Bone age prediction from carpal radiographic images using deep learning

Predição de idade óssea a partir de imagens radiográficas do carpo usando aprendizado profundo

Predicción de la edad ósea a partir de imágenes radiográficas del carpo usando aprendizaje profundo

Rafael Guimarães Malanga[1], Viviane Rodrigues Botelho[2], Thatiane Alves Pianoschi[2], Jose Rodrigo Mendes Andrade[3], Guilherme Ribeiro Garcia[4], Rochelle Lykawka[3], Alexandre Bacelar[3], Carla Diniz Lopes Becker[2]
1 Master's Student, Federal University of Health Sciences of Porto Alegre – UFCSPA, Porto Alegre (RS), Brazil.
2 Ph.D., Federal University of Health Sciences of Porto Alegre – UFCSPA, DECESA, Porto Alegre (RS), Brazil.
3 M.Sc, , Hospital of Clinics of Porto Alegre – HCPA, Porto Alegre (RS), Brazil.
4 Bachelor of Physics, Hospital of Clinics of Porto Alegre – HCPA, Porto Alegre (RS), Brazil.

Corresponding author: Rafael Guimarães Malanga
E-mail: rg.malanga@gmail.com

Abstract

Biological age, a crucial indicator of human development, reflects the physical and mental changes associated with aging. Estimating bone age, a common method in clinical practice that seeks information about biological age, can be subjective and imprecise. **Objective:** This study proposes methods based on deep learning techniques to estimate skeletal age from hand X-ray images. **Methods:** We used datasets divided by gender and age to train and test the models. **Results:** The results show promising estimates, with mean errors of 10.808 months in a public dataset and 15.548 months in a private dataset. The developed tool, with its intuitive graphical interface, offers practical use for medical professionals and researchers. **Conclusion:** This study applies deep learning to predict bone age, which can aid in assessing skeletal development in fields like pediatrics and orthopedics.

Keywords: Bone Age; Deep Learning; Radiodiagnosis

DOI: 10.59681/2175-4411.v16.iEspecial.2024.1361

Resumo

A idade biológica, indicador crucial do desenvolvimento humano, reflete as mudanças físicas e mentais associadas ao envelhecimento. A estimativa da idade óssea, um método comum na prática clínica que busca informações sobre idade biológica, pode ser subjetiva e imprecisa. **Objetivo:** Este estudo propõe métodos baseados em técnicas de aprendizado profundo para estimar a idade esquelética a partir de imagens de raios-X da mão. **Método:** Utilizamos conjuntos de dados divididos por gênero e idade para treinar e testar os modelos. **Resultados:** Os resultados demonstram promissoras estimativas, com erros médios de 10,808 meses em um conjunto de dados públicos e 15,548 meses em um conjunto privado. A ferramenta desenvolvida, com sua interface gráfica intuitiva, oferece uma utilização prática para profissionais médicos e pesquisadores. **Conclusão:** Este estudo aplica aprendizado profundo para prever a idade óssea, o que pode auxiliar na avaliação do desenvolvimento esquelético em áreas como pediatria e ortopedia.

Descritores: Idade óssea; Deep Learning; Radiodiagnóstico

Resumen

La edad biológica, un indicador crucial del desarrollo humano, refleja los cambios físicos y mentales asociados con el envejecimiento. La estimación de la edad ósea, un método común en la práctica clínica que busca información sobre la edad biológica, puede ser subjetiva e imprecisa. **Objetivo:** Este estudio propone métodos basados en técnicas de aprendizaje profundo para estimar la edad esquelética a partir de imágenes de rayos X de la mano. **Método:** Utilizamos conjuntos de datos divididos por género y edad para entrenar y probar los modelos. **Resultados:** Los resultados muestran estimaciones prometedoras, con errores medios de 10,808 meses en un conjunto de datos público y 15,548 meses en un conjunto de datos privado. La herramienta desarrollada, ofrece un uso práctico para profesionales. **Conclusión:** Este estudio aplica aprendizaje profundo para predecir la edad ósea, lo que puede ayudar en la evaluación del desarrollo esquelético en áreas como pediatría y ortopedia.

Descriptores: Edad ósea; Aprendizaje profundo; Radiodiagnóstico

Introduction

Growth and maturation processes are not always linear. Chronological age, based solely on date of birth, is a simplistic way to describe a person. However, hormonal levels and other factors can lead to variations in development, even among children born on the same date. Considering factors like hormones, nutrition, and environment can provide a more accurate profile. Biological age, defined by bodily changes over time, offers another perspective. Skeletal maturation or skeletal age is one method used to assess biological age. (1,11)

The evaluation of skeletal maturity or skeletal age of an individual and its comparison with chronological age is extremely important for two main reasons. First, from a medical standpoint, this evaluation is crucial for the diagnosis and treatment of pediatric disorders related to endocrinology, orthodontics, and orthopedics, as well as providing estimates of final height for the individual. Second, from a legal standpoint, the assessment of skeletal age plays a crucial role in determining the minority status of an individual when verified documents are not available. (2,12)

The estimation of skeletal age is performed through radiographic images, with hand radiography being the most used due to easy access to the anatomical site, which is an extremity of the body. Furthermore, during the image acquisition examination, there is no exposure of vital organs to the primary radiation beam. (3)

Hand radiography provides an analysis of long bones, such as the metacarpals and phalanges. Long bones have three distinct areas: epiphysis (the end region), diaphysis (the elongated central region), and metaphysis (the intermediate region between the epiphysis and diaphysis). Skeletal maturation is directly related to the calcification of the epiphysis, occurring as the individual grows. Initially, the epiphysis is an expanded joint structure that is separated from the bone by cartilage, but it later fuses, forming the bone in its final shape as the individual grows. (2)

In a radiographic examination, image formation occurs through the attenuation difference of the radiation beam in different media. Denser media attenuate radiation more, resulting in a brighter appearance on the image, while less dense media attenuate radiation less, appearing darker. Therefore, it is possible to analyze the contrast difference in the images, with bones appearing brighter and soft tissues appearing darker. (4) This contrast distinction allows for the analysis of calcification points in the epiphysis of long bones, enabling the classification of skeletal age based on the degree of these calcifications.

In clinical practice, the analysis of skeletal age is performed by the radiologist through visual examination of the radiograph. According to Olivete Júnior and Rodrigues (2010), there are several possible methods for estimating a patient's skeletal age, with the most classic and important ones being the Greulich-Pyle (GP), Tanner-Whitehouse (TW), and Elizur-Ramon (ER) methods. The first two methods utilize pattern atlases, such as the "Atlas of Skeletal Maturation of the Hand" (5), while the latter uses measurements of dimensions of specific ossification centers to estimate skeletal age. However, visual image analysis is subjective, as it depends on the physician who visually compares the patient's examination with an equivalent image in a pattern atlas, resulting in operator-dependent outcomes. (6) To optimize analysis and reduce subjectivity in skeletal age prediction, it is advantageous to implement methodologies capable of analyzing images in an automated and standardized manner.

In this context, approaches that use deep learning for predicting skeletal age from hand X-ray images have shown promising results. An example is the work by Zulkifley et al. (2021), which proposed a predictive model using deep learning and achieved an MAE of 7.699 months, demonstrating the potential of deep learning techniques for skeletal age prediction. (8) Additionally, An (2017) proposed a methodology for estimating bone age through the segmentation of ossification centers in carpal radiographs using convolutional neural networks, achieving classification accuracies of 81.38% for males and 78.17% for females within a tolerance of up to 2 years. (9) Another work is by Lee et al. (2017), which proposed a deep learning pipeline to segment the region of interest, standardize, and preprocess radiographs using a pre-trained and fine-tuned convolutional neural network, achieving accuracies of 75.59% for females and 75.54% for males within a tolerance of 1 year. (10)

The aim of this study is the development of methodologies based on Deep Learning techniques and image processing to assist professionals in determining the skeletal age of patients using combined carpal radiographic images and biological sex.

Methods

This research project proposes a methodology to assist in the prediction of bone age in patients using Deep Learning (DL) techniques, through the construction of deep learning networks. To achieve this objective, the programming language Python was employed. Figure 1 summarizes the methodology developed in this study.

Figure 1 – Flowchart of the methodology

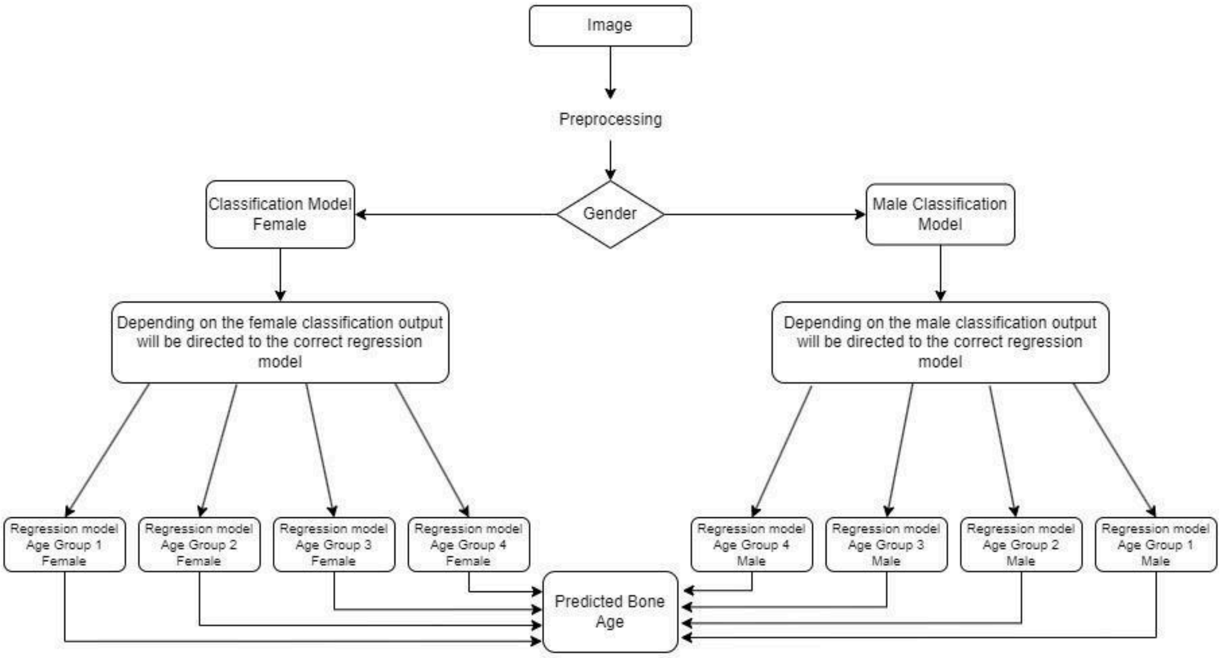


Image
Preprocessing
Gender
Classification Model Female
Male Classification Model
Depending on the female classification output will be directed to the correct regression model
Depending on the male classification output will be directed to the correct regression model
Regression model Age Group 1 Female
Regression model Age Group 2 Female
Regression model Age Group 3 Female
Regression model Age Group 4 Female
Predicted Bone Age
Regression model Age Group 4 Male
Regression model Age Group 3 Male
Regression model Age Group 2 Male
Regression model Age Group 1 Male

Database

In this research project, we utilize three different databases to develop and enhance methodologies based on Deep Learning techniques and image processing for determining the skeletal age of patients using carpal radiographic images combined with biological sex. These databases include a public database, a private database, and an atlas database.

The public database used in this research project is the repository from the article "The RSNA Pediatric Bone Age Machine Learning Challenge" [(7)], which contains a dataset of 14036 carpal radiographs. Each image is associated with its corresponding bone age and biological sex, in a .csv file format, totaling over 10 GB of data. This database is intended to be used for training and validation phases of the tool.

The private database, comprising 717 hand radiographic images, patient's biological sex, and bone age, was acquired through collaboration with HCPA (Hospital of Clinics of Porto Alegre). This database was specifically utilized for the testing phase of the tool. It is noteworthy that the use of this database for testing was approved by the Research Ethics Committee (CEP) of HCPA (CAAE: 51351321.8.3001.5327).

The Atlas database was obtained from a set of bone development atlases. The use of this database aims to reduce the bias of subjectivity. Unlike the public database, which contains images with bone ages predicted through medical evaluation based on an atlas, the images obtained from the atlas dataset have bone ages directly related to a specific atlas, without the extra subjectivity of the radiologist. This database was used only in the training and validation phases of the models. This set of atlases comprises 283 images, with 148 male and 135 female images.

Image Pre-processing

Regardless of the database used, the images have different dimensions, contrasts, and artifacts, requiring pre-processing. The following steps were performed on the imported images:

- **Background Removal**: The images were subjected to background removal using the rembg function, an automated tool that utilizes advanced machine learning techniques to remove the background and create a transparent background.
- **Object Selection:** A custom function based on mathematical morphology was used to select the object of interest and make the background black. This step ensured that only the desired object (hand) remained in the image.
- Contrast Adjustment: Histogram equalization was applied to adjust the contrast of the images, enhancing their visual quality.
- **Image Resizing**: The resulting images were resized to a dimension of 500 pixels by 500 pixels, ensuring consistency in size across the dataset.

Methodology for Model Training and Evaluation

The data was divided in a way that ensured no overlap between the sets. The public database and the atlas database were initially combined, and then an 80-20 random split was performed using the train_test_split function from the sklearn library. This split reserved 80% of the data for model training and 20% for validation. The database from HCPA was solely used for testing the models. This ensured that the three datasets had no correlation with each other.

To improve the distribution of data, the bone ages, ranging from 0 to 288 months, were segmented into four distinct age groups. This segmentation was done separately for males and females. The four age groups were as follows: Group 1: 0- 57 months, Group 2: 58- 114 months, Group 3: 115- 171 months, and Group 4: 172- 228 months.

The models employed in this study utilized a common architecture based on transfer learning, which enables the application of pre-existing knowledge from models trained on large datasets to solve similar tasks. Additional layers were incorporated, and individual hyperparameter adjustments were made for each specific model.

The EfficientNetV2L network was selected as the base architecture for the models due to its superior performance in preliminary tests compared to other models like ResNet50 and the EfficientNetB family. EfficientNetV2L demonstrated better adaptability to smaller datasets. For both classification and regression tasks, EfficientNetV2L was used with frozen layers, and additional layers were added to refine the final model.

For the classification model, after the base EfficientNetV2L layers, a Global Average Pooling layer was added to reduce the spatial dimensions of the feature maps. This was followed by a Dense layer with 1024 neurons and ReLU activation to learn complex patterns, and a Dropout layer with a rate of 0.2 to prevent overfitting. Another Dense layer with 512 neurons and ReLU activation was added, followed by another Dropout layer with the same rate. The output layer consisted of a Dense layer with 4 neurons and softmax activation, corresponding to the four age groups.

Similarly, for the regression model, the base EfficientNetV2L layers were followed by a Global Average Pooling layer. A Dense layer with 1024 neurons and ReLU activation was added, followed by a Dropout layer with a rate of 0.2. Another Dense layer with 512 neurons and ReLU activation was included, with the output layer being a Dense layer with a single neuron, corresponding to the predicted age.

Hyperparameters were meticulously tuned for optimal performance. The learning rate was adjusted using a cosine annealing function, and the batch size was optimized to balance data representativeness and processing limitations. The number of epochs was determined using the EarlyStopping method from Keras.

A delicate hyperparameter search process was conducted to identify the optimal combination. This involved variation of the hyperparameter values, training the model with the training dataset, and evaluating its performance with the validation dataset. The results of this iterative process guided the selection of the most effective hyperparameter configurations.

The evaluation of model performance employed utilized a comprehensive set of metrics. For regression models, mean absolute error (MAE) and mean squared error (MSE) were used as loss metrics, quantifying the average difference between predicted and actual ages. For classification models' accuracy, sensitivity and specificity were used as metrics and Categorical cross-entropy was used as the loss function.

The models' performance was assessed using the k-fold cross-validation method, which divided the dataset into five folds. In each iteration, one fold was selected as the validation set, while the remaining folds were used for training. This process was repeated five times, allowing each fold to act as the validation set. The application of k-fold cross-validation contributed to the reliability of the results, providing a more comprehensive evaluation of performance and a robust analysis of the models' generalization capabilities.

After completing the k-fold training process, we select the fold that exhibited the best performance during the previous iterations. This final training step refines the model using this best-performing fold and aims to achieve a model that generalizes well to unseen data, leading to more precise and reliable predictions. Following this final training, the ultimate metrics are then recorded, providing a definitive evaluation of the model's performance.

Tool Development

The tool was designed to provide an intuitive and user-friendly graphical interface. Users can import images directly from a selected directory and specify the corresponding gender. If the user knows the age range of the image, they can select it directly, skipping the classification step and directing it to the appropriate regression model, reducing the uncertainty of the classification models. Gender information is used to direct the image to one of the available classification models with the aim of determining the appropriate age range or estimating the exact age of the image. On completion of the tool, there is an option to save the image with the bone age and sex data for future training and improvement of the models.

Results and Discussion

The outcomes of each classification and regression model were obtained using training and validation data. After the 5-fold stage, the weights of the most optimal folds were saved and employed as initial weights in the retraining process, thus enabling the attainment of the final results.

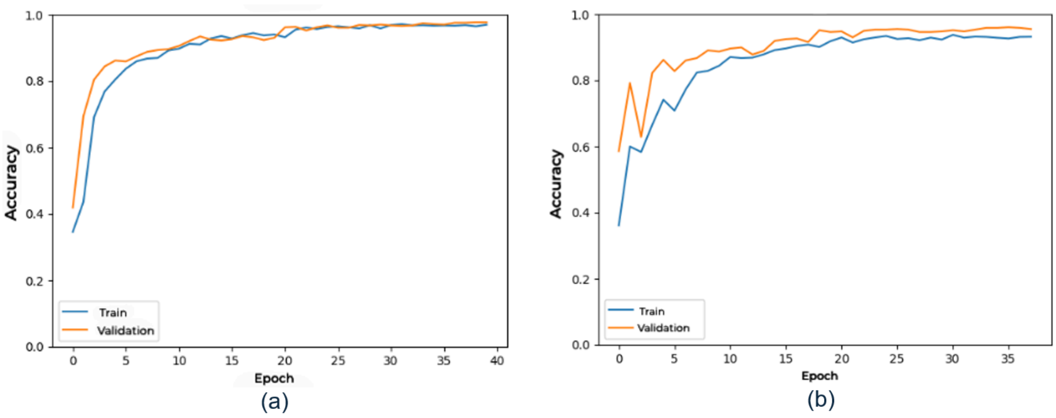
Table 1 shows the average results of the average 5-folds and the final training results for the classification models and Figure 2 shows the learning curves. It can be observed that all the learning curves (male and female), exhibit good behavior for both the training and validation data.

Table 1 – Accuracy Results for each and final training Classification Model.

Classification Model	Average of folds (Accuracy)	Final Training
Male	0.834 ± 0.034	0.814
Female	0.822 ± 0.041	0.804

The findings demonstrate that both models exhibited comparable performance in classifying the images, with little variation between the folds. In addition to accuracy, sensitivity and specificity were also calculated to provide a more comprehensive evaluation of the models' performance. The average sensitivity across all classes for the female model was 0.858, and the average specificity was 0.935. For the male model, the average sensitivity was 0.796, and the average specificity was 0.910. These metrics indicate that both models are effective at correctly identifying positive instances (high sensitivity) and accurately recognizing negative instances (high specificity), providing a comprehensive evaluation of their classification capabilities A comparison of the models reveals that the male classification model attained slightly superior results in terms of average accuracy compared to the female classification model. However, it is important to note that the differences between the models are relatively modest, indicating a similar performance between them.

Figure 2 – Accuracy Curves for the Male (a) and Female (b) Classification Model.



(a) (b)

Table 2 shows the average Mean Absolute Error (MAE) results obtained in the 5-fold stage and the final training results for the regression models. Analyzing the results in Table 2, we can observe that the regression models showed different performances regarding age ranges and gender.

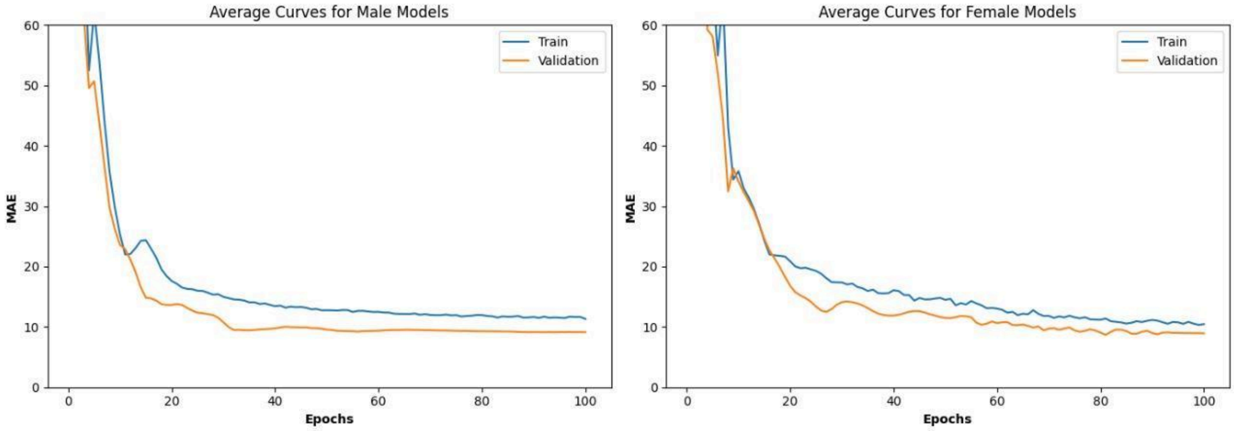
The results of the regression models' retraining indicate the difference in accuracy in estimating bone age for each age group and gender. Overall, the regression models showed relatively low Mean Absolute Error (MAE) values, exhibiting good behavior for both the training and validation data (see Figure 3 for an illustration of the average learning curves obtained with the regression models.), which suggests a good ability to estimate bone age based on the characteristics of radiographic images.

Table 2 – MAE Results for Each fold and final training Regression Model

Regression Model	Average of folds (months)	Final training (months)
Age Group 1 – Male	5.289 ± 0.146	6.991
Age Group 2 – Male	7.739 ± 0.319	8.451
Age Group 3 – Male	16.567 ± 0.295	18.077
Age Group 4 – Male	9.504 ± 1.219	15.085
Age Group 1 – Female	3.604 ± 0.181	5.656
Age Group 2 – Female	10.346 ± 0.578	12.217
Age Group 3 – Female	11.517 ± 0.298	15.809

Age Group 4 – Female	11.423 ± 0.898	15.085

Figure 3 – Average MAE Curves for Female (a) and Male (b) Regression Model for Age.

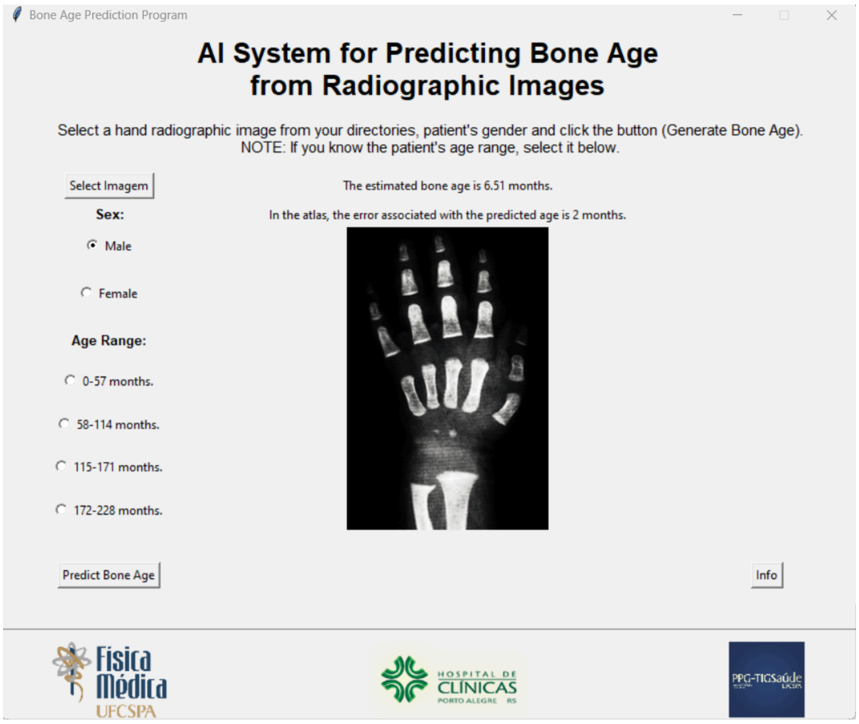


These findings suggest that specific folds demonstrated superior performance for different age groups and genders in the regression models. These results highlight the importance of considering different age groups and genders when evaluating the performance of regression models. It is important to note that the variation in MAE values across age groups is a consequence of the difference in the quantity and characteristics of available images for each model. This emphasizes the need for diverse and representative datasets that encompass a wide range of ages and genders to ensure accurate and reliable regression model performance. By considering these factors, researchers and practitioners can obtain more meaningful insights and make informed decisions based on the specific demographics they aim to target.

By applying the methodology as a whole, which involves directing the image first through the classification models and then forwarding it to the regression model recommended by the classification, we obtained notable results. In the public dataset, we achieved a final average MAE of 10.808 months, while in the private dataset, the final average MAE was 15.548 months. The observed variation in MAE can be attributed to the differences between datasets. Each dataset contains distinct images, and since the models were trained on the public dataset, it is expected that the performance would be better when applied to images with characteristics closer to those in the training data.

The Figure 4 illustrates the primary interface, which includes an exemplar of a processed image and the displayed results of bone age prediction. The developed tool offers practical use for medical professionals and researchers. The direct import of images, along with the selection of gender and corresponding age range, allows for personalized analysis.

Figure 4 – Interface image.



Bone Age Prediction Program

AI System for Predicting Bone Age from Radiographic Images

Select a hand radiographic image from your directories, patient's gender and click the button (Generate Bone Age).
NOTE: If you know the patient's age range, select it below.

The estimated bone age is 6.51 months.

In the atlas, the error associated with the predicted age is 2 months.

Sex:
● Male
○ Female

Age Range:
○ 0-57 months.
○ 58-114 months.
○ 115-171 months.
○ 172-228 months.

Comparing with the state of the art, it is important to consider certain criteria. Firstly, only methodologies that used the same database should be compared. The results found by the state-of-the-art methodologies were based on the same database, meaning they trained and tested with the same database (unlike the current study, which was tested with HCPA data). Thus, the results related to the validation of the present study were compared with the state of the art.

Applying the validation images in the final flow of the proposed methodology resulted in a final average Mean Absolute Error (MAE) of 10.808 months. Compared to the state of the art of 7.699 months (Zulkifley et al., 2021), it is noted that the result found demonstrates a lower performance than that obtained by the state of the art. This result was a consequence of the accuracies close to 80% of the classification models added to the MAE associated with each regression model. It is evident that besides the accuracy close to 80% in the classification models, in the regression step, the higher MAE of the older age groups elevated the final average result of the methodology.

In comparison to An (2017), who achieved classification accuracies of 81.38% for males and 78.17% for females within a tolerance of up to 2 years, our model's performance is similar, with the male model achieving an accuracy of 0.834 ± 0.034 and the female model achieving an accuracy of 0.822 ± 0.041. However, our methodology provides the added benefit of segmentation by age group, allowing for a more precise analysis. Comparing with Lee et al. (2017), who achieved accuracies of 75.59% for females and 75.54% for males within a tolerance of 1 year, our models demonstrated better performance with accuracies above 80%. This suggests that our approach, while having higher MAE in the older age groups, offers improved accuracy in classification.

It is important to note that, in the younger age groups, both for males (with a found error of 4.993) and for females (with a found error of 3.151), the proposed methodology presented a superior performance compared to the state of the art, which registered an average error of 7.699. Additionally, unlike the state of the art, the methodology proposed considered the separation of data by age group, which provides a more precise and specific analysis for each age group, allowing for future adjustments to the methodology. A possible explanation for this difference is that in older age groups, images have more subtle variations compared to younger age groups, where there is greater diversity and variation in characteristics.

Conclusion

In this study, it was developed and evaluated deep learning models for estimating bone age in carpal radiographs using classification and regression techniques. The methodology achieved an accuracy of over 80% and a mean absolute error ranging from 5 to 18 months for private data, with the lowest errors in younger age groups. However, performance was lower in older age groups due to lower variation in image characteristics. The complete methodology, involving classification and recommended regression models, showed promising results, with an average MAE of 10.808 months in the public dataset and 15.548 months in the private dataset.

A significant innovation of our work is the development of distinct models for different age groups, recognizing that bone maturation is not uniform. There is a notable disparity in the early age groups, while the later age groups exhibit minimal differences even across several years. This stratified approach allows for the optimization of models tailored to each specific age range, thereby creating a framework of models that could have a profound impact in clinical settings in the future.

A practical solution for medical professionals and researchers was developed, enabling personalized analysis through the direct import of images and selection of gender and age range. In comparison the state-of-the-art methodologies by Zulkifley et al. (2021), which achieved an MAE of 7.699 months, our approach demonstrated slightly lower performance with a higher MAE. However, compared to An (2017), who achieved classification accuracies of 81.38% for males and 78.17% for females, and Lee et al. (2017), who achieved accuracies of 75.59% for females and 75.54% for males, our models demonstrated better performance with accuracies above 80%. In particular, the younger age groups exhibited superior performance, with an MAE of 4.993 for males and 3.151 for females, in comparison to the state-of-the-art average error of 7.699.

The findings demonstrate that our methodology offers a promising and approach by segmenting data by age group. This provides a more precise and specific analysis that could lead to significant improvements in clinical applications. Future work includes exploring different neural network architectures, preprocessing techniques, and expanding the dataset to enhance model performance.

References

1. Prokop-Piotrkowska M, Marszałek-Dziuba K, Moszczyńska E, Szalecki M, Jurkiewicz E. Traditional and new methods of bone age assessment-an overview. J Clin Res Pediatric Endocrinology. 2021;13:251.

2. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS One. 2019;14

3. Delorme AL. Automatic methodology for bone age estimation using shape analysis in carpal radiographs [dissertação de mestrado]. São Carlos: School of Engineering of São Carlos, University of São Paulo; 2010. [citado em 13 fev 2023].

4. Vrbaški S, Ito M, Moyano LG, de Santana VF. Characterization of breast tissues in density and effective atomic number basis via spectral X-ray computed tomography. Physics in Medicine & Biology. 2023;68(14):145019.

5. Todd TW. Atlas of Skeletal Maturation. The C.V. Mosby Company; 1937. p. 37.

6. Olivete Júnior C, Rodrigues ELL. Bone maturity: estimation by simplifications of the Eklof and Ringertz method. Radiol Bras. 2010;43.

7. Halabi SS, et al. The RSNA pediatric bone age machine learning challenge. Radiology. 2019;290:498-503.

8. Zulkifley MA, Mohamed NA, Abdani SR, Kamari NAM, Moubark AM, Ibrahim AA. Intelligent bone age assessment: an automated system to detect a bone growth problem using convolutional neural networks with attention mechanism. Diagnostics. 2021;11(5):765.

9. An DY. Bone age estimation using mosaics of ossification centers from carpal radiographs as input images for Deep Learning [dissertação de mestrado]. Espírito Santo: Federal Institute of Espírito Santo; 2017.

10. Lee H, et al. Fully Automated Deep Learning System for Bone Age Assessment. Boston: Springer; 2017. p. 30, 427-441.

11. Tuma CESN, et al. Assessment of bone age in children aged 9 to 12 years in the city of Manaus-AM. Dental Press J Orthod. 2011;16(3):63-69.

12. Pinto VCM, et al. Relationship of bone age and hormonal markers with the physical capacity of adolescents. J Hum Growth Dev. 2017;27(1):77-83.

DOI: 10.59681/2175-4411.v16.iEspecial.2024.1361