Deep Learning Approaches for Pediatric Bone Age Prediction from Hand Radiographs

Umut Berk Cakmakci†

***Abstract*—In the field of pediatric medicine, accurate and timely prediction of bone age is essential for assessing growth and development, and informing treatment decisions. This study addresses the RSNA Pediatric Bone Age Challenge using deep learning and image analysis techniques and presents experimental results and analysis from two alternative model architectures. The challenge aims to advance artificial intelligence (AI) in medical imaging by encouraging the development of accurate machine learning (ML) models for determining skeletal age from pediatric hand radiographs. In this paper, two models are presented: a transfer learning model using Xception, enhanced with the AdamW optimizer and the preprocess_input image pre-processing algorithm, and a ConvMixer model, also using AdamW with CLAHE. Both models incorporate categorical, numerical, and visual data to improve prediction accuracy. Additional techniques, including batch normalization and regularization, are applied to stabilize intermediate activations, prevent overfitting, and improve robustness. The significance of our results lies in its potential to improve pediatric bone age prediction accuracy, streamline diagnostic workflows, and reduce the workload on healthcare professionals. Overall, this research offers valuable insights into pediatric bone age assessment and presents opportunities for further exploration in this important area of child health.**

***Index Terms*—Bone Age Prediction, Deep Learning, Pediatric Radiology, Xception Model, ConvMixer Architecture, Medical Image Analysis**

I. Introduction

In pediatric healthcare, accurately assessing bone age is crucial for diagnosing and monitoring growth disorders, enabling early intervention and informed treatment decisions. Traditional bone age estimation methods rely on radiologists manually comparing X-ray images to standard references, such as the Greulich-Pyle atlas [1]. This approach, while widely accepted, is time-intensive and prone to inter-observer variability, which can result in inconsistent diagnoses and treatment plans. Consequently, there is a pressing need for automated and reliable solutions that streamline this process and improve diagnostic accuracy.

The recent emergence of deep learning techniques offers promising alternatives, particularly in medical imaging tasks where large datasets and sophisticated models can significantly improve prediction accuracy and efficiency. In this context, the RSNA Pediatric Bone Age Challenge [2] has encouraged the development of machine learning (ML) solutions for estimating skeletal age from pediatric hand X-rays, promoting advancements in artificial intelligence (AI) applications in healthcare. Several studies [3] have applied deep learning to this problem, achieving varying degrees of success. However, challenges persist, particularly with the integration of multi-dimensional data and optimizing model robustness and interpretability for clinical settings.

This paper addresses these challenges by presenting two novel approaches for pediatric bone age estimation using deep learning: a transfer learning model based on Xception [4] architecture used with AdamW [5] optimizer and preprocess_input pre-processor, and a custom ConvMixer [6] model that combines convolutional and mixer layers, also used with AdamW and CLAHE [7] pre-processor. Both models utilize a combination of categorical, numerical, and visual data to improve prediction accuracy, with techniques like batch normalization and regularization to reduce overfitting and enhance model robustness. Our research holds potential to advance the accuracy and efficiency of bone age assessment, offering a scalable solution that could be adapted for clinical workflows, reducing radiologist workload and providing more consistent diagnostic support.

The main contribution of this paper is to include ConvMixer architecture, a simple yet efficient model for integrating convolutional and mixed layers, into a regression task. This model effectively improves feature extraction in pediatric hand X-ray data, leads to increased discriminative power of the network and more accurate predictions. Additionally, we incorporate multi-input data fusion by combining categorical, numerical, and visual data inputs, which significantly enhances prediction accuracy and highlights the value of multi-dimensional data integration in bone age estimation.

This paper is structured as follows: Section II reviews related work in automated bone age estimation. Section III presents the general model architectures, while Section IV details the dataset preparation and processing. Section V outlines the detailed model architectures and training strategy and Section VI provides a comprehensive evaluation of results. Finally, Section VII offers concluding remarks and future research directions.

II. Related Work

Bone age assessment (BAA) has been a crucial tool in pediatric healthcare for decades, traditionally carried out using manual techniques like the Greulich-Pyle (GP) [8] and Tanner-Whitehouse (TW2/TW3) [9] [10] methods. However, manual assessments are time-consuming, subjective, and prone to intra- and inter-observer variability. Over the years, automated approaches have sought to address these limitations. The

†Department of Information Engineering, University of Padova, umutberk.cakmakci@studenti.unipd.it

goal of this paper is to perform automatic skeletal bone age assessment (BAA) using deep-learning methods. Thus, we will first review existing automated bone age assessment methods analyzing their advantages and limitations and then deep-learning based approaches for medical images.

Early automated systems aimed to mimic the manual GP and TW2/TW3 methods, leveraging image processing techniques. One of the pioneering efforts was by Thodberg et al. [11], who developed the BoneXpert system, automating BAA by extracting features from hand radiographs and applying statistical models to predict bone age. Harmsen et al. [12] expanded on this work by refining the morphological feature extraction process. Similarly, Somkantha et al. [13] utilized edge detection and shape analysis to identify bone structures. These traditional methods laid the foundation for automated BAA but fell short in addressing variability in radiograph quality and required significant manual intervention during the feature extraction phase.

To motivate the development of high quality and more general BAA methods, the Radiological Society of North America (RSNA) organized the Pediatric Bone Age Machine Learning Challenge. Several teams participating in the RSNA Challenge have explored advanced deep learning architectures, such as InceptionNet and ResNet. In the winner model [3], InceptionV3 network was used for feature extraction on the input gray image, and gender input was transformed into a 32-dimensional feature vector through the fully connected layer. Finally, the concatenation of the two feature vectors allowed the network to learn from images and gender information to accurately predict the bone age. The result for this model was mean absolute error (MAE) of 4.2 months.

Using this dataset, a large number of automated BAA methods have been proposed. Larson et al. [14] were among the first to introduce a regression-based deep convolutional neural network (CNN) model for bone age prediction, training their model end-to-end on the RSNA dataset. Their work demonstrated the feasibility and superiority of deep learning for BAA, achieved state-of-the-art performance (MAE of 6.24 months), and drastically reducing human involvement in the prediction process. Building on this, Spampinato et al. [15] enhanced model robustness by incorporating advanced data augmentation techniques, allowing the model to generalize better across different radiograph qualities. This study improved prediction consistency, reached the MAE of 9.48 months. Gonzalez et al. [16] took a step forward by integrating attention mechanisms into their CNN model, enabling the system to focus on the most relevant regions of the hand radiograph. This approach improved accuracy, reached the MAE of 6.34 months.

Our research improves upon these deep learning methods by not only leveraging ConvMixer architecture for bone age prediction but also harnessing the power of pre-trained model using transfer learning with Xception. For image processing, CLAHE and preprocess_input for different types of networks were used to enhance image quality by improving contrast and standardizing pixel values. Finally, we employ the AdamW optimizer for its advanced regularization through weight decay, enhancing the model's stability and generalization performance.

Our main contributions can be summarized as:

- We introduce a simple yet efficient ConvMixer architecture in our model. ConvMixer's patch-based approach allows efficient spatial and channel mixing without self-attention, making it ideal for extracting features from radiographic images.
- We leverage the power of transfer learning by utilizing the pre-trained Xception model. This enables us to benefit from the knowledge and features learned from a large dataset and adapt it to our specific image regression task.
- To enhance image quality and model performance, preprocessing methods such as CLAHE and preprocess_input were employed. CLAHE improves contrast by locally adjusting image intensities. The preprocess_input function standardizes pixel values to align with the model's expectations.
- We used the AdamW optimizer, which improves upon the traditional Adam optimizer by including weight decay directly in the optimization process. AdamW's adaptive learning rates and decay mechanism make it especially well-suited for complex deep learning models, enhancing stability during training.

III. Processing Pipeline

This section mainly focuses on the processing pipeline developed for BAA from hand radiographs using two distinct deep learning architectures: Xception and ConvMixer. The pipeline integrates image preprocessing, data augmentation, model architecture design, model training and model evaluation stages to improve prediction accuracy and robustness. Both models trained on the training set to learn the patterns in the data which are compared later.

A. ConvMixer

The core architecture, built around a ConvMixer-256/8 model (see Fig. 1), processes x-ray images through a series of convolutional and depthwise convolutional layers. The model uses a ”conv stem” block to preprocess and patchify images, followed by multiple ConvMixer blocks that progressively capture image features through both spatial mixing and feature mixing. These blocks use residual connections and activation functions to effectively extract features from the input images, even in complex cases. Additionally, the model incorporates gender data as a separate input at the end of the GlobalAveragePooling2D layer, utilizing a custom generator to preprocess and encode gender information, enhancing predictive accuracy. Gender information is introduced as a one-dimensional input, processed by a dense layer, and concatenated with the main feature stream before the final regression layers. This integration enables the model to learn both image-based and demographic correlations for accurate bone age prediction.

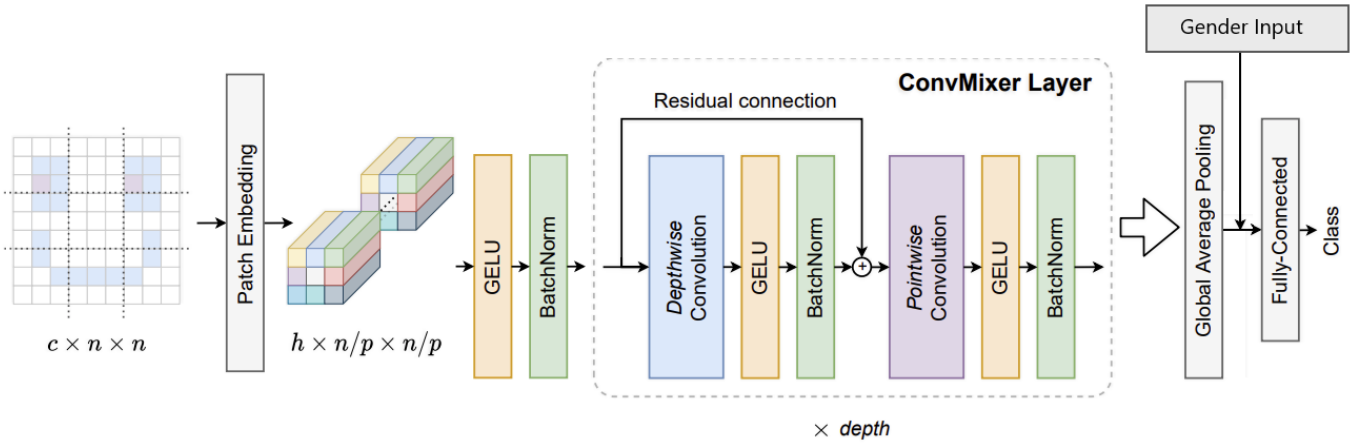

Fig. 1: ConvMixer Model Architecture

B. Xception

The pipeline integrates the Xception model's image processing capabilities, which pre-trains on a large image dataset and extracts high-level features through its convolutional layers, each refining the image representations. The Xception model architecture (see Fig. 2) is built around depth-wise seperable convolutions, which enable it to capture intricate patterns within the radiographs. Using pretrained weights from ImageNet helps the model leverage prior knowledge for efficient feature extraction. The output from the Xception model is globally averaged and processed through dense layers to enhance feature extraction further. Similar to ConvMixer model, gender information is added as an auxiliary input to enhance the model's accuracy by adding demographic context, further improving prediction accuracy. This gender encoding is passed through a dense layer to be transformed into a suitable feature space before merging with the main image stream. After merging, the combined features pass through dense layers and ultimately output a single regression prediction for bone age. By using both visual and demographic data, this model pipeline leverages image features while accounting for gender-based variations in bone age, providing a robust prediction framework.

IV. Signals and Features

Both models were implemented using the TensorFlow framework (version 2.11.0) and trained on Tesla P100-PCIE-16GB (Kaggle online compiler), which provides 16 GB of memory. It operates with driver version 550.90.07 and CUDA version 12.4. The system also has 33.7 gigabytes of available high-RAM runtime, providing ample memory capacity for working with large datasets and complex model architectures. For the final versions, both models individually took 7-9 hours to train but the experiments and comparisons to determine the hyperparameters took over two months.

The dataset of RSNA challenge consists of X-ray scans of hands (Fig. 3) for people from ages between 0 and 19. It is composed of 14236 images obtained from 2 different

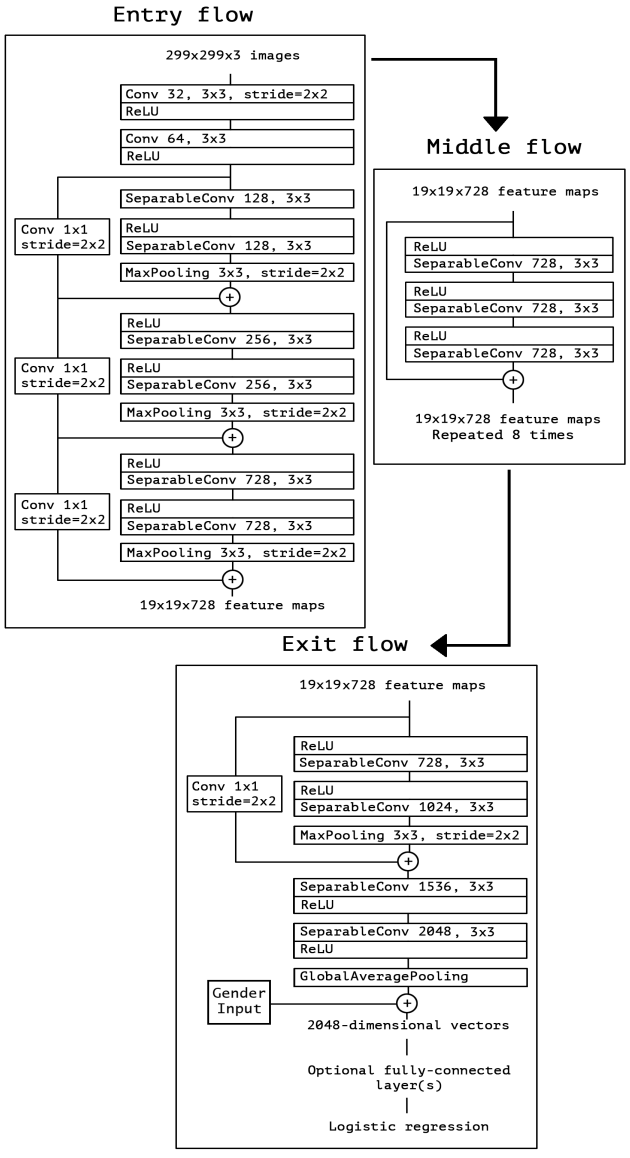

Fig. 2: Xception Model Architecture

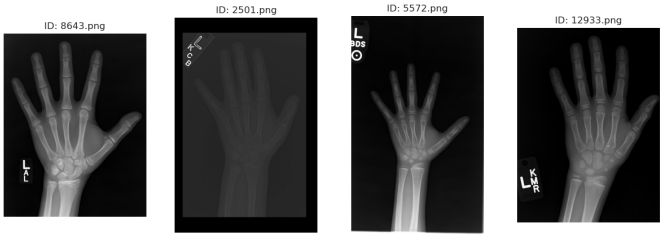

Fig. 3: Random Samples from RSNA Dataset

sources, and it is divided into 3 subsets; training, validation and test sets. Training set contains 12611 images, validation set contains 1425 images and test set contains 200 images. Each image in the training and validation set labeled by its owner's age and gender information and saved into csv file. Images in the test set has only one label information, which is gender. Distribution of the datasets is shown in Fig. 4. The age of the samples is mostly around 144 months (12 years), while only a few samples fall near the two edges (see Fig. 5). The distribution of genders is almost balanced for training and validation sets, while it perfectly balances for test set.

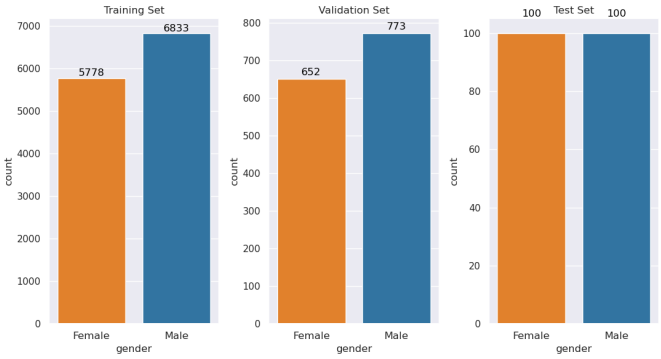

Fig. 4: Dataset Distribution

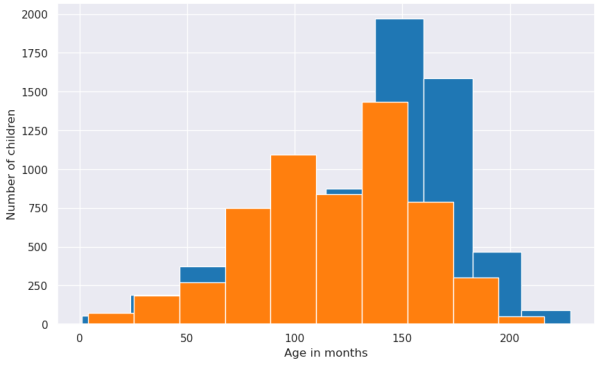

Fig. 5: Dataset Visualization

The Mean Absolute Error (MAE) value between ground-truth age and predicted age is reported to evaluate model performance, with the unit of measurement in months. Training uses Mean Squared Error (MSE) as the loss function for this regression task, while both MAE in months and R^2 (coefficient of determination) are tracked as primary performance metrics to assess accuracy and the proportion of variance explained by the model.

Before starting to work on either model, we needed to organize the dataset for training phase. The gender labeling in the dataset was based on true/false valuation of male sex. First, we transformed it to 'male' and 'female' labeling and then we applied mapping one more time to transform the labels into numerical arrays, with 0 represents 'female' and 1 represents 'male' for training phase. We also adjust the column names for all sets to perform operations on them easily.

After dataset adjustments, pre-processing is performed as an initial step to prepare the medical images for further processing. In the ConvMixer model, Contrast Limited Adaptive Histogram Equalization (CLAHE) technique is used to enhance contrast. This technique helps highlight bone structures in the grayscale radiographs, which are then returned to RGB format for further processing. The pre-processing step include resizing all the images to the same dimension of 256x256. The batch size is 8 because the model is too large for bigger batch size. In the Xception model, we used the $preprocess\_input$ function from Keras for image pre-processing. This function prepares input images by scaling pixel values to a range suitable for Xception, aligning with weights pre-trained on ImageNet. We applied the same data augmentation techniques for this model as well. After numerous experiments, the optimum batch size is found as 16. For this model, the image size is resized to the dimension of 500x500.

Training, validation, and test sets are managed using TensorFlow ImageDataGenerator, which applies the CLAHE-enhanced pre-processing (for ConvMixer model) or preprocess_input (for Xception model) and data augmentation steps to the input images. These additional data augmentation steps applied only on training set and includes random transformations like rotation, zooming, horizontal flip, width and height shifts, and shearing. These transformations diversify the training dataset, which aids in the model's capacity to recognize bones in varied orientations and scale.

V. Learning Framework

In this section, we describe the learning strategies/algorithms that we conceived and used to solve this problem at stake. The first model we implemented is ConvMixer. This model works as it operates directly on patches as input, separates the mixing of spatial and channel dimensions and maintains equal size and resolution throughout the network. It uses simple standard convolutions to achieve mixing steps. Our second model, the Xception, can be seen as an 'extreme' version of Inception module which the Inception modules have been replaced with depthwise separable convolutions. Xception architecture is a linear stack of depthwise separable convolution layers with residual connections. This model significantly outperforms the InceptionV3 module on a very large dataset.

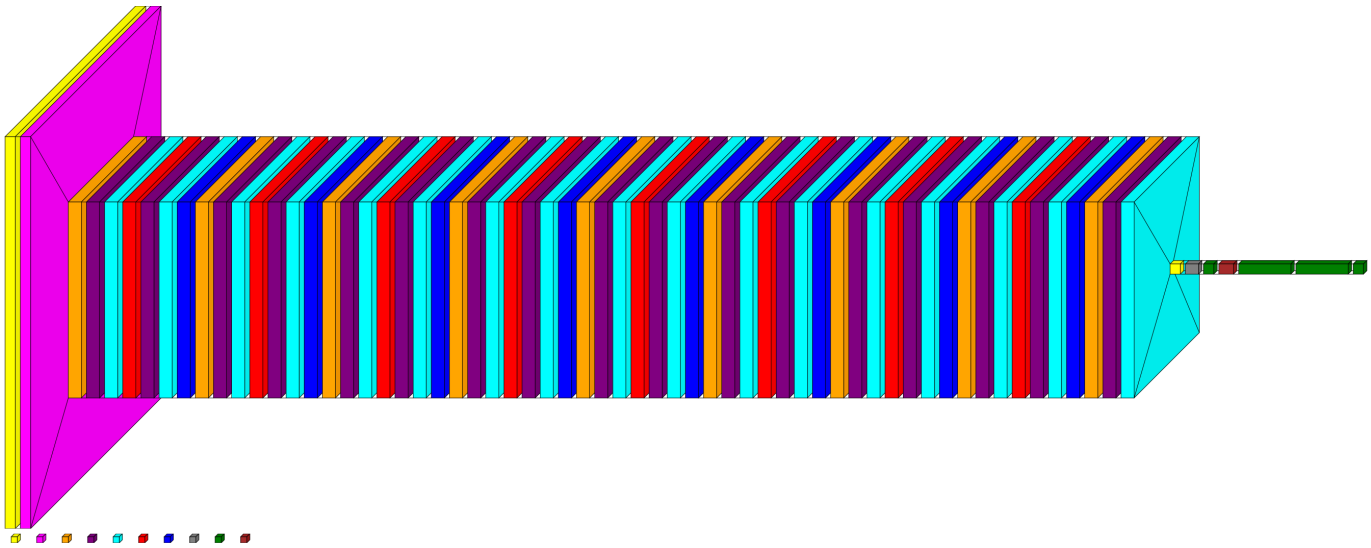

Fig. 6: Layer Architecture of ConvMixer Model

A. *ConvMixer*

This model follows a structured processing pipeline (see Fig. 6), beginning with an input layer that takes images of shape (256, 256, 3) (colored with yellow). The images undergo a rescaling operation, normalizing pixel values to be between [0,1] for consistent input (colored with magenta). In the following, a series of Conv2D (colored with light orange) and DepthwiseConv2D (colored with red) layers with 'GeLU' activation functions (colored with purple) are applied, capturing and processing spatial hierarchies in the images. Each Conv2D layer (orange) with shape of (128, 128) and 256 filters is followed by 'GeLU' activation function and batch normalization (colored with cyan) to improve model convergence and stability. After this step, DepthwiseConv2D layer is applied with output shape of (128, 128) and 256 filters enhancing efficiency by separating spatial and depth dimensions. Same activation function and batch normalization steps are applied in here as well. After that, the residual connections coming from both batch normalization processes are implemented together through add layer (colored with blue), allow gradients to propagate smoothly across layers, promoting robust learning. This convolution block repeats 7 more times, creating the middle flow of our model. After these blocks, a GlobalAveragePooling2D layer with shape of 256 (colored with gray) applied to reduce the spatial dimensions, resulting in a compact feature representation. At this point, we introduce the gender input layer (colored with yellow) with dimension of 1 for additional gender feature. This layer is processed by a dense layer of 32 units (colored with green) and 'ReLU' activation function is used. The outputs of the GlobalAveragePooling2D layer and Gender dense layer are concatenated (colored with brown), creating a combined feature representation for the subsequent fully connected layers. The concatenated features are passed through two dense layers (colored with green), each with 1024 units, capturing intricate patterns and relationships within the data. Finally, the model output is a single dense layer (colored with green) and used the 'linear' activation function to predict the bone age as a regression output. The model has 1,946,945 parameters, 1,938,241 of which are trainable.

As an optimizer, we used AdamW to improve training stability and convergence speed. AdamW is a variant of the Adam optimizer, with an added weight decay term that helps in regularizing the model by preventing overfitting. This is particularly valuable in our pipeline, as weight decay explicitly penalizes large weights, which helps the model focus on the most essential features within the radiographs without overemphasizing certain areas. After evaluating the performance of different combinations, we were determined to use the learning rate of $1*10^{-3}$ together with weight decay of $1 * 10^{-5}$. Additional optimizations, such as early stopping and learning rate scheduling, are employed to improve model convergence and prevent overfitting. The ReduceLROnPlateau callback dynamically adjusts the learning rate to promote better convergence. It monitors the validation MAE (mean absolute error) in months, and if there is no improvement after three epochs (as specified by patience=3), it reduces the learning rate by a factor of 0.1. This approach helps the model make finer updates by decreasing step size when improvements stall. The minimum learning rate is set to $1 * 10^{-5}$. The EarlyStopping callback is designed to halt training once further improvements in validation MAE cease. If the validation MAE does not improve for seven epochs (patience=7), training stops, and the model's weights revert to those of the best-performing epoch. This callback prevents overfitting by stopping training early if no progress is detected for an extended period. The model was trained for 50 epochs, and the best weights were loaded from the model checkpoint. The performance of the model was assessed through metrics like MAE, MSE, R^2 score, and the loss curve over training

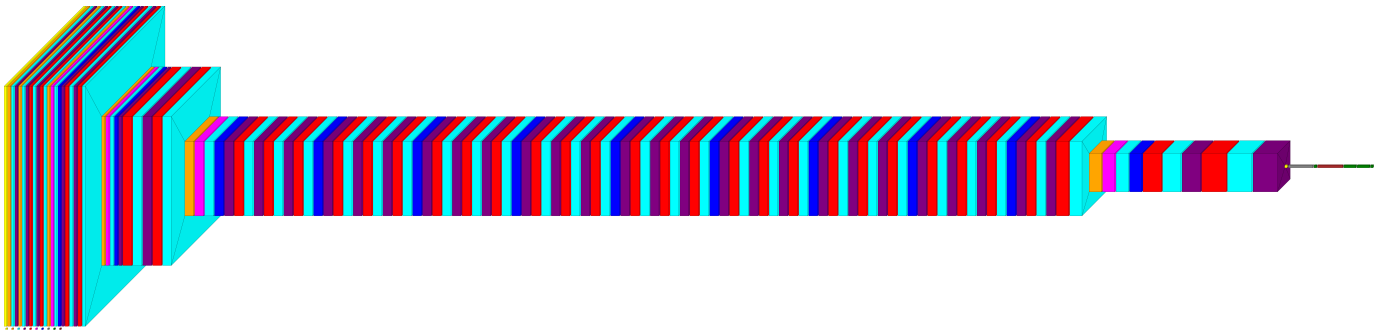

Fig. 7: Layer Architecture of Xception Model

and validation sets. After training, the model is tested on a separate test dataset to evaluate its predictive performance on new, unseen data, an essential step for confirming the model's reliability.

B. Xception

The model (see Fig. 7) starts with an input layer designed to handle images of size (500, 500, 3), which are then processed through multiple convolutional and batch normalization layers. The initial convolutional layer, block1_conv1, has 32 filters, a kernel size of 5x5, a stride of 1x1, and uses ReLU activation. This layer captures basic patterns in the data, followed by a batch normalization layer to maintain stable activations. An activation layer is then applied, preparing data for the next layer, block1_conv2, which has 64 filters, increasing the complexity of learned features. Each Conv2D layer is followed by a batch normalization layer and an activation layer to streamline the training process. The model proceeds with block2 and block3 layers, introducing separable convolutions to reduce computational costs while capturing intricate patterns. Each of these blocks also includes max pooling layers, which downsample spatial dimensions, and additional batch normalization layers to stabilize output distributions. As processing advances to deeper layers, block4 through block12 further refine the feature maps using separable convolutions, max-pooling, and batch normalization. Each block systematically reduces spatial dimensions while increasing the depth of learned features, allowing the model to capture more complex patterns. In block13, separable convolution layers with 1024 filters are applied, with batch normalization and max-pooling to manage feature map sizes and provide regularization. The block14 layers further upscale filter size to 1536 and then 2048, learning high-level abstractions with batch normalization and activation layers to ensure smooth gradients. A global average pooling layer is applied to the final convolutional output, transforming it into a 2048-dimensional feature vector. In the pipeline's final stage, gender data is processed through a dense layer with 32 units before being concatenated with the image features. This concatenated vector is then passed through dense layers with 1024 units each, ending with a single-unit dense layer representing the bone age output. This configuration yields 24,043,113 total parameters, where 23,988,585 are trainable, providing the capacity to capture a wide range of patterns and relationships in the dataset. The final dense layer uses a single output neuron and 'linear' activation function to predict the continuous target value for bone age.

As an optimizer, we also used AdamW to improve training stability and convergence speed. After evaluating the performance of different combinations, we were determined to use the learning rate of $1 * 10^{-4}$ together with weight decay of $1 * 10^{-5}$. Similar to ConvMixer, early stopping and learning rate reduction strategies are applied, optimizing the training process to avoid underfitting and overfitting. The ReduceLROnPlateu callback dynamically adjusts the learning rate by monitoring the validation MAE in months, and if there is no improvement after three epocs, it reduces the learning rate by a factor of 0.1. The minimum learning rate is set to $1 * 10^{-6}$. The EarlyStopping callback monitors the validation MAE regularly, and if the validation MAE does not improve for eleven epochs, training stops and the model's weights revert to those of the best-performing epoch. Also similar to previous model, this model's predictive capabilities are tested on a separate dataset after training to evaluate its effectiveness on new data, confirming its potential for real-world clinical applications.

VI. Results

To evaluate the effectiveness of our models in predicting pediatric bone age, we report four main regression performance metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), and Loss. Table 1 provides a comparison between the ConvMixer and Xception models. These metrics allow us to measure each model's accuracy and error in predicting bone age

Model	MSE	MAE	R^2	Loss
ConvMixer	142	9.01	0.9216	0.0842
Xception	100	7.63	0.9411	0.0606

TABLE 1: Model Performance Summary

From Table 1, it is evident that the Xception model outperforms ConvMixer in terms of both MSE and MAE, achieving a lower error and a higher R^2 score, suggesting a closer fit

Optimizer	Learning Rate	Batch Size	Dataset Splitting	Image Size	Framework	Pre-processing	Result
AdamW	$1 * 10^{-4}$	8	Train (Val) + Test	224x224	Convmixer+alt	preprocess_input	10,56
AdamW	$1 * 10^{-3}$	8	Train (Val + Test)	224x224	Convmixer+alt	preprocess_input	10,37
AdamW	$1 * 10^{-3}$	8	Train + Val + Test	256x256	Convmixer+alt	CLAHE	9,78
AdamW	$1 * 10^{-3}$	8	Train + Val + Test	256x256	Convmixer+winner	CLAHE	9,01

TABLE 2: Different hyper-parameter results for ConvMixer model

to the true bone age values. Additionally, the lower Loss value for Xception further supports its higher performance over ConvMixer. Figure 8 and Figure 9 present the training and validation MAE and loss graphs for both the ConvMixer and Xception models, respectively, offering insight into model convergence and potential overfitting, and allowing for a clearer understanding of each model's training progression. The results show that the Xception model converges faster and maintains lower training and validation loss throughout, indicating better generalization. ConvMixer, while slower to converge, demonstrates steady progress with no significant overfitting.

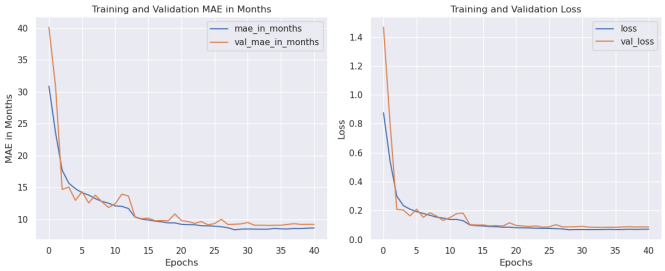

Fig. 8: Training and Validation Results for ConvMixer

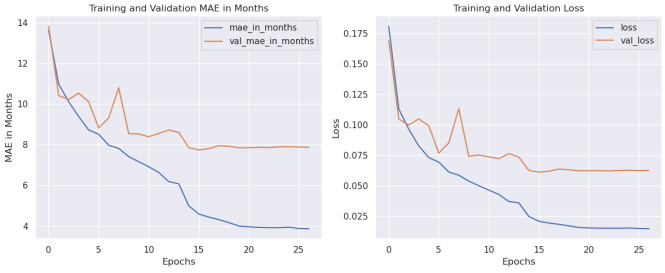

Fig. 9: Training and Validation Results for Xception

To explore the impact of model architecture and hyper-parameters on performance, we varied the learning rate, batch size, dataset splitting, image size, model frameworks, and optimizer configurations across both models. Tables 2 and 3 show the MAE as result of these parameters, respectively, allowing for a clear visualization of the effects on model performance. In the framework section, 'alt' represents alternative framework (which is model+GAP2D+Densed[2]+Flatten+Dropout+GenderInput +Densed[2]), while 'winner' represents 2017 RSNA challange winner model framework (which is model+GAP2D+GenderDensed+Densed[2]). The results indicate that both models consistently achieve lower error rates across most parameter combinations, with minimal overfitting. Both models also show slightly higher variance in error rates across parameter settings, highlighting a sensitivity to hyperparameters that may require careful tuning in practical applications.

Figure 10 displays the total prediction distribution graphs for the ConvMixer and Xception models, respectively. These graphs illustrate the distribution of predicted bone ages compared to actual values, which can be realized that the Xception model's predictions show tighter clustering around true values, suggesting more consistent accuracy. ConvMixer's distribution is broader compared to Xception, this indicates slightly more variation in its predictions, which aligns with its higher MAE and MSE values.

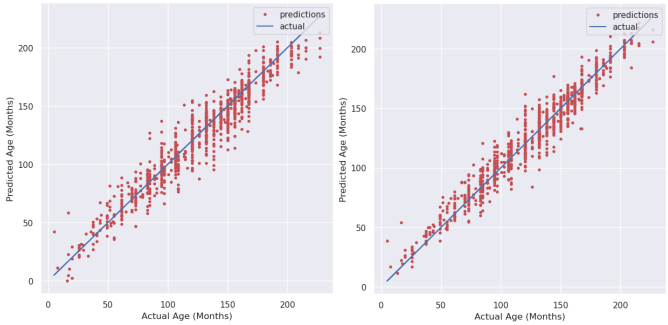

Fig. 10: Total Prediction Distributions for ConvMixer and Xception models, respectively

A critical aspect of our analysis is the trade-off between model accuracy and computational complexity, which is essential for deploying bone age assessment models in real-world scenarios with varying resource constraints. We evaluated both models in terms of training and inference time, memory usage, and overall computational demands. Table 4 highlights that, Xception model delivers higher predictive accuracy and faster inference time. However, this higher accuracy and efficiency come with a cost, which is higher memory demand and longer training time. This could impact scalability in memory-constrained environments. Conversely, ConvMixer is slightly less accurate with a shorter training time and lower memory requirements. Its slightly slower inference time is a minor trade-off because reduced memory usage and shorter training time makes it a viable option for scenarios with limited resources.

Optimizer	Learning Rate	Batch Size	Dataset Splitting	Image Size	Framework	Pre-processing	Result
Adam	$1 * 10^{-5}$	8	Train (Val) + Test	224x224	Xception	no	11,26
Adam	$1 * 10^{-4}$	32	Train (Val) + Test	256x256	Xception+alt	no	10,41
Adam	$1 * 10^{-4}$	16	Train (Val) + Test	256x256	Xception+alt	no	10,28
AdamW	$1 * 10^{-4}$	16	Train (Val) + Test	256x256	Xception+alt	no	10,15
AdamW	$1 * 10^{-4}$	16	Train (Val) + Test	256x256	Xception+alt	preprocess_input	9,28
AdamW	$1 * 10^{-4}$	16	Train + Val + Test	256x256	Xception+alt	preprocess_input	9,01
AdamW	$1 * 10^{-4}$	16	Train + Val + Test	500x500	Xception+winner	preprocess_input	7,63

TABLE 3: Different hyper-parameter results for Xception model

Model	Training Time (s)	Inference Time (ms)	Memory Usage (MB)	R^2
ConvMixer	25724	21.2	5783	0.9216
Xception	30905	18.6	9762	0.9411

TABLE 4: Complexity-Performance Trade-off

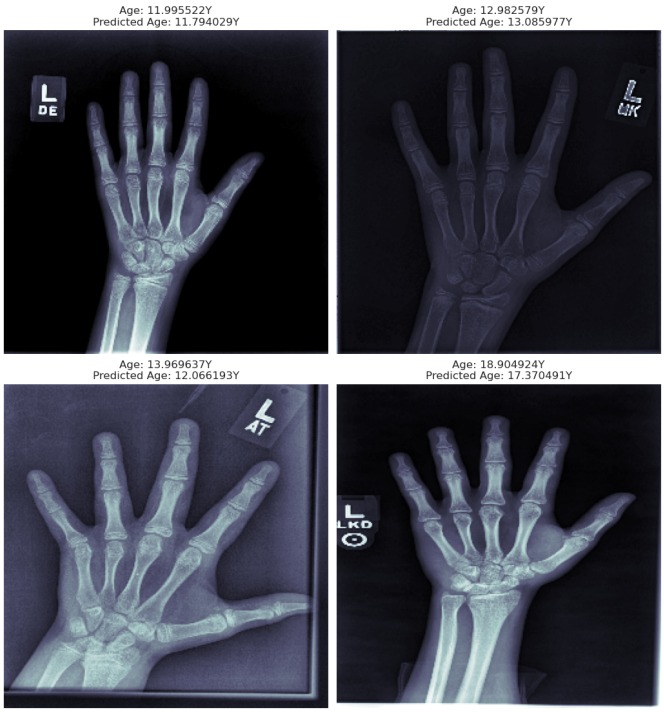

Fig. 11: Prediction Results for ConvMixer Model

In summary, the Xception model is ideal for high-performance applications requiring high predictive accuracy and rapid inference but demands more computational resources, including memory and training time. On the other hand, the ConvMixer model, with its lower memory usage and faster training time, offers a more resource-efficient alternative while maintaining competitive accuracy. This makes it ideal alternative for scenarios where computational efficiency is vital. These results show the trade-offs between performance and computational demands, underlining the importance of leveling model selection with the operational requirements of pediatric bone age assessment applications.

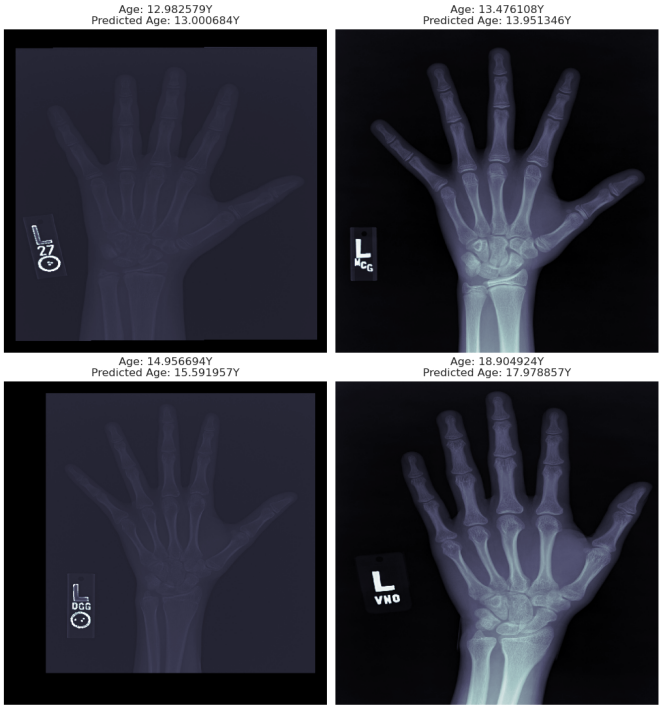

Fig. 12: Prediction Results for Xception Model

VII. Concluding Remarks

In this study, we developed and evaluated deep learning models to predict bone age from hand radiographs using the RSNA Pediatric Bone Age dataset. We implemented and compared two architectures -Xception and ConvMixer- each trained with data pre-processing techniques such as CLAHE or preprocess_input and various data augmentation strategies. We have developed an algorithm (custom generator) to concatenate the gender input with the hand x-ray images and applied it to mixed data, which includes categorical, numerical and visual data. Results indicate that both models were able to predict bone age with a reasonable degree of accuracy (less than 10 months), though the Xception model generally showed better performance in terms of MAE and model consistency across different hyper-parameter settings. The significance of our findings lies in the potential application of these models within clinical settings. By accurately predicting bone age, our models demonstrate promise in supplementing radiologists'

assessments, potentially leading to faster and more objective evaluations in pediatric endocrinology and orthopedics.

While our models achieved promising results, several areas for future work remain. To advance this study, future research could involve implementing Computer Vision algorithms on data pre-processing part to focus specific areas (Region of Interest-based models) where the bone age is predicted. Due to computational constraints, specifically working with an NVIDIA GeForce GTX 1050 or Kaggle online compiler (Tesla P100), we were unable to experiment with larger models or more advanced architectures. Using new generation NVIDIA RTX 4000 series (especially 4080 or 4090) could potentially improve prediction accuracy. Exploring alternative architectures and hyper-parameter tuning on more powerful hardware would also allow for a deeper examination of model potential and robustness. Additionally, while our pre-processing methods proved effective, additional features such as patient ethnicity or chronological age could enhance predictions by providing context beyond image data alone.

Throughout this project, we have learned extensively about the intricacies of image regression tasks and the nuances of model evaluation in a clinical context. We also learned to work with mixed data types. Working with deep learning in medical imaging presented unique challenges, particularly in balancing model complexity and performance with computational limitations. Despite these challenges, this experience underscored the significant impact that machine learning can have on healthcare, pushing forward the potential for automation in medical diagnostics.

REFERENCES

[1] A. Mughal, N. Hassan, and A. Ahmed, "The applicability of the greulich pyle atlas for bone age assessment in primary school-going children of karachi, pakistan," *Pakistan journal of medical sciences*, vol. 30, pp. 409–11, 03 2014.

[2] "Rsna pediatric bone age challenge (2017)." https://www.rsna.org/rsnai/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017.

[3] S. Halabi, L. Prevedello, J. Kalpathy-Cramer, A. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. Pereira, R. Sousa, N. Abdala, F. Kitamura, H. Thodberg, L. Chen, G. Shih, K. Andriole, M. Kohli, B. Erickson, and A. Flanders, "The rsna pediatric bone age machine learning challenge," *Radiology*, vol. 290, p. 180736, 11 2018.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," pp. 1800–1807, 07 2017.

[5] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[6] A. Trockman and J. Z. Kolter, "Patches are all you need?," 2022.

[7] S. Pizer, R. Johnston, J. Ericksen, B. Yankaskas, and K. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," pp. 337–345, 1990.

[8] W. Greulich and S. Pyle, *Radiographic Atlas of Skeletal Development of the Hand and Wrist*. Stanford University Press, 1959.

[9] J. Tanner, R. Whitehouse, N. Cameron, W. Marshall, M. Healy, and H. Goldstein, *Assessment of Skeletal Maturity and Prediction of Adult Height: TW2 Method*. Academic Press Inc, 1983.

[10] J. Tanner, N. Cameron, M. Healy, and H. Goldstein, *Assessment of Skeletal Maturity and Prediction of Adult Height: TW3 Method*. W B Saunders Co, 2001.

[11] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The bonexpert method for automated determination of skeletal maturity," *IEEE Transactions on Medical Imaging*, vol. 28, no. 1, pp. 52–66, 2009.

[12] M. Harmsen, B. Fischer, H. Schramm, T. Seidl, and T. M. Deserno, "Support vector machine classification based on correlation prototypes applied to bone age assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 190–197, 2013.

[13] K. Somkantha, N. Theera-Umpon, and S. Auephanwiriyakul, "Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression," *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, vol. 24, pp. 1044–58, 02 2011.

[14] D. Larson, M. Lungren, S. Halabi, N. Stence, and C. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, p. 170236, 11 2017.

[15] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in x-ray images," *Medical Image Analysis*, vol. 36, pp. 41–51, 2017.

[16] C. Gonzalez, M. Escobar, L. Daza, F. Torres, G. Triana, and P. Arbelaez, "Simba: Specific identity markers for bone age assessment," 2020.