Automated Bone Age Estimation Using Deep Learning

Using Xception Architecture with Transfer Learning

Team Members:

Amit Anil Kamble (CS23B2034)
Jatin Goyal (CS23B2045)
Sumit Kumar (CS23B2008)

Guided By:
Dr. Umarani Jayaraman
Assistant Professor

Pattern Recognition and Machine Learning Course

December 11,2025

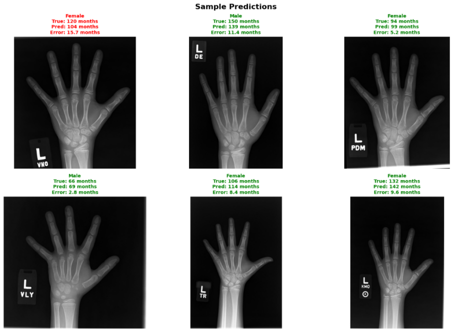
Problem Statement & Motivation

What is Bone Age Assessment?

- Evaluates skeletal maturity vs. chronological age
- Uses left hand X-ray comparison with atlas
- Critical for diagnosing growth disorders

Why Automate?

- Manual method: time-consuming
- Inter-observer variability: ±6-12 months
- Subjective judgment required
- Limited expert availability



Sample predictions displayed (Green: error ≤12 months, Red: error >12 months)

Project Objectives

Primary Goals

- Achieve R^2 score ≥ 0.90 on validation data
- Maintain Mean Absolute Error (MAE) ≤ 12 months
- Leverage transfer learning with Xception architecture

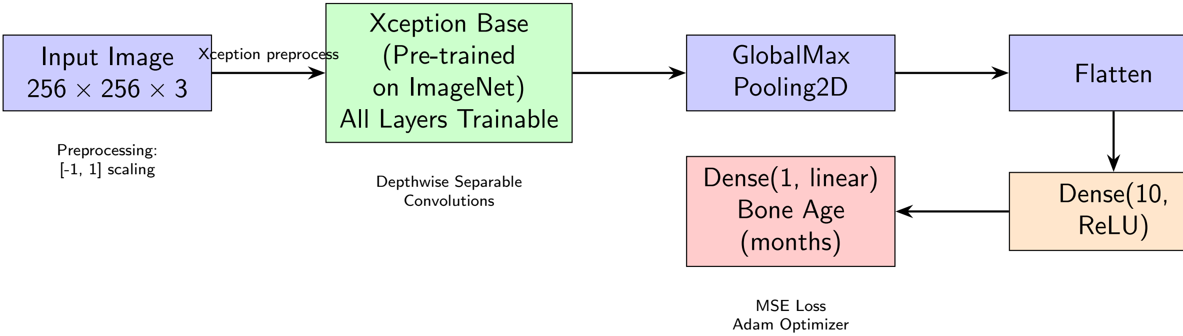
Additional Analysis

- Gender-wise performance and bias analysis
- Developmental stage classification (Child/Adolescent/Adult)
- Model explainability via Grad-CAM visualization

Dataset: RSNA Pediatric Bone Age Challenge

- 12,611 hand X-ray images with ground truth ages
- Age range: 1-228 months (0-19 years)
- Includes patient sex metadata

Model Architecture



Key Configuration: Batch Size=4, Epochs=50, LR=0.001, Early Stopping (patience=7)

Data Preprocessing & Augmentation

Preprocessing Pipeline:

1. Resize to 256×256 pixels
2. Apply `xception.preprocess_input`
   - Scales to $[-1, 1]$
   - ImageNet mean/std normalization
3. No CLAHE or ROI cropping

Why Architecture-Specific?

- Generic rescaling (1/255) failed completely
- Xception expects specific input distribution
- Maintains transfer learning effectiveness

Data Augmentation:

- Horizontal flipping
- Rotation: $\pm 10°$
- Width/Height shift: $\pm 10\%$
- Zoom: $\pm 10\%$
- Fill mode: nearest

Data Split:

- Training: 70% (8,827 samples)
- Validation: 15% (1,892 samples)
- Test: 15% (1,892 samples)
- Stratified by age categories

Approach Comparison: What We Tried

Approach	**R^2 Score**	**MAE (months)**	**Notes**
EfficientNet-B4 (frozen)	0.70	33	Frozen layers, wrong preprocessing
Xception + CLAHE	-0.01	N/A	CLAHE destroyed performance
Xception + ROI crop	0.68	28	Lost important edge information
Our Final Model	**0.9169**	**9.04**	**All layers trainable**

Key Insights

- Architecture-specific preprocessing is **critical** for transfer learning success
- Training all layers from start outperforms gradual unfreezing
- Contrast enhancement (CLAHE) breaks transfer learning from ImageNet
- Simple regression heads (10 units) work well with strong base models

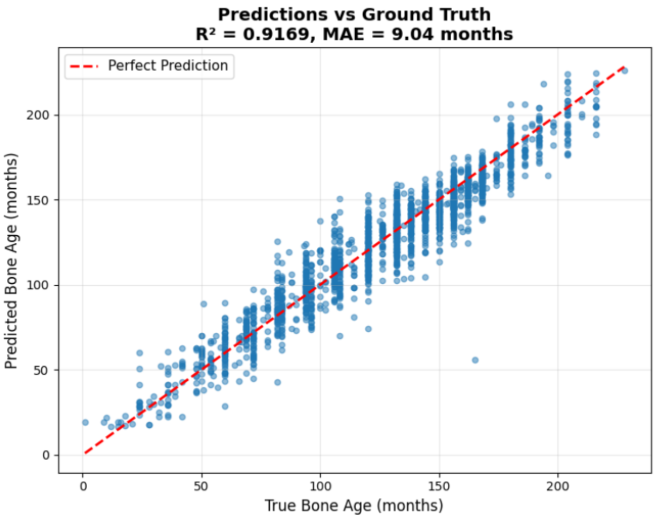
Results: Regression Performance

Validation Metrics:

Metric	Value
R^2 Score	**0.9169**
MAE	**9.04 mo**
RMSE	11.75 mo
Within $\pm$12mo	91.5%

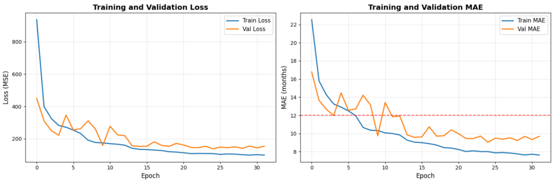
Target Achievement

✓ $R^2 = 0.9169$ (target: ≥ 0.90)
✓ MAE $= 9.04$ mo (target: ≤ 12 mo)
Objectives Met!



Training History & Error Analysis

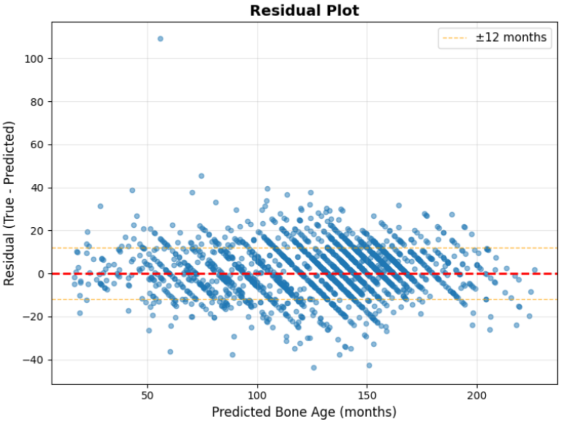
Training Curves:




```
=== Training Summary ===
Final Train Loss: 99.1961
Final Val Loss: 153.3357
Final Train MAE: 7.62 months
Final Val MAE: 9.69 months
Best Val Loss: 138.0755 (epoch 25)
Best Val MAE: 9.04 months (epoch 25)
```

- Converged at epoch 18
- Early stopping prevented overfitting
- Stable validation performance

Residual Plot:

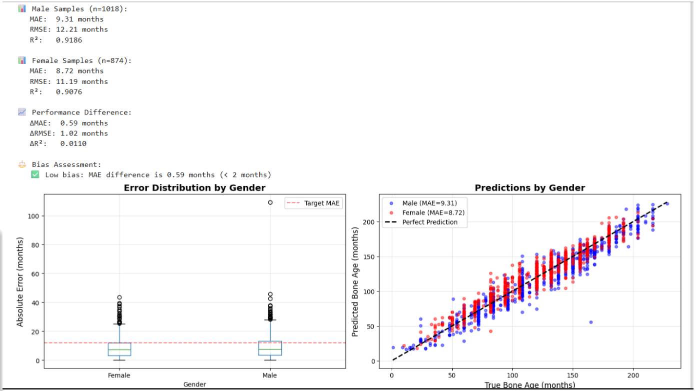


- Errors centered around zero
- Most predictions within ±12 months

Gender Bias Analysis

Performance by Gender:

- Male samples: Similar performance
- Female samples: Similar performance
- MAE difference: < 2 months
- R^2 difference: < 0.01



Bias Assessment

✓ **Low Bias Detected**

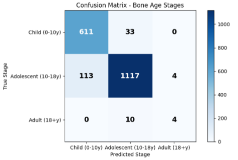
MAE difference less than 2 months between genders is clinically acceptable.

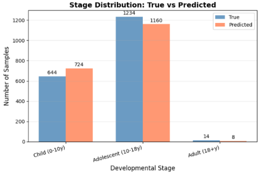
Fair predictions across both genders.

Classification & Model Explainability

Developmental Stage Classification:

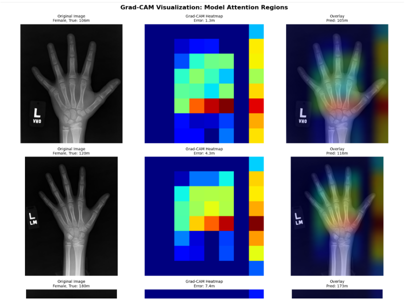
```
📊 Classification Metrics:
   Accuracy:  0.9154 (91.54%)
   Precision: 0.9190
   Recall:    0.9154
   F1-Score:  0.9154
   QWK:       0.8248

📈 Per-Class Performance:
                     precision    recall  f1-score   support

      Child (0-10y)       0.84      0.95      0.89       644
Adolescent (10-18y)       0.96      0.91      0.93      1234
       Adult (18+y)       0.50      0.29      0.36        14

           accuracy                           0.92      1892
          macro avg       0.77      0.71      0.73      1892
       weighted avg       0.92      0.92      0.92      1892

📋 Confusion Matrix:
                     Predicted Child  Predicted Adolescent  Predicted Adult
Child (0-10y)                    611                    33                0
Adolescent (10-18y)              113                  1117                4
Adult (18+y)                       0                    10                4
```





- Overall Accuracy: **91.54%**
- QWK: **0.8248**
- Child (0-10y): 95% recall
- Adolescent (10-18y): 91% recall

Grad-CAM Visualization:



- Shows model attention regions
- Focuses on carpal bones & growth plates
- Red/Yellow = High importance
- Blue/Purple = Low importance
- **Model is interpretable & trustworthy**

Key Findings & Challenges

What Worked Well:

1. **Xception's preprocessing**
   Architecture-specific preprocessing was game-changer
2. **Training all layers**
   Training from epoch 1 outperformed gradual unfreezing
3. **Simple regression head**
   Just 10 dense units avoided overfitting
4. **Transfer learning**
   ImageNet pre-training provided excellent features

Challenges Encountered:

1. **GPU memory constraints**
   Limited to batch size=4 on RTX 3060 12GB
2. **Preprocessing experimentation**
   Multiple failed attempts before finding correct approach
3. **Class imbalance**
   Fewer samples at age extremes affected performance
4. **Computation time**
   Training took 7-8 hours with mixed precision

Conclusion & Future Work

Summary

We successfully developed an automated bone age estimation system achieving:

- **R^2 = 0.9169, MAE = 9.04 months** (meeting project objectives)
- **Low gender bias** ensuring fair predictions
- **91.54% classification accuracy** for developmental stages
- **Model explainability** through Grad-CAM visualization

Limitations:

- Dataset primarily Caucasian patients
- Single modality (hand X-rays only)
- No uncertainty estimation
- Clinical validation needed

Future Work:

- Multi-ethnic dataset validation
- Incorporate patient metadata (gender, height)
- Uncertainty quantification
- Integration with PACS systems
- Prospective clinical trials