



A Novel Dual-Output Deep Learning Model Based on InceptionV3 for Radiographic Bone Age and Gender Assessment

Baraa Rayed¹ · Hakan Amasya^{2,3,5} · Mana Sezdi^{4,5}

Received: 30 April 2025 / Revised: 26 June 2025 / Accepted: 17 July 2025
© The Author(s) under exclusive licence to Society for Imaging Informatics in Medicine 2025

Abstract

Hand-wrist radiographs are used in bone age prediction. Computer-assisted clinical decision support systems offer solutions to the limitations of the radiographic bone age assessment methods. In this study, a multi-output prediction model was designed to predict bone age and gender using digital hand-wrist radiographs. The InceptionV3 architecture was used as the backbone, and the model was trained and tested using the open-access dataset of 2017 RSNA Pediatric Bone Age Challenge. A total of 14,048 samples were divided to training, validation, and testing subsets with the ratio of 7:2:1, and additional specialized convolutional neural network layers were implemented for robust feature management, such as Squeeze-and-Excitation block. The proposed model achieved a mean squared error of approximately 25 and a mean absolute error of 3.1 for predicting bone age. In gender classification, an accuracy of 95% and an area under the curve of 97% were achieved. The intra-class correlation coefficient for the continuous bone age predictions was found to be 0.997, while the Cohen's κ coefficient for the gender predictions was found to be 0.898 ($p < 0.001$). The proposed model aims to increase model efficiency by identifying common and discrete features. Based on the results, the proposed algorithm is promising; however, the mid-high-end hardware requirement may be a limitation for its use on local machines in the clinic. The future studies may consider increasing the dataset and simplification of the algorithms.

Keywords X-ray · Deep learning · Automatic bone age · Convolutional neural network (CNN)

Introduction

Radiographic skeletal maturity indicators can be found in the bones of the hand and wrist, cervical vertebrae, knee, hip, and foot. Several types of radiographic data contain such maturity indicators on the image to be used in skeletal maturity assessment, based on the imaging method [1]. Skeletal maturation is important in orthopedics, pediatrics, and orthodontics, and more. Evaluating growth and development may be critical in clinical decisions regarding treatment alternatives [2]. Similarly, forensic sciences may benefit from skeletal maturity assessment to obtain evidence for tasks such as victim identification [3].

Hand-wrist maturation (HWM) assessment is a widely accepted method for evaluating skeletal maturation using specific maturity indicators of the hand and wrist bones in hand-wrist radiographs (HWRs). Several HWM methods were reported in the literature, based on two main approaches: the atlas method, in which sample images are

already provided for each maturation stage for an overall comparison, or the region of interest method, suggesting investigation of the specific bone maturity indicators on specific bones. Greulich and Pyle (GP) and Tanner-Whitehouse (TW) reported separate atlases for HWM in 1959 and 1962, respectively, and the latter was further revised as TW2 in 1975 and TW3 in 2001 [4, 5]. Alternatively, the Fishman method focuses on specific bones in the first, third, and fifth phalanges and the radius bone to determine the maturation stage based on the morphology of the specific indicators [6]. Cervical vertebra maturation (CVM) using lateral cephalograms is another method of radiographic skeletal maturation assessment [7, 8].

Traditional radiographic bone age analysis is performed in the clinic by human observers. Analyses require specially trained personnel, and the results may be influenced by several factors such as the operator experience and the subjectivity. Due to inter-observer variance, the repeatability and the reproducibility of the radiographic bone age determination methods can be considered questionable [9–11]. In addition to the radiographic techniques, various clinical

Extended author information available on the last page of the article

and laboratory tests can be performed in the analysis of skeletal maturation. However, in reliability studies of the relevant methods, it is commonly recommended that they should not be used alone but combined with other techniques as adjunctive. Although methods based on HWRs are commonly accepted, it is difficult to interpret due to the complex structure of the hand bones. For this reason, in addition to radiography equipment for image acquisition, the need for trained clinicians to evaluate the image is also a limitation of the technique to be used with different tests [12, 13]. Increasing standardization in radiographic bone age analysis has been a common goal in research for years.

Modern artificial intelligence (AI) technologies introduced in the 1950s with a common question of the health professionals and the engineers: “Can machines think?” The earliest systems were developed as rule-based expert systems and further progressed with advances in computer technologies [14, 15]. Machine learning (ML) algorithms are trained with provided datasets to establish a mathematical model between input and output values. Such data mining tools can be used to process digital medical data and provide suggestions based on the model architecture [16]. Tajmir et al. conducted a study to evaluate the effect of computer-aided skeletal bone age assessment and reported that radiographic bone age assessment with software assistance was found to be superior when compared to a single observer or the software alone [17].

Research on the use of relevant technologies in radiographic bone age analysis has been ongoing for a long time. In 1992, Tanner and Gibbons developed a computer-assisted system (CASAS) for HWM assessment. The design included digitizing analog radiographs using a video camera operated according to a template, and then computing the similarities with the atlas images. Total processing time, including the image capturing and the computing, was reported to be between 5 and 15 min, depending on the number of bones, and was suggested as useful for inexperienced users in HWM [18]. The Radiological Society of North America (RSNA) organized the Pediatric Bone Age Machine Learning Challenge in 2017 and provided 14,236 HWRs labeled according to the GP method. Accordingly, the highest-ranking models’ architectures were based on InceptionV3, ResNet-50, and a custom convolutional neural network (CNN) algorithm developed with Ice Module [19–21].

Materials and Methods

This section outlines the materials, datasets, and methodological framework employed to develop and evaluate a multi-output prediction model for skeletal maturity and

gender assessment using radiographic images. Leveraging advanced deep learning techniques and robust datasets, the study integrates established architectures and data processing strategies to achieve reliable and accurate predictions. The following subsections detail the specific approaches, including the use of transfer learning with InceptionV3, the dataset utilized, and the overall methodology.

InceptionV3 Integrated with Transfer Learning

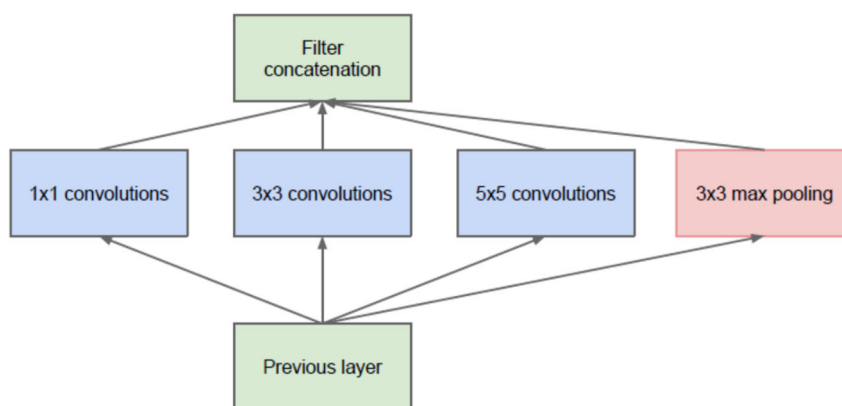
Before delving deeply into the framework of InceptionV3 powered by transfer learning, a brief literature review is presented.

InceptionV3 combined with transfer learning has attracted the attention of researchers in recent years due to its ability to achieve high performance on a small range of datasets. Zang [22] achieved a high-precision classification of five representative snakes, and Lin et al. [23] completed the classification of the German Traffic Sign Recognition Standard (GTSRB). Li et al. [24] successfully established the classification of lymph node metastasis in colorectal cancer. Meanwhile, Mednikov et al. [25] adopted the related architecture for an effective classification of the breast lumps.

InceptionV3

The Inception algorithm has a similar architecture to its predecessor, the GoogLeNet algorithm proposed by Google in 2014, which not only reduces the amount of network parameters, but also increases the network depth. Hence, it is widely used in image classification tasks. As the core of the GoogLeNet network is similar to the Inception network structure, the GoogLeNet network is also called the Inception network [26]. GoogLeNet architecture has several variants, which are mainly divided into InceptionV1 (2014), InceptionV2 (2015), InceptionV3 (2015), InceptionV4 (2016), and Inception-ResNet (2016).

Typically, the Inception module has one maximum pooling and three distinct convolution layers. Following the convolution process, the channel is aggregated for the network output of the preceding layer, as a nonlinear fusion is subsequently carried out. By doing this, overfitting may be avoided, and the network’s ability to expression and adaptability to various scales can be enhanced. The Inception network structure is displayed in Fig. 1. In contrast to Inception V1 and V2, the InceptionV3 network structure splits huge volume integrals into smaller convolutions using a convolution kernel splitting technique. A 3*3 convolution, for instance, can be divided into 3*1 and 1*3 convolutions (Fig. 2) [27]. The splitting method can be used to reduce the number of parameters, which will speed up network training

Fig. 1 Inception network structure

by reducing the computational cost and improve the process of extracting the spatial features.

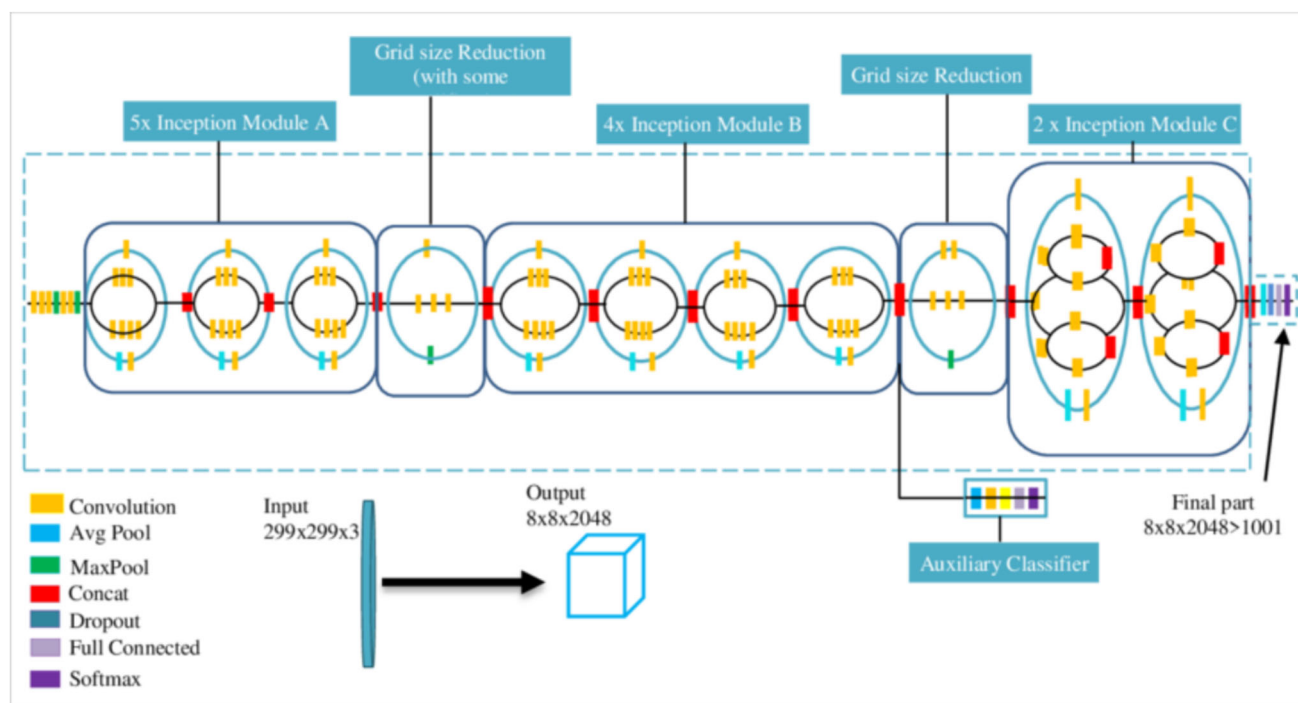
Transfer Learning

The main characteristics of deep learning models are that model development requires a high volume of data and, in case of insufficient data, under-fitting can occur. Transfer learning has been offered by scholars as a way to train with deep learning on a small dataset [28]. With just a little dataset, precise and effective picture classification is possible thanks to transfer learning's learning capabilities.

The transfer learning method is used to pre-train a strong performance model on ImageNet (1.2 million images with 1000 categories) and learn the characteristics of the ImageNet

dataset in the event that there is not enough training data for an upcoming task. In order to guarantee that your model receives the previously learnt features and produces superior results, it is recommended that you initialize the weight parameters that ImageNet pre-trained in your model. After exploring the methods discussed earlier, this study focuses on constructing a network model using InceptionV3 to predict bone age and gender from radiographic images. The model is initialized with pre-trained weights from ImageNet and then trained on the 2017 RSNA Pediatric Bone Age Challenge dataset. This strategy leverages the pre-trained features to achieve strong classification performance on the relatively small RSNA dataset.

A similar approach was explored by Lee et al. [29], who developed a fully automated pipeline for bone age assessment

**Fig. 2** InceptionV3 network structure (Source: [27])

using transfer learning with ImageNet-pre-trained CNN. Their model incorporated preprocessing, segmentation, and fine-tuning steps, achieving high accuracy for both male and female cohorts on radiographic images. Unlike our method, which jointly predicts bone age and gender in a multi-output model, their system handled gender by splitting the data into male and female cohorts rather than directly predicting gender. Nonetheless, both studies highlight the effectiveness of transfer learning in medical imaging tasks with limited data. InceptionV3 is supported by its demonstrated ability to efficiently extract multi-scale features and maintain competitive accuracy while requiring reasonable computational resources, making it well-suited for multi-output prediction in this domain.

Data

In this study, the open-access source of [2017 RSNA Pediatric Bone Age Challenge](#) was utilized, which contains pediatric HWRs for bone age assessment, along with data labels of bone age and gender information for each sample [19]. The dataset was divided into three subsets: 70% for training (9472 images), 20% for validation (3152 images), and 10% for testing (1424 images). Gender outputs were converted to a binary format, with males represented as 0 and females as 1. To prevent overfitting, data augmentation was applied to the training and validation subsets using TensorFlow's ImageDataGenerator. The augmentation pipeline included a range of transformations such as random rota-

tion (up to 10 degrees), horizontal flipping, width and height shifting (up to 20%), shearing (up to 20%), and zooming (up to 20%). These operations helped introduce variability and improve model robustness. As a result, the number of training images increased from 9472 to 10,472, and the validation set increased from 3152 to 4152. The test set remained unchanged at 1424 images. This augmentation approach improves the model's generalization ability and sustains robust performance on unseen data. Additionally, preprocessing using InceptionV3's preprocessing function was employed to normalize pixel values, ensuring consistency with the pre-trained model's input expectations. Since data augmentation is beneficial only for training, the test set remained unaltered to ensure an unbiased evaluation of model performance. Furthermore, a custom image generator was implemented to handle the dual-output prediction format, ensuring the correct structure of bone age regression and gender classification labels.

Prediction Methodology

The InceptionV3 architecture was used as the backbone for the developed multi-output prediction model, which utilizes pre-trained weights from Google's ImageNet database. The transfer learning approach allows to leverage the extensive feature extraction capabilities of InceptionV3, which has been trained on a vast array of images, thereby providing a solid foundation for this specific task of predicting both bone age and gender from radiological images Fig. 3.

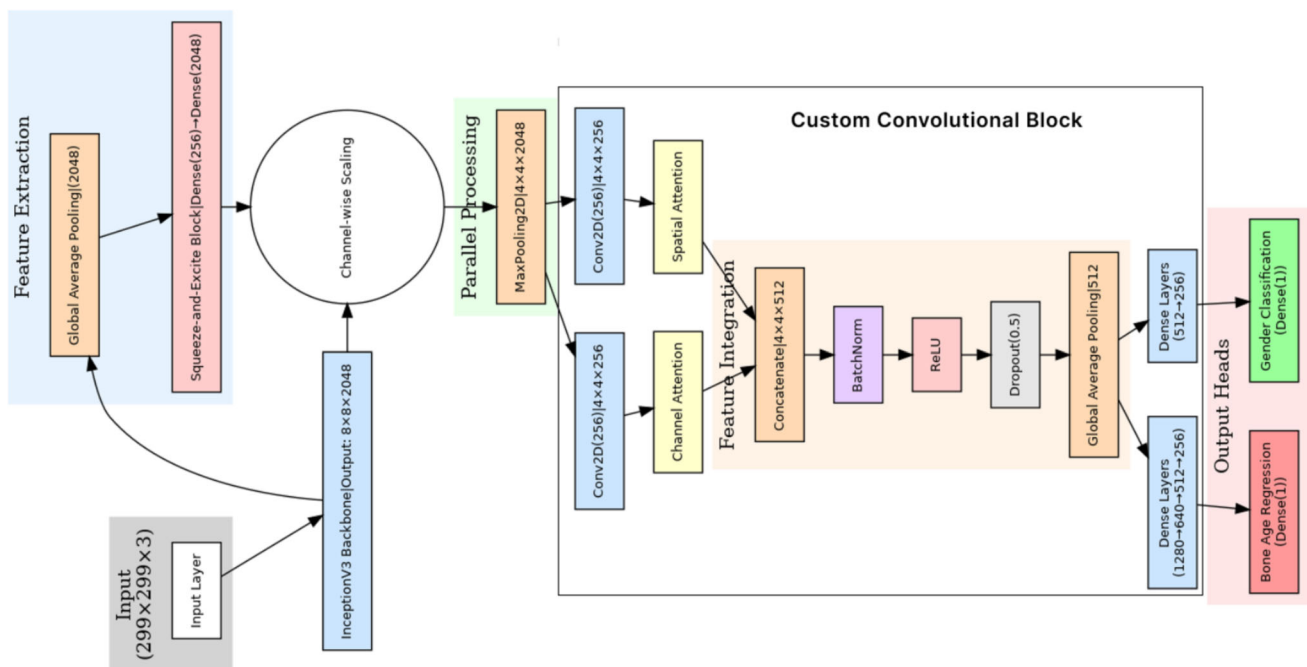


Fig. 3 Multi-output model structure for predicting both bone age and gender

To adapt the InceptionV3 model for dual-output predictions, the architecture was augmented with specialized CNN layers tailored for enhanced feature extraction. A key enhancement is the integration of a Squeeze-and-Excitation (SE) block, which performs channel-wise recalibration via a bottleneck dense layer (256 units) followed by an expansion to 2048 units. This channel attention mechanism enables dynamic feature weighting, allowing the network to prioritize the most informative channels for bone age estimation and gender classification.

Following the SE block, the recalibrated feature maps undergo parallel convolutional processing, including dilated convolutions (dilation rate = 2) to expand the receptive field without increasing computation. These are encapsulated within custom convolutional blocks, which include two attention mechanisms: a Channel Attention block that learns to emphasize important feature channels using global average pooling and dense layers and a Spatial Attention block that localizes salient regions across the spatial dimensions using convolutional filters. These attention-enhanced paths enable the model to extract both global and localized discriminative features.

The outputs of these parallel branches are then integrated through concatenation, forming a rich, multi-perspective feature representation. This combined tensor is passed through a batch normalization layer, a ReLU activation, and dropout for regularization, followed by a final global average pooling layer that compresses the spatial dimensions into a compact vector. This integrated feature vector feeds into two separate dense heads for bone age regression and gender classification, enabling efficient and accurate multi-task learning.

Given that the model is tasked with generating two distinct predictions from the same input image, it is essential that it excels in feature extraction while effectively balancing the unique requirements of each output. This necessitates a careful consideration of the challenges associated with multi-output modeling. Specifically, the need for setting layer conductivities that facilitate effective shared learning between the two outputs was addressed. The architecture incorporates a shared backbone that extracts core features from the radiological images, followed by two separate branches. Each branch is fine-tuned to optimize performance for its respective prediction task—one dedicated to bone age estimation and the other to gender classification. To ensure stable training, the last 110 layers of InceptionV3 was fine-tuned while keeping batch normalization layers frozen, which helps maintain consistent feature distributions and mitigates internal covariate shifts.

To further enhance the model's performance, hyperparameter optimization techniques were implemented to fine-tune architectural and training parameters. A combination of manual search and grid-based exploration was used to iteratively adjust key hyperparameters, including the learn-

ing rate, filter kernel sizes, batch size, number of filters in convolutional layers, and dropout rates. The search space was constrained to practical ranges—for instance, learning rates between 1e-5 and 1e-3, kernel sizes of 3×3 and 5×5, batch sizes of 16, 32, and 64, and dropout rates ranging from 0.1 to 0.5.

The optimization process involved evaluating training and validation metrics across multiple runs, with each configuration trained for up to 50 epochs using early stopping based on validation loss (patience = 5 epochs). This allowed for efficient pruning of suboptimal configurations while avoiding overfitting. In particular, targeted dropout strategies were refined for each output head: a dropout rate of 0.35 was applied in the gender classification branch, and 0.2 in the bone age regression branch. These values were selected based on repeated experimentation and their observed impact on generalization, achieving a balance between model capacity and regularization.

The Huber loss was used for bone age regression, which is robust to outliers by combining mean squared error (MSE) and mean absolute error (MAE) as follows:

$$L_{\text{Huber}}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{for } |y - \hat{y}| \leq \delta \\ \delta \cdot (|y - \hat{y}| - \frac{1}{2}\delta), & \text{for } |y - \hat{y}| > \delta \end{cases} \quad (1)$$

where y is the true bone age, \hat{y} is the predicted bone age, and δ is a threshold parameter that determines the point at which the loss transitions from quadratic to linear.

Binary focal crossentropy loss was employed for gender classification, which modifies standard binary crossentropy by introducing a focusing factor to reduce the relative impact of easy-to-classify examples:

$$L_{\text{Focal}}(p, \hat{p}) = -\alpha \cdot (1 - \hat{p})^\gamma \cdot y \log(\hat{p}) - (1 - \alpha) \cdot \hat{p}^\gamma \cdot (1 - y) \log(1 - \hat{p}) \quad (2)$$

where y is the true label (0 for male, 1 for female), \hat{p} is the predicted probability, γ is the focusing parameter ($\gamma = 1.2$ in our case), and α is a balancing factor. This loss function helps the model focus on harder examples, improving classification performance, particularly in imbalanced datasets. To stabilize training, we use AdamW as the optimizer, incorporating a learning rate of 0.00004, weight decay (0.0003), and gradient clipping (global norm = 0.3), ensuring efficient and stable learning dynamics.

By carefully calibrating these architectural enhancements and optimization strategies, we maximize the predictive power of our model while ensuring robust performance across both outputs. The integration of transfer learning, custom convolutional layers, attention mechanisms, and carefully tuned hyperparameters allows our model to effec-

tively handle the complexity of radiological images while maintaining high generalization capability on unseen data.

To enhance the statistics, the continuous bone age prediction performance was investigated by calculating the intra-class correlation coefficients (ICCs) and the gender output as the binary categorical data was tested with Cohen's κ coefficients. The statistical significance threshold was considered as $p=0.05$.

Results

In this section, we present the results using detailed graphs that illustrate the performance of our multi-output model. The model was trained and evaluated on a system equipped with a Tesla P100-PCIE GPU, 16 GB VRAM, and 30 GB RAM. The inference process was optimized for batch processing, with a typical inference time of approximately 30 milliseconds per image, enabling near real-time predictions. These hardware requirements suggest that the model is computationally feasible for deployment on cloud systems with scalable GPU resources. The bone age branch is evaluated using MAE and MSE, while the gender branch is assessed based on accuracy and area under the curve (AUC).

As shown in Fig. 4, after performing inference on the validation dataset over three epochs, the model achieved a MSE of approximately 25 and a MAE of 3.1 for predicting bone age. These results indicate strong performance in bone age prediction. Additionally, the model demonstrated impressive results in gender classification, achieving an accuracy of 95%

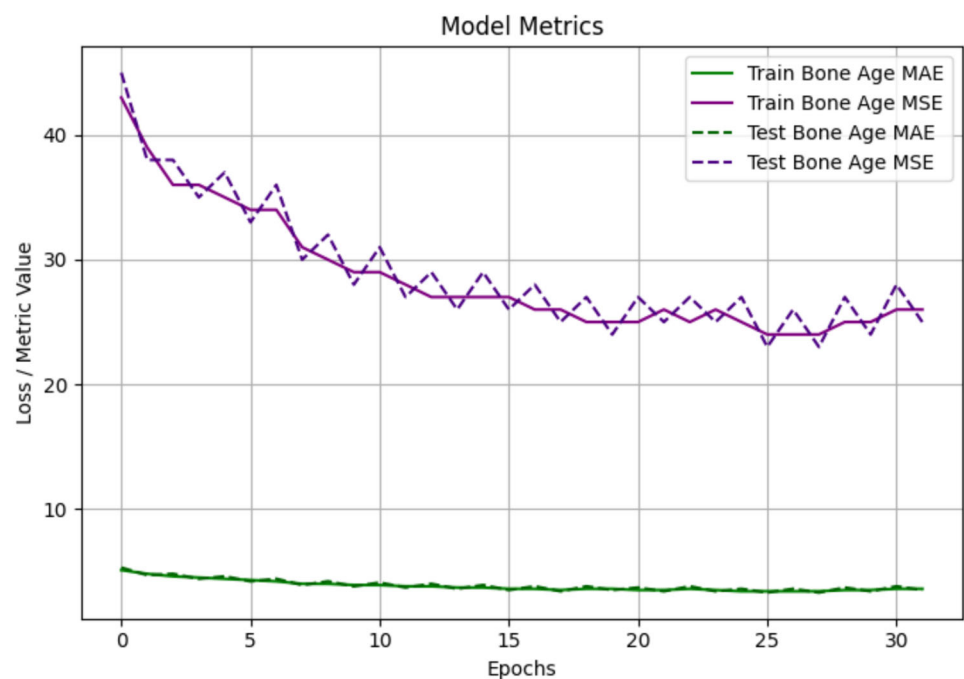
and an AUC of 97% (Fig. 5). These metrics further confirm the model effective performance across both tasks.

The scatter plot (Fig. 6) was examined to further analyze the model's performance in predicting bone age, which visualizes the relationship between the predicted and actual bone age values. Ideally, a well-performing model should exhibit points closely aligned along the diagonal, indicating minimal deviation between predictions and ground truth values.

According to the results, the scatter plot reveals a strong correlation, with most predictions clustering around the diagonal line. This suggests that the model effectively captures the underlying patterns in the data. Notably, the model maintains a consistent trend across the entire age spectrum, including both lower (< 50 months) and higher (> 180 months) age ranges. While a slightly wider spread is observed in these extreme regions, the predictions remain generally centered around the ideal diagonal line, demonstrating the model's ability to generalize even in less-represented age segments. This performance indicates that the model has successfully learned critical developmental cues across all ages. Its stable predictions in both young and older age groups underscore its robustness and clinical applicability for diverse patient populations. Despite these small variations, the overall trend demonstrates that the model generalizes well across different age groups, further reinforcing its reliability in bone age estimation.

The ICC for the bone age prediction was found to be 0.997 ($p < 0.001$). The Cohen's κ coefficient for the gender prediction was found to be 0.898 ($p < 0.001$).

Fig. 4 Graph representing the results of MSE and MAE for bone age prediction



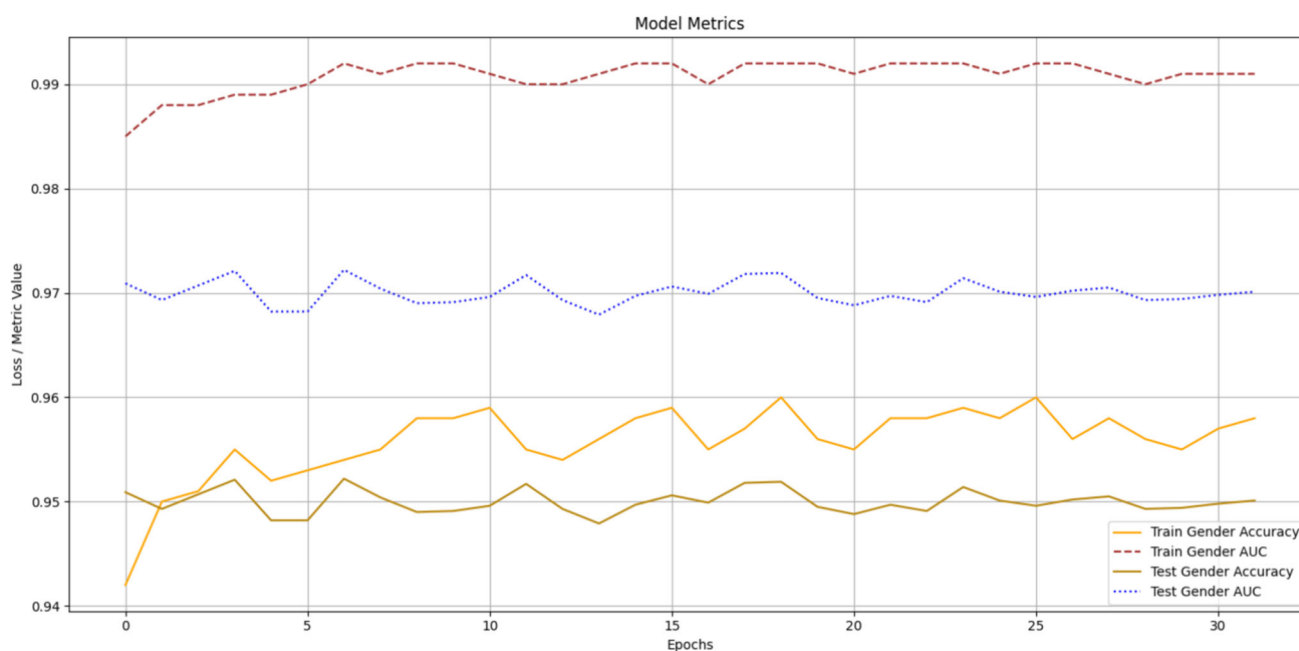


Fig. 5 Graph represents the results of the AUC and accuracy of gender prediction

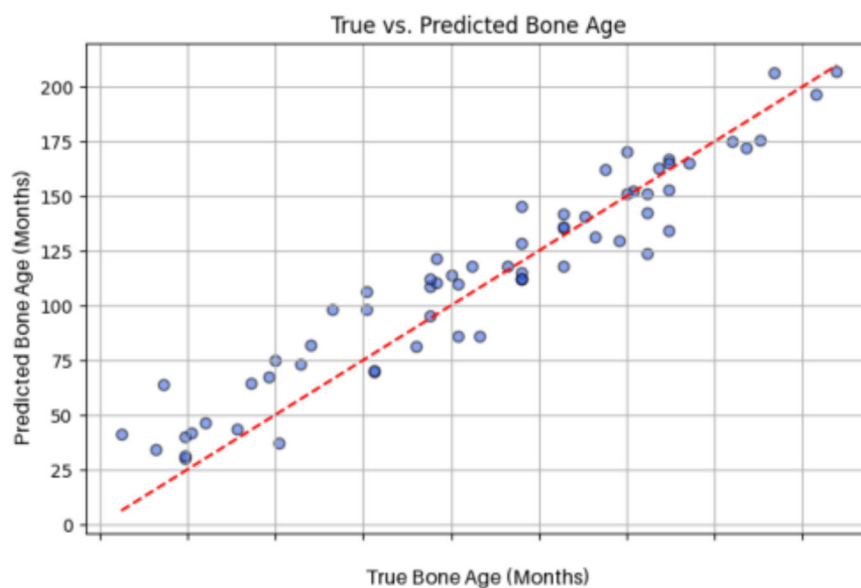
Discussion

The multi-output model with the InceptionV3 architecture as the backbone performed an accuracy of 95% for the gender predictions and an ICC value of 0.997 ($p < 0.001$) for bone age prediction. The concept of computer-aided radiographic bone age and gender estimation has been of interest to scientists from various disciplines.

In addition to common performance metrics, Cohen's kappa and ICC statistics were calculated for the present study. The kappa coefficient is used to assess the agreement between

two independent observers' scores for nominal/categorical variables. Negative values indicate that the observed agreement is less than that expected from chance alone, while +1 indicates perfect agreement. Several types of kappa coefficients are available. The weighted kappa coefficient can be used if the magnitude of the difference between raters is important, while Fleiss kappa is suitable when there are three or more raters. Since gender information was binary in the present study, Cohen's kappa statistic was utilized because if any, the magnitude of the difference between observers was equal. On the other hand, ICC is used to the evaluation of

Fig. 6 Graph shows the scatter plot of performing the model in bone age predicting on test dataset



differences between interval or ratio variables. The purpose of use is similar to kappa coefficients, is relatively more flexible as the same formula can be used for two or more raters, and can be used in case of missing variables. In this study, ICC statistics were used to evaluate the bone age prediction performance, which is a continuous data type [30].

Several studies are focused on gender estimation. A study investigated sex estimation with CNN using the patella magnetic resonance image slices. The dataset was consisted of 6710 magnetic resonance slices of 696 patients and EfficientNetB3, MobileNetV2, Visual Geometry Group 16 (VGG16), ResNet50, and DenseNet121 architectures were adopted in model development. A rectified linear unit (ReLU) was used for activation in convolution layers, while Adamax optimizer was used for model optimization. The researchers reported the average accuracy of the developed models as 85.70%, while the highest performance was achieved with the ResNet50 model (88.88%) [31]. In that study, magnetic resonance imaging of the patella was utilized as a method of sex determination, whereas in this study, the HWM method was adopted. While the first method has an advantage in terms of patient dose due to the absence of ionizing radiation, the need for a magnetic resonance system with special requirements for image acquisition, which is an important limitation of the first method. On the other hand, HWRs can be obtained on-site with mobile X-ray devices in the case

of forensic events. In a study investigating the reliability of VGG16 model for gender prediction using orthopantomographs (OPGs), 1050 OPGs were included, root mean square propagation was used as optimizing algorithm, and binary cross entropy was adopted as loss function. The authors reported an overall accuracy of 89% for the developed model, while the F1-scores changed between 0.88 and 0.90, according to age and gender of the subgroups [32]. OPGs, used in that study, are a popular imaging technique in dentistry in which the complete dental arch and jaw bones are captured on a planar image in a single rotation. It is a frequently used technique for radiographic assessment of the overall condition as a supplement to the clinical examination. However, the acquisition of images is not as standardized as HWR or cephalograms, and patient- and operator-related variances are expected. In this aspect, the imaging modality in the present study is advantageous in terms of repeatability and reproducibility compared to the related study. Another study compared various machine learning classifiers in sex estimation based on combined anatomical measurements of the long bones. A total of 2141 individuals were included, with 18 measurements recorded from the radius, humerus, femur, and tibia. Five machine learning classifiers were adopted as linear discriminant analysis, penalized logistic regression, random forest, support vector machine, and artificial neural network. The authors reported the highest accuracy (92%) for

Table 1 A summary of the relevant studies mentioned in the article

Author, year	Data type	Sample size	Algorithm	Assessment	Performance
Proposed model	Hand-wrist radiographs	14,048	InceptionV3 Multi-Output	Bone age & Gender	Bone age MAE: 3.1 MAD: 3.33; Gender accuracy: 95%; Gender AUC: 97%
Halabi et al., 2018	Hand-wrist radiographs	14,236	Deep learning and CNN-based methods	Bone age	The best five MAD were 4.2, 4.4, 4.4, 4.5, and 4.5 months
Larson et al., 2017	Hand-wrist radiographs	14,236 + 1377	Deep residual network	Bone age	MAE of 0.5 year; RMSE: 0.63 and 0.73
Cavlak et al., 2025	Magnetic resonance sagittal patella image slices	6710	EfficientNetB3, MobileNetV2, VGG16, ResNet50, DenseNet121	Gender	Best accuracy of 88.88% (ResNet50)
Pereira et al., 2025	Orthopantomography	1050	VGG16	Gender	Overall accuracy of 89%
Knecht et al., 2023	18 measurements from 4 long bones (radius, humerus, femur, tibia)	2141	LDA, penalized logistic regression, random forest, SVM, ANN	Gender	Accuracy: 90–92% (all bones), 83.3–90.3% (isolated bones)
Yilmaz et al., 2025	Orthopantomography	1914	17 DL models (Xception, ResNet, ShuffleNet, InceptionV3 etc.)	Bone age	Polygon area metric of 0.8828
Matthijs et al., 2024	Panoramic radiographs	4000	DenseNet201	Dental age	Accuracy: 0.53, MAE: 0.71, Cohen's kappa: 0.71, ICC: 0.89
Li et al., 2022	Hand-wrist radiographs	12,611 + 1709	MobileNetV3, MLP (1 hidden layer)	Bone age	MAE of 6.2 months

Note: MAE, mean absolute error; MAD, mean absolute distance; CNN, convolutional neural networks; ICC, intra-class correlation coefficient

random forest and lowest accuracy (90%) for linear discriminant analysis when all the bone measurements are included. For the isolated experiments which was based on the measurements of a single bone, the highest accuracy (90.3%) was reported with humerus bone and random forest algorithm, while the lowest accuracy (83.3%) was reported for radius bone and penalized logistic regression algorithm [33]. In that study, existing numerical anthropometric data were classified by machine learning without radiographic image acquisition, whereas in this study, HWR was utilized. The clinical use of this simple approach, which can be useful in fields such as archaeology, is limited due to the inability to directly measure bones or the need for volumetric image acquisition. The model developed in this study classified 94.93% of the gender labels correctly. This is slightly higher than previous studies and provides support for the evidence that the methodology in this study can be utilized for gender prediction.

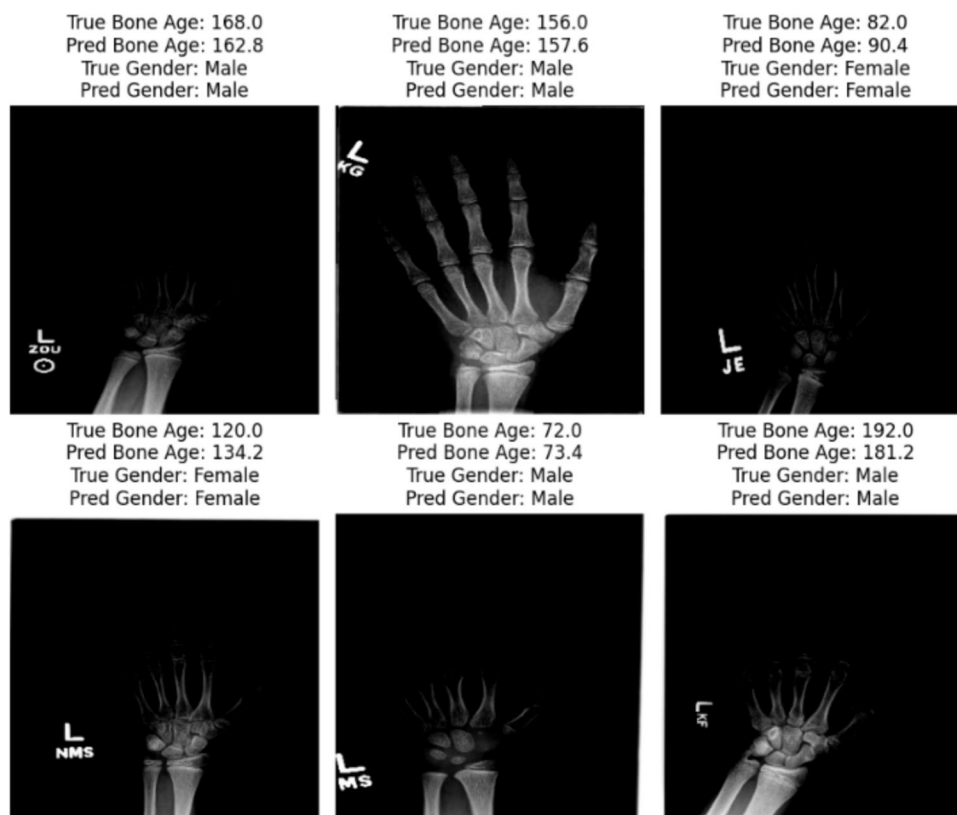
Some studies are focused on bone age prediction. In 2019, a systematic review investigated the studies on bone age assessment with various machine learning techniques. A total of 26 studies were included, and regression-based methods (13 studies) were found to be the most common, followed by artificial neural networks (eight studies) and support vector machines (five studies). As a result of the meta-analysis conducted with seven studies included in the systematic review, the average performance was determined as 9.96

MAE (months), while it was stated that there were differences among the age groups in the relevant studies which should be considered. Moreover, the importance of techniques such as magnetic resonance imaging that do not require ionizing radiation was also highlighted by the authors [34].

A study investigated the analysis of whether an individual in panoramic images is under or over 12 years old by developing eighteen different machine learning models. Panoramic radiographs of 1941 pediatric patients between 5 and 15 years were used in the development of algorithms based on Xception, ResNet, ShuffleNet, InceptionV3, DarkNet, NasNet, DenseNet, EfficientNet, MobileNet, ResNet18, GoogleNet, SqueezeNet, and AlexNet, and more. The researchers reported accuracy values between 0.7 and 0.94, with the highest score obtained using the Forensic Xception model. In addition, novel polygon area metric values were given, and the highest value was again found with the Forensic Xception model (0.88) [35]. In that study, OPGs were utilized, and age outputs were binary categorical as “under and over 12 years of age,” whereas in the present study, gender output was binary, but bone age estimates were conducted as a continuous data type using regression approach.

Another study focused in developing automated models to analyze the development of all mandibular tooth types and comparing the models’ performances. The authors adopted a modified Demirjian staging technique for evaluating the

Fig. 7 Figure shows the model’s performance on real X-ray radiology



incisors, canines, premolars, molars, and third molars in the left side of the mandible. A deep, densely connected DenseNet201 CNN architecture with 12, 24, 96, and 32 convolutional layers was developed, and the authors reported accuracy values of 0.30, 0.55, 0.51, 0.71, and 0.57 for the teeth numbers 31, 33, 34, 37, and 38, respectively [36]. In the relevant study, the dataset was manually segmented into different types of tooth regions in the mandible in OPGs, while the model output was categorical data type to predict tooth developmental stages, not bone age directly. In the present study, the image features were extracted using algorithms, and manual segmentation was not performed. In 2021, a study used the RSNA dataset with an additional dataset of 1709 samples to develop a deep learning-based system for bone age assessment. Given the study design, bounding boxes were first obtained using a CNN model for object localization. The features clustered according to the system, MobileNetV3 with pre-trained weights on ImageNet algorithm as the backbone, are classified with multiple layer perceptron with one hidden layer model for future bone age prediction. Gender data was used as an additional input, and the authors reported a mean absolute error of approximately 6.2 months on the RSNA dataset, and 5.1 months with the inclusion of the additional dataset [37]. This study used the same open-source dataset as the current research. In this study, gender was specified as one of the two outputs in the model architecture, while in the other research, gender was used as an additional input. The model designed in this study was reached to a MAE of 3.1 for predicting bone age, which is a lower error when compared to the relevant studies. Scatter plot (Fig. 6) shows the fit of the regression bone age outputs with the label data, and the distribution may show that the early ages may be resulted to be older, and this trend reverses with the increase in age. The algorithms in this study are not available to end-users and clinicians in an accessible interface. Conversely, providing clinicians with such information may benefit patients by enabling rational use of relevant systems. A summary of the relevant studies mentioned in the article can be found in Table 1. The proposed model features an atlas-based bone age analysis with fully automated approach, while several other studies focusing on specific bones are also reported [38, 39].

This study has several limitations. First, although the dataset is composed of data from multiple centers, the dataset used may not be fully representative of the broader population, potentially limiting the generalizability of the results. In the case of similar data sets from centers in different parts of the globe, studies on the generalizability of the models or their optimization for local use can be planned.

Second, the computational complexity of the proposed model presents a challenge. Due to its deep architecture and large number of parameters, the model requires significant computational resources and may be difficult to run effi-

ciently without access to a high-performance GPU. Future studies can be considered with a simpler architecture and fewer features to increase model success while reducing the need for computational sources.

Third, the quality and resolution of the input images may affect prediction accuracy, especially in cases where images are noisy or of low resolution. These factors could limit the model's applicability in real-time clinical settings or in resource-constrained environments. Indeed, image artifacts are a fact of life, and algorithms that feed on multiple data sources and compensate for each other in terms of lack of data may be the subject of future studies to overcome this issue.

Conclusion

In this study, we developed a deep learning-based multi-output model for automatic bone age prediction and gender classification. Our model successfully predicts bone age with a MSE of approximately 25 and a MAE of 3.1. Additionally, it achieves high classification performance for gender, with an accuracy of 95% and an AUC of 97%. These results demonstrate that despite the challenges posed by radiograph quality, our approach achieves robust performance, surpassing existing automated models. As shown in Fig. 7, the model's performance on real X-ray radiology data further validates its effectiveness in practical scenarios.

Author Contributions B. Rayed: conceptualization, modeling, and writing original draft. H. Amasya: writing original draft and editing. M. Sezdi: writing—reviewing and supervision.

Data Availability The dataset used in this study was released as open access by the Radiology Society of North America (RSNA) in 2017, and the relevant dataset can be accessed from the link: <https://www.rsna.org/rsnai/aiimage-challenge/rsna-pediatric-bone-age-challenge-2017>

Declarations

Ethical Approval Ethical approval was not required.

Informed Consent This article does not contain any studies involving human subjects.

Conflict of Interest The authors declare no competing interests.

References

1. Olivares, LAF, De León, LG, Fragoso, ML. Skeletal age prediction model from percentage of adult height in children and adolescents. *Scientific Reports*. 2020;10(1):15768. <https://doi.org/10.1038/s41598-020-72835-5>
2. Ferrillo, M, et al. Reliability of cervical vertebral maturation compared to hand-wrist for skeletal maturation assessment in growing



- subjects: A systematic review. *Journal of Back and Musculoskeletal Rehabilitation*. 2021;34:925-936. <https://doi.org/10.3233/BMR-210003>
3. Schmeling, A, et al. Forensic age estimation: methods, certainty, and the law. *Dtsch Arztebl International*. 2016;113(4):44-50. <https://doi.org/10.3238/arztebl.2016.0044>
 4. Greulich, WW, Pyle, SI. Radiographic atlas of skeletal development of the hand and wrist. Stanford University Press; 1959.
 5. Pinchi, V, et al. Skeletal age estimation for forensic purposes: A comparison of GP, TW2 and TW3 methods on an Italian sample. *Forensic Science International*. 2014;238:83-90. <https://doi.org/10.1016/j.forsciint.2014.02.030>
 6. Fishman, LS. Radiographic evaluation of skeletal maturation: a clinically oriented method based on hand-wrist films. *The Angle Orthodontist*. 1982;52(2):88-112.
 7. Baccetti, T, Franchi, L, McNamara, JA. The cervical vertebral maturation (CVM) method for the assessment of optimal treatment timing in dentofacial orthopedics. *Seminars in Orthodontics*. 2005;11(3):119-129. <https://doi.org/10.1053/j.sodo.2005.04.005>
 8. McNamara, JA, Franchi, L. The cervical vertebral maturation method: A user's guide. *The Angle Orthodontist*. 2018;88(2):133-143.
 9. Sella Tunis, T, Masarwa, M, Finkelstein, T, et al. The reliability of a modified three-stage cervical vertebrae maturation method for estimating skeletal growth in males and females. *BMC Oral Health*. 2024;24:1255. <https://doi.org/10.1186/s12903-024-05028-5>
 10. Schoretsaniti, L, Mitsea, A, Karayianni, K, Sifakakis, I. Cervical vertebral maturation method: Reproducibility and efficiency of chronological age estimation. *Applied Sciences*. 2021; 11(7):3160. <https://doi.org/10.3390/app11073160>
 11. Szemraj, A, Wojtaszek-Słomińska, A, Racka-Pilszak, B. Is the cervical vertebral maturation (CVM) method effective enough to replace the hand-wrist maturation (HWM) method in determining skeletal maturation?—A systematic review. *European Journal of Radiology*. 2018;102:125-128. <https://doi.org/10.1016/j.ejrad.2018.03.012>
 12. Ibrahim, RSM, Shaker, CW, Mira, MF, et al. Clinical, laboratory and radiological assessment of skeletal maturation in children and adolescents with obesity. *Egypt Pediatric Association Gaz* 2020;68:13. <https://doi.org/10.1186/s43054-020-00024-0>
 13. Khade, D.M., Bhad, W.A., Chavan, S.J., Muley, A., Shekokar, S. Reliability of salivary biomarkers as skeletal maturity indicators: A systematic review. *International Orthodontics*. 2023;21(1). <https://doi.org/10.1016/j.ortho.2022.100716>
 14. Turing, A.M. Can a machine think. *The World of Mathematics*. 1956;4:2099–2123.
 15. Orhan, K., Amasya, H. Artificial intelligence from medicine to dentistry. In: Orhan, K., Jagtap, R., editors. *Artificial Intelligence in Dentistry*. Springer International Publishing; 2023. pp. 33–42. https://doi.org/10.1007/978-3-031-43827-1_3
 16. Wang, S., Summers, R.M. Machine learning and radiology. *Medical Image Analysis*. 2012;16(5):933–951. <https://doi.org/10.1016/j.media.2012.02.005>
 17. Tajmir, S.H., et al. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. *Skeletal Radiology*. 2019;48(2):275–283. <https://doi.org/10.1007/s00256-018-3033-2>
 18. Tanner, J.M., Gibbons, R.D. Automatic bone age measurement using computerized image analysis. 1994;7(2):141–146.
 19. Halabi, S.S., et al. The RSNA pediatric bone age machine learning challenge. *Radiology*. 2019;290(2):498–503. <https://doi.org/10.1148/radiol.2018180736>
 20. Pan, I., et al. Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge. *Radiology: Artificial Intelligence*. 2019;1(6):e190053. <https://doi.org/10.1148/ryai.2019190053>
 21. Larson, D.B., et al. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. 2018;287(1):313–322. <https://doi.org/10.1148/radiol.2017170236>
 22. Zhang, H. Snake image recognition based on InceptionV3 model. *Electronic Technology and Software Engineering*. 2019;10:58–61.
 23. Zhao, J.D., Bai, Z.M., Chen, H.B. Research on road traffic sign recognition based on video image. In: *10th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. IEEE; 2017. pp. 110–113.
 24. Li, J., et al. Transfer learning of pre-trained Inception-v3 model for colorectal cancer lymph node metastasis classification. In: *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE; 2018. pp. 1650–1654.
 25. Mednikov, Y., et al. Transfer representation learning using Inception-V3 for the detection of masses in mammography. In: *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2018. pp. 2587–2590.
 26. Abadi, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. <https://doi.org/10.48550/arXiv.1603.04467>
 27. Iparraguirre-Villanueva O., Orozco-Arias S., Muñoz W., et al. Convolutional neural networks with transfer learning for pneumonia detection. *International Journal of Advanced Computer Science and Applications*. 2022;13(9):544–551. <https://doi.org/10.14569/IJACSA.2022.0130963>
 28. Idress, W.M., Zhao, Y., Abouda, K.A., Ahmed, A.A., Hassan, W., Abdalla, O., & Elhindi, T. Enhanced onychomycosis diagnosis using a dynamic weighted ensemble classifier: Integrating light GBM and transfer learning with a game theory approach. *15th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2025;1209–1216.
 29. Lee, H., Tajmir, S., Lee, J., et al. Fully automated deep learning system for bone age assessment. *Journal of Digital Imaging*. 2017;30:427–441. <https://doi.org/10.1007/s10278-017-9955-8>
 30. Gisev, N., Bell, J.S., Chen, T.F. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*. 2013;9(3):330–338. <https://doi.org/10.1016/j.sapharm.2012.04.004>
 31. Cavlak, N., Cinarer, G., Erkoç, M.F., et al. Sex estimation with convolutional neural networks using the patella magnetic resonance image slices. *Forensic Science, Medicine and Pathology*. 2025. <https://doi.org/10.1007/s12024-025-00943-7>
 32. Pereira, C.P., Correia, M., Augusto, D., et al. Forensic sex classification by convolutional neural network approach by VGG16 model: accuracy, precision and sensitivity. *International Journal of Legal Medicine*. 2025;139:1381–1393. <https://doi.org/10.1007/s00414-025-03416-2>
 33. Knecht, S., Santos, F., Ardagna, Y., et al. Sex estimation from long bones: a machine learning approach. *International Journal of Legal Medicine*. 2023;137:1887–1895. <https://doi.org/10.1007/s00414-023-03072-4>
 34. Dallora, A.L., Anderberg, P., Kvist, O., Mendes, E., Diaz, R.S., Sanmartin, B.J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. *PLoS ONE*. 2019;14(7):e0220242. <https://doi.org/10.1371/journal.pone.0220242>
 35. Yilmaz, E., Görürgöz, C., Kış, H.C., et al. Forensic dental age estimation with deep learning: a modified Xception model for panoramic X-Ray images. *Forensic Science, Medicine and Pathology*. 2025. <https://doi.org/10.1007/s12024-025-00962-4>

36. Matthijs, L., Delande, L., De Tobel, J., et al. Artificial intelligence and dental age estimation: development and validation of an automated stage allocation technique on all mandibular tooth types in panoramic radiographs. *International Journal of Legal Medicine*. 2024;138:2469–2479. <https://doi.org/10.1007/s00414-024-03298-w>
37. Li, S., Liu, B., Li, S., et al. A deep learning-based computer-aided diagnosis method of X-ray images for bone age assessment. *Complex & Intelligent Systems*. 2022;8:1929–1939. <https://doi.org/10.1007/s40747-021-00376-z>
38. Lashin H.I., Sharif A.F., Ghaly M.S., et al. Bridging gaps in age estimation: a cross-sectional comparative study of skeletal maturation using Fishman method and dental development using Nolla method among Egyptians. *International Journal of Legal Medicine*. 2025;139:695–714. <https://doi.org/10.1007/s00414-024-03394-x>
39. Amasya H., Aydogan T., Cesur E., et al. Using artificial intelligence models to evaluate envisaged points initially: A pilot study. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*. 2023;237(6):706–718. <https://doi.org/10.1177/09544119231173165>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Baraa Rayed¹  · Hakan Amasya^{2,3,5}  · Mana Sezdi^{4,5} 

✉ Baraa Rayed
baraarayad@ogr.iuc.edu.tr

Hakan Amasya
hakanamasya@iuc.edu.tr

Mana Sezdi
mana@iuc.edu.tr

¹ Biomedical Engineering Department, Institute of Graduate Studies, Istanbul University-Cerrahpasa, 34320 Istanbul, Turkey

² Department of Oral and Dentomaxillofacial Radiology, Faculty of Dentistry, İstanbul University-Cerrahpaşa, 34098 İstanbul, Turkey

³ CAST (Cerrahpaşa Research, Simulation and Design Laboratory), Istanbul University-Cerrahpaşa, 34098 İstanbul, Turkey

⁴ Biomedical Device Technology Program, Vocational School of Technical Sciences, Istanbul University-Cerrahpasa, 34500 İstanbul, Turkey

⁵ Health Biotechnology Joint Research and Application Center of Excellence, 34220 Istanbul, Turkey