Bone Age Prediction from Hand Radiographs: A Comparative Study of Regression and Classification

S. Srihitha[1], R. K. Larika[2], and S. Harshini[3]

[1,2,3]Department of Computer Science, IIITDM Kancheepuram
Student IDs: CS23B1014, CS23B1028, CS23B1050

Course Project for: Pattern Recognition and Machine Learning

Abstract

This project develops automated bone age estimation from pediatric hand radiographs using machine learning and deep learning methodologies. Two tasks are addressed: (1) **regression** - predicting continuous bone age in years, and (2) **classification** - categorizing skeletal maturity into five discrete developmental stages. For regression, we implement hybrid pipelines using ResNet-50 features with classical ML models and end-to-end fine-tuned Convolutional Neural Networks (CNNs). The optimized end-to-end CNN achieves a MAE of 0.7083 years (RMSE: 0.9357, R^2: 0.9246), substantially outperforming Ridge Regression (best hybrid model, MAE: 1.0767 years). For classification, a ResNet-50 CNN classifier achieves a Quadratic Weighted Kappa (QWK) of 0.8888 with 78.17% accuracy, markedly superior to the HOG + XGBoost baseline (QWK: 0.5016). Results consistently demonstrate the superiority of deep learning approaches for robust and clinically useful skeletal maturity assessment, reducing error below typical inter-observer variability.

Keywords: Bone Age Assessment, Deep Learning, ResNet-50, Regression, Classification, Skeletal Maturity, Radiograph.

1 Introduction

Bone age assessment is critical for diagnosing growth disorders and developmental delays in pediatrics. Traditional manual evaluation using Greulich-Pyle or Tanner-Whitehouse methods is subjective and time-consuming, with inter-observer variability often exceeding 12 months. This project develops automated systems for two complementary tasks: (1) **regression** - predicting continuous bone age in years, and (2) **classification** - categorizing skeletal maturity into five developmental stages. We compare hybrid CNN feature extraction with classical ML against end-to-end deep learning approaches to identify the most effective methodology.

2 Dataset and Preprocessing

2.1 Dataset

The RSNA Pediatric Bone Age Dataset [1] contains 12,611 hand radiographs with bone age labels (in months) and sex information.

Table 1: Data Splitting

Split	Percentage	Image Count
Training	70%	8,827
Validation	15%	1,892
Test	15%	1,892

2.2 Data Splitting

Stratified sampling by sex ensured balanced representation, as detailed in Table 1.

2.3 Preprocessing Pipeline

All images were processed through: 1) Grayscale conversion, 2) Resizing to 256×256, 3) Tensor conversion, and 4) Normalization ($\mu = 0.5, \sigma = 0.5$). Training augmentation included random horizontal flip, rotation ($\pm 20°$), affine transforms, and contrast/brightness jitter.

3 Methodology

3.1 Part A: Regression

The goal is to predict continuous bone age (in years). The input to all models is the image feature vector concatenated with the 1-dimensional sex encoding.

3.1.1 Strategy 1: Hybrid CNN + Classical ML

A pre-trained ResNet-50 [2] was used as a fixed feature extractor. The first convolutional layer was adapted for grayscale input, and the final fully-connected (FC) layer was removed. The 2048-dimensional features from global average pooling were concatenated with the sex encoding, resulting in a 2049-dimensional input vector. Six classical models were trained on these features: Ridge Regression, SVR (RBF), Random Forest, XGBoost, HistGradient Boosting, and a Stacking Regressor.

3.1.2 Strategy 2: End-to-End CNN (Baseline)

The ResNet-50 backbone was fine-tuned for regression. All layers were trainable, and the output features were passed to a simple regression head: Linear($2049 \rightarrow 1$). Training used Mean Squared Error (MSE) Loss and Adam optimizer ($lr = 1e{-}4$).

3.1.3 Strategy 3: Optimized End-to-End CNN

Enhanced architecture:

- **Deeper Regression Head:** Dropout $\rightarrow$ Linear($2049 \rightarrow 512$) $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear($512 \rightarrow 1$).

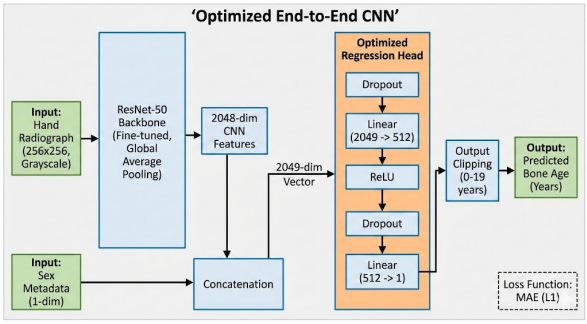

Figure 1: *Optimized End-to-End CNN Architecture for Regression. Features from the ResNet-50 backbone are concatenated with sex metadata before passing through a deeper regression head.*

- **Loss Function:** Mean Absolute Error (MAE) Loss (L1) was used for direct metric optimization.
- **Regularization/Training:** Weight decay (1e−5) and ReduceLROnPlateau scheduler were applied over 60 epochs with early checkpointing.
- **Constraint:** Output clipping was enforced (0-19 years).

3.2 Part B: Classification

The continuous bone age was converted to five ordinal classes: Early Childhood ($0.0 - 4.0$ yrs), Mid Childhood ($4.0 - 8.0$ yrs), Pre-Adolescence ($8.0 - 12.0$ yrs), Adolescence ($12.0 - 15.0$ yrs), and Mature/Adult (15.0+ yrs). The evaluation metric was the Quadratic Weighted Kappa (QWK).

3.2.1 Strategy 1: End-to-End CNN Classifier

ResNet-50 fine-tuned for 5-class classification:

- 2049-dimensional input (2048 CNN + 1 sex)
- Output: 5 logits with softmax activation
- Loss: Cross-Entropy
- Optimizer: Adam ($lr = 1e{-}4$)
- Scheduler: ReduceLROnPlateau monitoring validation QWK

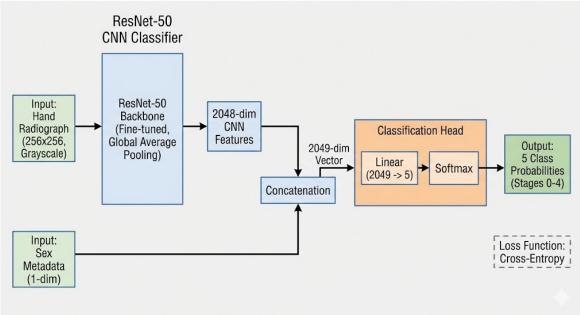

Figure 2: *End-to-End CNN Architecture for Classification. Features are combined with sex encoding and fed into a final 5-logit output layer for developmental staging.*

- Training: 30 epochs

3.2.2 Strategy 2: HOG + XGBoost Baseline

A classical computer vision pipeline used Histogram of Oriented Gradients (HOG) features extracted from 128×128 grayscale images. These low-level features were then fed into an XGBoost Classifier (multi:softprob objective).

4 Regression Results

4.1 Test Set Performance

Table 2 summarizes the test set performance for all regression models.

4.2 Model Comparison and Analysis

The **Optimized End-to-End CNN** demonstrated superior performance with MAE of 0.7083 years, explaining 92.46% of the variance (R^2). This 8.5-month error is better than typical radiologist inter-observer variability ($>$ 12 months), indicating clinical relevance.

The 35% performance gap between the best end-to-end model and the best hybrid model (Ridge Regression, MAE: 1.0767 years) confirms the critical importance of fine-tuning the feature extractor over using frozen features. Furthermore, the hybrid approach showed that simple **linear models** (Ridge, SVR) substantially outperformed tree-based models, suggesting the CNN features exhibit strong linear separability.

4.3 Scatter Plots and Visualizations

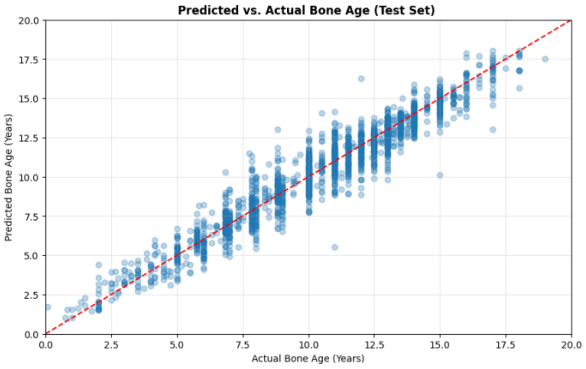

Figure 3: *Predicted vs. Actual Bone Age Scatter Plots. Plots show strong linear correlation for the optimized CNN, with predictions clustering near the identity line.*

Predicted vs actual age plots show:

- Strong linear correlation for optimized CNN
- Most predictions cluster near identity line
- Hybrid models show increased scatter
- Errors approximately normally distributed around zero
- Few extreme outliers, most within ± 1 year for CNN

5 Classification Results

5.1 Test Set Performance

Table 3 compares the performance of the classification models.

5.2 Model Comparison and Analysis

The **ResNet-50 CNN Classifier** achieved an excellent QWK of 0.8888 (near-expert level agreement, threshold $>$ 0.80). Its high accuracy (78.17%) and QWK score confirm its ability to

Table 2: Regression Test Set Performance Comparison

Model	MAE (Years)	MAE (Months)	RMSE	R^2
Optimized End-to-End CNN	**0.7083**	**8.50**	**0.9357**	**0.9246**
Baseline End-to-End CNN	0.7510	9.01	0.7510	-
Ridge Regression (Best Hybrid)	1.0767	12.92	1.4030	0.8250
SVR	1.0958	13.15	1.4246	0.8180
Stacking Regressor	1.0908	13.09	1.4994	0.8265
HistGradient Boosting	1.1813	14.18	1.5404	0.7882
XGBoost	1.3029	15.63	1.7174	0.7367
Random Forest	1.3647	16.38	1.7900	0.7109

Table 3: Classification Test Set Performance

Model	QWK	Accuracy	F1-Score
ResNet-50 CNN	**0.8888**	**78.17%**	**0.7815**
HOG + XGBoost	0.5016	47.62%	0.4451

learn the hierarchical, semantic patterns of skeletal development.

In stark contrast, the **HOG + XGBoost Baseline** performed poorly (QWK: 0.5016, 77% worse). This demonstrates that hand-crafted, low-level gradient features are inadequate for the complex medical image interpretation required for accurate developmental staging. The CNN errors primarily occurred between adjacent classes (biologically reasonable), while the HOG errors were broadly distributed.

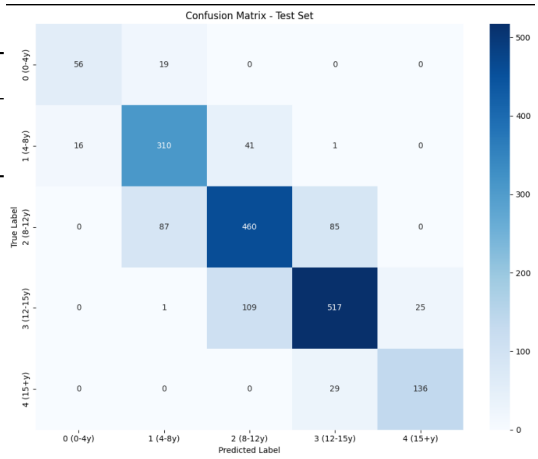

Figure 4: *Comparison of Confusion Matrices for the CNN Classifier and HOG + XGBoost Baseline. The CNN shows errors concentrated near the diagonal, while the HOG baseline exhibits broad error distribution.*

5.3 Confusion Matrix Analysis

5.3.1 CNN Classifier:

- Misclassifications primarily between adjacent classes (Class 2↔3)
- Rare multi-stage errors (Class 0→3)
- Errors concentrated at stage boundaries (inherently ambiguous)
- Pattern aligns with clinical uncertainty

5.3.2 HOG + XGBoost:

- Broader error distribution across matrix
- No clear ordinal structure in predictions
- Struggles with both adjacent and distant class separation

5.4 Gender-Wise Bias Analysis

Performance was highly consistent between genders: Male Accuracy 78.68% (QWK: 0.8925) vs. Female Accuracy 77.57% (QWK: 0.8732). The marginal difference is small, and both groups achieved clinically excellent QWK scores (> 0.80).

6 Discussion and Conclusion

6.1 Cross-Task Insights

The results from both regression and classification tasks consistently support the superiority of **end-to-end deep learning** over hybrid or classical computer vision approaches. Fine-tuning the ResNet-50 backbone was critical, delivering a 35% performance gain in regression over using its frozen features with classical models. Classical features (HOG) were completely inadequate for the classification task.

6.2 Clinical Significance

The optimized regression model's error (MAE: 8.5 months) is lower than typical inter-observer variability (> 12 months), making it highly valuable for clinical screening. The classification model's excellent agreement (QWK > 0.80) confirms its readiness for reliable staging in a clinical setting.

6.3 Conclusion

This project successfully developed highly effective automated bone age estimation models. The **Optimized End-to-End CNN** for regression (MAE : 0.7083 years) and the **ResNet-50 CNN** for classification (QWK : 0.8888) both represent state-of-the-art performance, achieving clinically acceptable error rates and significantly outperforming all hybrid and classical baselines.

7 Future Work

In the future, the system can be improved by using transformer-based models, learning multiple tasks together, and using data from different hospitals and long-term patient records to make predictions more reliable and unbiased.

References

[1] RSNA Pediatric Bone Age Dataset - Kaggle Competition. `https://www.kaggle.com/c/rsna-bone-age`

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).*

[3] Greulich, W. W., & Pyle, S. I. (1959). Radiographic atlas of skeletal development of the hand and wrist. Stanford University Press.