End-to-end vs. human-defined feature extraction: comparing deep learning approaches for age classification using mandibular third molars

Witsarut Upalananda[1], Arnon Charuakkra[2], Sitthichok Chaichulee[3]

[1] Department of Oral Diagnostic Sciences, Faculty of Dentistry, Prince of Songkla University, Songkhla, Thailand, [2] Department of Oral Biology and Oral Diagnostic Sciences, Faculty of Dentistry, Chiang Mai University, Chiang Mai, Thailand, [3] Department of Biomedical Sciences and Biomedical Engineering, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand

Corresponding author:
sitthichok.c@psu.ac.th

The authors declare that they have no conflict of interest.

KEYWORDS

Dental age estimation,
Deep learning,
Mandibular third molar,
Forensic odontology

J Forensic Odontostomatol
2025. Dec; (43): 3 -20:30
ISSN :2219-6749
DOI: doi.org/10.5281/zenodo.17776415

ABSTRACT

Accurate age classification using mandibular third molar radiographs is crucial for legal and forensic applications. This study evaluated different methods for classifying age as under or over 18 years in a Thai population. We compared three approaches: (i) a traditional human-based method using a modified Demirjian classification adapted for mandibular third molars, (ii) an end-to-end deep learning model in which a convolutional neural network (CNN) directly predicts age group, and (iii) a human-defined feature extraction approach, where a CNN estimates tooth developmental stages that are subsequently used for age classification. The dataset included 3,407 images of individuals aged 14–23 years. The results indicated that the traditional human-based method achieved high specificity (0.99) and a strong Bayes' post-test probability (0.99), but it exhibited low sensitivity (0.45). In comparison, the end-to-end deep learning models showed higher sensitivity (0.65 to 0.74) than the traditional method, along with a specificity of 0.91 to 0.95 and Bayes' post-test probability of 0.93 to 0.95. The human-defined feature extraction approach, which used developmental stages for age determination, achieved an accuracy of 0.88 to 0.92 in developmental stage classification. For age classification, the models demonstrated higher specificity (0.95 to 0.97) and Bayes' post-test probability (0.95 to 0.97) than the end-to-end deep learning method, along with sensitivity ranging from 0.51 to 0.56. Our results indicate that although traditional methods excel in specificity, the human-defined feature extraction approach provides a balanced solution with high specificity and interpretability, suggesting its potential value in clinical practice for age estimation.

INTRODUCTION

Age estimation is crucial in legal contexts, with 18 years widely recognised as the threshold for adulthood, as defined by international standards such as the United Nations Convention on the Rights of the Child.[1] Dental age estimation is an integral part of forensic odontology, providing a reliable method for determining an individual's age by evaluating dental development.[2] Unlike skeletal age estimation, dental age estimation is less affected by environmental factors, enhancing its reliability.[3]

Dental radiographs are a non-invasive, straightforward, and

cost-effective method, providing valuable insights into the maturation process of teeth.[4] By the age of 15 years, human teeth, except the third molars, are fully developed,[5] making third molars crucial indicators for age estimation beyond this age. The Demirjian method is a commonly accepted technique for dental age estimation. This involves the analysis of the tooth development patterns using panoramic radiographs to predict an individual's age.[6] Despite its reliability, this method can suffer from observer errors, necessitating rigorous training to minimise inaccuracies.[7]

Deep learning is increasingly used in biomedical imaging, including widespread applications in dental age estimation.[8] Using artificial intelligence (AI) for mandibular third molar classification in legal contexts, previous studies have investigated the application of AI in assessing developmental stages of the mandibular third molar,[9] as well as performing binary age classification at legal thresholds of 14, 16, and 18 years.[10] Although developmental stage classification offers valuable insights, it does not directly provide a dental age. In contrast, end-to-end binary classifiers often function as "black boxes," limiting interpretability and clinical acceptance.[11] Given the legal implications of age estimation, enhancing automation, accuracy, and interpretability is essential for forensic use. A commonly used architecture is the convolutional neural network (CNN), a core component of deep learning. CNNs operate through a hierarchical cascade of layers that automatically extract and refine important features from raw image data. This makes them well-suited for efficient and accurate classification tasks in medical imaging, including radiographic analysis.[12]

This study thus aimed to explore and compare various deep learning approaches for predicting whether individuals are under or over 18 years of age, using mandibular third molar radiographs. We evaluated three approaches: (i) a traditional method using expert assessment of mandibular third molar development based on a modified Demirjian classification, (ii) an end-to-end deep learning model where a CNN directly predicts age group from the radiograph, and (iii) a human-defined feature extraction approach, where a CNN first predicts the tooth's developmental stage, which is then used to classify the age group.

MATERIALS AND METHOD

Data collection

The study was approved by the Faculty of Dentistry Human Experimentation Committee of the Faculty of Dentistry, Chiang Mai University (approval no. 60/2022) and the Faculty of Medicine Human Research Ethics, Prince of Songkla University (approval no. REC.66-235-38-2). It was registered with the Thai Clinical Trial Registry (TCTR identification number: TCTR20230519002).

We used a subset of the dataset from the previous study.[9] Digital panoramic radiographs of 1,872 patients (831 male and 1,041 female) were randomly collected from the database of the Dental Hospital of the Faculty of Dentistry, Chiang Mai University, between 2012 and 2019. Patients aged 14–23 years who underwent radiographic examination and had available data on their birth date and date of radiographic examination were included in the study. Patients were excluded whose radiographic images were of poor quality, who had missing or malaligned mandibular third molars (severe buccoversion or linguoversion), or who had developmental anomalies, jawbone pathology, or syndromes affecting the dental development.

Mandibular third molar images were obtained from radiographic examinations, cropped, and rotated to align along the tooth axis at 224 × 336 pixels. Age was calculated by subtracting the birth date from the examination date and used as ground truth. Images were categorised into two groups: patients aged under 18 years and those aged 18 years or older. The data were randomly split into training, validation, and test sets (70:15:15). The training set was used to develop deep learning and human-based prediction models, the validation set to identify the optimal deep learning models and the test set to evaluate both methods. The study workflow and sample distribution are presented in Figure 1 and Table 1.

Traditional human-based approach

The mandibular third molars' developmental stage was assessed by an expert who reported almost perfect intra- and inter-observer agreement in previous studies.[9] The assessment followed a modified version of the Demirjian method,[6] in which the eight developmental stages (Stages A to H), originally proposed for the mandibular first and second molars, were applied to the mandibular third molar, as previously

implemented by Duangto et al. (2017) in a Thai population.[13] However, since the patients in our study were aged 14 to 23 and the development of mandibular third molars begin at 7 to 8 years, we observed only the later stages (Stages D to H) (Figure 2). Individuals with mandibular third molars at Stage H were considered 18 years or older, while those at lower stages were considered younger.[14] We evaluated the method's performance using the test set.

Figure 1. Overall workflow of the study (A) data collection workflow (B) research workflow

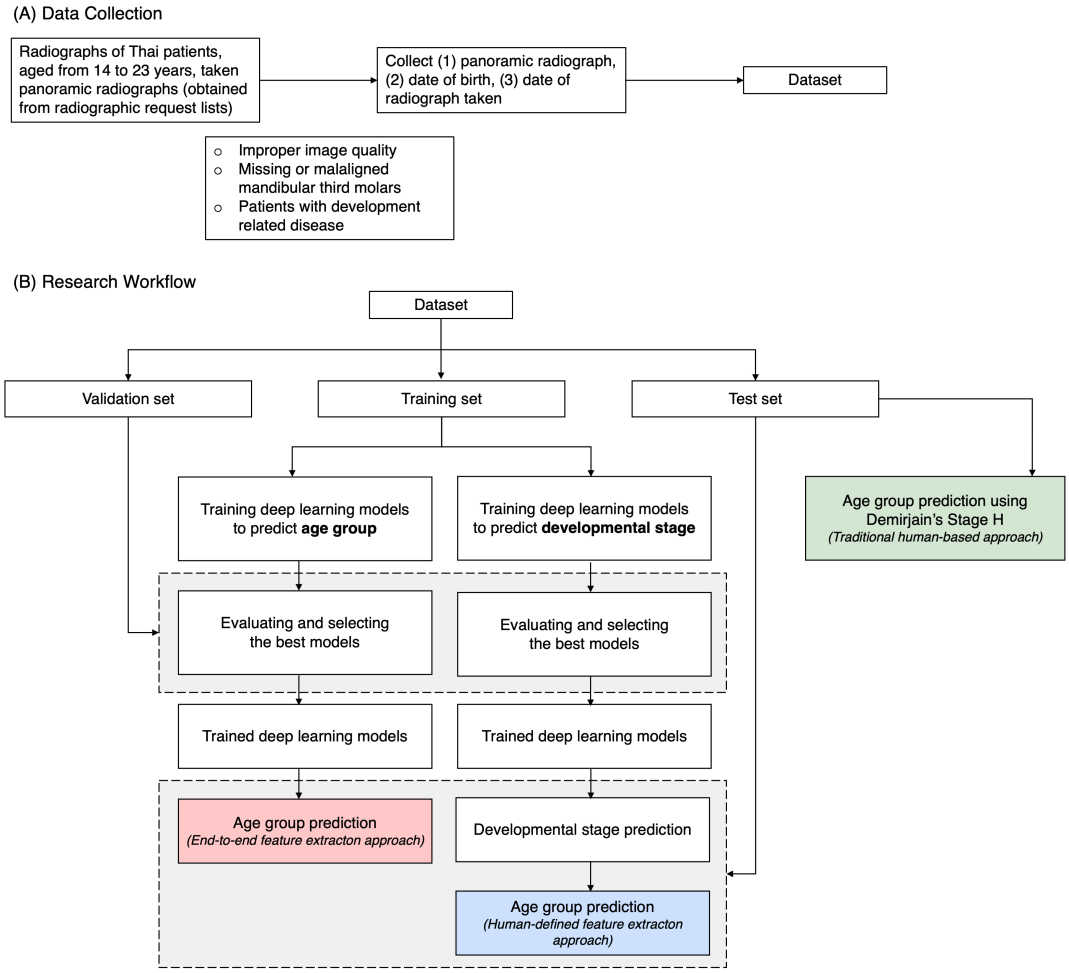
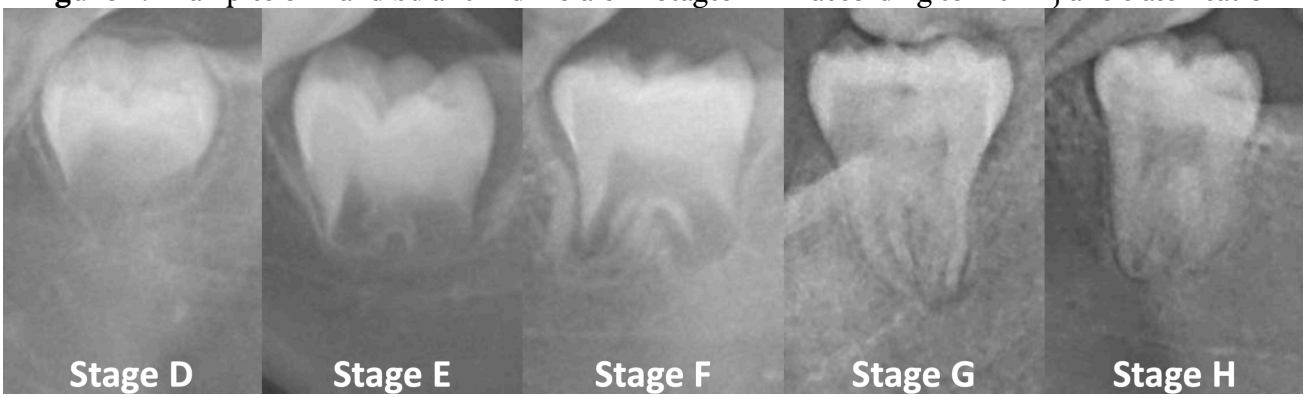


Table 1. Number of mandibular third molar images in the training, validation, and test set

	Training (n=2,385)	**Validation (n=511)**	**Test (n=511)**
	n (%)	n (%)	n (%)
Mandibular third molar			
Left	1,180 (49.48)	267 (52.25)	251 (49.12)
Right	1,205 (50.52)	244 (47.75)	260 (50.88)
Age group			
Less than 18 years old	1,259 (52.79)	294 (57.53)	279 (54.60)
More than 18 years old	1,126 (47.21)	217 (42.47)	232 (45.40)
Developmental stage			
Pre-H Stages	1819 (76.27)	384 (75.15)	404 (79.06)
Stage H	566 (23.73)	127 (24.85)	107 (20.94)

Figure 2. Examples of mandibular third molars in Stages D–H according to Demirjian's classification



End-to-end feature extraction approach

For the end-to-end approach, deep learning models were trained to classify individuals as either under or over 18 years of age directly from mandibular third molar radiographs. Several CNN architectures were evaluated for this task: ResNet (ResNet-50 and ResNet-101),[15] DenseNet (DenseNet-121 and DenseNet-169),[16] and EfficientNet (EfficientNet-B0 and EfficientNet-B2),[17] were used for end-to-end age group classification. Age groups were labelled as 0 for individuals under 18 years and 1 for individuals 18 years or older. Training included data augmentation (random rotation, horizontal flipping, zooming and translation), with cross-entropy loss and the ADAM optimiser. Training parameters were initial learning rate 0.00005, weight decay 0.0001, 100 epochs, and batch size of six. In forensic age estimation, avoiding misclassification of minors as adults is critical.[18] Thus, precision was prioritised. The best model was selected based on the highest precision score on the validation set.

Human-defined feature extraction approach

This approach used the same deep learning models and training parameters as the end-to-end method. Data augmentation techniques, the loss function, and the optimiser were similarly applied. However, in contrast to direct age classification, these models were trained to predict the developmental stage of the mandibular third molar. In this approach, labels were assigned based on human evaluation: 0 for "pre-H stages" and 1 for "Stage H." Unlike the end-to-end method, the models' primary objective was accurate classification of mandibular third molar developmental stages. The predicted stage was subsequently used to determine age group, with Stage H indicating individuals 18 years or older, and pre-H stages indicating individuals under 18 years. Model selection on the validation set was based on the highest F1-score. An overview of the age classification approaches is presented in Figure 3.

Experimental setup

The deep learning–based models were developed using Python 3.8.10, PyTorch 1.12.0 and MONAI 0.10.dev2229 with CUDA 11.7 and CuDNN 8.6.0. All experiments were performed on a workstation equipped with a 4-core processor, with 16 GB of RAM and an NVIDIA RTX 2080 8GB graphics card.

Performance evaluation

The models estimated age groups at the 18-year threshold on the test set. Predictions were classified as true positives (TP) for correctly identified individuals older than 18 years, true negatives (TN) for those correctly identified as younger than 18 years, false positives (FP) for individuals under 18 incorrectly predicted as older and false negatives (FN) for those over 18 incorrectly predicted as younger.

These matrices were used to calculate the performance matrices of the validation and testing procedures in terms of accuracy, sensitivity, specificity, precision, and F1-score. These metrics were calculated as follows (Equations 1–5):

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

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

$$\text{Specificity} = \frac{TN}{TN + FP}$$

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

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

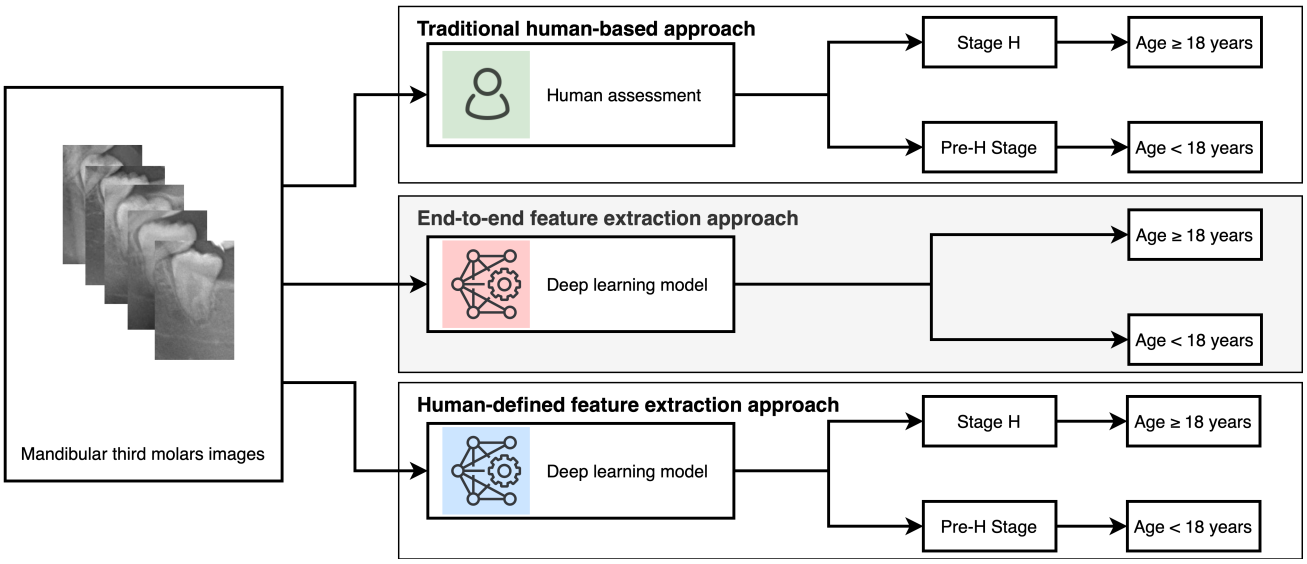
We also calculated the Bayes' post-test probability. This measures the likelihood of a condition after testing, considering pre-test probability and test performance. In legal age classification, a high Bayes' post-test probability ensures reliable predictions for individuals being 18 years or older, which is crucial for legal accuracy.[19] Minimising false positives is essential to avoid wrongful adult classification, ensuring correct legal responsibilities and protections. The calculation of Bayes' post-test probability is as follows (Equation 6):

$$\text{Bayes' post-test probability} = \frac{\text{Sensitivity} \times p_0}{\text{Sensitivity} \times p_0 + (1\text{-Specificity})(1-p_0)}$$

When calculating the Bayes post-test probability, p_0 represents the likelihood that an individual belongs to the age range of 18 to 23 years, given that their age falls between 14 and 23 years. To determine this probability, we computed p_0 by assessing the proportion of individuals aged 18 to 23 years among those within the 14 to 23 years age bracket in the Thai population. Our investigation, based on data sourced from the Thai National Statistical Office (http://statbbi.nso.go.th), unveiled that the calculated p_0 for this specific demographic subgroup stands at 0.61.

In the human-defined feature extraction approach, models classified individuals into developmental stages. True positives were those correctly identified as Stage H, true negatives as pre-H stages, false positives as pre-H stages misclassified as Stage H, and false negatives as Stage H misclassified as pre-H stages. Performance metrics were calculated using the same criteria as before.

Figure 3. Overview of different approaches to age classification including traditional human-based approach, end-to-end deep learning models, and human-defined feature extraction approaches



RESULTS

The results of all approaches to classify age groups at the 18-year threshold are shown in Table 2. The traditional human-based approach achieved a specificity score of 0.99, indicating that nearly all individuals identified as 18 years or older were correctly classified. However, the sensitivity was low at 0.45, indicating that only 45% of individuals aged 18 or older were correctly identified as adults, and the remaining 55% were misclassified as minors. Considering Bayes' post-test probability, this method achieved a value of 0.99, indicating a high probability that individuals predicted to be 18 years or older are so.

The end-to-end deep learning approach, an AI-based method with a straightforward idea, achieved higher scores in accuracy and F1-score than the traditional method. This indicates that this approach can successfully classify individuals

as younger or older than 18 years. Among all models, DenseNet-121 achieved the highest accuracy score and F1-score of 0.85 and 0.82, respectively. However, considering the legal context where false positives are the most serious cases to avoid, we further examined the precision score, which represents the accuracy of positive predictions. Although this approach showed improved balance, its lower precision (0.87–0.91) resulted in a higher risk of misclassifying minors, as reflected in slightly reduced Bayes' post-test probabilities (0.93–0.95) relative to the traditional method.

Table 2. Number of mandibular third molar images in the training, validation, and test set

Approach	Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	p
Traditional		0.75	0.45	0.99	0.98	0.62	0.99
End-to-end feature extraction	EfficientNet-B0	0.84	0.72	0.93	0.89	0.80	0.94
	EfficientNet-B2	0.79	0.62	0.92	0.87	0.73	0.93
	DenseNet-121	0.85	0.76	0.92	0.89	0.82	0.94
	DenseNet-169	0.83	0.74	0.91	0.88	0.80	0.93
	ResNet-50	0.83	0.69	0.95	0.91	0.79	0.95
	ResNet-101	0.80	0.65	0.92	0.87	0.74	0.93
Human-defined feature extraction	EfficientNet-B0	0.76	0.51	0.97	0.93	0.66	0.96
	EfficientNet-B2	0.77	0.56	0.96	0.91	0.69	0.95
	DenseNet-121	0.76	0.53	0.95	0.90	0.67	0.95
	DenseNet-169	0.78	0.55	0.97	0.93	0.69	0.96
	ResNet-50	0.77	0.53	0.97	0.93	0.68	0.96
	ResNet-101	0.76	0.51	0.97	0.94	0.66	0.97

p: Bayes post-test probability

The performance of the human-defined feature extraction approach in classifying dental development stages (pre-H stages vs Stage H) is shown in Table 3. The accuracy scores for this task ranged from 0.88 to 0.92 among all models, with DenseNet-169 and ResNet-101 achieving the highest accuracy at 0.92. However, considering the F1-score, which represents the balance of false positives and false negatives, DenseNet-169 outperformed the other models with an F1-score of 0.84. When applying the classified developmental stages from the deep learning model for classifying age groups at the 18-year threshold, all models achieved precision scores ranging from 0.91 to 0.94, with the ResNet-101 model performing the best in this metric. Considering a Bayes' post-test probability, ResNet-101, the best model in this approach, achieved the probability score of 0.97. Although this score was still lower than that of the traditional approach, it was higher than those in the end-to-end feature extraction approach. The results of all approaches to classify age groups at the 18-year threshold are shown in Table 2. The traditional human-based approach achieved a specificity score of 0.99, indicating that nearly all individuals identified as 18 years or older were correctly classified. However, the sensitivity was low at 0.45, indicating that only 45% of individuals aged 18 or older were correctly identified as adults, and the remaining 55% were misclassified as minors. Considering Bayes' post-test probability, this method achieved a value of 0.99, indicating a high probability that individuals predicted to be 18 years or older are so.

The end-to-end deep learning approach, an AI-based method with a straightforward idea, achieved higher scores in accuracy and F1-score than the traditional method. This indicates that this approach can successfully classify individuals as younger or older than 18 years. Among all

models, DenseNet-121 achieved the highest accuracy score and F1-score of 0.85 and 0.82, respectively. However, considering the legal context where false positives are the most serious cases to avoid, we further examined the precision score, which represents the accuracy of positive predictions. Although this approach showed improved balance, its lower precision (0.87–0.91) resulted in a higher risk of misclassifying minors, as reflected in slightly reduced Bayes' post-test probabilities (0.93–0.95) relative to the traditional method.

The performance of the human-defined feature extraction approach in classifying dental development stages (pre-H stages vs Stage H) is shown in Table 3. The accuracy scores for this task ranged from 0.88 to 0.92 among all models, with DenseNet-169 and ResNet-101 achieving the highest accuracy at 0.92. However, considering the F1-score, which represents the balance of false positives and false negatives, DenseNet-169 outperformed the other models with an F1-score of 0.84. When applying the classified developmental stages from the deep learning model for classifying age groups at the 18-year threshold, all models achieved precision scores ranging from 0.91 to 0.94, with the ResNet-101 model performing the best in this metric. Considering a Bayes' post-test probability, ResNet-101, the best model in this approach, achieved the probability score of 0.97. Although this score was still lower than that of the traditional approach, it was higher than those in the end-to-end feature extraction approach.

Table 3. Performance of deep learning models in classifying developmental stages at the Stage H threshold on the test set

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score
EfficientNet-B0	0.91	0.87	0.92	0.73	0.79
EfficientNet-B2	0.88	0.88	0.88	0.67	0.76
DenseNet-121	0.90	0.91	0.90	0.71	0.80
DenseNet-169	0.92	0.95	0.92	0.75	0.84
ResNet-50	0.91	0.92	0.91	0.74	0.82
ResNet-101	0.92	0.91	0.93	0.77	0.83

DISCUSSION

Accurate age estimation is greatly significant in legal and forensic contexts, including criminal proceedings, immigration processes, human trafficking concerns, and the age-specific rights and responsibilities of individuals.[20] However, to date, the only existing study in deep learning-based dental age estimation in Thai populations from panoramic radiographs has shown wide prediction errors of up to five years.[21] Given the 18-year legal threshold, using specific cut-off points for age groups may provide more precise results. In recent years, AI has been increasingly applied in forensic odontology, particularly for age and sex estimation from maxillofacial radiographs.[22] Despite its potential, real-world implementation remains challenging due to the limited transparency and interpretability of AI models. This is especially critical in forensic contexts, where explainability is essential for clinical acceptance and legal defensibility.[23] To address these challenges, we compared the performance of traditional, end-to-end deep learning, and human-defined feature extraction approaches for age classification using mandibular third molars in a Thai population. The inclusion of a stage-based model aimed to enhance interpretability while maintaining high precision, reflecting the need for AI systems that are not only accurate but also aligned with established forensic reasoning. Our results demonstrate notable differences in performance metrics across these methods, highlighting the strengths and weaknesses of each approach in the context of legal age classification.

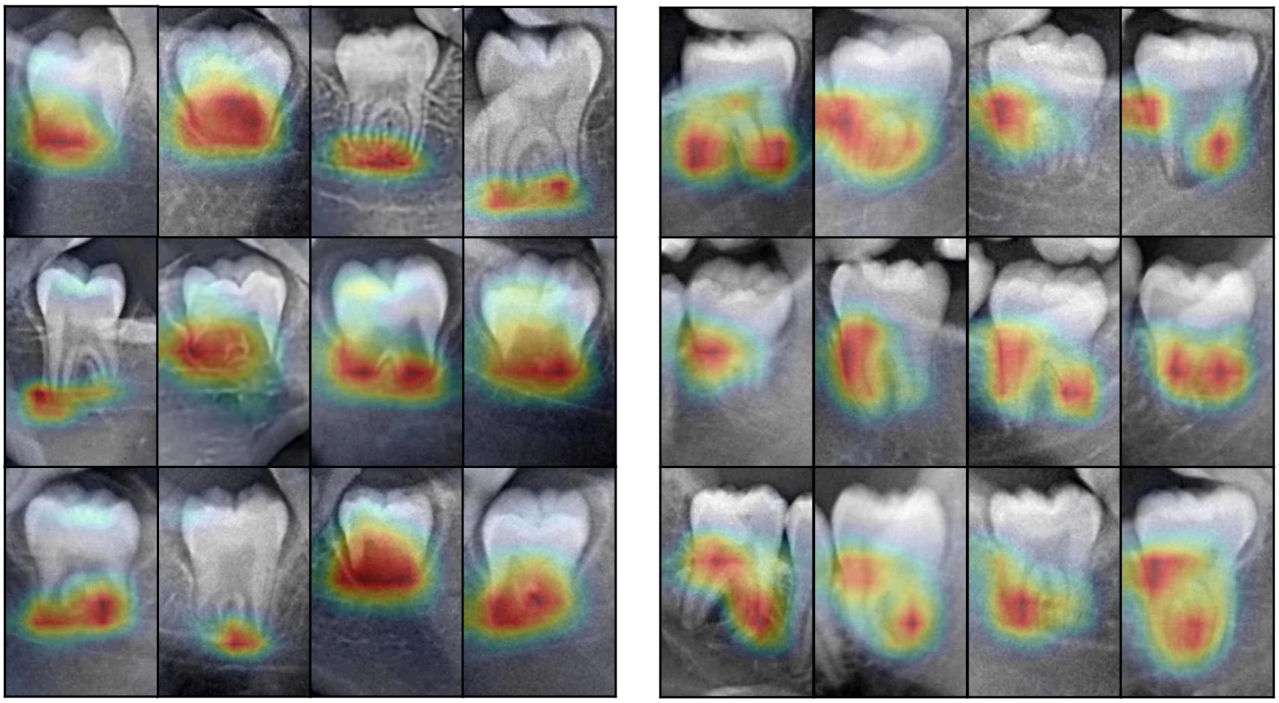
A recent systematic review demonstrated that AI methods for dental age estimation range from expert-guided approaches to fully automated deep learning models, targeting both numerical and categorical age prediction, including legal age thresholds.[8] Han et al. compared human-based, stage-based, and end-to-end deep learning methods for dental age estimation, finding that the end-to-end model performed best with a

mean absolute error of 0.83 years.[24] Similarly, Guo et al. reported superior performance of deep learning over manual methods in classifying ages at 14-, 16- and 18-year thresholds.[25] However, at the 18-year cutoff, deep learning showed lower specificity than Demirjian-based staging, suggesting that end-to-end models may not always be optimal for threshold-based classification tasks.

In forensic contexts, where precise age classification is crucial to avoid misclassifying minors as adults, the precision of these methods becomes particularly significant.[18] Our study aimed to develop a deep learning model optimised for precision to minimise false adult classifications. The traditional method demonstrated excellent precision (0.98) and Bayes' post-test probability (0.99), effectively ruling out false positives but with low sensitivity (0.45), potentially missing some minors. Although the end-to-end deep learning model achieved higher accuracy and F1-scores, its lower precision (0.87–0.91) and post-test probability (0.93–0.95) reduce its reliability in forensic screening. In contrast, the human-defined feature extraction approach balanced performance and interpretability, with improved precision (up to 0.94) and post-test probability (0.97), aided by staging indicators such as Stage H. This approach complements the other methods by enhancing precision and clinical transparency, reflecting the common trade-off between sensitivity and specificity.[26]

To understand the areas of an image that are most important for a model's classification decision, we employed the gradient-weighted class activation mapping (Grad-CAM) technique[27] and analysed the heatmap images of a test group. The area with the greatest influence on the classification was marked in red, whereas the area with the least influence was mapped in blue. Green and yellow represented other portions of the intermediate region. Figure 4 illustrates examples of Grad-CAM of successfully classified age-group cases from the ResNet-50 model, which achieved the highest precision scores in the end-to-end approach. The Grad-CAM revealed that for the under-18 age group, the model highly focuses on the apical structure of the tooth. However, for the over-18 age group, the model's attention is more heterogeneous, with some focusing on the apical region while others pay attention to the mid-root or periodontal ligament space at the cervical region, which is not reasonably understandable by human knowledge. These findings align with a previous study by Gou et al., suggesting that the end-to-end deep learning model might extract more complex and comprehensive features to effectively classify age groups.[25]

Figure 4. Grad-CAM visualisations of successfully classified age-group cases from the ResNet-50 model, which achieved the highest precision scores in the end-to-end approach

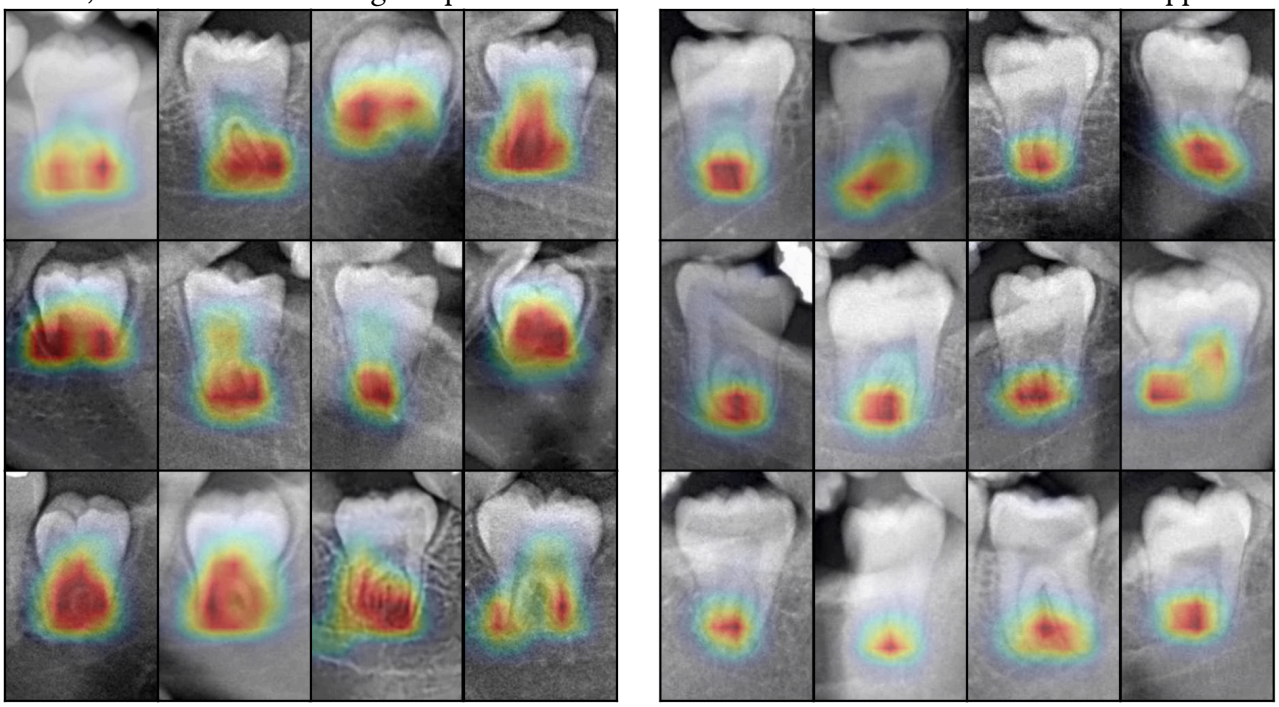


(A) Under 18 years (B) Over 18 years

Figure 5 conversely illustrates examples of Grad-CAM of correctly classified age-group cases from the ResNet-101 model, which achieved the highest precision scores in the human-defined feature extraction approach. In both correctly classifying the pre-H stages and the Stage H, the model focused on the apex of the root, aligning with Demirjian's dental development classification concept that emphasises root development. These results suggested that the human-defined feature extraction approach offered more interpretable results that align with established dental development knowledge.

This study has certain limitations. First is its generalisation. Ethnicity plays a crucial role in wisdom tooth mineralisation.[28] A meta-analysis reported that using the fully mature third molar to predict adulthood at the 18-year threshold yields an overall diagnostic accuracy of 71%.[29] In this study, the Demirjian method for age group prediction in the Thai population yielded an accuracy of 86.5%. Variations in accuracy may be caused by genetic diversity among ethnic groups, which affects dental development. These findings emphasise the need to use a population-specific method to estimate the dental age. Therefore, the results of this study should be interpreted cautiously since the analysis was based solely on data from a Thai population and may not be generalisable to other populations. To address this limitation, future studies should validate the model across various ethnic groups to ensure its accuracy and reliability for different populations. This will enhance the model's applicability and effectiveness in universally estimating dental age.

Figure 5. Grad-CAM visualisations of successfully classified age-group cases from the ResNet-101 model, which achieved the highest precision scores in the human-defined feature extraction approach



(A) Pre-H Stages (B) H Stage

Data imbalance is a key limitation, especially in developmental stages. Although the age groups were fairly balanced (53.25% under 18 years), pre-H stages dominated the dataset (76.47%) compared to Stage H (23.53%). This imbalance may bias the model against accurately classifying Stage H.[30] Synthetic data, such as that generated by generative adversarial networks in brain tumour classification to improve balance and accuracy,[31] could be a solution. However, no evidence supports this approach in dental radiographs, making it currently unfeasible. Collecting more real data to balance the stages remains the most viable option to improve model performance and reliable classification of Stage H.

Although this study introduced deep learning-based methods for accurate age group classification, the

process remained semi-automatic, requiring manual tooth segmentation. Establishing automated detection and segmentation of the tooth of interest is essential for achieving full automation. However, legal and ethical concerns must be addressed. In forensic cases, misclassifying minors can have serious consequences. Human oversight remains essential to ensure that AI predictions are carefully interpreted, with AI serving to support rather than replace expert judgment in sensitive contexts.

CONCLUSION

The traditional method minimised false positives with high precision and strong post-test probability but had lower sensitivity. The end-to-end deep learning model achieved higher accuracy and F1-scores but lower precision, reducing its effectiveness in critical forensic contexts. The human-defined feature extraction approach offered a balanced performance with improved precision and interpretability, making it a useful adjunct for clinicians. However, clinical expertise remains essential to ensure accurate and responsible age classification.

ACKNOWLEDGEMENT

We thank the Faculty of Dentistry, Chiang Mai University, for providing the data used in this study.

REFERENCES

1. McGoldrick D. The United Nations convention on the rights of the child. Int J Law Policy Family. 1991;5(2):132-69. doi:10.1093/lawfam/5.2.132
2. Willems G. A review of the most commonly used dental age estimation techniques. J Forensic Odontostomatol. 2001;19(1):9-17.
3. Manjunatha BS, Soni NK. Estimation of age from development and eruption of teeth. J Forensic Dent Sci. 2014;6(2):73-6. doi:10.4103/0975-1475.132526
4. Panchbhai A. Dental radiographic indicators, a key to age estimation. Dentomaxillofac Radiol. 2011;40(4):199-212. doi:10.1259/dmfr/19478385
5. AlQahtani SJ, Hector MP, Liversidge HM. Brief communication: The London atlas of human tooth development and eruption. Am J Phys Anthropol. 2010;142(3):481-90. doi:10.1002/ajpa.21258
6. Demirjian A, Goldstein H, Tanner JM. A new system of dental age assessment. Hum Biol. 1973:211-27.
7. Pillai JP, Nilendu D, Thomas N, Nagpal S, Sneha Nedunari LS. Inter-observer agreement in the radiographic interpretation of Demirjian's developmental stages in the mandibular second and third molars - A comparative study. J Oral Maxillofac Pathol. 2021;25(3):554-5. doi:10.4103/jomfp.jomfp_85_21
8. Vila-Blanco N, Varas-Quintana P, Tomás I, Carreira MJ. A systematic overview of dental methods for age assessment in living individuals: from traditional to artificial intelligence-based approaches. Int J Legal Med. 2023;137(4):1117-46. doi:10.1007/s00414-023-02960-z
9. Upalananda W, Wantanajittikul K, Na Lampang S, Janhom A. Semi-automated technique to assess the developmental stage of mandibular third molars for age estimation. Aust J Forensic Sci. 2023;55(1):23-33. doi:10.1080/00450618.2021.1882570
10. Franco A, Murray J, Heng D, Lygate A, Moreira D, Ferreira J, et al. Binary decisions of artificial intelligence to classify third molar development around the legal age thresholds of 14, 16 and 18 years. Sci Rep. 2024;14(1):4668. doi:10.1038/s41598-024-55497-5
11. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognit Comput. 2024;16(1):45-74. doi:10.1007/s12559-023-10179-8
12. Chen C, Mat Isa NA, Liu X. A review of convolutional neural network based methods for medical image classification. Comput Biol Med. 2025;185:109507. doi:10.1016/j.compbiomed.2024.109507
13. Duangto P, Iamaroon A, Prasitwattanaseree S, Mahakkanukrauh P, Janhom A. New models for age estimation and assessment of their accuracy using developing mandibular third molar teeth in a Thai population. Int J Legal Med. 2017;131(2):559-68. doi:10.1007/s00414-016-1467-4
14. Roberts GJ, McDonald F, Andiappan M, Lucas VS. Dental Age Estimation (DAE): Data management for tooth development stages including the third molar. Appropriate censoring of Stage H, the final stage of tooth development. J Forensic Leg Med. 2015;36:177-84. doi:10.1016/j.jflm.2015.08.013
15. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA: IEEE; 2016. 770-8. doi:10.1109/CVPR.2016.90
16. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ, editors. Densely connected convolutional networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI, USA: IEEE; 2017. 4700-8. doi:10.1109/CVPR.2017.243
17. Tan M, Le Q, editors. Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning; Long Beach, CA, USA: PMLR; 2019. 6105-14.

18. Sironi E, Gittelson S, Bozza S, Taroni F. Minor or adult? Introducing decision analysis in forensic age estimation. Sci Justice. 2021;61(1):47-60. doi:10.1016/j.scijus.2020.09.004
19. Sironi E, Gallidabino M, Weyermann C, Taroni F. Probabilistic graphical models to deal with age estimation of living persons. Int J Legal Med. 2016;130(2):475-88. doi:10.1007/s00414-015-1173-7
20. Franklin D, Flavel A, Noble J, Swift L, Karkhanis S. Forensic age estimation in living individuals: methodological considerations in the context of medico-legal practice. Research and Rep Forensic Med Sci. 2015;5:53-66. doi:10.2147/RRFMS.S75140
21. Pintana P, Upalananda W, Saekho S, Yarach U, Wantanajittikul K. Fully automated method for dental age estimation using the ACF detector and deep learning. Egypt J Forensic Sci. 2022;12(1):54. doi:10.1186/s41935-022-00314-1
22. Singh S, Singha B, Kumar S. Artificial intelligence in age and sex determination using maxillofacial radiographs: A systematic review. J Forensic Odontostomatol. 2024;42(1):30-7. doi:10.5281/zenodo.11088513
23. Vodanović M, Subašić M, Milošević DP, Galić I, Brkić H. Artificial intelligence in forensic medicine and forensic dentistry. J Forensic Odontostomatol. 2023;41(2):30-41.
24. Han M, Du S, Ge Y, Zhang D, Chi Y, Long H, et al. With or without human interference for precise age estimation based on machine learning? Int J Legal Med. 2022;136(3):821-31. doi:10.1007/s00414-022-02796-z
25. Guo Y-C, Han M, Chi Y, Long H, Zhang D, Yang J, et al. Accurate age classification using manual method and deep convolutional neural network based on orthopantomogram images. Int J Legal Med. 2021;135:1589-97. doi:10.1007/s00414-021-02542-x
26. Chu K. An introduction to sensitivity, specificity, predictive values and likelihood ratios. Emerg Med. 1999;11(3):175-81. doi:10.1046/j.1442-2026.1999.00041.x
27. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, editors. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceeding of the 2017 IEEE International Conference on Computer Vision (ICCV); Venice, Italy: IEEE; 2017. 618-26. doi:10.1109/ICCV.2017.74
28. Olze A, Schmeling A, Taniguchi M, Maeda H, van Niekerk P, Wernecke K-D, et al. Forensic age estimation in living subjects: the ethnic factor in wisdom tooth mineralization. Int J Legal Med. 2004;118(3):170-3. doi:10.1007/s00414-004-0434-7
29. Haglund M, Mörnstad H. A systematic review and meta-analysis of the fully formed wisdom tooth as a radiological marker of adulthood. Int J Legal Med. 2019;133(1):231-9. doi:10.1007/s00414-018-1842-4
30. Ghosh K, Bellinger C, Corizzo R, Branco P, Krawczyk B, Japkowicz N. The class imbalance problem in deep learning. Mach Learn. 2024;113(7):4845-901. doi:10.1007/s10994-022-06268-8
31. Moshe YH, Buchsweiler Y, Teicher M, Artzi M. Handling Missing MRI Data in Brain Tumors Classification Tasks: Usage of Synthetic Images vs. Duplicate Images and Empty Images. J Magn Reson Imaging. 2024;60(2):561-73. doi:10.1002/jmri.29072