Bone Age Prediction Using Convolutional Neural Networks

Sebastiano Monti† Valentina Tonazzo‡

***Abstract*—This project revolves around the replication and enhancement of a neural network model originally employed in a successful bone age assessment contest using hand radiography. The overarching goal was to explore innovative architectural approaches that could simultaneously improve predictive performance and reduce computational complexity. One of the main aspects of this project was the exploration of three prominent pretrained architectures: InceptionV3, Xception, and a customized variant of ResNet50 featuring overlapping patches. The comparative analysis of these architectures during the testing phase yielded invaluable insights into their strengths and weaknesses. Additionally, an exploration into the role of patch dimensions revealed a compelling finding: larger overlapping patches may play a crucial role in enhancing the model's overall understanding of the image. This revelation underscores the importance of patch size in medical image analysis and its potential implications for future research. To further leverage the strengths of each architecture, an ensemble approach was employed achieving more robust predictions and less overfitting. The project's findings are poised to pave the way for advancements in healthcare through the integration of artificial intelligence.**

***Index Terms*—Convolutional Neural Networks, Bone Age Prediction, Computer Vision, Image Classification.**

I. Introduction

During the growth of a child, bones change in size and shape and a deviation of the bone age from the chronological age may address the presence of a growth problem. Bone age assessment is a common clinical practice that estimates the maturity of skeletal system and is used in order to diagnose endocrine disorders in children and adolescents. The adopted standard for bone age assessment is based on two most relevant techniques. The Greulich-Pyle method [1] and the Tanner Whitehouse method [2]. Being these human supervised techniques, they require a significant amount of time and workload from experienced medical personnel to be conducted. Also, they can face the influence of doctors with different standards. The problem of bone age assessment using automated approaches received therefore increasing attention in the last years, in that it could for sure minimize judgment time, but also standardize the evaluation removing the human variability component. Of course still requiring doctoral supervision and expertise.

Many studies have shown that deep learning techniques can be successfully used in the domain of medical imaging [3], [4]. In particular, in the field of bone age assessment, studies have shown that convolutional neural network (CNN) based techniques can achieve predictions whose accuracy is similar to that of an expert radiologist [5].

The purpose of this report is to show the different approaches that have been followed in facing the problem of bone age prediction and compare the implemented models in order to evaluate if one is able to outperform the others. The adopted techniques have in common the usage of CNNs, differing however in the processing of the image dataset. One of the implemented approaches regarded the age prediction over the radiographs in their entirety, while the other consisted in processing multiple patches of the original images. Moreover, the sex information about the candidates was added to the radiographs, hoping that this would improve the accuracy of the predictions.

II. Related Work

In the domain of pediatric radiology, accurate assessment of skeletal maturity, often determined by bone age estimation, plays a pivotal role in clinical practice. Traditionally, this task has relied on manual assessment by radiologists, which is subject to inter-observer variability and time-consuming procedures. The emergence of deep learning techniques has introduced the potential for automating and enhancing the precision of skeletal maturity assessments through the analysis of pediatric hand radiographs. The *"Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs"* [5] presents a study aimed to compare the performance of a deep-learning-based bone age assessment model using hand radiographs with that of expert radiologists and existing automated models. A total of 14,036 clinical hand radiographs and corresponding reports were obtained from two children's hospitals for model training and validation. In the first test set comprising 200 examinations, the model's bone age estimates were compared with the mean estimates from clinical reports and three human reviewers.The assessment of the model's performance included comparing the root mean square (RMS) and mean absolute difference (MAD) between the model's estimates and the reference standard bone ages.

The *"RSNA Pediatric Bone Age Machine Learning Challenge"* [6] aimed to showcase the role of AI in medical imaging. Participants were tasked with creating ML algorithms to determine pediatric skeletal age from hand radiographs. The challenge attracted 260 participants, with the top-performing models achieving a mean absolute distance (MAD) of 4.2

†Department of Physics and Astronomy, University of Padova, email: sebastiano.monti@studenti.unipd.it

‡Department of Physics and Astronomy, University of Padova, email: valentina.tonazzo.1@studenti.unipd.it

to 4.5 months. This collaborative approach demonstrated the potential of ML in improving diagnostic accuracy and patient care. In particular the winning approach utilized both pixel and sex information within a single neural network, using 500x500-pixel images. It employed the Inception V3 architecture for pixel data and concatenated it with sex information. In this paper is shown a computational implementation of the latter model, reaching reasonable results, and setting it as a baseline model for further comparison.

This work mainly differ from the previous ones since the neural network was then modified in order to make it computationally less expensive but at the same time more performative: in order to do this it was explored the Neural network called Xception, first presented in the *"Xception: Deep Learning with Depthwise Separable Convolutions"* [7]. The second-place approach involved training sex-specific models using contrast-enhanced image patches of size 224x224 pixels, which were subdivided into 49 overlapping patches. Transfer learning and fine-tuning of ResNet-50 architectures pretrained on ImageNet were used in this approach. Also in this case the role of patches was further explored and analyzed to achieve a reacher comprehension of the patch utility. Finally a tecnique of ensamble voting was used to enhance the overall performances and draw a comparison between all the implemented methods.

III. Processing Pipeline

A first step of this project consist in downloading the image RSNA train, validation and test sets and associated metadata from the official website, which includes pediatric hand radiographs, gender information, and age labels. The dimension of the train set were over 10 GB of memory so it was necessary to reduce the images dimensions and make the dataset more tractable. Firstly, each radiographic image was resized to a consistent size of 500x500 pixels. Simultaneously, the metadata have been processed. It is essential to format this data appropriately for model training and evaluation. Data generators play a pivotal role in organizing and batching the images, so they were used to prevent the memory overusage; additionally, a specialized data generator was designed to extract patches from each image, which was useful for patch-based ResNet50 model. Inside the dataloader images are augmented to enrich the dataset's diversity. This project involves the implementation and training of three distinct neural network models, based on different pretrained deep neural networks. During the training process, *ReduceOnPlatou* and *Earlystopping* callbacks were used to improve training efficiency and prevent overfitting. A learning rate scheduler was also implemented, considering an exponential decay from a value of 10^{-3} to 10^{-5} within each epoch, without however displaying a real improvement over the losses and resulting instead in slower performance improvements. For this reason the learning rate was initially fixed to 10^{-3} and eventually automatically reduced with the *ReduceOnPlatou* callback in case the validation loss stopped to improve. Each model was trained for 7 epochs, and the validation set serves as a critical reference to monitor training progress. In the compiling step it was chosen Adam as optimizer, Mean Squared Error as loss and Mean Absolute Error in months as accuracy. Each model was then evaluated using the designated test set. Essential evaluation metrics were computed, including mean absolute error (MAE), root mean square error (RMSE), and correlation coefficients, to gauge the models' accuracy in age predictions.

IV. Signals and Features

The dataset that have been considered for this project is the *2017 RSNA Pediatric Bone Age Challenge Dataset*. It contains a total of 14236 `.png` pediatric hand radiographs subdivided as shown in Table 1.

Set Type	Number of Samples	Female %	Average Age
Training	12611	0.46	127
Validation	1425	0.46	127
Test	200	0.5	132

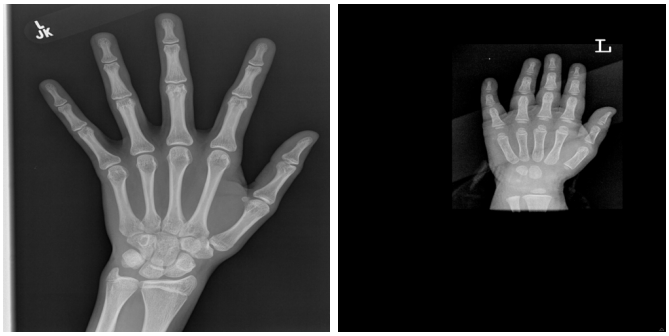
TABLE 1: Summary of dataset subdivision.

Each dataset is accompanied by a `.csv` file containing for each image, the sex of the candidate and his or her ground truth bone age expressed in months, spanning from a minimum of 1 months to a maximum of 228 months.

The same preprocessing pipeline was applied to all the subsets and involved the following steps. Firstly, in order to have images with same dimensions, but also to reduce computation times and memory expense, all images have been resized to a dimension of 500×500, following the same choice of RSNA challenge's winners [6]. Two examples from the resized training set can be seen in Figure 1.

By examining the distribution of the data in terms of bone age and sex of the candidates, it is possible to observe a similar trend between all the datasets. Histograms of these distributions for the training dataset are shown in Figure 2a.

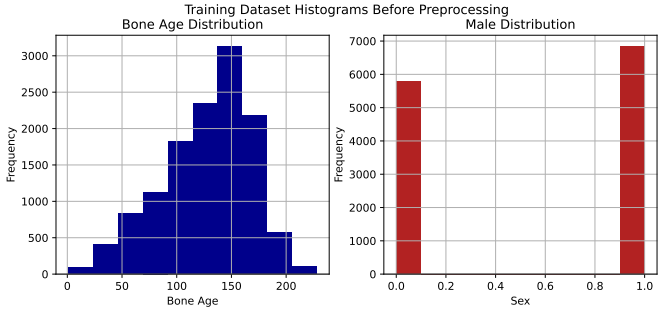
Preliminary trials showed that, as one may expect, the model learns to predict with higher accuracy the bone age from samples that are present with a higher frequency, with respect to the ones that instead lays on the tails of the



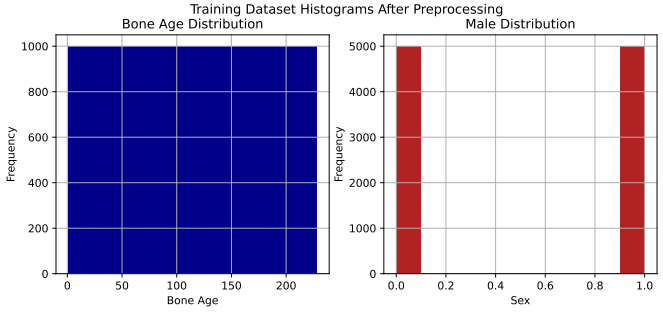
(a) Bone age = 180 (b) Bone age = 33

Fig. 1: Examples of images after resizing. Dimensions are 500×500.

histogram. For this reason, it was decided to randomly re-sample the images from each dataset, in order to obtain a uniform distribution both in the bone age and in the sex of the candidates. The training set histograms after re-sampling are depicted in Figure 2b. This allowed to obtain a significant improvement of the loss in each of the implemented architectures.



(a)



(b)

Fig. 2: Training dataset's bone age and sex distributions, before and after preprocessing.

Having this done, a normalized version of the bone age was calculated by subtracting to each age value, the average over the re-distributed dataset and dividing by its standard deviation. Then, a dataloader was created by following two different approaches.

- **Image data augmentation:** the dataloader generates randomly shuffled batches of 8 images, respectively associating to each of them the information about sex of the candidate and the bone age as label. Images are converted from a single channel to RGB, then transformed from a representation in interval $[0, 255]$ to one in $[-1, 1]$ and finally augmented using the augmentation parameters summarized in table 2.
- **Patches data augmentation:** images are converted to RGB and randomly shuffled, then 3 patches of dimensions 224×224 are randomly extracted from each of them. Each patch is associated to the respective sex of the candidate and to the bone age label. Data augmentation is also applied, although in this case just a rotation range of $5°$ and horizontal flip were considered. In this case, generated batches contained a total amount of 24 images, corresponding to 8 different candidates (i.e. 3 patches $\times$ 8 candidates).

Augmentation Parameter	Range
Rotation	$5°$
Width shift	0.1
Height shift	0.1
Shear	0.01
Zoom	0.25
Horizontal flip	True

TABLE 2: Image data augmentation parameters.

V. Learning Framework

A. Inception-V3 model

The proposed neural network model is based on the RSNA boneage challenge winning model: the input is formed by two different layers:

- Gender Input: This layer receives the gender information as a single value (0 for male, 1 for female) and is designed to accommodate batch processing.
- Image Input: accepts RGB images of size 500x500 pixels, representing pediatric hand radiographs, also images are processed in batches for efficiency.

A dense layer with 32 units and a sigmoid activation function is applied to the gender input to capture and encode the sex-related information. The model utilizes the InceptionV3 architecture, pretrained on the ImageNet dataset, as a feature extractor; transfer learning is employed, with the layers of the InceptionV3 base frozen to retain the prelearned features. The global average pooling layer aggregates the feature maps generated by the InceptionV3 base, it was choosen as the most popular layer after pretrained inceptionV3. Finally two densely connected layers with 1000 units and a hyperbolic tangent (tanh) activation function further processes the fused features. The final prediction layer consists of a single unit with a linear activation function, producing the estimated skeletal age as a continuous numeric value. Between the last trainable layers it were added some dropout layers in order to prevent overfitting.

B. Xception model

In a similar way to the one described for the previous network, the first input layer of the Xception-based model is dedicated to gender-related data, accepting a single numerical value representing the individual's male or female variable. This input is processed through a dense layer with sigmoid activation, allowing the network to learn gender-specific features. The image input layer again, accepts images with a resolution of 500x500 pixels and 3 color channels (RGB), since these images serve as the primary source of visual information for age estimation. Notably, in contrast with the previous network now it has been used the pre-trained Xception model, initialized with ImageNet weights, to

extract features from the images. Xception, short for "Extreme Inception", [ref] is a neural network architecture that was introduced as an evolution of the Inception architecture. It was designed to achieve state-of-the-art performance in computer vision tasks, particularly image classification and object recognition. Xception draws inspiration from the Inception architecture, particularly the idea of using multiple parallel convolutional filters with different receptive field sizes. However, instead of traditional convolutions, it primarily utilizes depthwise separable convolutions, which are more efficient . This means it separates the spatial convolution (depthwise convolution) and the pointwise convolution (1x1 convolution) into two separate operations. This helps reduce the number of parameters and computation, making the network faster and more efficient. Fine-tuning of the Xception model is disabled by setting its layers to non-trainable. As in the previous network, Xception typically uses global average pooling (GAP) to reduce spatial dimensions. The extracted image features and the processed gender-related information are concatenated to form a fused feature representation. Then, the fused feature representation is passed through a flatten layer and subsequently through a dense layer with 10 units and ReLU activation. These intermediate layers help the network capture complex relationships between the input data. As a main difference with respect to InceptionV3-based model, here there are no Droupout Layers, reducing the complexity of the neural network, and parameters. The final prediction is obtained through a dense layer with a single output unit and linear activation. This output represents the estimated age of the individual. In complex the Neural network is formed by a total of 20.882.365 (79.66 MB) parameters, which is quite similar to the amount obtained for the Inception-V3 model, but just 20.885 (81.58 KB) of them are trainable, making the whole network less computational expensive.

C. Xception with attention Layer

In order to enhance performances of the Xception model it was included a simple spatial attention mechanism. The multi-head attention mechanism is a fundamental component of modern deep learning models, particularly in the field of computer vision. Its main purpose to capture complex relationships and dependencies within data, enabling models to focus on different aspects or patterns within the input simultaneously. This mechanism has proven to be highly effective in improving the performance of various machine learning tasks, such as image analysis. Multi-head attention consists of multiple "heads" or attention channels. Each head is a distinct attention mechanism that learns a unique set of attention weights. These attention heads work in parallel and capture different patterns or features from the input data. Within each attention head, a set of learnable parameters is The weights determine the importance of each element in the input data concerning the other elements. In this project the number of attention heads has been fixed to 8 while the kernel size which determines the size of the convolutional kernels used within each attention head has been fixed to equal to 3. After processing all attention heads, the layer concatenates the head outputs along the last axis. This concatenation combines the information learned by each attention head, creating a richer representation of the input feature maps.

D. ResNet50 model

In a similar way as the previous presented models, here it has been explored the ResNet50 architecture: ResNet50 is a deep neural network with 50 layers. These layers are organized into blocks, with each block consisting of multiple convolutional layers. The central innovation of ResNet50 is the use of residual blocks, which include shortcut connections (skip connections or identity shortcuts) that skip one or more layers. Whitin a residual block, the output of a layer is added to the output of a previous layer, effectively creating an identity mapping which ensures that the network can always learn the identity function as a minimum requirement. Any learned transformation is applied as an adjustment to this identity. In this way the architecture helps address the vanishing gradient problem, making it easier to train very deep networks. The main difference with the previous presented models is in the input of this network: the images are (224x224) overlapping patches of the original augmented images: by analyzing small, localized regions, the network can extract local features and patterns, which can be crucial for recognizing complex structures. The output from the ResNet50 base model is passed through a Global Average Pooling 2D layer, which helps reduce the spatial dimensions while The gender information input and the processed image features are concatenated together, in this case the gender information is not previously processed before the concatenation. The concatenated features are then flattened, followed by a dense layer with 14 units and a ReLU activation function. Finally, a single dense layer with a linear activation function produces the output prediction.

Model Type	**Total parameters**	**trainable parameters**	**no-trainable parameters**
InceptionV3	24.885.849	3.083.065	21.802.784
Xception	20.882.365	20.885	20.861.480
ResNet50	23.616.427	28.715	23.587.712

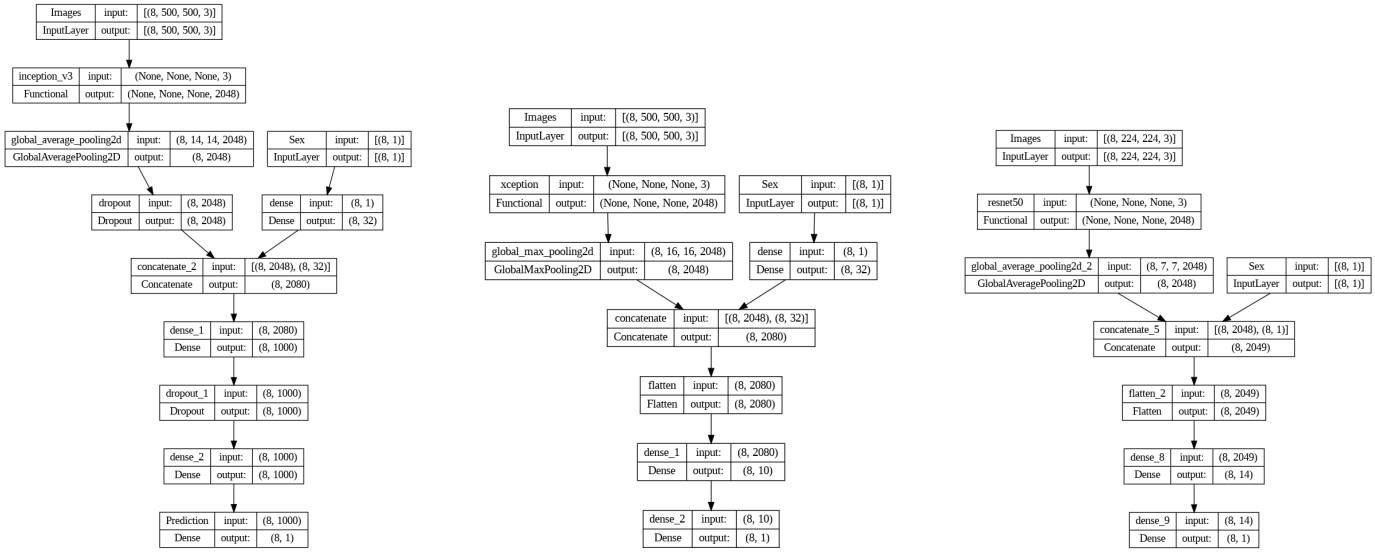
TABLE 3: Comparison of different architecture's parameters.

VI. Results

In this section, the obtained results over the training of the implemented models are summarized. The main tested conditions regard the variation of the dataloader (i.e. analysis of complete augmented images or analysis of augmented patches of the original radiographs) or the variation of the core models (Inception-V3, Xception, Xception with attention layer, or ResNet50). In Table 4 the main training configurations are summarized.

A. Analysis of complete radiographs

The batch and validation losses during the training process for each implemented model are respectively showed



(a) InceptionV3-based architecture (b) Xception-based architecture (c) Resnet50-based architecture

Fig. 3: Detailed description of the implemented architectures.

Data Size	Core model
500×500 Images	Inception-V3
500×500 Images	Xception
500×500 Images	Xception with attention
224×224 Patches	ResNet50

TABLE 4: Summary of considered training configurations.

in Figures 4a, 4b in blue, orange and green. From the plots it can be observed that the validation loss exhibits an oscillatory behaviour, apart probably for the Exception model, where performances would may have benefited from a longer training.

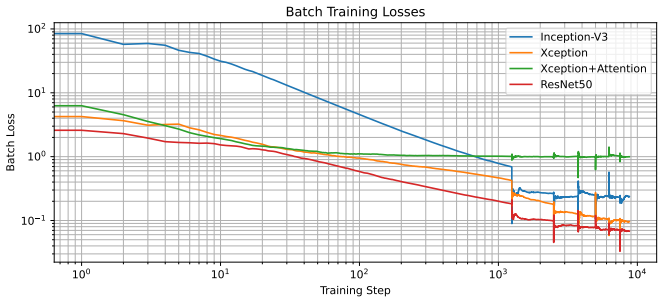
B. Analysis of radiographs' patches

Batch and validation losses for the training process of radiographs' patches are displayed in red in Figures 4a, 4b. In this case the validation loss displays a constant decreasing trend, leading to the best result obtained so far. Probably, also in this case performances would have benefited from a longer training.

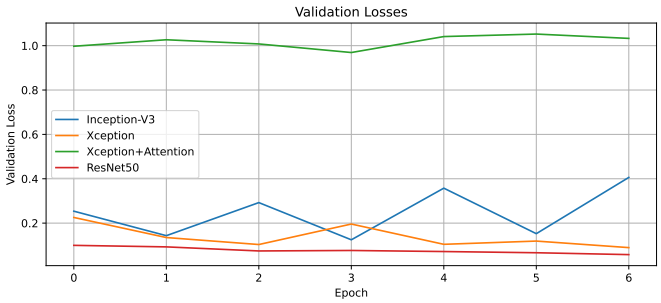
Since analyzing patches gives more freedom in the adjustment of patch size and number of extracted patches from each radiography, a fine tuning test was done in order to assess which combination of these two parameters led to the best performances on the training loss over one training epoch. In

Number of patches	Patch size
3	224×224
9	128×128
36	64×64
150	32×32

TABLE 5: Tested fine tuning parameters for patches' analysis.



(a)



(b)

Fig. 4: Batch and validation losses for each implemented model.

Table 5, the considered combinations are summarized, while the results of the training are showed in Figure 5.

The observation that it is possible to take from this plot is that when generating a high number of small patches, the model is not able to generalize as well as when generating a small number of bigger patches. This justifies also the initial choice of setting the patch number and size respectively to 3 and 224×224.

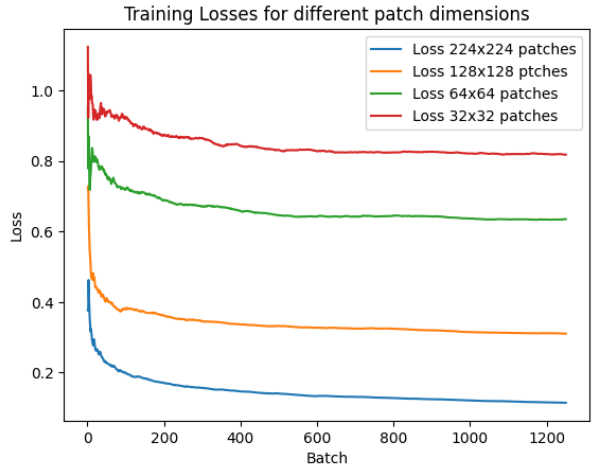


Fig. 5: Plot of the losses over one train epoch for the fine tuning over patch number and size.

C. Comparison between the models

To better compare the implemented architectures, their performances have been assessed over their output predictions on the test set and then compared to the real bone age values. Results of the Xception model with attention mechanism were not included in the analysis, since its predictions significantly deviates from the ones of the other architectures. Furthermore, in order to derive the average behaviour of the implemented models, an ensemble prediction was computed by performing a weighted average over the corresponding bone age predictions from each of the architectures. The weights of each model have been fixed to: 0.5 for the ResNet based model and 0.1 for the Inception-V3 and Xception-based models. The comparison between the results is visualized in Figure 6.

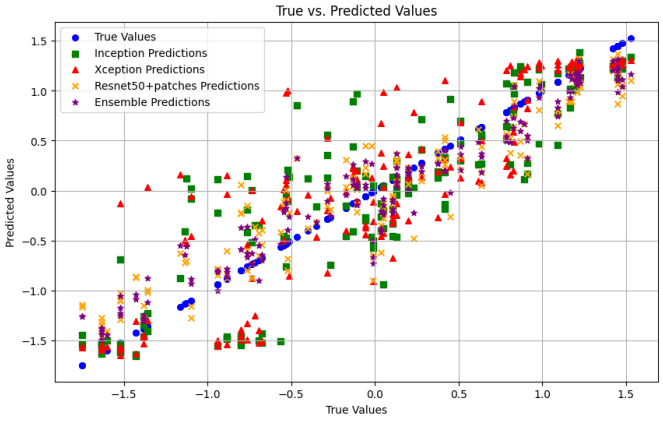


Fig. 6: Plot of the predictions of each model over the test set.

From this graph it is possible to observe that all the models are approximately equally able to predict the bone age around the edge areas (i.e. in small intervals around -1.5 and 1.5). However, the ResNet50-based model is the best in predicting also the central area of the bone age axis. This suggests that the architecture based on patches of the original radigraphs is better at generalizing over all bone ages with respect to the other models. Probably because it is more able to focus on local features of the images. In table 6, the results of a statistical analysis are presented, including the calculation of Mean Squared Error, Root Mean Squared Error and P-value between the predictions of each model and the ground truth:

Model	MSE	RMSE	P-value
InceptionV3	0.19	0.44	0.89
Xception	0.18	0.42	0.91
ResNet50	0.06	0.24	0.97

TABLE 6: Comparison of different architecture's parameters.

VII. Concluding Remarks

In conclusion, the replication of the original winning model from RSNA bone age challenge was successfully achieved; the introduction of a novel architecture using Xception-based model was crucial to improve the overall performances and reduce computational costs. However the model do not seems to benefit from the usage of attention mechanism approach. Additional improvements has been accomplished with the introduction of the patch-based approach which outperformed all models both in the stability and generalization of predictions. Further investigations into fine-tuning the attention mechanism, as well as the exploration of more complex attention structures, could potentially bridge the performance gap between attention and patch-based approaches. Finally also a general fine-tuning, in particular regarding the learing rate, for all the models proposed could promise better performances.

References

[1] T. Widek, P. Genet, T. Ehammer, T. Schwark, M. Urschler, and E. Scheurer, "Bone age estimation with the greulich-pyle atlas using 3t mr images of hand and wrist," *Forensic Science International*, vol. 319, p. 110654, 2021.

[2] A. K. Poznanski, "Assessment of skeletal maturity and prediction of adult height (tw2 method)," *American Journal of Diseases of Children*, vol. 131, no. 9, pp. 1041–1042, 1977.

[3] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pp. 737–744, Springer, 2018.

[4] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," *Scientific reports*, vol. 8, no. 1, p. 1727, 2018.

[5] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, 2018.

[6] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, *et al.*, "The rsna pediatric bone age machine learning challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, 2019.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016.