79.00
DenseNet201	20.2 M	77.30
Xception	22.9 M	79.00
InceptionV3	23.9 M	77.00
ResNetRS101	29.4 M	77.09
InceptionResNetV2	55.9 M	80.30

dense layer with 256 units, utilizing the rectified linear unit (ReLU) activation function. To enhance the model's generalization capabilities and prevent overfitting, a dropout layer was incorporated next. The culmination of this architecture was a final dense layer equipped with a SoftMax activation function.

By meticulously integrating these custom layers into the pretrained architectures, we effectively adapted these powerful models to our specific task of classifying the angulation, class, and position of impacted teeth. This allowed us to harness the depth and breadth of learned features from diverse image data to enhance the precision and reliability of the model prediction.

Training process

The training was conducted with a workstation equipped with an NVIDIA GeForce RTX 3090Ti GPU (24 GB memory), utilizing Python 3.6 and TensorFlow 2.4 on Ubuntu. We used the Adam optimizer and categorical cross-entropy loss function, which are suitable for multiclass classification tasks. The training spanned 300 epochs, with both training and validation data used for continuous model adjustment. The key techniques included model checkpointing, where the model was saved at its highest validation accuracy, and a dynamic learning rate scheduler.

Crucially, data augmentation played a vital role in our training process. The training dataset was enriched through random horizontal flips and random erasing techniques [18], introducing essential variability. This approach not only combats overfitting but also allows the models to learn diverse features, which are important for accurate classification in complex scenarios. These collective strategies ensure a balanced and effective training regimen, optimizing model performance while efficiently managing computational resources.

Evaluation metrics

After the training phase was finished, the performance of each model on the test dataset was evaluated. The results were primarily summarized via two key performance metrics, accuracy and area under the curve (AUC), which are part of the receiver operating characteristic (ROC) analysis. Accuracy measures the proportion of total correct predictions, whereas the AUC reflects the model's ability to differentiate between classes. In conjunction with expert insights, these metrics were essential for selecting the best-performing model for further analysis.

In our multiclass predictive modeling, we used the one-vs.-all (OvA) strategy, where true positives (TPs)

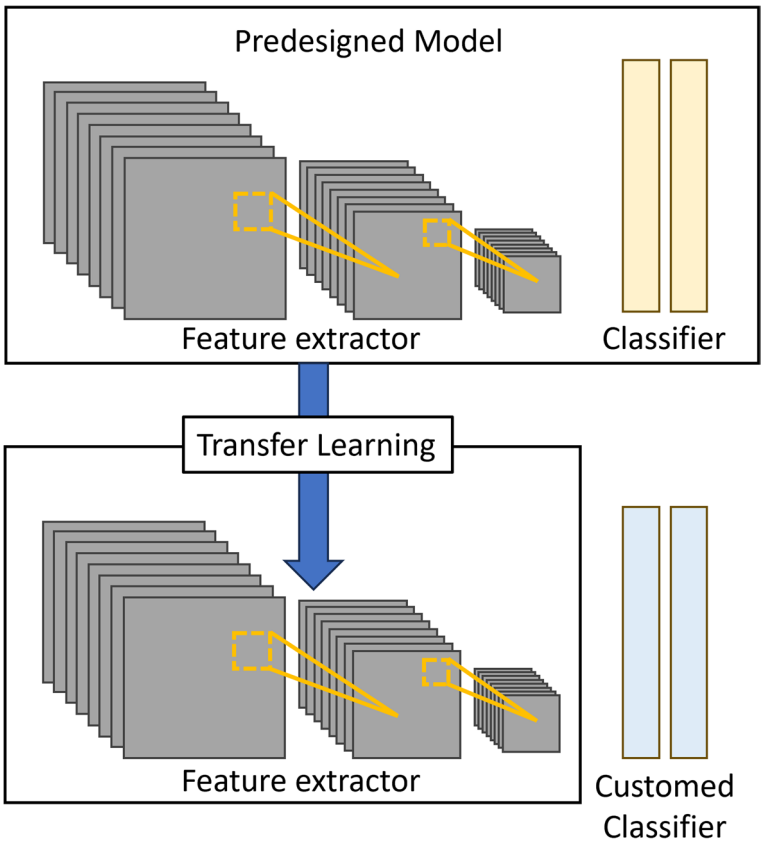

Fig. 3 Customized Architecture for Impacted Tooth Classification via Transfer Learning

represented correct identifications of a specific class, and true negatives (TNs) denoted correct rejections of the other classes. False positives (FPs) are instances where other classes are incorrectly identified as the specific class, whereas false negatives (FNs) occur when the specific class is mistakenly labeled as one of the other classes.

In addition to conventional evaluation metrics, in this study, we incorporated Score-CAM [19]. This advanced technique enhanced our understanding of CNN decision-making processes by creating heatmaps that indicated important areas in an image influencing the model's decisions. Crucial areas are highlighted in red, whereas less significant areas are marked in blue.

Part B: Comparison of model and human performance

The second phase of analysis involved a comparative evaluation between human experts and the best-performing CNN model in terms of accuracy and AUC using a set of 50 unseen images. These images were classified into angulation, class, and position by the best-performing CNN model from part A. The time used was also recorded. These same datasets were sent to DSs and GPs who still perform the ILTM as part of their routine dental practice. The correct answers were previously determined by human experts, representing the gold standard method.

All images for ILTM classification were sent to DSs and GPs via Google Forms as a questionnaire. There were two sets of questionnaires: the first consisted of a questionnaire without the best-performing CNN assistance, whereas Part II included the questionnaire with the best-performing CNN assistance, where the predictive responses of angulation, class, and position from the best-performing AI were shown. The duration from the first questionnaire set to the second questionnaire with CNN assistance was two weeks.

The participants were asked to classify the angulation, class, and position of the ILTM, both without (part I) and with (part II) the best-performing CNN model assistance. The time used for each image classification was recorded in the form. The DI of ILTM from the AI, GP and DS methods was calculated automatically after the questions were answered. The Pederson DI is shown in Table 3.

After receiving the questionnaire responses, the answers for ILTM angulation, class, and position and the DI for ILTM with and without AI assistance according to the Pederson index and the time used for both groups

Table 3 Pederson difficulty index for the removal of impacted lower third molars

Class	Difficulty index value
Angulation	
Mesioangular	1
Horizontal	2
Vertical	3
Distoangular	4
Class	
I	1
II	2
III	3
Position	
A	1
B	2
C	3

Table 4 Summary of the image distributions across the angulation, class, position, and DI categories of ILTM for the questionnaire

Class	N (%)
Angulation	
Mesioangular	18 (36)
Horizontal	17 (34)
Vertical	5 (10)
Distoangular	10 (20)
Class	
I	6 (12)
II	34 (68)
III	10 (20)
Position	
A	16 (32)
B	22 (44)
C	12 (24)
Difficulty index	
Minimally difficult	7 (14)
Moderately difficult	23 (46)
Very difficult	20 (40)

with and without best-performing CNN assistance were analyzed statistically. A summary of the image distributions across the angulation, class, position, and DI categories of ILTM for the questionnaire is shown in Table 4.

Statistical analysis

The accuracy and agreement evaluated by Cohen's kappa score compared with the gold standard in categorizing the angulation, class, position, DI of ILTM, and time used by the GPs ($N=35$) and DSs ($N=35$), both without and with the best-performing CNN model assistance, were calculated and reported as means and standard deviations.

Normal distribution and homogeneity of variances were assessed via the Shapiro–Wilk test and the Levene test, respectively. The differences in accuracy, kappa score, and time duration between the GP and 6th-year DS groups were examined via independent t tests, whereas the differences in these performance tests and time duration within each group (without and with CNN assistance) were examined via paired t tests. All analyses were performed with IBM SPSS Statistics version 29.0 (IBM). A p value < 0.05 was considered to indicate a statistically significant difference.

Results

Best performing model for ILTM classification

The selection of the best-performing models for the forthcoming comparative evaluation with human experts was meticulously carried out, considering both accuracy and AUC. For the Angulation classification, InceptionResNetV2 was chosen because of its superior accuracy, whereas RegNetY032 was also selected because of its impressive AUC. In the Class category, Xception was identified as the leading model for both accuracy and AUC values. With respect to the Position category, despite both InceptionResNetV2 and Xception achieving the highest AUC score, InceptionResNetV2 was preferred on the basis of its overall performance, particularly its marginally higher accuracy. A summary of the key performance metrics for each CNN is shown in Table 5.

The results reveal a notable trend in model accuracy and AUC, where InceptionResNetV2 and Xception generally maintain high levels of accuracy and AUC across all categories. For example, in the Angulation classification, InceptionResNetV2 leads with an accuracy of 0.88 and an AUC of 0.97, closely followed by RegNetY032 with slightly lower accuracy but a higher AUC of 0.98. In the Class category, Xception emerges as the top model, achieving the highest scores in both accuracy (0.78) and AUC (0.87). In the Position category, InceptionResNetV2 and Xception attained the highest AUC score of 0.96, with InceptionResNetV2 having a slightly higher accuracy (0.92) than Xception did (0.90).

Score-CAM visualizations, which provide insight into the decision-making process of the best-performing CNN models in categorizing ILTM, are shown in Fig. 4. In the Score-CAM images, the areas highlighted in red indicate the regions that the models prioritize when making classification decisions.

Table 5 Summary of the key performance metrics for the CNN models

Model	Accuracy	AUC	Processing time (seconds)
Angulation			
RegNetY032	0.82	**0.98**	3.08 ± 0.10
DenseNet201	0.76	0.96	5.57 ± 0.07
Xception	0.72	0.94	1.77 ± 0.07
InceptionV3	0.66	0.89	2.52 ± 0.08
ResNetRS101	0.28	0.54	5.14 ± 0.10
InceptionResNetV2	**0.88**	0.97	6.01 ± 0.04
Class			
RegNetY032	0.68	0.85	3.08 ± 0.10
DenseNet201	0.74	0.86	5.57 ± 0.07
Xception	**0.78**	**0.87**	1.77 ± 0.07
InceptionV3	0.60	0.74	2.52 ± 0.08
ResNetRS101	0.38	0.61	5.14 ± 0.10
InceptionResNetV2	0.62	0.83	6.01 ± 0.04
Position			
RegNetY032	0.76	0.87	3.08 ± 0.10
DenseNet201	0.60	0.85	5.57 ± 0.07
Xception	0.90	**0.96**	1.77 ± 0.07
InceptionV3	0.80	0.92	2.52 ± 0.08
ResNetRS101	0.60	0.83	5.14 ± 0.10
InceptionResNetV2	**0.92**	**0.96**	6.01 ± 0.04

In the Angulation category, the InceptionResNetV2 model focuses primarily on the space between the third and second molars. This observation was important because the angulation of impacted teeth can vary from vertical to horizontal, and this area is key to accurately classifying the angulation of ILTM.

In the Class category, the Xception model focused its attention mainly on the third molar, including the area between the occlusal plane of the second molar. This approach was consistent with classification methods that consider how deep the third molar is compared to the ramus.

In the Position category, the InceptionResNetV2 model examined a broader region that included the area up to the second molar. This matched the method of assessing the location of the impacted tooth from the Pell and Gregory classification criteria.

As shown by the performance metrics in Table 5 and the visualizations in Fig. 4, the models were capable of identifying key areas in radiographs that were significant for classifying ILTM, similar to the methods used by dentists. These insights assured us that the best-performing models were well suited for comparison with human experts in the task of automating this classification process.

A series of Score-CAM visualizations that provide insight into the decision-making process of the best-performing CNN models in categorizing impacted teeth is shown in Fig. 4. In these images, the areas highlighted in red indicate the regions that the models prioritize when making classification decisions.

Comparison of performance between DSs and GPs

A comparison of the performance tests, including accuracy and agreement evaluated by Cohen's kappa score compared with the gold standard, according to the angulation, class, position, DI of ILTM, and time duration of the GPs and DSs, both without and with the best-performing CNN model assistance, is shown in Table 6.

In both groups, the within-group comparison without and with the best-performing CNN model assistance revealed that the accuracy and kappa score with CNN assistance were significantly higher than those without CNN assistance (Fig. 5).

For between-group comparisons, performance tests without CNN assistance in all categories revealed no statistically significant difference between the GP and DS groups. However, GPs used a significantly shorter time duration for the inspection of class and total time than DSs did, which is consistent with the time duration when using CNN assistance.

With CNN assistance, the GPs showed significantly higher accuracy and kappa scores for the class category than did the DS group ($p = 0.035$ and 0.010, respectively). In contrast, the DS group with CNN assistance had higher accuracy and kappa scores for the position category than did the GPs ($p < 0.001$) (Fig. 6).

Discussion

In this study, we present the first data comparing AI in evaluating the class, angulation, and position of the ILTM, along with the overall assessment, with the Pederson index for assessing tooth difficulty and human performance. Additionally, we compared the outcomes of readings when best-performing AI assistance was present versus when it was not, within both the DS and GP groups actively involved in wisdom tooth extraction procedures.

In our study, some of the images in the distoangular, Class III, and Position C classes are relatively small compared with those in the other classes in the same category. Despite their small representation in the dataset, these classes provide important information because of their distinct features. Their inclusion ensures that our models are trained on a spectrum of scenarios, potentially leading to better accuracy and robustness. To

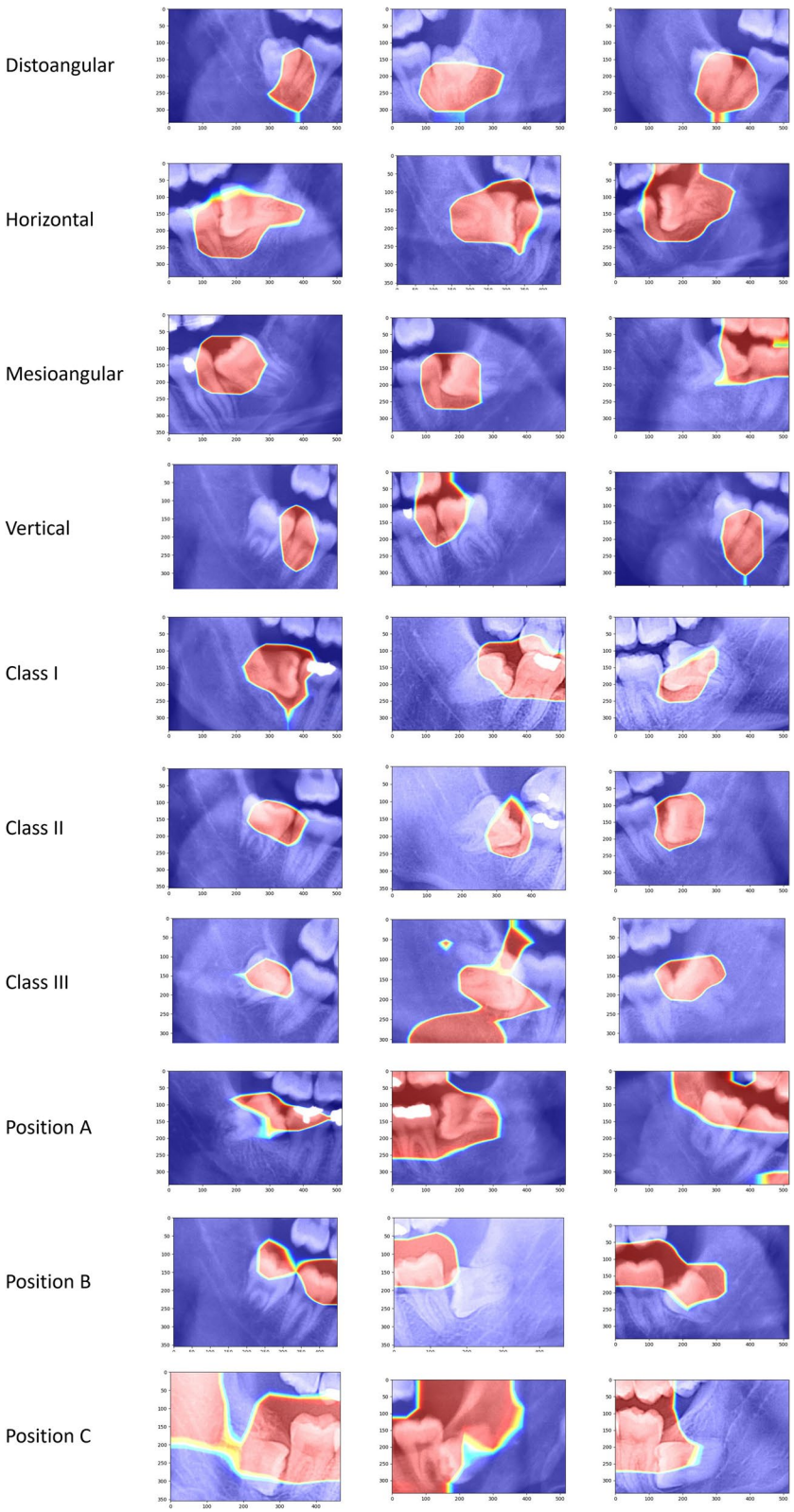

Fig. 4 Visualization of key regions via the score-cam for model decision analysis, categorized by angulation, class, and position, arranged from left to right

Table 6 Comparison of the performance tests according to the angulation, class, position, difficulty index of the ILTM, and time duration for GPs and DSs, with and without the best-performing CNN model assistance

Performance test, mean (SD)	Best CNN model	Without CNN-assistance	With CNN-assistance	*P* value[a]	Without CNN-assistance	With CNN-assistance	*P* value[a]	Without CNN-assistance	With CNN-assistance
		GPs, (*N* = 35)			DSs, (*N* = 35)			*P* value (GP vs. DS)[b]	
Accuracy									
Angulation	0.94	0.88 (0.05)	0.95 (0.03)	< 0.001	0.87 (0.05)	0.94 (0.05)	< 0.001	0.426	0.412
Class	0.94	0.67 (0.10)	0.86 (0.08)	< 0.001	0.67 (0.10)	0.81 (0.11)	< 0.001	0.922	0.035
Position	0.90	0.71 (0.05)	0.79 (0.05)	< 0.001	0.73 (0.07)	0.85 (0.06)	< 0.001	0.262	< 0.001
Difficulty index	0.96	0.74 (0.05)	0.88 (0.06)	< 0.001	0.73 (0.05)	0.86 (0.07)	< 0.001	0.849	0.364
Kappa score									
Angulation	0.92	0.83 (0.07)	0.93 (0.04)	< 0.001	0.82 (0.07)	0.92 (0.07)	< 0.001	0.525	0.437
Class	0.87	0.40 (0.12)	0.72 (0.14)	< 0.001	0.39 (0.12)	0.61 (0.20)	< 0.001	0.807	0.010
Position	0.85	0.54 (0.09)	0.68 (0.08)	< 0.001	0.58 (0.11)	0.77 (0.09)	< 0.001	0.143	< 0.001
Difficulty index	0.93	0.58 (0.08)	0.81 (0.09)	< 0.001	0.58 (0.08)	0.78 (0.11)	< 0.001	0.873	0.319
Duration (min)									
Angulation	0.10	4.01 (0.89)	2.59 (0.59)	< 0.001	4.44 (1.30)	2.84 (0.73)	< 0.001	0.118	0.124
Class	0.03	4.16 (1.05)	3.05 (0.79)	< 0.001	4.86 (1.49)	3.80 (1.06)	< 0.001	0.027	0.001
Position	0.10	3.70 (0.81)	2.61 (0.59)	< 0.001	4.02 (1.25)	2.68 (0.95)	< 0.001	0.209	0.713
Total	0.23	11.88 (1.73)	8.25 (1.46)	< 0.001	13.31 (2.81)	9.32 (2.40)	< 0.001	0.013	0.029

A statistically significant difference is indicated in bold (p value < 0.05)

[a] Differences between groups were analyzed by paired t tests

[b] Differences within groups were analyzed by independent t tests

develop the model used for classification, the best model that is reliable must include every aspect of the disease category.

To address the smaller sample sizes of certain classes, data augmentation has played a significant role. By artificially enhancing our dataset through techniques such as horizontal flipping and random erasing [10], we have been able to simulate a broader range of dental scenarios. This augmentation not only compensates for the lack of data in underrepresented classes but also helps prevent overfitting, resulting in models that are better generalized and more robust in their predictive capabilities.

There are promising avenues we have yet to explore that could further enrich our dataset and enhance model performance. The use of generative adversarial networks (GANs) is one such option. GANs have the potential to generate new, synthetic images of impacted teeth, providing a more extensive and varied dataset for model training [12–14]. This approach could be particularly beneficial for augmenting underrepresented classes, offering a novel solution to the challenge of data scarcity.

The assessment of the position, class, and angulation of the ILTM is crucial in evaluating the difficulty of the surgical removal procedure. It plays a significant role in determining the appointment time for the patient and is highly beneficial in making preliminary assessments regarding whether a referral to an oral and maxillofacial surgeon is necessary.

In a setting without many patients, assessing using the Pederson index can be relatively easy, but it is time-consuming, especially when dealing with a substantial patient volume. This is due to the need for angle measurements to evaluate various parameters. The incorporation of AI in assessing the Pederson index can significantly expedite the process, making it much more efficient.

The development of AI from this study, particularly the use of CNNs with high accuracy, can be applied to various user-friendly applications. Dentists, for example, can upload images into the application and use the AI-generated values to expedite the assessment of the Pederson index. Our study findings indicate that the accuracy and kappa score with CNN assistance were significantly greater than those with assessments without CNN assistance. This suggests the potential for improved and more reliable results when using CNNs in dental assessments.

Our study achieved accuracy values for class, position, and angulation that are close to those reported by Kim et al.: 78.91% for position, 82.03% for class, and 90.23% for angulation [20]. However, it is worth noting that those authors did not compare these values with human

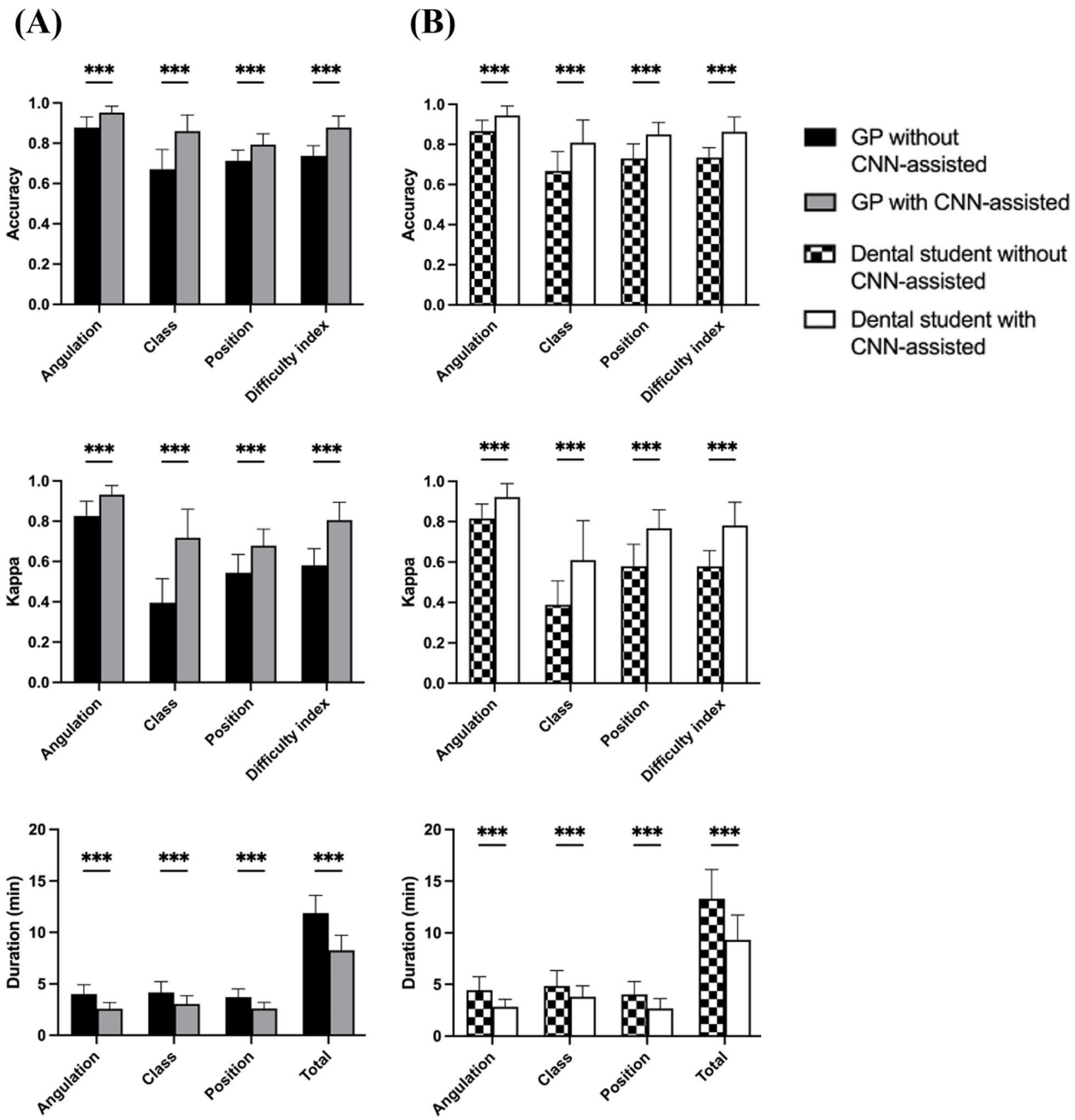

Fig. 5 Performance tests and time duration comparisons between patients without and with the best-performing CNN model assistance, according to (**A**) GPs and (**B**) DSs. Statistically significant differences were assessed via paired *t* tests. *** $p < 0.001$

assessments in each category, and there was no evaluation of the Pederson index derived from reading class, position, and angulation.This suggests that while the accuracy values are similar, our study provides additional insights by comparing results with human assessments and evaluating the Pederson index, contributing valuable information to the field.

Although several studies have shown that the Pederson index may not be the best DI for ILTM, it is widely used, relatively simple, and easy to apply. Other clinical considerations, such as trismus, tongue size, angulation of the external oblique ridge, cheek flexibility [21], bone density [22], and patient age [23], are used in other indices, which are more complicated and not practical. The Pernambuco index, which incorporates clinical information

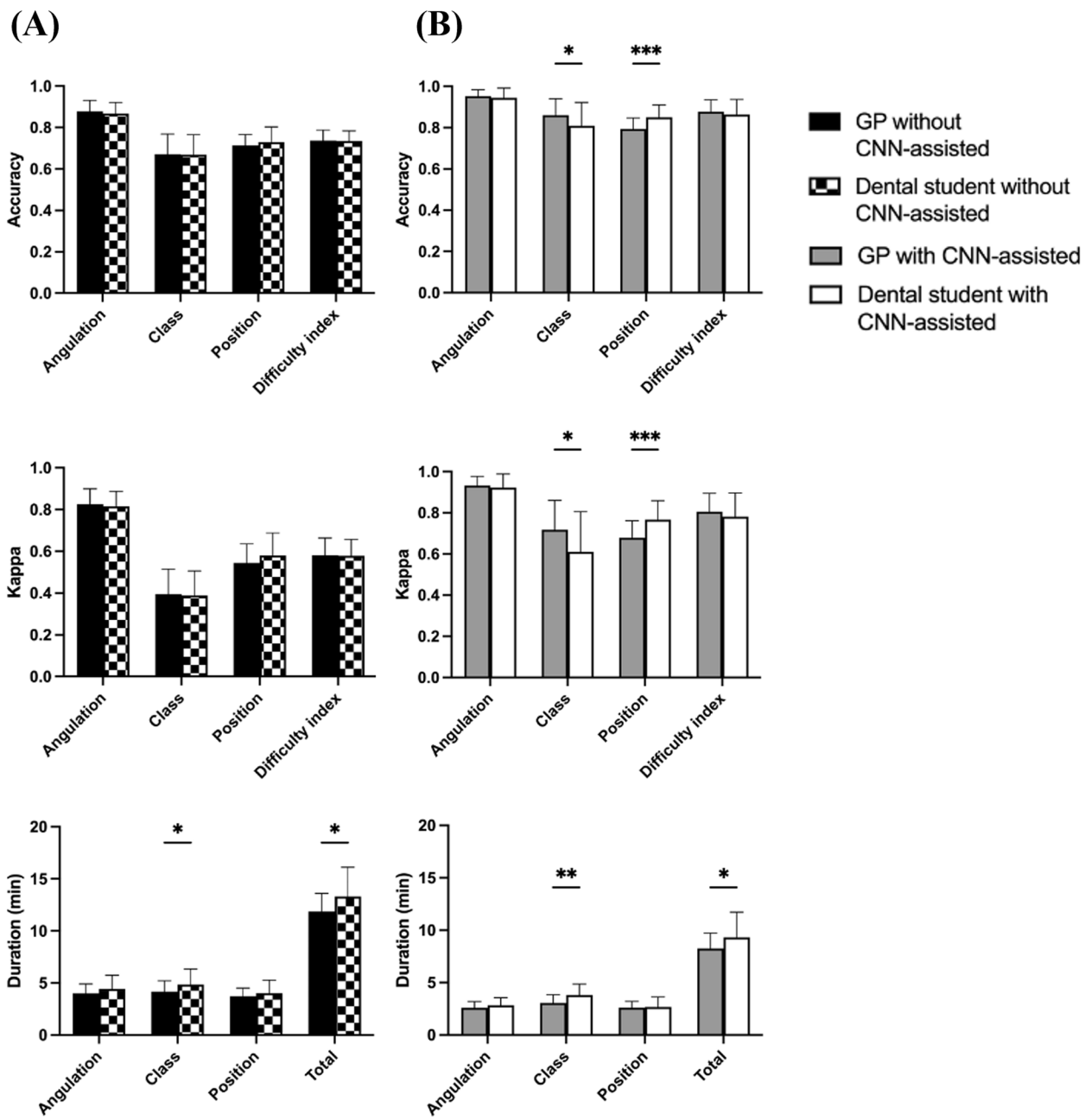

Fig. 6 Performance tests and time duration comparisons between the GP and DS groups: (**A**) without CNN assistance and (**B**) with CNN assistance. Statistically significant differences were assessed by independent *t* tests. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. 0.01; *** $p < 0.001$

such as the number of roots, root curvature, relationship to the second molar, patient age, and the patient's body mass index (BMI), is an interesting index that could be explored for experimentation with various parameters in CNNs [24].

The Pederson index did not significantly differ between the DSs and GPs. However, when AI assistance was introduced, compared with DSs, GPs completed the task in a significantly shorter amount of time. This suggests that more experienced individuals were able to benefit more from AI, as it contributed to increased accuracy and efficiency in their assessments. Although our study demonstrated that AI can accurately assess DI to a satisfactory extent, it does not imply that AI will replace humans in DI assessment, radiological image interpretation, or

histopathology reading. The suitable role for AI lies in being an auxiliary technique in medical practices.

Several studies indicate that the use of AI as an assistive tool accelerates and enhances the accuracy of medical tasks. For example, Choo et al. [25] reported that 80% of clinicians changed their initial predictions at least once after the model's prediction was conveyed to them. Moreover, 90.53% of these changes influenced the accuracy of the outcome [25].

The study conducted by Habib et al., which involved the assessment of images of tympanic membrane perforation, revealed a clear reduction in diagnostic error rates with the use of AI assistance [26]. Additionally, the study by Yang et al. on oral squamous cell carcinoma (OSCC), which compared junior and senior pathologists, demonstrated that AI assistance led to an improvement in the F1 score. Specifically, it increased from 0.9221 to 0.9566 for junior pathologists and from 0.9361 to 0.9463 for senior pathologists [7]. This affirms that DL can increase the accuracy and speed of diagnosis.

Conclusion

The CNN achieved classification accuracies between 87 and 96% for ILTM. With CNN support, both DSs and GPs demonstrated notably improved accuracy in ILTM classification. Furthermore, GPs required significantly less time for class inspection and overall evaluation compared to DSs, whether or not the DSs utilized CNN assistance.

Acknowledgements

We thank Princess Srisavangavadhana College of Medicine, Chulabhorn Royal Academy, Faculty of Dentistry Chulalongkorn University and College of Dental Medicine, Rangsit University, for their support.

Disclosures

This study was approved by the Human Research Ethics Committee, Rangsit University, and was conducted in accordance with the Declaration of Helsinki and adhered to the CONSORT 2010 statement.
This study was funded by Rangsit University, Pathum Thani, Thailand.

Authors' contributions

Paniti Achararit: First author, Conceptualization, Study design, Methodology, Software, Validation, Data analysis, Investigation, Resources, Writing–original draft. Chawan Manaspon: Conceptualization, Study design. Chavin Jongwannasiri: Conceptualization, study design. Promphakkon Kultanaamondhita: Data analysis, Investigation, Resources, Writing–original draft. Chumpot Ittichaisri: Conceptualization, Study design, Validation, Data analysis, Investigation, Resources, Writing–original draft. Soranun Chandrangsu: Conceptualization, Study design, Methodology, Software, Validation, Data analysis, Investigation, Resources, Writing–original draft. Thanaphum Osathanon: Conceptualization, Study design, Methodology, Validation, Data analysis, Investigation, Resources, Writing–original draft. Ekarat Phattarataratip: Conceptualization, Study design, Methodology, Validation, Data analysis, Investigation, Resources, Writing–original draft, Corresponding author. Kraisorn Sappayatosok: Conceptualization, Study design, Methodology, Collection of data, Software, Validation, Data analysis, Investigation, Resources, Writing–original draft, Corresponding author.

Funding

This study was funded by Rangsit University, Pathum Thani, Thailand.

Data availability

The datasets generated and/or analyzed during the present study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study was approved by the Human Research Ethics Committee of Rangsit University, was conducted in accordance with the Declaration of Helsinki and adhered to the CONSORT 2010 statement. The informed consent to participate was obtained from all of the participants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Princess Srisavangavadhana Faculty of Medicine, Chulabhorn Royal Academy, Bangkok 10210, Thailand. [2]Biomedical Engineering Institute, Chiang Mai University, Chiang Mai 50200, Thailand. [3]Bangkok Hospital Dental Center Holistic Care and Dental Implant, Bangkok Hospital, Bangkok 10310, Thailand. [4]College of Dental Medicine, Rangsit University, PathumThani 12000, Thailand. [5]Department of Oral Pathology, Faculty of Dentistry, Chulalongkorn University, Bangkok 10330, Thailand. [6]Center of Excellence for Dental Stem Cell Biology, Department of Anatomy, Faculty of Dentistry, Chulalongkorn University, Bangkok 10330, Thailand.

Received: 20 September 2024 Accepted: 3 January 2025
Published online: 28 January 2025

References

1. Cankurtaran CZ, Branstetter BFt, Chiosea SI, Barnes EL, Jr.: Best cases from the AFIP: ameloblastoma and dentigerous cyst associated with impacted mandibular third molar tooth. Radiographics 2010, 30(5):1415–1420.
2. Kanneppady SK. Balamanikandasrinivasan, Kumaresan R, Sakri SB: A comparative study on radiographic analysis of impacted third molars among three ethnic groups of patients attending AIMST Dental Institute. Malaysia Dent Res J (Isfahan). 2013;10(3):353–8.
3. GW. P: Oral Surgery. Philadephia: WB Saunders; 1988.
4. GB W: Impacted mandibular third molar. St Louis: American Medial Book Co.; 1926.
5. Pell GJGG. Impacted mandibular third molars: classification and modified techniques for removal. Dent Dig. 1933;39:330–8.
6. Duron L, Ducarouge A, Gillibert A, Laine J, Allouche C, Cherel N, Zhang Z, Nitche N, Lacave E, Pourchot A, et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. Radiology. 2021;300(1):120–9.
7. Yang SY, Li SH, Liu JL, Sun XQ, Cen YY, Ren RY, Ying SC, Chen Y, Zhao ZH, Liao W. Histopathology-Based Diagnosis of Oral Squamous Cell Carcinoma Using Deep Learning. J Dent Res. 2022;101(11):1321–7.
8. Achararit P, Manaspon C, Jongwannasiri C, Phattarataratip E, Osathanon T, Sappayatosok K. Artificial Intelligence-Based Diagnosis of Oral Lichen Planus Using Deep Convolutional Neural Networks. Eur J Dent. 2023;17(4):1275–82.
9. Celik ME: Deep Learning Based Detection Tool for Impacted Mandibular Third Molar Teeth. Diagnostics (Basel) 2022, 12(4).
10. Sukegawa S, Matsuyama T, Tanaka F, Hara T, Yoshii K, Yamashita K, Nakano K, Takabatake K, Kawai H, Nagatsuka H, et al. Evaluation of multi-task learning in deep learning-based positioning classification of mandibular third molars. Sci Rep. 2022;12(1):684.

11. Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P: Designing network design spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 2020; 2020: 10428–10436.
12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition: 2017; 2017: 4700–4708.
13. Szegedy C, Ioffe S, Vanhoucke V, Alemi A: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence: 2017; 2017.
14. Bello I, Fedus W, Du X, Cubuk ED, Srinivas A, Lin T-Y, Shlens J. Zoph BJAiNIPS: Revisiting resnets: Improved training and scaling strategies. 2021;34:22614–27.
15. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition: 2016; 2016: 2818–2826.
16. Chollet F: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition: 2017; 2017: 1251–1258.
17. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition: 2009: Ieee; 2009: 248–255.
18. Zhong Z, Zheng L, Kang G, Li S, Yang Y: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence: 2020; 2020: 13001–13008.
19. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Mardziel P, Hu X: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops: 2020; 2020: 24–25.
20. Kim BS, Yeom HG, Lee JH, Shin WS, Yun JP, Jeong SH, Kang JH, Kim SW, Kim BC: Deep Learning-Based Prediction of Paresthesia after Third Molar Extraction: A Preliminary Study. Diagnostics (Basel) 2021, 11(9).
21. Roy I, Baliga SD, Louis A, Rao S. Importance of Clinical and Radiological Parameters in Assessment of Surgical Difficulty in Removal of Impacted Mandibular 3rd Molars: A New Index. J Maxillofac Oral Surg. 2015;14(3):745–9.
22. Sammartino G, Gasparro R, Marenzi G, Trosino O, Mariniello M, Riccitiello F. Extraction of mandibular third molars: proposal of a new scale of difficulty. Br J Oral Maxillofac Surg. 2017;55(9):952–7.
23. Zhang X, Wang L, Gao Z, Li J, Shan Z: Development of a New Index to Assess the Difficulty Level of Surgical Removal of Impacted Mandibular Third Molars in an Asian Population. J Oral Maxillofac Surg 2019, 77(7):1358 e1351–1358 e1358.
24. de Carvalho RWF, Vasconcelos BC. Pernambuco index: predictability of the complexity of surgery for impacted lower third molars. Int J Oral Maxillofac Surg. 2018;47(2):234–40.
25. Choo H, Yoo SY, Moon S, Park M, Lee J, Sung KW, Cha WC, Shin SY, Son MH. Deep-learning-based personalized prediction of absolute neutrophil count recovery and comparison with clinicians for validation. J Biomed Inform. 2023;137: 104268.
26. Habib AR, Wong E, Sacks R, Singh N. Artificial intelligence to detect tympanic membrane perforations. J Laryngol Otol. 2020;134(4):311–5.

Publisher's Note