ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2026.0013

A framework to integrate deep learning and an ethnicity-specific reference data set for reliable dental age estimation from digital panoramic radiographs

Kuen Wai MA[1,2], Jiajun ZHU[3,4]✉, Hai Ming WONG[1]✉

[1]*Faculty of Dentistry, The University of Hong Kong, Hong Kong 999077, China*
[2]*Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China*
[3]*Stomatology Hospital, School of Stomatology, Zhejiang University School of Medicine, Zhejiang Provincial Clinical Research Center for Oral Diseases, Zhejiang Key Laboratory of Oral Biomedical, Hangzhou 310000, China*
[4]*Engineering Research Center of Oral Biomaterials and Devices of Zhejiang Province, Hangzhou 310000, China*

Abstract: Automated dental age estimation is essential for clinical dentistry. However, traditional methods like Demirjian's tooth development stage (TDS) require time-consuming manual annotation by trained experts. Moreover, population differences in dental development patterns can significantly affect the accuracy of age estimation, highlighting the need for an ethnicity-specific reference data set (RDS). This study aims to use deep neural networks (DNNs) for automated estimation of Demirjian's TDS of molars on digital panoramic radiographs, integrating the most complete southern Chinese RDS for ethnicity-appropriate dental age assessment. Panoramic radiographs from individuals aged 2 to 25 years are annotated for molar TDS and used to train eight DNN architectures, including AlexNet, DenseNet-201, and ResNet-50. DenseNet-201 achieves the highest accuracy of 93% in classifying molar TDS. Most misclassifications involve adjacent stages. The integrated mean dental age (IMDA) is obtained by mapping the predicted molar TDS using the southern Chinese RDS. The mean difference and correlation coefficient between the estimated IMDA from the best-performing DNN (AlexNet) and chronological age are −0.063 years (−3.3 weeks) and r=0.898 (p< 0.001), respectively. These findings demonstrate that combining DNN for TDS estimation with ethnic-specific RDS enables accurate and reliable dental age assessment.

Key words: Deep neural networks (DNNs); Dental age estimation; Panoramic radiographs; Southern Chinese; Tooth development

1 Introduction

Dental age estimation is a critical component across various sectors, including clinical practice, forensic investigations, and pediatric dentistry. The accurate assessment of dental age not only plays a pivotal role in identifying individuals with uncertain or unknown ages but also serves as a fundamental tool in legal and medical contexts, shaping decisions related to criminal responsibility, age-based healthcare, and overall well-being (Franco et al., 2024). Tooth development stage (TDS) has been extensively studied since the 1960s, with the number of stages described ranging from 4 to 32 (Nolla, 1960). TDS classifications are one of the conventional approaches in dental age estimation, determining an individual's chronological age based on the development of his/her teeth. In particular, Demirjian's classification (Demirjian et al., 1973; Demirjian and Goldstein, 1976) has proven to be most effective and reliable for diagnosis (Olze et al., 2005; Maber et al., 2006) because it uses well-defined objective criteria of the eight stages (A–H) that illustrate tooth images in a clear-cut manner. Demirjian et al. (1973) and Demirjian and Goldstein (1976) proposed a series of crude averaged percentile curves for dental age estimation by scoring the TDS based on a French–Canadian sample. However, TDS classifications are often labor-intensive, relying on manual annotations by trained experts, and several studies have reported that Demirjian's dental age assessment method is less accurate when applied to other ethnic groups (Lee et al., 2008; Cruz-Landeira et al., 2010). Moreover, only seven lower left teeth (central and lateral incisors, canine, first and second bicuspids, and first and second molars) are assessed in Demirjian's dental age estimation method, making the estimation of ages of legal importance impossible because the third molars are the only remaining teeth

✉ Jiajun ZHU, zhujj@zju.edu.cn
Hai Ming WONG, wonghmg@hku.hk
Kuen Wai MA, https://orcid.org/0000-0001-7784-0700
Hai Ming WONG, https://orcid.org/0000-0003-3411-6442
CLC number: TP183; R783.1

Received: Jan. 16, 2026; Revision accepted: May 28, 2026;
Crosschecked: June 2, 2026; Published online: July 17, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd.
This is an open access article distributed under the terms of the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

to have not yet reached maturity in young people aged 15 to 22 years (AlQahtani et al., 2010). Therefore, this study aims to develop an accurate and reliable dental age assessment by combining the deep learning technique in TDS classification of molars and the systematic integration of data from the most complete southern Chinese reference data set (RDS).

Leveraging the power of artificial intelligence, applications of deep neural networks (DNNs) have demonstrated remarkable capabilities in image recognition, natural language processing, and medical imaging analysis (Ahn et al., 2022; Wodzinski et al., 2022; Kim M et al., 2023; Migliorelli et al., 2023; Shang et al., 2024; Wang P et al., 2025). The evolution of DNNs has revolutionized dental age estimation by automating TDS classification and age estimation. After systematically searching for the keywords "neural network" and "dental age" across electronic bibliographic databases including PubMed, the Institute for Scientific Information (ISI) Web of Science, ScienceDirect, and Scopus (Table 1), it is found that DNN applications in dental age estimation using digital panoramic radiographs (DPRs) primarily fall into four prediction categories: binary legal-threshold decisions (Franco et al., 2024), TDS classification (de Tobel et al., 2017; Banar et al., 2020; Han et al., 2022; Pintana et al., 2022; Matthijs et al., 2024; Milani et al., 2024; Ong et al., 2024), age group prediction (Mualla et al., 2020; Vila-Blanco et al., 2020; Kim S et al., 2021; Wang XD et al., 2022; Kim YR et al., 2023; Wang J et al., 2023; Oliveira et al., 2024; Sivri et al., 2024; Koch et al., 2025), and dental age estimation (Kahaki et al., 2020; Merdietio Boedi et al., 2020; Sharifonnasabi et al., 2022; Aljameel et al., 2023; Baydoğan et al., 2023; Fan et al., 2023; Yeom et al., 2023; Büyükçakır et al., 2024; Mohammad et al., 2024; Şahin and Kölüş, 2024; Sathyavathi and Baskaran, 2024; Shi et al., 2024; Ozlu Ucan et al., 2025; Yavuz et al., 2025). Some studies target narrow objectives, such as binary legal-age determinations (Franco et al., 2024) or single-tooth analysis (Pintana et al., 2022). Although TDS classification studies have evolved from analysing single teeth (Ong et al., 2024) to mandibular (lower) teeth (Han et al., 2022; Matthijs et al., 2024) and eventually to full dentition (Ong et al., 2024), these studies typically stop at classifying developmental stages and do not extend their results to direct dental age estimation. Early DNN-based dental age estimation often relies on either the entire panoramic radiograph (Mualla et al., 2020; Vila-Blanco et al., 2020) as input or TDS classification of a single tooth (Merdietio Boedi et al., 2020). Although a high accuracy (>80%) of age ranges (such as 10–19 years) is reported, the error for the estimated age can be as large as 9 years (Vila-Blanco et al., 2020). Therefore, using TDS classification as input is considered a more reliable approach for DNN-based dental age estimation (Shi et al., 2024), as it provides more detailed and comprehensive information about dental structure and development.

Despite significant advancements in automated dental age estimation facilitated by DNN applications, challenges remain in achieving precise and reliable results across diverse populations. Variations in dental development patterns among different ethnic groups have been shown to significantly impact the accuracy of age estimation models, highlighting the necessity for constructing ethnicity-specific RDS to improve the performance of DNN-based approaches. Extensive research has documented these variations in dental development (Lai et al., 2008; Wong et al., 2016a, 2017, 2019; Mallineni et al., 2019). For example, studies have demonstrated that the United Kingdom Caucasian data set underestimates the age of southern Chinese individuals by up to 3.5 years (Jayaraman et al., 2012, 2018), and Demirjian's data set overestimates the age of southern Chinese by up to 4.1 years (Jayaraman et al., 2011;Jayaraman and Roberts, 2016). Systematic reviews and meta-analyses have further indicated that Demirjian's dental age estimation method is not applicable to any modern population groups (Jayaraman et al., 2013). Analysis of DPRs from random samples of southern Chinese children born in 1981 and 2001 reveals a secular trend in the development of permanent teeth, with the 2001 cohort exhibiting more advanced dental maturation compared to the 1981 group (Jayaraman et al., 2013). These findings underscore the need for applying an up-to-date and ethnicity-specific RDS for the accurate dental age estimation of modern population groups.

To address current dental, medical, legal, and forensic demands (Jayaraman et al., 2016a; Roberts et al., 2017), the first southern Chinese RDS has been established using 4000 panoramic radiographs (Jayaraman et al., 2016b). This RDS overcomes limitations of previous methods and enables the development of a valid and reliable approach to dental age estimation (Jayaraman et al., 2016b, 2019; Wong et al., 2016b). In this new method, only the first part of Demirjian's method, 8-stage TDS classification from initial calcification (stage A) to root completion (stage H), has been adopted. The calculation of integrated mean dental age (IMDA) in this study will be based on the systematic integration of the data of the predicted molar TDS from the most complete southern Chinese RDS. By combining this ethnicity-specific RDS with state-of-the-art DNN technology, this study aims to develop an accurate and reliable framework (DNN-IMDA) for automated estimation of Demirjian's TDSs on DPRs, thereby enabling reliable dental age estimation for southern Chinese individuals.

2 Methods

2.1 Southern Chinese RDS

Southern Chinese RDS is developed from 2306 DPRs collected from the Department of Paediatric Dentistry & Orthodontics, Prince Philip Dental Hospital, Hong Kong, China, covering individuals aged 2–25 years with balanced gender distribution (Jayaraman et al., 2016b). Ethical approval for this study was obtained from the Institutional Review Board of the University of Hong Kong West Cluster Hospital Authority (IRB No. UW 12-280). The inclusion criteria are healthy subjects of the target population and age group with proper dental health-related documentation and written consent. The exclusion criteria are as follows:

1. systemic diseases or syndromes that would affect skeletal or dental development,

2. localised oral pathology or anomalies that would affect dental development,

3. severe malocclusion with or without previous/current orthodontic treatment,

4. congenitally missing or premature loss of targeted teeth, and

5. a poor-quality DPR, in which one or more targeted teeth cannot be identified.

Individual details, including the hospital number, date of birth, date of radiograph, gender and ethnicity, along with TDS scoring, are

Table 1 Summary of dental age estimation related studies using DNNs and DPRs

Prediction task	Year	Image data set size	Input size	DNN	TDS classification	Age range (years)	Population	Reference
Binary decisions (thresholds of legal interest)	2024	11 640	224×224	DenseNet-121	Nil	Nil	Brazilian	Franco et al. (2024)
TDS classification	2024	5133	640×640	YOLOv5	Demirjian's method	4–16	Korean	Ong et al. (2024)
	2024	3422	192×192	EfficientNet-B0	Demirjian's method (third molars)	2–97	Brazilian	Milani et al. (2024)
	2024	1639	224×224	DenseNet-201	Demirjian's method (lower tooth)	Nil	Uzbek Leuven	Matthijs et al. (2024)
	2022	10 257	224×224	VGG-16 and ResNet-101	Demirjian's method (lower tooth)	5–24	Chinese	Han et al. (2022)
	2022	1000	224×224	ResNet-50	Demirjian's method (lower left third molar, D-H)	15–23	Thai	Pintana et al. (2022)
	2020	400	224×224	DenseNet-201	Demirjian's method (third molars)	7–24	Uzbek Leuven	Banar et al. (2020)
	2017	400	240×390	AlexNet	Demirjian's method (lower left molar)	Nil	Uzbek Leuven	de Tobel et al. (2017)
Age group prediction	2025	21 814	224×224	VGG-16, Inception-v3, ResNet-50, etc.	Nil	1–25	German	Koch et al. (2025)
	2024	12 827	299×299	Inception-v4	Nil	2–97	Brazilian	Oliveira et al. (2024)
	2024	5898	256×256	AlexNet, VGG-16, ResNet-152, etc.	Demirjian's method	4–17	Antalya	Sivri et al. (2024)
	2023	10 023	256×512	Own model	Nil	Nil	Korean	Kim YR et al. (2023)
	2023	9586	224×224	VGG-16 and ResNet-101	Nil	6–20	Chinese	Wang J et al. (2023)
	2022	1903	447×447	SSR-Net	Nil	3–85	Chinese	Wang XD et al. (2022)
	2021	1586	151×112	ResNet-152	Nil	>50	Korean	Kim S et al. (2021)
	2020	2289	256×128	DANet	Nil	4–90	Spanish	Vila-Blanco et al. (2020)
	2020	1429	224×224 227×227	AlexNet and ResNet	Nil	>70	Kuwaitis	Mualla et al. (2020)
Dental age estimation	2025	1169	224×224	ResNet-50	Nil	5–16	Turkish	Yavuz et al. (2025)
	2025	622	224×224	1D-DCNN 2D-DCNN	Nil	6–15	Turkish	Ozlu Ucan et al. (2025)
	2023	9663	384×384	ResNet-50	Nil	3–79	Korean	Yeom et al. (2023)
	2024	7667	227×227 256×256 229×229	SqueezeNet, DarkNet-53, Xception	Nil	5–21	Turkish	Şahin and Kölüş (2024)
	2024	3896	400×250	EfficientNet-B4, DenseNet-201, etc.	Nil	<25	Uzbek Leuven	Büyükçakır et al. (2024)
	2023	15 195	256×256	DASE-Net	Nil	16–51	Chinese	Fan et al. (2023)
	2024	4008	224×224	ResNet-50	Nil	Nil	Vihana	Sathyavathi and Baskaran (2024)
	2023	933	224×224	AlexNet, ResNet-50, etc.	Nil	2–21	Turkish	Baydoğan et al. (2023)
	2022	1922	224×224	Own model	HCNN-KNN	15–23	Malaysian	Sharifonnasabi et al. (2022)
	2023	529	224×224	Xception, VGG-16, etc.	Nil	6–11	Saudi Arabian	Aljameel et al. (2023)
	2024	673	224×224	ResNet-50, Inception-v3, etc.	Demirjian's method (full dentition excluding the third molars)	3–15	Chinese	Shi et al. (2024)
	2024	10 892	–	F-DentEst Lab	Willems method (lower premolars)	5–16	Malaysian	Mohammad et al. (2024)
	2019	400	240×390	DenseNet-201	Demirjian's method (lower left third molar)	Nil	Uzbek Leuven	Merdietio Boedi et al. (2020)
	2019	456	227×227	AlexNet	Demirjian's method (first to third molars)	1–17	Malaysian	Kahaki et al. (2020)

Lower teeth: mandibular teeth; 1D-DCNN: one-dimensional deep convolutional neural network; 2D-DCNN: two-dimensional deep convolutional neural network; HCNN-KNN: hybrid convolutional neural networks and *K* nearest neighbors; –: input size was not reported in the corresponding reference

recorded in a Microsoft Access database. Chronological age is calculated as the difference between date of birth and radiograph date, with data grouped by gender due to observed differences in dental maturity. RDS is constructed by scoring teeth on the left side of both arches and the two right third molars for TDS (stages A–H) using Demirjian's classification (Demirjian et al., 1973; Demirjian and Goldstein, 1976). Summary statistics (see Table S1 in the supplementary materials) including the mean, standard deviation (SD), standard error (SE), median, percentiles, and 99% confidence intervals (CIs) for each TDS and tooth, are computed and stored in structured spreadsheets. A separate validation set of 484 individuals is used to confirm the RDS accuracy, with mean differences between dental age and chronological age of less than 0.1 years and no significant difference found (Jayaraman et al., 2016b).

2.2 Image data set for training and validation of DNN

All DPRs in the RDS are digitized using a scanner at 300 dpi resolution and anonymized before scoring. The scanned images are viewed on a widescreen monitor under standard magnification of 160% without image editing. Regions of interest containing molars of the radiographs are manually labelled by two calibrated examiners (Fig. 1a). A total of 400 molar images (50×8 TDSs) with annotated TDS are randomly selected to construct an image data set for DNN training and validation. All images are resized to match the input size (e.g., 224×224×3 for AlexNet (Krizhevsky et al., 2017)) required for each DNN architecture.

A 5-fold cross-validation approach is implemented for model training and validation (Fig. 1b). The data set is split using 5-fold cross-validation with 80% for training and 20% for validation in each fold. The training process of each DNN architecture will run five times with the first training using the first 20% of the input data set as the first validation set (and the remaining 80% as the training set), the second training using the second 20% of the image data set as the second validation set, and so on. Data augmentation artificially expands the training data set through label-preserving transformations applied in real time during training, while the validation set remains without augmentation to ensure unbiased performance evaluation. Seven transformation types are employed: zoom (1.0×–1.3×), rotation (±45°), shear (±0.5), horizontal translation (±30 pixels and ~13% of 227-pixel width), vertical translation (±30 pixels and ~13% of 227-pixel height), horizontal flip, and vertical flip. These generate 128 possible combinations $\left(\sum_{k=0}^{7}\frac{7!}{k!(7-k)!}\right)$, enlarging the training data set from 320 images to about 40 000 annotated variants to reduce overfitting risks.

2.3 Architecture of DNN

Eight different DNN architectures are evaluated for their performance in TDS classification (Table 2). AlexNet (Krizhevsky et al., 2017) is a pioneering DNN architecture that demonstrates the power of deep learning in image classification tasks. A non-saturating linear rectification function (ReLU) is introduced to achieve high performance (Krizhevsky et al., 2017). For example, AlexNet consists of five convolutional (conv) layers, three max-pooling (MP) layers, three fully-connected (FC) layers, and one softmax layer (Fig. 1c). In each conv layer, the specific features of the input are learned by sequential convolution kernels and nonlinear activation layers:

$$\boldsymbol{Z}_{i,j} = \max\left(\boldsymbol{0}, \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}_{i,j} + \boldsymbol{b}\right), \tag{1}$$

where $\boldsymbol{Z}_{i,j}$, $\boldsymbol{X}_{i,j}$, $\boldsymbol{W}$, $\boldsymbol{b}$, and $\boldsymbol{y}$+=max($\boldsymbol{0}$, $\boldsymbol{y}$) are the convolution feature at a location (i, j), input centred at (i, j), kernel weights, kernel bias, and ReLU activation function (Gu et al., 2018), respectively. The output of a conv layer is a feature map to store the extracted features at locations. In each MP layer, a feature map is downsized by extracting the most informative features in small regions. In each FC layer, high-level features are stored in the artificial neurons that are fully connected to all neurons in the previous layer. The output layer is the softmax layer to provide the probabilities of the eight TDSs, $P(\mathrm{TDS_{A–H}})$. TDS of a tooth will be classified as the stage with the highest probability.

DenseNet-201 (Huang et al., 2017) introduces a dense connectivity pattern, where each layer is directly connected to every other layer in a feed-forward manner. This design encourages feature reuse and helps address the vanishing gradient problem. GoogLeNet (Szegedy et al., 2015), known for its inception modules, applies computation-efficient strategies such as 1×1 convolutions and parallel paths for feature extraction. Inception-v3 (Szegedy et al., 2016) improves this architecture by adding factorized convolutions and global average pooling. ResNet-18 and ResNet-50 belong to the ResNet family, which uses residual connections to make training very deep networks more effective (He et al., 2016). VGG16, recognized for its simplicity, relies on the consistent use of 3×3 convolutional filters and has shown strong performance in image classification tasks (Simonyan and Zisserman, 2015). Xception, based on depthwise separable convolutions, offers a more parameter-efficient alternative to standard convolutional layers (Chollet, 2017).

2.4 Transfer learning implementation

The pretrained DNN of the eight architectures is loaded by using MATLAB, R2025a software (MathWorks, Natick, USA). The transfer learning approach involves extracting all layers except the final three classification layers. These transferred layers serve as feature extractors for the molar images. The architecture is then modified by adding new classification layers specifically designed for the 8-stage TDS classification. A new FC layer is added with the number of outputs equal to eight TDS, followed by a softmax layer for probability distribution and an output layer for final classification. The newly added FC layer is configured with enhanced learning rates (WeightLearnRateFactor=20, BiasLearnRateFactor=20) to accelerate the learning of task-specific features while maintaining the pretrained convolutional features.

Moreover, the training process is configured using stochastic gradient descent with momentum optimization. Key hyperparameters are carefully selected based on empirical testing, including a mini-batch size of 80 images per iteration to balance computation-efficient and gradient stability, a maximum of 4000 epochs to ensure sufficient training iterations, an initial learning rate of 1×10^{-3} for optimal convergence, a momentum coefficient of 0.95 to accelerate convergence and reduce oscillations, L2 regularization ($\lambda=1\times10^{-4}$) for weight decay, and a dropout of 0.5 for preventing overfitting in FC layers. A piecewise learning rate schedule is implemented with a learning rate reduction every 400 epochs by a factor of 0.6 to facilitate fine-tuning at the later training stage. The training data set is shuffled at every

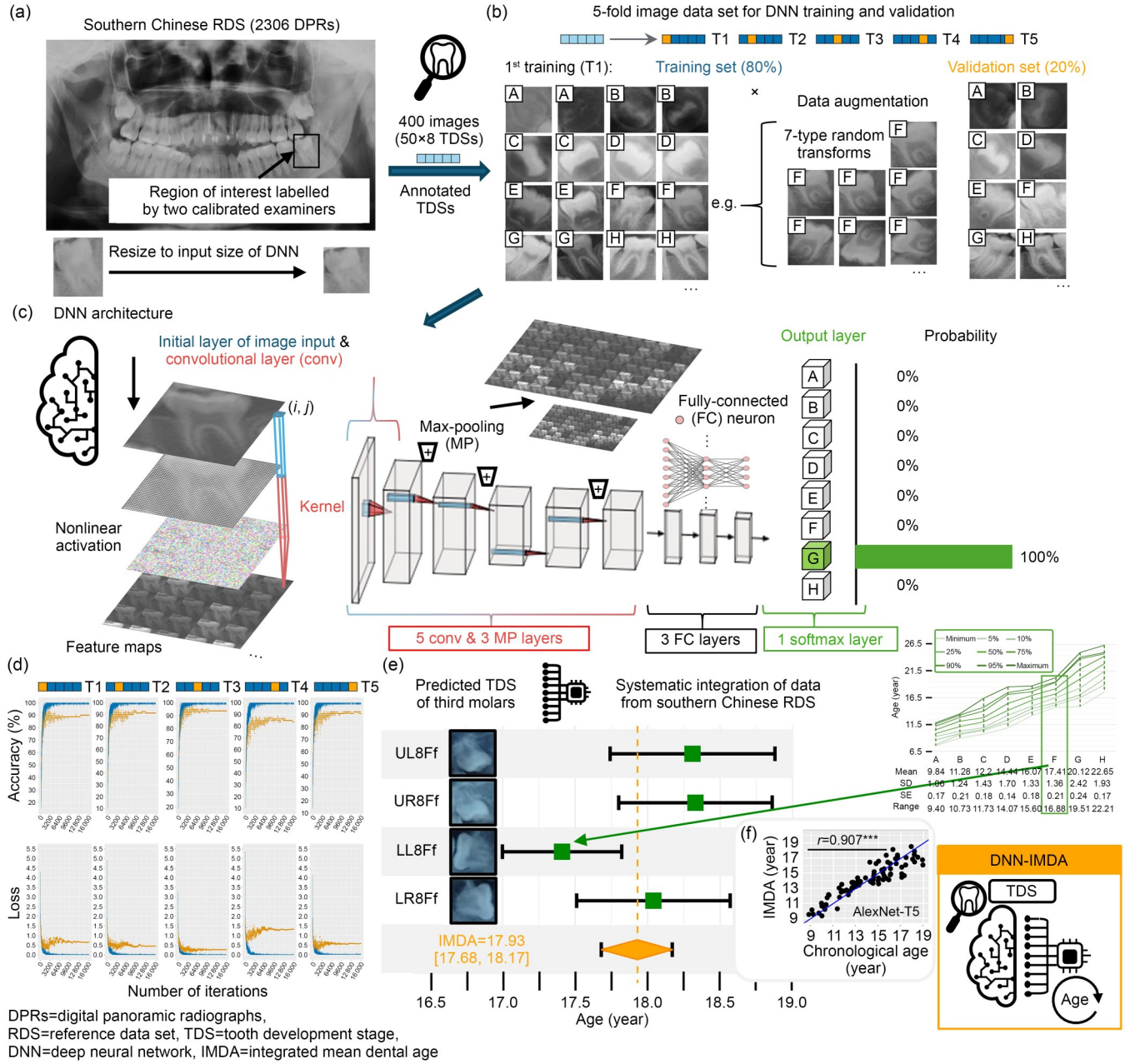

Fig. 1 Schematic of the development of the DNN-IMDA framework for automated deep learning-based classification of Demirjian's TDSs on DPRs and ethnicity-appropriate dental age assessment from IMDA calculation: (a) preparation of a 5-fold image data set for training and validation of DNN; (b) procedure of 5-fold cross-validation training and data augmentation; (c) DNN architecture of AlexNet as an example for TDS classification; (d) training performance curves showing classification accuracy and prediction loss across iterations in the 5-fold training (blue spots) and validation (orange spots) of the AlexNet models; (e) an example of IMDA calculation by mapping the predicted molar TDS and systematically integrating the data from the southern Chinese RDS (error bar in the forest plot indicates the 95% CI of each predicted molar TDS and IMDA; orange dashed line: IMDA; green box: mean age of the predicted molar TDS); (f) scatterplot of IMDA obtained from AlexNet-T5 against chronological age (blue line represents the line of identity with $y=x$ and IMDA=chronological age)

Table 2 Summary of the eight architectures of DNNs applied in the study for TDS classification

Architecture	Depth	Input size	Input value range	Number of parameters ($\times 10^6$)	Activation function	Pooling	Reference
AlexNet	8	227×227×3	[0, 255]	61.0	ReLU	Max-pooling	Krizhevsky et al. (2017)
DenseNet-201	201	224×224×3	[0, 255]	20.0	ReLU	Average pooling	Huang et al. (2017)
GoogLeNet	22	224×224×3	[0, 255]	7.0	ReLU	Average pooling	Szegedy et al. (2015)
Inception-v3	48	299×299×3	[0, 255]	23.9	ReLU	Average pooling	Szegedy et al. (2016)
ResNet-18	18	224×224×3	[0, 255]	11.7	ReLU	Max-pooling	He et al. (2016)
ResNet-50	50	224×224×3	[0, 255]	25.6	ReLU	Max-pooling	He et al. (2016)
VGG-16	16	224×224×3	[0, 255]	138	ReLU	Max-pooling	Simonyan and Zisserman (2015)
Xception	71	299×299×3	[0, 255]	22.9	ReLU	Average pooling	Chollet (2017)

epoch to prevent overfitting to data order patterns. Although validation patience is set to infinity to allow complete 4000-epoch (16 000 iterations) training cycles (no early stopping), validation monitoring of model performance is continuously conducted every 12 iterations to detect potential overfitting during training. Real-time monitoring is achieved by plotting the accuracy and loss curves in 5-fold training across 16 000 iterations, enabling early identification of train-validation divergence to reduce overfitting risks.

2.5 Comparison of performance metrics of the DNN models

During training, both training and validation performance metrics (Sokolova and Lapalme, 2009) are continuously monitored and recorded to assess model convergence and prevent overfitting (Fig. 1d). A prediction error will be returned by the loss function:

$$\text{Loss}=\frac{1}{N_{\text{batch}}}\sum_{b=1}^{N_{\text{batch}}}\text{Loss}_b, \tag{2}$$

with

$$\text{Loss}_b=\frac{1}{n}\cdot\sum_{i=1}^{n}\sum_{j=1}^{8}\left(\text{TDS}_{i,j}\cdot\ln P\left(\text{TDS}_{i,j}\right)\right), \tag{3}$$

where n is the number of samples in each batch b, $\text{TDS}_{i,j}$ is 1 if sample i belongs to TDS_j, N_{batch} is the number of mini-batches in each epoch, and $P(\text{TDS}_{i,j})$ is the predicted probability of TDSs (A–H). The network weights of the layers will be updated per batch to obtain a smaller prediction error. Four mini-batches will be included in each epoch, a completed loop of the whole training data set.

The overall accuracy and loss of each DNN architecture will be the mean of the classification accuracy and prediction loss of the final models in the 5-fold cross-validation training after 4000 epochs, respectively. The macro recall and macro F1 score of each DNN architecture are calculated from the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) of prediction for each TDS.

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{4}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{5}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{6}$$

$$\text{F1_score}=2\cdot\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{7}$$

$$\text{Macro_recall}=\sum_{i=1}^{8}\text{Recall}_i, \tag{8}$$

$$\text{Macro_F1_score}=\sum_{i=1}^{8}\text{F1_score}_i. \tag{9}$$

2.6 Calculation of chronological age and IMDA

The chronological age of individuals is calculated in decimal form using [(date of panoramic radiograph)−(official date of birth)]/365.25. Eighty DPRs with third molars in the validation set (80/484) of the RDS (Jayaraman et al., 2016b) are randomly selected for the calculation of IMDA. IMDA is calculated using a meta-analysis approach that systematically integrates TDS data from multiple teeth. The ground-truth TDS annotations from the original validation set of the RDS are used for all teeth except third molars, where the TDS classification from each trained DNN model is substituted. The meta-analysis framework treats each developing tooth (stages A–G) as an independent study, following the established methodology for the southern Chinese RDS (Fig. 1e). For each individual, the corresponding mean age and SE for each scored TDS are extracted from the southern Chinese RDS using the following systematic approach. Each TDS is weighed inversely proportional to the SE of its mean age:

$$w_i=\frac{1}{\text{SE}_i^2}, \tag{10}$$

where w_i is the weight for tooth i and SE_i is the SE of the mean age for that specific TDS. IMDA is calculated by

$$\text{IMDA}=\sum_{i=1}^{m}w_i\cdot\bar{x}_i\Big/\sum_{i=1}^{m}w_i, \tag{11}$$

where m is the number of developing teeth (stages A–G) present in the individual, and $\bar{x}_i$ is the mean age for TDS of tooth i from the southern Chinese RDS. The 95% CI for IMDA is calculated using the SE of the weighted mean:

$$\text{IMDA}\pm 1.96\cdot\sqrt{1\Big/\sum_{i=1}^{m}w_i}. \tag{12}$$

2.7 Comparison between chronological age and IMDA

All the data from the statistical analyses are coded and analysed using the commercial package SPSS, version 23.0 (IBM Corp., Armonk, NY, USA). All statistical tests are two-tailed tests with a significance level of 0.05. The normality of the data is checked for the selection of suitable statistical tests. The nonparametric tests, such as Mann–Whitney U tests and Spearman's rank correlation tests, will be applied for the non-normally distributed data. The two-tailed paired t-test and Pearson's correlation coefficient (r) (Fig. 1e) are used to compare IMDA and the chronological age (gold standard) for the normally distributed data. The DNN models will be regarded as valid (within the range of accuracy) for dental age estimation if the p value is greater than 0.05 in the paired t-test, and r is greater than 0.8 (p<0.001).

3 Results

3.1 Model performance in molar TDS classification

All eight DNN architectures achieve high final validation accuracy (81.3%–96.3%) and low training loss (0.0001–0.0092) after 4000 epochs, indicating effective model convergence across the 5-fold cross-validation training process (Table 3). Among these architectures, the values of the overall final validation accuracy, macro recall, and macro F1 score of DenseNet-201 models are the highest (93.0%±2.1%, 0.985±0.008, and 0.984±0.008). The value of the overall validation loss of DenseNet-201 models is also the lowest (0.39±0.23). The results demonstrate the superior ability of DenseNet-201 models to correctly identify all molar TDS and maintain balanced

Table 3 Performance metrics of eight DNN architectures for molar TDS classification

Architecture	Model	Final training accuracy (%)	Final validation accuracy (%)	Final training loss	Final validation loss	Macro recall	Macro F1 score
AlexNet	T1	100	90.0	0.0001	0.65	0.978	0.977
	T2	100	91.3	0.0092	0.46	0.983	0.983
	T3	100	93.8	0.0006	0.26	0.988	0.988
	T4	100	83.8	0.0025	1.33	0.968	0.968
	T5	100	91.3	0.0006	0.57	0.980	0.980
	Overall[#]	100	90.0 (±3.8)	0.0026 (±0.0038)	0.66 (±0.41)	0.979 (±0.007)	0.979 (±0.007)
DenseNet-201	T1	100	90.0	0.0000	0.51	0.975	0.975
	T2	100	92.5	0.0001	0.31	0.993	0.992
	T3	100	95.0	0.0001	0.21	0.988	0.987
	T4	100	92.5	0.0000	0.73	0.978	0.978
	T5	100	95.0	0.0001	0.18	0.990	0.990
	Overall[#]	100	93.0 (±2.1)	0.0001 (±0.000 03)	0.39 (±0.23)	0.985 (±0.008)	0.984 (±0.008)
GoogLeNet	T1	100	93.8	0.0001	0.46	0.983	0.982
	T2	100	88.8	0.0001	0.34	0.975	0.975
	T3	100	91.3	0.0000	0.72	0.978	0.977
	T4	100	87.5	0.0000	1.08	0.968	0.968
	T5	100	96.3	0.0000	0.11	0.993	0.992
	Overall[#]	100	91.5 (±3.6)	0.000 04 (±0.000 04)	0.54 (±0.37)	0.979 (±0.009)	0.979 (±0.009)
Inception-v3	T1	100	92.5	0.0003	0.32	0.975	0.975
	T2	100	87.5	0.0005	0.47	0.965	0.965
	T3	100	83.8	0.0001	0.57	0.965	0.965
	T4	100	83.8	0.0020	0.89	0.948	0.947
	T5	100	90.0	0.0004	0.49	0.963	0.963
	Overall[#]	100	87.5 (±3.9)	0.0006 (±0.0008)	0.55 (±0.21)	0.963 (±0.010)	0.963 (±0.010)
ResNet-18	T1	100	88.8	0.0000	0.45	0.978	0.977
	T2	100	91.3	0.0001	0.44	0.985	0.985
	T3	100	85.0	0.0001	0.74	0.978	0.978
	T4	100	90.0	0.0000	0.44	0.968	0.968
	T5	100	90.0	0.0000	0.41	0.970	0.970
	Overall[#]	100	89.0 (±2.4)	0.0001 (±0.000 04)	0.49 (±0.14)	0.976 (±0.007)	0.976 (±0.007)
ResNet-50	T1	100	88.8	0.0000	0.69	0.975	0.975
	T2	100	87.5	0.0000	0.59	0.978	0.978
	T3	100	92.5	0.0000	0.44	0.988	0.987
	T4	100	85.0	0.0006	0.95	0.975	0.975
	T5	100	90.0	0.0004	0.46	0.988	0.987
	Overall[#]	100	88.8 (±2.8)	0.0002 (±0.0003)	0.63 (±0.21)	0.981 (±0.006)	0.980 (±0.006)
VGG-16	T1	100	90.0	0.0025	0.66	0.978	0.977
	T2	100	93.8	0.0000	0.36	0.985	0.985
	T3	100	91.3	0.0001	0.40	0.983	0.982
	T4	100	92.5	0.0001	0.64	0.973	0.973
	T5	100	93.8	0.0003	0.26	0.985	0.985
	Overall[#]	100	92.3 (±1.6)	0.0006 (±0.0011)	0.46 (±0.18)	0.981 (±0.005)	0.981 (±0.005)
Xception	T1	100	82.5	0.0016	0.57	0.965	0.965
	T2	100	87.5	0.0005	0.56	0.960	0.960
	T3	100	90.0	0.0023	0.30	0.970	0.970
	T4	100	88.8	0.0035	0.53	0.953	0.952
	T5	100	81.3	0.0003	0.37	0.953	0.952
	Overall[#]	100	86.0 (±3.9)	0.0016 (±0.0013)	0.47 (±0.12)	0.960 (±0.008)	0.960 (±0.008)

[#] Mean (±SD) of the final values of all models in 5-fold cross-validation training (T1–T5) after 4000 epochs, calculated based on the data before rounding

performance. In contrast, Xception models record the lowest values of the overall final validation accuracy, macro recall, and macro F1 score (86.0%±3.9%, 0.960±0.008, and 0.960±0.008), suggesting difficulty in capturing the full range of molar TDS.

Using the performance of the AlexNet models as an example, Fig. 2 illustrates the training and validation accuracy and loss curves for the models across 5-fold cross-validation training (T1–T5). In all folds, the model achieves perfect final training accuracy (100%),

indicating complete convergence on the training data. The final validation accuracy shows some variability, ranging from 83.8% in the T4 fold to 93.8% in the T3 fold, reflecting differences in generalization performance depending on the data split. The training loss consistently approaches zero in all folds, while the final validation loss values vary from 0.26 to 1.33.

The classification performance of the AlexNet models (AlexNet-T1 to -T5) in each molar TDS classification (stages A–H) is demonstrated by their confusion matrices (Fig. 3). The matrices reveal consistently high accuracy across all folds. Most predictions are correctly classified along the diagonal elements, as indicated by the high true positive rates shown in green percentages. Most misclassifications involve adjacent stages. For instance, stages B and C, as well as stages F and G, show occasional confusion due to their transitional nature and subtle visual differences. The overall performance remains excellent across all the AlexNet models. Macro recall values consistently exceed 90%.

3.2 IMDA calculated from the predicted molar TDS

The DNN application on IMDA calculation is illustrated in Fig. 4, using an 18.14-year-old girl as an example. A trained DNN (AlexNet-T5) is applied to classify the TDS of all third molars (UL8, UR8, LL8, and LR8) of the girl (Fig. 4a). The individual tooth predictions reveal high confidence classifications, with two teeth (UR8 and LR8) showing a 100% probability of stage F predictions. The other two teeth (UL8 and LL8) show mixed classifications between stages F and G (e.g., UL8: 99% stage F and 1% stage G; LL8: 62% stage F and 38% stage G). Despite this uncertainty, both teeth are ultimately assigned to stage F based on the highest probability classification. The systematic integration results in an IMDA of 17.93 years with a 95% CI of [17.68, 18.17]. This IMDA value shows close alignment with the chronological age of 18.14 years, with a difference of approximately 0.21 years. The narrow CI demonstrates the reliability of the meta-analysis approach in integrating multiple TDSs. The weighted contribution of each tooth (UL8Ff, UR8Ff, LL8Ff, and LR8Ff) to IMDA ranges from 19.2% to 36.6%, reflecting the varying precision of age estimation for different TDSs. This example illustrates the practical application of the DNN-IMDA framework in providing reliable dental age estimation that accounts for the uncertainty in individual TDS classifications.

3.3 Model performance in dental age assessment

The model performance of the 40 trained DNNs (8 architectures× 5 cross-validation folds, Fig. 5) is evaluated by comparing the calculated IMDA and the gold standard (chronological age) of the 80 individuals aged 8.8 to 18.9 years (the mean age=14.2±2.7 years). The normality of IMDA and chronological age is assessed using the Shapiro–Wilk test. The results indicate that the data do not significantly deviate from a normal distribution (p>0.05). Therefore, parametric statistical tests are considered appropriate for further analysis. Each subplot displays the correlation between IMDA and chronological age, with the blue line indicating the line of identity (y=x). Across all models, there is a significant positive correlation (p< 0.001), with r consistently above 0.85 for most models and reaching as high as 0.907 and 0.905 for AlexNet-T5 and DenseNet-201-T3 models, respectively. The differences in ages between IMDA and

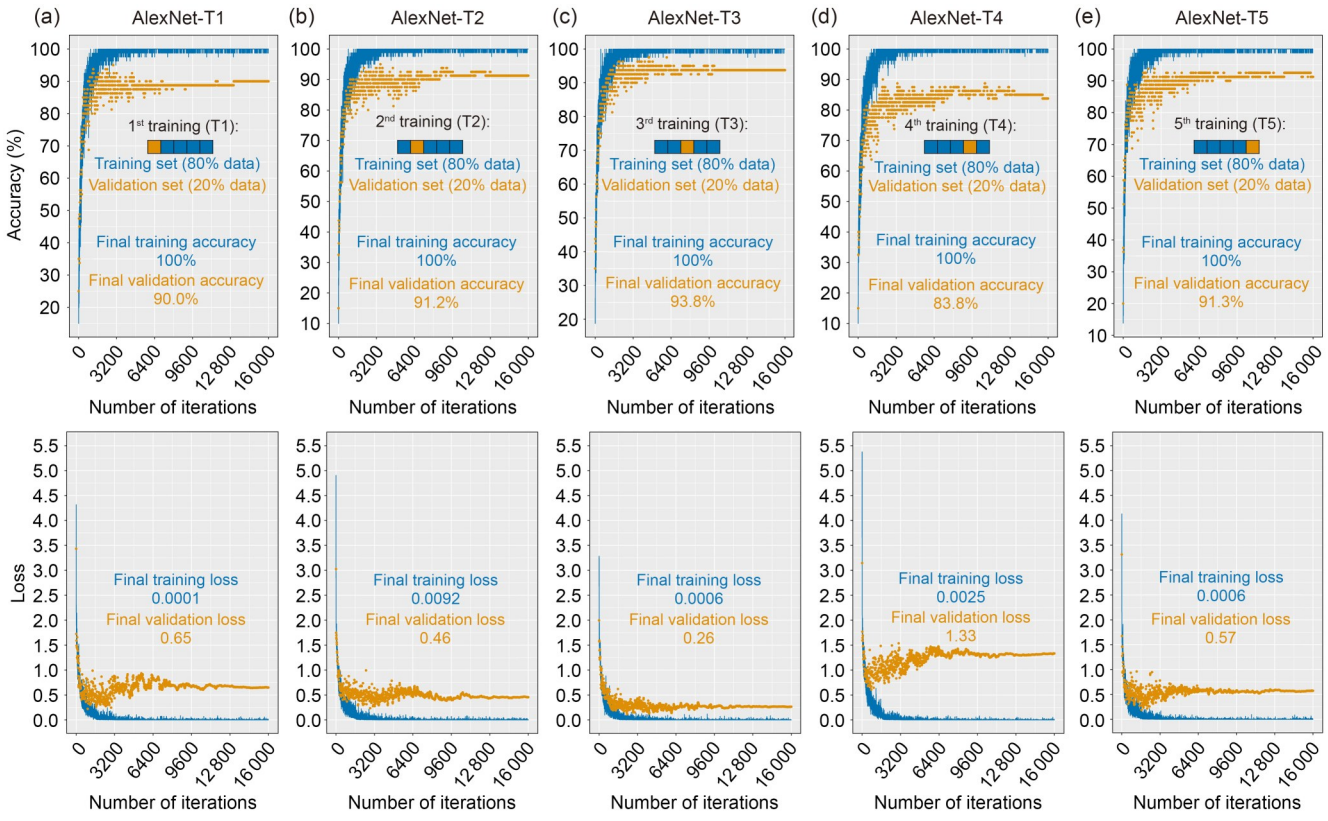

Fig. 2 Training and validation performances of AlexNet models across 5-fold cross-validation training (T1–T5) after 16 000 iterations in 4000 epochs: (a) AlexNet-T1; (b) AlexNet-T2; (c) AlexNet-T3; (d) AlexNet-T4; (e) AlexNet-T5 (blue: metrics of model performance in training; orange: metrics of model performance in validation; accuracy: classification accuracy; loss: prediction loss for 8-stage prediction)

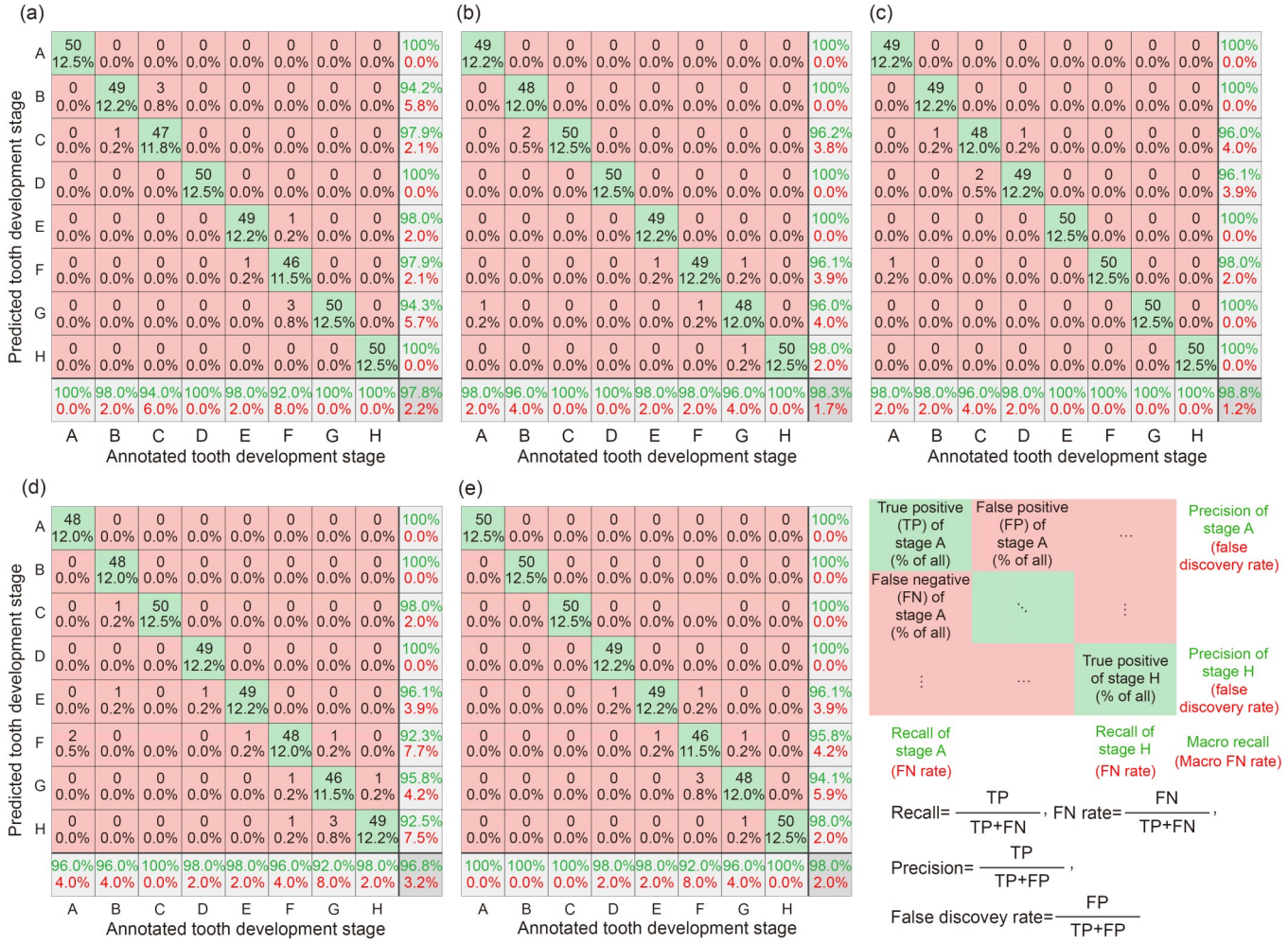

Fig. 3 Confusion matrices for classification performance of the AlexNet models in each molar TDS classification: (a) AlexNet-T1; (b) AlexNet-T2; (c) AlexNet-T3; (d) AlexNet-T4; (e) AlexNet-T5 (green diagonal cells: the counts that are correctly classified for that class; red off-diagonal cells: the counts that are misclassified)

chronological age are not statistically significantly different for all the DNN models (p>0.05). The mean differences between IMDA from the AlexNet models and chronological age are generally small, typically within ±0.25 years, and SDs range from approximately 1.30 to 1.24 years.

4 Discussion

This study demonstrates that DNN architectures, particularly DenseNet-201 and AlexNet, offer significant advantages in the automated classification of Demirjian's molar TDS and dental age estimation when combined with an ethnicity-specific southern Chinese RDS (Jayaraman et al., 2016b). DenseNet-201 achieves the highest overall validation accuracy (93.0%±2.1%) and the lowest overall validation loss (0.39±0.23) in molar TDS classification. The overall validation accuracies of VGG-16 (92.3%±1.6%), GoogLeNet (91.5%±3.6%), and AlexNet (90.0%±3.8%) are higher than 90%. The comparative analysis across eight DNN architectures further reveals that model complexity and connectivity patterns can affect model performance in dental image analysis (Table 3). DenseNet-201's (Huang et al., 2017) dense connectivity facilitates feature reuse and mitigates vanishing gradient issues, likely contributing to its high accuracy and balanced performance metrics (macro recall and F1 score). In contrast, lighter models like Xception (Chollet, 2017), while computationally efficient, struggle to capture the subtle morphological differences across adjacent TDS, resulting in lower and more variable performance (the overall validation accuracy ranging from 82.5% to 90.0%). AlexNet (Krizhevsky et al., 2017), an early but influential architecture, demonstrates consistent convergence on training data but shows variability in validation performance depending on data partitioning (Fig. 2), with overall validation accuracy ranging from 83.8% (T4 fold) to 93.8% (T3 fold) across folds. This highlights the importance of 5-fold cross-validation protocols to ensure model generalizability and avoid overfitting specific data splits.

Data augmentation, including random horizontal reflections and spatial translations, is implemented to enhance model robustness and prevent overfitting. These strategies effectively increase the diversity of the training data set, enabling the models to generalize better to unseen data while preserving the semantic integrity of the TDS annotation. The augmentation strategy includes random horizontal reflection and random spatial translations in both x and y directions within a pixel range of [−30, 30]. These transformations are applied dynamically during training to create variations in the

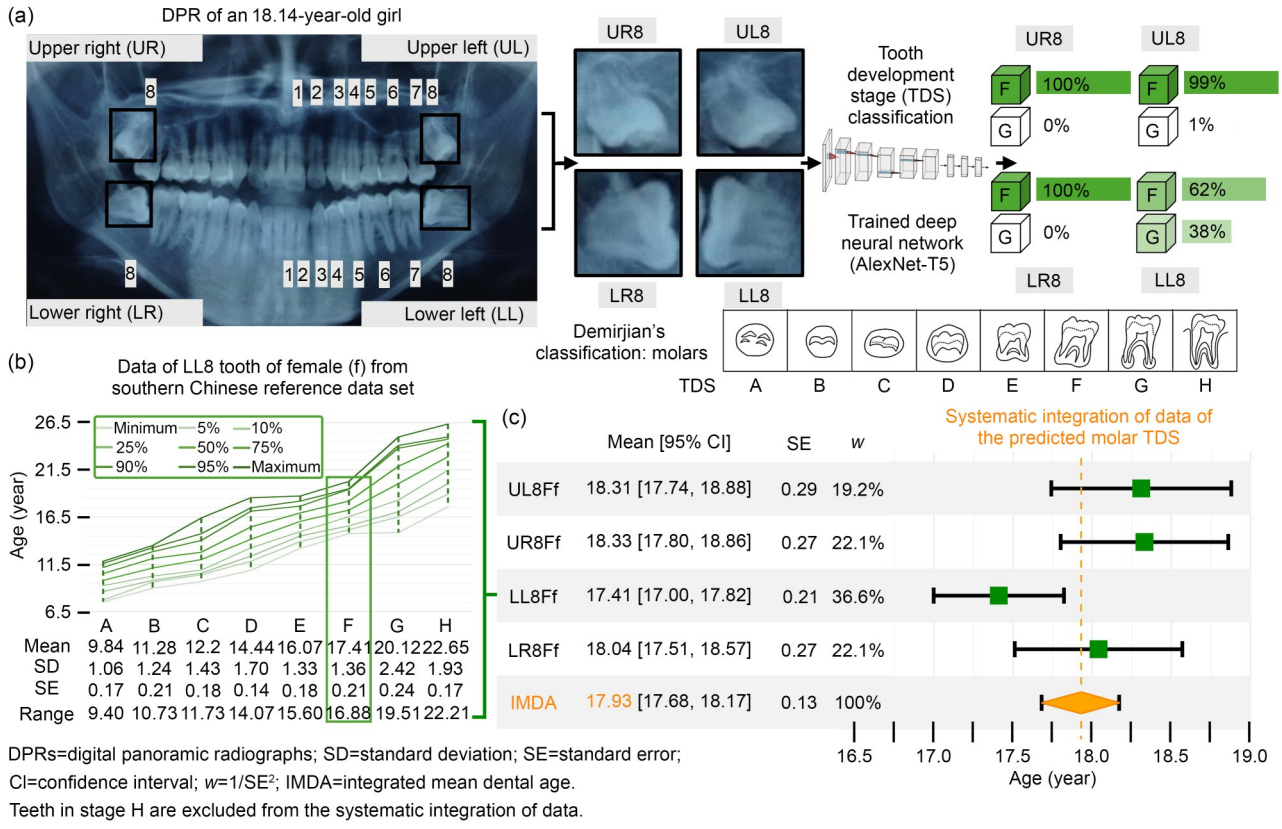

Fig. 4 IMDA calculation example for an 18.14-year-old girl using the DNN-IMDA framework integrating the trained DNNs and the southern Chinese RDS from her DPR: (a) TDS classification of the third molars (UL8, UR8, LL8, and LR8) using the trained DNN and AlexNet-T5 (green bar is the predicted probability of TDSs (stage F or G)); (b) statistical data of LL8Ff from the southern Chinese RDS (see Table S17 in the supplementary materials); (c) systematic integration of data of the predicted molar TDS (error bar in the forest plot indicates the 95% CI of each predicted molar TDS and IMDA; orange dashed line: IMDA; green box: mean age of the predicted molar TDS)

input images while preserving the semantic content and TDS annotation. The augmented training data set is prepared using an image augmenter that applies these transformations randomly to each input image, effectively expanding the training data set size and improving model generalization capabilities. Introducing such augmentations also simulates the types of image distortions and variability commonly encountered in real clinical settings, such as shifts in patient positioning or minor movement during panoramic radiograph acquisition. This ensures that the model performs reliably even when presented with clinically realistic variations in input data. Analysis of confusion matrices reveals that most misclassifications occur between adjacent TDSs, particularly in intermediate TDSs (e.g., B vs. C and F vs. G), sharing subtle morphological features (Fig. 3). In contrast, stages A and H, representing the earliest and latest developmental phases, achieve near-perfect classification due to their distinct radiographic characteristics. This pattern mirrors clinical reality, where dental professionals often face challenges in distinguishing borderline cases. It therefore supports the clinical relevance of the automated approach.

The case example of the 18.14-year-old girl demonstrates the practical effectiveness of the DNN-IMDA framework in real-world applications (Fig. 4). There is close alignment between the estimated age (17.93 years) and chronological age (18.14 years), with a difference of only 0.21 years. The mixed classifications observed in some teeth (e.g., LL8: 62% stage F and 38% stage G) reflect the inherent biological variability in tooth development and the transitional nature of adjacent TDSs. This uncertainty pattern aligns with the established clinical observations that consecutive TDS share morphological similarities, making precise classification challenging even for trained experts. The decision to assign teeth to the highest probability stage mirrors standard clinical practice in borderline cases and demonstrates the model's ability to handle diagnostic uncertainty systematically. The meta-analysis approach using weighted TDS predictions reveals differential reliability in age estimation based on tooth position and development characteristics. The narrow 95% CI [17.68, 18.17] in the case example demonstrates a key advantage of the meta-analysis approach over single-tooth assessments. By systematically integrating multiple TDSs with appropriate weighting based on their individual reliability, the framework provides more robust age estimation with quantified uncertainty. This approach addresses a significant limitation in traditional dental age estimation methods, which often lack explicit uncertainty quantification. Since TDS prediction accuracy is crucial for IMDA calculation, the performance of dental age estimation is directly associated with model performance in molar TDS classification. For instance, Xception's lower performance in TDS classification (validation accuracy of Xception-T1=82.5%) results in correspondingly lower correlation coefficients between IMDA and chronological age (r=0.677, p< 0.001, in T1). Conversely, the higher performances of AlexNet (r= 0.907, p<0.001, in T5) and DenseNet-201 (r=0.905, p<0.001, in T3) reflect their enhanced classification capabilities (validation accuracy of AlexNet-T5=91.3%; validation accuracy of DenseNet-201-T3=

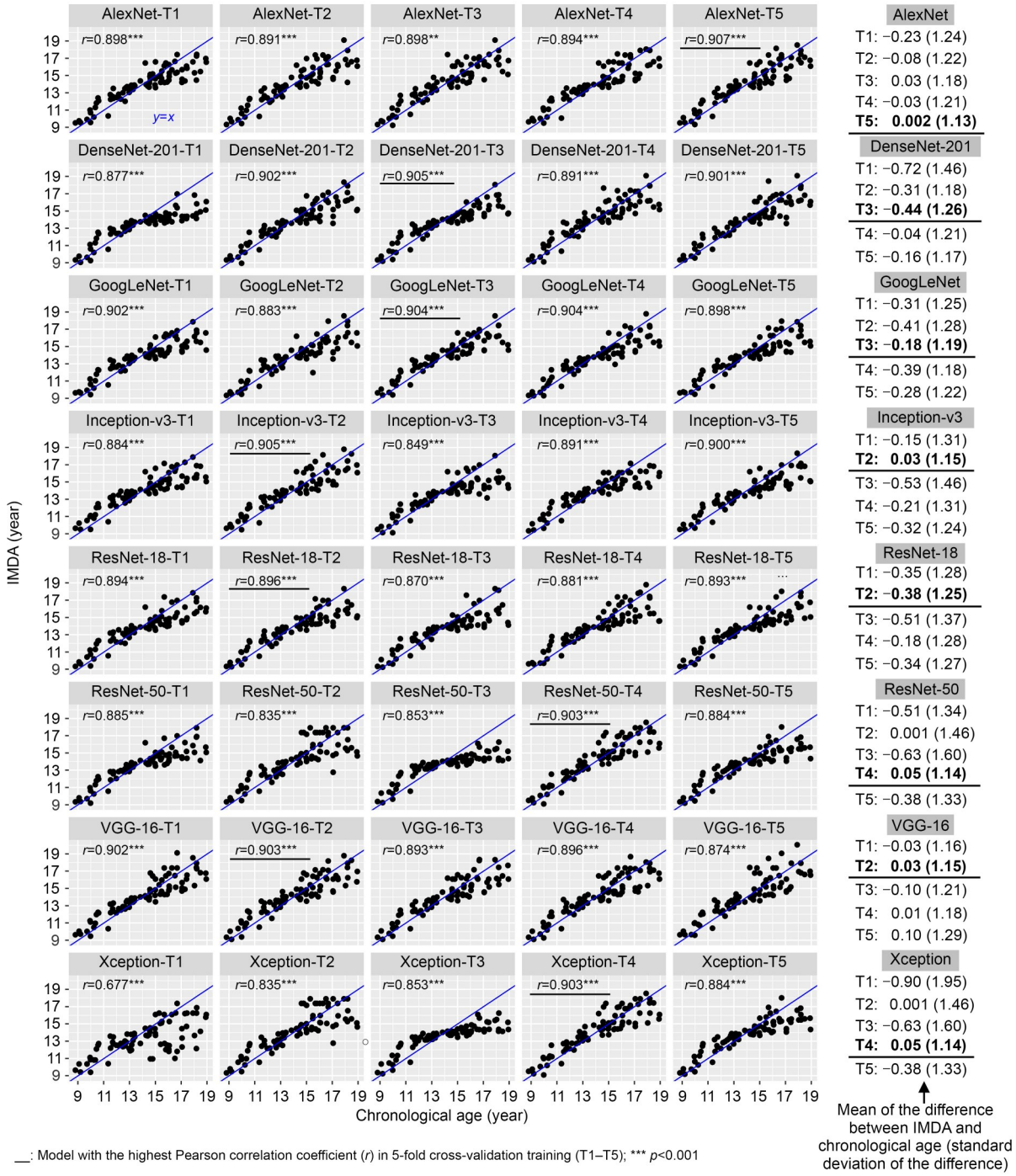

Fig. 5 Scatterplots of IMDA obtained from the 40 trained models (8 deep neural network×5 cross-validation folds) against chronological age (blue line represents the line of identity with $y=x$ and IMDA=chronological age)

95.0%). The consistent statistical results ($p<0.001$ in Pearson correlation tests and $p>0.05$ in paired t-tests) across all DNN models validate the fundamental relationships between TDS predictions and chronological age (Fig. 5). The mean difference between IMDA and chronological age is found to be the smallest with −0.063 years (−3.3 weeks) and their mean correlation coefficient is the highest with $r=0.898$ ($p<0.001$). Notably, although AlexNet is not the model with the highest overall validation accuracy in molar TDS classification, it achieves the best performance in dental age estimation with the smallest mean difference, highlighting the complex relationship between classification accuracy and age estimation precision. Moreover, the achieved correlation coefficients (consistently above 0.85, reaching 0.907) align with or exceed those reported in recent literature for automated dental age estimation systems. Several possible explanations are considered for this discrepancy, including differences in prediction confidence distributions, distinct misclassification patterns, and feature extraction biases that may affect IMDA mapping. However, detailed model-specific error analysis (confusion matrices, softmax distributions, and error propagation studies) is required for definitive interpretation. Since the current study's primary focus is DNN-IMDA framework validation, comprehensive model dissection should be systematically investigated in future work through

confusion matrix comparisons, prediction confidence analysis, and error propagation studies to elucidate the black-box behavior across different DNN architectures and improve methodological transparency. Previous studies using DNN approaches have reported correlations ranging from 0.50 to 0.90 (Oliveira et al., 2024), with the present DNN-IMDA framework achieving values at the upper end of this range. SDs (approximately 1.3–1.24 years) are comparable to those reported in the southern Chinese RDS, indicating that the automated approach maintains the precision of traditional manual methods (Jayaraman et al., 2016b), while offering significant advantages in terms of speed, consistency, and scalability.

The integration of DNN-based molar TDS classification with an ethnicity-specific RDS addresses a longstanding limitation in dental age estimation: the impact of ethnicity-specific dental development patterns on model accuracy. Previous studies have shown that applying RDSs from different ethnicities can lead to substantial over- or underestimation of age (Lee et al., 2008; Cruz-Landeira et al., 2010). Southern China (the region south of the Huai River and the Qin Mountains) has been established as the place of origin for many ethnic groups throughout east Asia and the Pacific (Stoneking and Delfin, 2010). Recent genetic studies, based on mitochondrial DNA and Y chromosome allelic markers, have suggested that the migration of human populations started in southern China and expanded slowly to other regions in east Asia, with people carrying similar gene subsets as they migrated (Li H et al., 2007; Normile, 2009). East Asian population groups share 90% of the major haplotypes and exhibit phenotypic traits common to southern Chinese, as evidenced by their physical appearance, such as the characteristic epicanthic fold and shovel-shaped incisors (Li JZ et al., 2008; Abdulla et al., 2009). The genetic similarity among east Asians presents an opportunity to compare the dental maturity patterns between the southern Chinese and other east Asians.

The proposed DNN-IMDA framework sets a new standard for ethnicity-appropriate and automated dental age assessment. It can minimize the need for expert annotation, significantly reduce assessment time, and achieve high agreement with chronological age from the ethnicity-specific RDS. The results make the DNN-IMDA framework suitable for clinical, forensic, and legal ethnicity-based applications. The integration of probabilistic outputs with CIs provides clinicians and forensic experts with essential information for decision-making in contexts where age determination carries legal or clinical consequences. The framework's ability to handle classification uncertainty transparently addresses a critical need in forensic applications, where courts increasingly demand explicit quantification of measurement uncertainty. The weighted contribution approach also offers insights into which teeth provide the most reliable age information in specific cases, potentially informing clinical decision-making when only partial dental information is available. This feature is particularly valuable in forensic contexts, where dental evidence may be incomplete or damaged. The DNN-IMDA framework is also adaptable, allowing for integration with other ethnicity-specific RDS as they become available. Future research should explore multicenter and multi-ethnic validation studies and the integration of explainable artificial intelligence techniques to enhance interpretability and clinical trust.

The DNN models for TDS classification demonstrate good performance and architectural generalizability, while the IMDA calculation requires ethnicity-specific RDS like the southern Chinese RDS (Jayaraman et al., 2016b) due to well-documented population differences in dental development (Jayaraman et al., 2013). International readers should note that the framework's DNN component is transferable, but reliable dental age estimation necessitates population-matched reference data for accurate TDS-to-IMDA mapping. Clinical implementation in other populations requires only ethnicity-appropriate RDS integration with the validated DNN models of TDS classification.

5 Conclusions

In short, this study demonstrates that the DNN-IMDA framework, which combines advanced DNN architectures with an ethnicity-specific RDS, enables accurate and reliable dental age estimation for southern Chinese individuals. The first part of the DNN-IMDA framework is the application of DNNs for the automated classification of Demirjian's TDS of molars on DPRs. Eight DNN architectures are evaluated for their performance in Demirjian's molar TDS classification. Among the architectures, DenseNet-201 achieves the highest mean validation accuracy of 93.0%±2.1%, with macro recall and macro F1 scores of 0.985±0.008 and 0.984±0.008, respectively, and the lowest validation loss (0.39±0.23), demonstrating excellent performance in molar TDS classification. Most misclassifications occur between adjacent stages, reflecting the inherent challenge of distinguishing subtle transitional features in dental development. The trained DNN models can serve as artificial experts that can quickly and automatically classify Demirjian's molar TDS, saving time compared to the time-consuming manual annotation by trained experts in the traditional way.

The second part of the DNN-IMDA framework is the ethnicity-appropriate dental age estimation method in terms of IMDA. The IMDA calculation is based on the meta-analysis approach for the statistics in an ethnicity-specific RDS. In this study, the DNN-predicted TDS is mapped to the data from the most complete southern Chinese RDS. The results of the IMDA method show strong agreement with the gold standard (chronological age). The maximum Pearson correlation coefficients between IMDA and chronological age in 5-fold cross-validation training are found to exceed 0.90 (p<0.001), except for ResNet-18 (r=0.896, p<0.001). These results meet the validity criteria for dental age estimation and highlight the reliability and precision of the DNN-IMDA approach. The lowest mean difference and highest mean correlation coefficient between the estimated IMDA from the best-performing DNN (AlexNet) and chronological age are −0.063 years (−3.3 weeks) and r=0.898 (p<0.001), respectively. The population differences in dental development patterns affecting the accuracy of age estimation could be effectively addressed by remapping the predicted TDS to ethnicity-specific RDSs. The DNN-IMDA framework minimizes the need for expert annotation, significantly reduces assessment time, and provides accurate, and reproducible dental age estimation suitable for clinical, forensic, and legal applications. This approach sets a new standard for ethnicity-appropriate dental age assessment and has the potential for broader applications across other populations, pending further validation.

Supplementary information

The online version contains supplementary materials available at https://doi.org/10.1631/ENG.ITEE.2026.0013.

Acknowledgments

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. 17102921).

Author contributions

Kuen Wai MA and Hai Ming WONG designed the research. Kuen Wai MA processed the data. Kuen Wai MA drafted the paper. Jiajun ZHU and Hai Ming WONG helped organize the paper. Kuen Wai MA, Jiajun ZHU, and Hai Ming WONG revised and finalized the paper.

Compliance with ethics guidelines

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the University of Hong Kong-West Cluster Hospital Authority (IRB No. UW 12-280). Informed consent was obtained from all subjects involved in the study.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Abdulla MA, Ahmed I, Assawamakin A, et al., 2009. Mapping human genetic diversity in Asia. *Science*, 326(5959):1541-1545. https://doi.org/10.1126/science.1177074

Ahn J, Nguyen TP, Kim YJ, et al., 2022. Automated analysis of three-dimensional CBCT images taken in natural head position that combines facial profile processing and multiple deep-learning models. *Comput Methods Programs Biomed*, 226: 107123. https://doi.org/10.1016/j.cmpb.2022.107123

Aljameel SS, Althumairy L, Albassam B, et al., 2023. Predictive artificial intelligence model for detecting dental age using panoramic radiograph images. *Big Data Cogn Comput*, 7(1):8. https://doi.org/10.3390/bdcc7010008

AlQahtani SJ, Hector MP, Liversidge HM, 2010. Brief communication: the London atlas of human tooth development and eruption. *Am J Phys Anthropol*, 142(3):481-490. https://doi.org/10.1002/ajpa.21258

Banar N, Bertels J, Laurent F, et al., 2020. Towards fully automated third molar development staging in panoramic radiographs. *Int J Leg Med*, 134(5):1831-1841. https://doi.org/10.1007/s00414-020-02283-3

Baydoğan MP, Baybars SC, Tuncer SA, 2023. Age-Net: an advanced hybrid deep learning model for age estimation using orthopantomograph images. *Trait Signal*, 40(4): 1553-1563. https://doi.org/10.18280/ts.400423

Büyükçakır B, Bertels J, Claes P, et al., 2024. OPG-based dental age estimation using a data-technical exploration of deep learning techniques. *J Forens Sci*, 69(3):919-931. https://doi.org/10.1111/1556-4029.15473

Chollet F, 2017. Xception: deep learning with depthwise separable convolutions. IEEE Conf on Computer Vision and Pattern Recognition, p.1800-1807. https://doi.org/10.1109/CVPR.2017.195

Cruz-Landeira A, Linares-Argote J, Martínez-Rodríguez M, et al., 2010. Dental age estimation in Spanish and Venezuelan children. Comparison of Demirjian and Chaillet's scores. *Int J Leg Med*, 124(2):105-112. https://doi.org/10.1007/s00414-009-0380-5

Demirjian A, Goldstein H, 1976. New systems for dental maturity based on seven and four teeth. *Ann Human Biol*, 3(5):411-421. https://doi.org/10.1080/03014467600001671

Demirjian A, Goldstein H, Tanner JM, 1973. A new system of dental age assessment. *Human Biol*, 45(2):211-227.

de Tobel J, Radesh P, Vandermeulen D, et al., 2017. An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. *J Forens Odontostomatol*, 35(2):42-54.

Fan F, Ke WC, Dai XH, et al., 2023. Semi-supervised automatic dental age and sex estimation using a hybrid transformer model. *Int J Leg Med*, 137(3):721-731. https://doi.org/10.1007/s00414-023-02956-9

Franco A, Murray J, Heng D, et al., 2024. Binary decisions of artificial intelligence to classify third molar development around the legal age thresholds of 14, 16 and 18 years. *Sci Rep*, 14(1):4668. https://doi.org/10.1038/s41598-024-55497-5

Gu JX, Wang ZH, Kuen J, et al., 2018. Recent advances in convolutional neural networks. *Patt Recogn*, 77:354-377. https://doi.org/10.1016/j.patcog.2017.10.013

Han MQ, Du SY, Ge YY, et al., 2022. With or without human interference for precise age estimation based on machine learning? *Int J Leg Med*, 136(3):821-831. https://doi.org/10.1007/s00414-022-02796-z

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Huang G, Liu Z, van der Maaten L, et al., 2017. Densely connected convolutional networks. IEEE Conf on Computer Vision and Pattern Recognition, p.2261-2269. https://doi.org/10.1109/CVPR.2017.243

Jayaraman J, Roberts G, 2016. Demirjian's method is unsuitable for dental age estimation. *Forens Sci Med Pathol*, 12(4):532-533. https://doi.org/10.1007/s12024-016-9811-z

Jayaraman J, King NM, Roberts GJ, et al., 2011. Dental age assessment: are Demirjian's standards appropriate for southern Chinese children? *J Forens Odontostomatol*, 29(2):22-28.

Jayaraman J, Roberts GJ, King NM, et al., 2012. Dental age assessment of southern Chinese using the United Kingdom Caucasian reference dataset. *Forens Sci Int*, 216(1-3):68-72. https://doi.org/10.1016/j.forsciint.2011.08.019

Jayaraman J, Wong HM, King NM, et al., 2013. The French–Canadian data set of Demirjian for dental age estimation: a systematic review and meta-analysis. *J Forens Leg Med*, 20(5):373-381. https://doi.org/10.1016/j.jflm.2013.03.015

Jayaraman J, Roberts GJ, Wong HM, et al., 2016a. Ages of legal importance: implications in relation to birth registration and age assessment practices. *Med Sci Law*, 56(1):77-82. https://doi.org/10.1177/0025802415590172

Jayaraman J, Wong HM, King NM, et al., 2016b. Development of a reference data set (RDS) for dental age estimation (DAE) and testing of this with a separate validation set (VS) in a southern Chinese population. *J Forens Leg Med*, 43:26-33. https://doi.org/10.1016/j.jflm.2016.07.007

Jayaraman J, Roberts GJ, Wong HM, et al., 2018. Dental age estimation in southern Chinese population using panoramic radiographs: validation of three population specific reference datasets. *BMC Med Imag*, 18(1):5. https://doi.org/10.1186/s12880-018-0250-z

Jayaraman J, Wong HM, Roberts GJ, et al., 2019. Age estimation in three distinct east Asian population groups using southern Han Chinese dental reference dataset. *BMC Oral Health*, 19(1):242. https://doi.org/10.1186/s12903-019-0942-y

Kahaki SMM, Nordin MJ, Ahmad NS, et al., 2020. Deep convolutional neural network designed for age assessment based on orthopantomography data. *Neur Comput Appl*, 32(13):9357-9368. https://doi.org/10.1007/s00521-019-04449-6

Kim M, Chung M, Shin YG, et al., 2023. Automatic registration of dental CT and 3D scanned model using deep split jaw and surface curvature. *Comput Methods Programs Biomed*, 233:107467. https://doi.org/10.1016/j.cmpb.2023.107467

Kim S, Lee YH, Noh YK, et al., 2021. Age-group determination of living individuals using first molar images based on artificial intelligence. *Sci Rep*, 11(1):1073. https://doi.org/10.1038/s41598-020-80182-8

Kim YR, Choi JH, Ko J, et al., 2023. Age group classification of dental radiography without precise age information using convolutional neural networks. *Healthcare*, 11(8):1068. https://doi.org/10.3390/healthcare11081068

Koch RM, Mentzel HJ, Heinrich A, 2025. Deep learning for forensic age estimation using orthopantomograms in children, adolescents, and young adults. *Eur Radiol*, 35(7):4191-4202. https://doi.org/10.1007/s00330-025-11373-y

Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90. https://doi.org/10.1145/3065386

Lai MC, King NM, Wong HM, 2008. Dental development of Chinese children with cleft lip and palate. *Cleft Palate Craniofac J*, 45(3):289-296. https://doi.org/10.1597/07-019

Lee SE, Lee SH, Lee JY, et al., 2008. Age estimation of Korean children based on dental maturity. *Forens Sci Int*, 178(2-3):125-131. https://doi.org/10.1016/j.forsciint.2007.12.010

Li H, Cai XY, Winograd-Cort ER, et al., 2007. Mitochondrial DNA diversity and population differentiation in southern East Asia. *Am J Phys Anthropol*, 134(4):481-488. https://doi.org/10.1002/ajpa.20690

Li JZ, Absher DM, Tang H, et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100-1104. https://doi.org/10.1126/science.1153717

Maber M, Liversidge HM, Hector MP, 2006. Accuracy of age estimation of radiographic methods using developing teeth. *Forens Sci Int*, 159(Suppl 1):S68-S73.

https://doi.org/10.1016/j.forsciint.2006.02.019

Mallineni SK, Jayaraman J, Wong HM, et al., 2019. Dental development in children with supernumerary teeth in the anterior region of maxilla. *Clin Oral Investig*, 23(7): 2987-2994. https://doi.org/10.1007/s00784-018-2709-2

Matthijs L, Delande L, de Tobel J, et al., 2024. Artificial intelligence and dental age estimation: development and validation of an automated stage allocation technique on all mandibular tooth types in panoramic radiographs. *Int J Leg Med*, 138(6):2469-2479. https://doi.org/10.1007/s00414-024-03298-w

Merdietio Boedi R, Banar N, de Tobel J, et al., 2020. Effect of lower third molar segmentations on automated tooth development staging using a convolutional neural network. *J Forens Sci*, 65(2):481-486. https://doi.org/10.1111/1556-4029.14182

Migliorelli L, Massini LS, Coccia M, et al., 2023. A deep learning-based telemonitoring application to automatically assess oral diadochokinesis in patients with bulbar amyotrophic lateral sclerosis. *Comput Methods Programs Biomed*, 242:107840. https://doi.org/10.1016/j.cmpb.2023.107840

Milani OH, Atici SF, Allareddy V, et al., 2024. A fully automated classification of third molar development stages using deep learning. *Sci Rep*, 14(1):13082. https://doi.org/10.1038/s41598-024-63744-y

Mohammad N, Ahmad R, Gaus MHA, et al., 2024. Accuracy of automated forensic dental age estimation lab (F-DentEst Lab) on large Malaysian dataset. *Forens Sci Int*, 361:112150. https://doi.org/10.1016/j.forsciint.2024.112150

Mualla N, Houssein EH, Hassan MR, 2020. Dental age estimation based on X-ray images. *Comput Mater Con*, 62(2):591-605. https://doi.org/10.32604/cmc.2020.08580

Nolla CM, 1960. The development of the permanent teeth. *J Dent Children*, 27:254-266.

Normile D, 2009. SNP study supports southern migration route to Asia. *Science*, 326(5959): 1470. https://doi.org/10.1126/science.326.5959.1470

Oliveira W, Santos MA, Burgardt CAP, et al., 2024. Estimation of human age using machine learning on panoramic radiographs for Brazilian patients. *Sci Rep*, 14(1): 19689. https://doi.org/10.1038/s41598-024-70621-1

Olze A, Bilang D, Schmidt S, et al., 2005. Validation of common classification systems for assessing the mineralization of third molars. *Int J Leg Med*, 119(1):22-26. https://doi.org/10.1007/s00414-004-0489-5

Ong SH, Kim H, Song JS, et al., 2024. Fully automated deep learning approach to dental development assessment in panoramic radiographs. *BMC Oral Health*, 24(1):426. https://doi.org/10.1186/s12903-024-04160-6

Ozlu Ucan G, Gwassi OAH, Apaydin BK, et al., 2025. Automated age estimation from OPG images and patient records using deep feature extraction and modified genetic-random forest. *Diagnostics*, 15(3):314. https://doi.org/10.3390/diagnostics15030314

Pintana P, Upalananda W, Saekho S, et al., 2022. Fully automated method for dental age estimation using the ACF detector and deep learning. *Egypt J Forens Sci*, 12(1): 54. https://doi.org/10.1186/s41935-022-00314-1

Roberts G, Lucas VS, McDonald F, et al., 2017. In our opinion. *Br Dent J*, 222(12): 918-921. https://doi.org/10.1038/sj.bdj.2017.537

Şahin TN, Kölüş T, 2024. Age and sex estimation in children and young adults using panoramic radiographs with convolutional neural networks. *Appl Sci*, 14(16):7014. https://doi.org/10.3390/app14167014

Sathyavathi S, Baskaran KR, 2024. Human age estimation using deep convolutional neural network based on dental images (orthopantomogram). *IETE J Res*, 70(2): 1585-1592. https://doi.org/10.1080/03772063.2023.2165177

Shang SN, Shi YQ, Zhang YJ, et al., 2024. Artificial intelligence for brain disease diagnosis using electroencephalogram signals. *J Zhejiang Univ-Sci B*, 25(10):914-940. https://doi.org/10.1631/jzus.B2400103

Sharifonnasabi F, Jhanjhi NZ, John J, et al., 2022. Hybrid HCNN-KNN model enhances age estimation accuracy in orthopantomography. *Front Publ Health*, 10:879418. https://doi.org/10.3389/fpubh.2022.879418

Shi YC, Ye ZL, Guo JX, et al., 2024. Deep learning methods for fully automated dental age estimation on orthopantomograms. *Clin Oral Invest*, 28(3):198. https://doi.org/10.1007/s00784-024-05598-2

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556

Sivri MB, Taheri S, Kırzıoğlu Ercan RG, et al., 2024. Dental age estimation: a comparative study of convolutional neural network and Demirjian's method. *J Forens Leg Med*, 103:102679. https://doi.org/10.1016/j.jflm.2024.102679

Sokolova M, Lapalme G, 2009. A systematic analysis of performance measures for classification tasks. *Inform Process Manag*, 45(4):427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Stoneking M, Delfin F, 2010. The human genetic history of East Asia: weaving a complex tapestry. *Curr Biol*, 20(4):R188-R193. https://doi.org/10.1016/j.cub.2009.11.052

Szegedy C, Liu W, Jia YQ, et al., 2015. Going deeper with convolutions. IEEE Conf on Computer Vision and Pattern Recognition, p.1-9. https://doi.org/10.1109/CVPR.2015.7298594

Szegedy C, Vanhoucke V, Ioffe S, et al., 2016. Rethinking the inception architecture for computer vision. IEEE Conf on Computer Vision and Pattern Recognition, p.2818-2826. https://doi.org/10.1109/CVPR.2016.308

Vila-Blanco N, Carreira MJ, Varas-Quintana P, et al., 2020. Deep neural networks for chronological age estimation from OPG images. *IEEE Trans Med Imag*, 39(7):2374-2384. https://doi.org/10.1109/TMI.2020.2968765

Wang J, Dou JW, Han JX, et al., 2023. A population-based study to assess two convolutional neural networks for dental age estimation. *BMC Oral Health*, 23(1):109. https://doi.org/10.1186/s12903-023-02817-2

Wang P, Bai YL, Xiao Y, et al., 2025. Aberrant network topological structure of sensorimotor superficial white-matter system in major depressive disorder. *J Zhejiang Univ-Sci B*, 26(1):39-51. https://doi.org/10.1631/jzus.B2300880

Wang XD, Liu YL, Miao XY, et al., 2022. DENSEN: a convolutional neural network for estimating chronological ages from panoramic radiographs. *BMC Bioinform*, 23(3): 426. https://doi.org/10.1186/s12859-022-04935-0

Wodzinski M, Daniol M, Socha M, et al., 2022. Deep learning-based framework for automatic cranial defect reconstruction and implant modeling. *Comput Methods Programs Biomed*, 226:107173. https://doi.org/10.1016/j.cmpb.2022.107173

Wong HM, Peng SM, Wen YF, et al., 2016a. A population survey of early factors associated with permanent tooth emergence: findings from a prospective cohort study. *Commun Dent Oral Epidemiol*, 44(5):467-475. https://doi.org/10.1111/cdoe.12236

Wong HM, Wen YF, Jayaraman J, et al., 2016b. Northern Chinese dental ages estimated from southern Chinese reference datasets closely correlate with chronological age. *Heliyon*, 2(12):e00216. https://doi.org/10.1016/j.heliyon.2016.e00216

Wong HM, Peng SM, Yang YQ, et al., 2017. Tooth eruption and obesity in 12-year-old children. *J Dent Sci*, 12(2):126-132. https://doi.org/10.1016/j.jds.2016.10.004

Wong HM, Peng SM, McGrath CPJ, 2019. Association of infant growth with emergence of permanent dentition among 12 year-aged southern Chinese school children. *BMC Oral Health*, 19(1):47. https://doi.org/10.1186/s12903-019-0737-1

Yavuz BS, Ekmekcioglu O, Ankarali H, 2025. Comparison of different dental age estimation methods with deep learning: Willems, Cameriere-European, London Atlas. *Int J Leg Med*, 139(4):1661-1672. https://doi.org/10.1007/s00414-025-03452-y

Yeom HG, Lee BD, Lee W, et al., 2023. Estimating chronological age through learning local and global features of panoramic radiographs in the Korean population. *Sci Rep*, 13(1):21857. https://doi.org/10.1038/s41598-023-48960-2