This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Uncertainty-Aware Bone Age Assessment: Analysis of Conformal and Bayesian Methods

Nadide Bilge Doğan*, Fuat Işıklan*, Mümin Emir Soydal*†, Tufan Kumbasar, Ilkay Oksuz, Hazım Kemal Ekenel, and Behçet Uğur Töreyin

Istanbul Technical University, Istanbul, Turkey
{dogann19,isiklanfuat,soydal25,kumbasart,oksuzilkay,ekenel,toreyin}@itu.edu.tr
*Equal Contribution †Corresponding Author

Abstract. Bone age assessment from pediatric hand radiographs is widely used to evaluate skeletal maturity, but most deep learning methods provide only point estimates without indicating reliability. This limitation is critical in clinical and forensic settings, where incorrect but overconfident estimates can affect downstream decisions. In this work, we present a unified benchmark of uncertainty quantification methods for bone age regression on the RSNA Pediatric Bone Age dataset. We train multi-input models that combine hand radiographs with sex information and evaluate three backbones: EfficientNet-B3, ConvNeXtV2-Tiny, and ViT-B/16, allowing us to assess whether uncertainty quantification performance is model-agnostic. Across ten random seeds, we compare six uncertainty quantification strategies: vanilla conformal prediction (CP), k-Nearest Neighbor-based CP (kNN-CP), Random-Forest CP (RF-CP), Conformalized Quantile Regression (CQR), Monte Carlo Dropout (MCD), and Bayesian Neural Network (BNN). Performance is assessed using mean absolute error, prediction interval coverage probability, prediction interval normalized average width, and coverage-width criterion at 90%, 95%, and 99% desired coverage. EfficientNet-B3 and ConvNeXtV2-Tiny provide the strongest point prediction accuracy, while conformal prediction methods offer the most reliable coverage-width trade-off. RF-CP provides the best coverage–width trade-off by producing adaptive prediction intervals with reliable empirical coverage. Vanilla CP and kNN-CP remain competitive, whereas MCD tends to under-cover, and BNNs produce overly conservative intervals. The source code is available at `https://github.com/bilgedogan/boneage_uncertainty`.

Keywords: Bone Age Assessment · Uncertainty Quantification · Conformal Prediction · Medical Image Regression

1 Introduction

Skeletal bone age assessment from hand radiographs is a fundamental task in medical imaging. Unlike chronological age, bone age reflects skeletal maturity. Pediatricians use bone age assessment to diagnose growth disorders, evaluate endocrine abnormalities, and monitor responses to growth hormone therapy [23].

Forensically, it plays a vital legal role in estimating chronological age when official records are missing or unreliable [31]. Both clinical and forensic use cases demand highly reliable bone age estimation frameworks.

Conventionally, bone age is estimated manually via the Greulich-Pyle atlas or Tanner-Whitehouse scoring [32], processes restricted by subjectivity and time constraints. To overcome these, deep learning approaches initially deployed convolutional neural networks (CNNs) for automated skeletal feature extraction [11]. Advanced iterations incorporated attention mechanisms to focus on diagnostically critical bone regions [30]. Most recently, multimodal transformer-based architectures [33] have emerged, jointly modeling radiographs and patient metadata within a unified network to maximize estimation accuracy. Yet, point predictions alone are insufficient for clinical and forensic deployment because they provide no measure of prediction reliability. Uncertainty quantification (UQ) through Prediction Intervals (PIs) is therefore essential to identify uncertain predictions and mitigate the risks of overconfident errors in decision-making.

Existing UQ paradigms address this challenge through different formulations. Bayesian Neural Networks (BNNs) learn distributions over model weights to estimate predictive uncertainty [3], while Monte Carlo Dropout (MCD) approximates Bayesian inference through stochastic dropout at test time [8]. In contrast, Conformal Prediction (CP) [21, 28] provides a distribution-free framework, producing PIs with finite-sample coverage guarantees under the exchangeability assumption. Vanilla CP calibrates the absolute residuals using a single global threshold; therefore, it generates constant-width PIs. To obtain adaptive PIs, CP variants such as k-Nearest Neighbor-based CP (kNN-CP) and Random Forest-based CP (RF-CP) introduce sample-dependent residual scaling [20, 22, 12, 4]. Conformalized Quantile Regression (CQR) extends CP by conformally calibrating conditional quantile estimates to guarantee finite-sample coverage [24], [19].

Systematic comparison of UQ methods has proven valuable across various fields. For example, BNNs, MCD, and ensemble methods have been compared for lifespan prognostics [1], while CP and quantile methods have been evaluated for physiological signals [2]. Closest to our work, CP was benchmarked across multiple CNNs for chronological age estimation from panoramic radiographs [27].

In this work, we present a unified benchmark of UQ methods for bone age assessment. We first develop multimodal point prediction models that combine hand radiographs with patients' sex data, using three backbone architectures: EfficientNet-B3 [25, 26], ConvNeXtV2-Tiny [29, 5], and ViT-B/16 [7]. These cover complementary design families, such as an efficient CNN, a modern convolutional architecture, and a vision transformer, letting us test whether the UQ methods remain model-agnostic rather than tied to a single feature extractor. Building upon these predictors, we systematically evaluate a diverse set of UQ methods, including BNNs, MCD, Vanilla-CP, kNN-CP, RF-CP, and CQR. We evaluate both predictive performance and UQ performance, providing a comprehensive assessment of the strengths and limitations of these approaches for clinically relevant bone age assessment.

The remainder of this paper is organized as follows. Section II describes the proposed methodology, including the prediction models and the UQ methods. Section III presents the quantitative and qualitative comparative results. Finally, Section IV concludes the paper and discusses the main findings.

2 Methodology

This section presents the benchmark framework used to evaluate UQ methods for pediatric bone age estimation. We first describe the multimodal point prediction models based on three deep learning backbones, followed by six UQ methods.

2.1 Bone Age Estimation Model Structure

Following prior multimodal bone age estimation pipelines that fuse hand radiographs with patient sex [15, 33], we adopt a similar design for our point prediction models, incorporating sex as an additional input alongside the radiograph. Specifically, we implement a multi-input regression model $f(x_i)$, as shown in Figure 1, where x_i consists of the hand radiograph and the corresponding binary sex information for sample i, and y_i denotes the ground-truth bone age. Image features are extracted using an ImageNet-pretrained backbone [6], while the sex information is projected to a 32-dimensional embedding through a fully connected layer. The backbone output and sex embedding are concatenated to obtain the joint representation z_i, which is then used to predict the normalized bone age $\hat{y}_i = f(x_i)$.

Each backbone is initialized with pretrained weights and trained end-to-end on our task. During training, all images are resized to $M \times M$, and are normalized with the ImageNet mean and standard deviation. To improve model generalization and reduce overfitting, standard data augmentation techniques are applied during training, including random rotations, horizontal flipping, affine transformations, and brightness and contrast jittering.

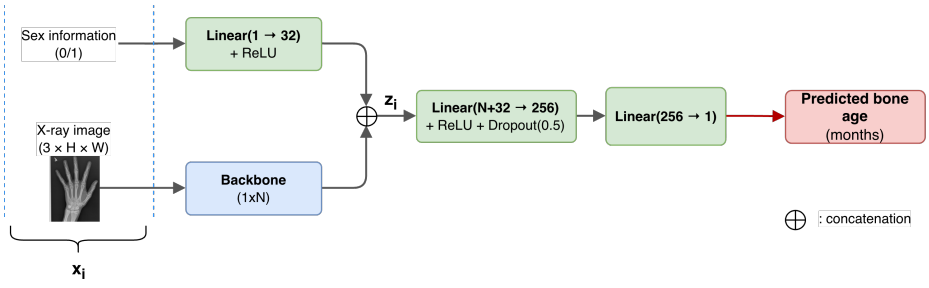

Fig. 1. Multi-input regression model. Image features (N-d) are concatenated with a 32-d sex embedding and regressed to bone age through two fully connected layers.

2.2 Benchmarked Uncertainty Quantification Methods

The benchmark comprises three categories of UQ methods. The first includes post-hoc CP methods, which construct PIs by calibrating pretrained point prediction models. The second is CQR, which first learns conditional quantiles and subsequently applies conformal calibration. The third consists of Bayesian approaches, namely BNNs and MCD, which estimate uncertainty directly from the predictive model. To ensure a fair comparison, all methods are evaluated using identical data splits, backbone architectures, and nominal coverage levels.
Conformal prediction– CP constructs PIs $\Gamma_\delta(\cdot)$ by calibrating a pretrained point predictor on the held-out calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$. For a target coverage level δ, calibration residuals (r_i) are used as nonconformity scores to estimate a finite-sample corrected quantile, which defines the $\Gamma_\delta(\cdot)$ [21]. In this study, we handle the standard vanilla CP, and two normalized variants.

Vanilla CP – Standard vanilla CP is used as the non-adaptive conformal baseline [18]. Given a $f(x_i)$ and $\mathcal{D}_{\text{cal}}$, we computed $\Gamma_\delta(x)$ as

$$\begin{aligned} r_i &= |y_i - f(x_i)|, \\ \Gamma_\delta(x) &= [f(x) - q_\delta(r), f(x) + q_\delta(r)] \end{aligned} \tag{1}$$

where $q_\delta(.)$ is the finite-sample corrected empirical δ-quantile and is computed from the sorted calibration residuals. Note that all PIs share a constant width, as q_δ reflects the overall error distribution.

kNN-CP – Unlike Vanilla CP, kNN-CP employs a sample-dependent difficulty estimate, $\sigma(\cdot)$, to normalize r_i before calibration [22]:

$$\tilde{r}_i = \frac{|y_i - f(x_i)|}{\sigma(z_i) + \epsilon} \tag{2}$$

where ϵ is a small constant for numerical stability. In the kNN variant, $\sigma(z)$ is computed as the mean absolute residual among the k nearest calibration samples in the embedding space. For a test sample, the calibrated normalized threshold q_δ is then rescaled by its estimated difficulty,

$$\Gamma_\delta(x) = \left[f(x) - q_\delta \left(\sigma(z) + \epsilon\right), f(x) + q_\delta \left(\sigma(z) + \epsilon\right)\right]. \tag{3}$$

This produces locally adaptive PIs that become wider in regions with larger calibration residuals.

RF-CP – The RF variant uses the same normalized conformal formulation in Eqs. (2)–(3), but replaces the kNN difficulty estimate with a bootstrap random-forest estimator. The RF is trained on the calibration set. Out-of-bag variability is used as the sample-dependent difficulty estimate $\sigma(z)$, following RF-CP approaches [4, 12]. This keeps the point predictor unchanged while allowing the final interval width to vary across test samples.
Conformalized quantile regression– CQR constructs adaptive PIs by combining quantile regression with conformal calibration [24]. Thus, CQR requires separate quantile training. For each target coverage level δ, the scalar regression

head is replaced with a two-output quantile head predicting the lower and upper conditional quantiles $\hat{q}_{i,\ell} = \hat{q}_{(1-\delta)/2}(x_i)$ and $\hat{q}_{i,u} = \hat{q}_{\delta/2}(x_i)$. The model is trained using the pinball loss [16], augmented with a quantile crossing penalty:

$$\mathcal{L}_{\text{CQR}} = \frac{1}{B}\sum_{i=1}^{B}\left[\rho_{(1-\delta)/2}(y_i - \hat{q}_{i,\ell}) + \rho_{\delta/2}(y_i - \hat{q}_{i,u}) + \lambda_{\text{cross}}\left[\hat{q}_{i,\ell} - \hat{q}_{i,u}\right]_+\right] \quad (4)$$

where B is the batch size, $\rho_\tau(u) = \max(\tau u, (\tau-1)u)$ is the pinball loss, $[.]_+ = \max(0, .)$, and $\lambda_{\text{cross}} = 10$. The preliminary quantile intervals are then calibrated on the $\mathcal{D}_{\text{cal}}$ using the CQR nonconformity score [24, 10]:

$$r_i^{\text{CQR}} = \max\left\{\hat{q}_{i,\ell} - y_i,\ y_i - \hat{q}_{i,u}\right\}. \quad (5)$$

In this study, the point prediction is calculated as $\hat{y}_i = (\hat{q}_{i,\ell} + \hat{q}_{i,u})/2$.

Monte Carlo Dropout – By leaving the dropout layer active during inference, repeated stochastic forward passes approximate samples from the posterior over the network's weights [8]. We run $T = 60$ stochastic forward passes per test input x, each yielding a prediction $\hat{y}_t = f(x; w_t), \quad t = 1, \ldots, T$, where w_t denotes the network weights realized under the t-th independently sampled dropout mask. The predictive mean and standard deviation are then given by

$$\mu(x) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t, \qquad \sigma(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t - \mu(x)\right)^2}. \quad (6)$$

Assuming approximate normality of the predictive distribution, the PI is constructed as

$$\Gamma_\delta(x) = [\mu(x) - z_\delta\,\sigma(x),\ \ \mu(x) + z_\delta\,\sigma(x)] \quad (7)$$

where z_δ is the two-sided standard normal critical value for the desired coverage δ (e.g., $z_\delta = 1.96$ for $\delta = 0.95$), obtained from the standard normal quantile table.

Bayesian Neural Networks – BNN is implemented via Bayes by Backprop [3, 17], which approximates the posterior over weights with a fully factorized Gaussian $q_\theta(w) = \mathcal{N}(w; \mu_\theta, \sigma_\theta^2)$ trained via the reparameterization trick. Only the regression head is replaced with variational layers, while the pretrained backbone remains deterministic. At each of $T = 50$ stochastic forward passes, weights are sampled as $w_t \sim q_\theta(w)$ and propagated through the radiograph and sex inputs to obtain $\hat{y}_t = f(x; w_t), \quad t = 1, \ldots, T$. As with MC Dropout, the resulting distribution of bone age predictions yields a predictive $\mu(x)$ and $\sigma(x)$, computed identically to Eq. (6), from which the $\Gamma_\delta(x)$ is obtained identically to Eq. (7).

3 Experimental Setting and Results

We evaluate the three backbone architectures and six UQ methods both quantitatively and qualitatively in terms of point prediction accuracy and PI quality.

3.1 Dataset, Experimental Setup and Evaluation Metrics

We use the 2017 RSNA Bone Age Challenge dataset [9], consisting of hand radiographs, patient sex, and bone age annotations. The dataset was released as part of the RSNA 2017 Pediatric Bone Age Machine Learning Challenge. As the official test labels are unavailable, the training set is partitioned into training (6,305), validation (3,153), and calibration (3,153) subsets using a stratified 50/25/25 split, while the official validation set (1,425 images) serves as the test set. Bone age targets are normalized to $[0, 1]$ using the maximum training age.

We evaluate three ImageNet-pretrained backbones: EfficientNet-B3 (M=300), ViT-B/16 (M=224), and ConvNeXtV2-Tiny (M=300). All models are optimized with Adam (learning rate 10^{-4}, batch size 32, 50 epochs). For the conformal methods, we set $k = 100$ for kNN-CP, while RF-CP uses 200 trees with a minimum leaf size of 5. To ensure statistical robustness, each backbone–UQ method combination is trained over 10 independent random seeds, with each seed determining both the network initialization and the data split.

The performance of all methods is evaluated on the common test set at desired coverage levels of 90%, 95%, and 99%. We evaluate coverage using PI coverage probability (PICP) while interval efficiency is measured by PI normalized average width (PINAW) [14, 13], where lower values indicate sharper intervals. We also report the Coverage Width-based Criterion (CWC) [14], which combines PICP and PINAW into a single metric by penalizing under-covered PIs (η=50). Point prediction performance is evaluated using Mean Absolute Error (MAE).

3.2 Comparative Point Prediction Results

Table 1 reports that EfficientNet-B3 and ConvNeXtV2-Tiny outperform ViT-B/16 in the point prediction task. However, applying CQR reduces the accuracy across all models. This drop in performance becomes much more severe when the target coverage is increased from 90% to 99%. At the 99% level, EfficientNet-B3 remains relatively stable, while the other models, especially ConvNeXtV2-Tiny, suffer a sharp decline in both accuracy and stability. The Bayesian-inspired methods, by contrast, leave point-prediction accuracy largely intact. MCD retains the deterministic weights and introduces stochasticity only at inference time, so averaging over the sampled predictions recovers the deterministic estimate and keeps its MAE close to the Deterministic baseline. BNN instead replaces the training objective with a variational one; since gradients are additionally propagated through the sampled weights, its accuracy penalty is accompanied by substantially larger variance across seeds.

3.3 Comparative UQ Results

Table 2 presents the evaluation metrics for all six methods across the three backbones. Among the evaluated techniques, MCD is the least effective method at every level. Its empirical coverage falls 7–11 points below the desired target

Table 1. Test-set MAE for bone age estimation.

Method	δ	EfficientNet -B3	ViT-B/16	ConvNeXtV2 -Tiny
Deterministic	–	7.52 ± 0.11	8.54 ± 0.49	7.59 ± 0.15
BNN	–	10.27 ± 0.67	11.02 ± 0.65	20.84 ± 8.61
MCD	–	7.58 ± 0.11	8.64 ± 0.48	9.66 ± 0.69
CQR	90%	8.06 ± 0.34	9.24 ± 0.21	9.57 ± 3.50
	95%	8.47 ± 0.38	10.46 ± 1.64	12.64 ± 5.99
	99%	10.71 ± 1.34	15.76 ± 3.22	26.78 ± 8.40

Table 2. UQ performance at 90%, 95%, and 99% desired coverage (A: ConvNeXtV2-Tiny; B: EfficientNet-B3; C: ViT-B/16). The nominal level δ is the ideal PICP value ($\rightarrow$); PINAW and CWC are minimised ($\downarrow$). **Bold** denotes the best-performing method for each backbone and coverage level.

		$\delta = 90\%$			$\delta = 95\%$			$\delta = 99\%$		
	Method	**PICP (%)** $\rightarrow$	**PINAW** $\downarrow$	**CWC** $\downarrow$	**PICP (%)** $\rightarrow$	**PINAW** $\downarrow$	**CWC** $\downarrow$	**PICP (%)** $\rightarrow$	**PINAW** $\downarrow$	**CWC** $\downarrow$
A	BNN	93.90 ± 4.80	0.510 ± 0.079	1.053 ± 1.557	96.20 ± 3.10	0.608 ± 0.095	1.309 ± 1.208	98.60 ± 1.50	0.799 ± 0.124	1.685 ± 1.120
	MCD	81.30 ± 0.02	$\mathbf{0.160 \pm 0.013}$	24.70 ± 38.07	86.10 ± 0.02	$\mathbf{0.191 \pm 0.015}$	23.60 ± 21.10	91.10 ± 0.01	$\mathbf{0.251 \pm 0.020}$	15.40 ± 8.920
	Vanilla-CP	89.10 ± 0.70	0.169 ± 0.008	0.437 ± 0.124	94.40 ± 0.60	0.204 ± 0.009	0.452 ± 0.145	98.70 ± 0.30	0.288 ± 0.012	$\mathbf{0.601 \pm 0.115}$
	kNN-CP	89.20 ± 0.80	0.197 ± 0.015	0.448 ± 0.186	94.50 ± 1.10	0.260 ± 0.025	0.525 ± 0.291	$\mathbf{99.00 \pm 0.30}$	0.482 ± 0.078	0.730 ± 0.244
	RF-CP	$\mathbf{90.30 \pm 0.90}$	0.224 ± 0.022	$\mathbf{0.348 \pm 0.177}$	$\mathbf{95.20 \pm 0.50}$	0.290 ± 0.033	$\mathbf{0.413 \pm 0.161}$	$\mathbf{99.00 \pm 0.30}$	0.476 ± 0.070	0.631 ± 0.247
	CQR	89.60 ± 0.40	0.207 ± 0.059	0.443 ± 0.161	94.70 ± 0.50	0.308 ± 0.103	0.585 ± 0.251	98.80 ± 0.20	0.749 ± 0.179	1.476 ± 0.569
B	BNN	95.60 ± 2.30	0.285 ± 0.049	$\mathbf{0.285 \pm 0.049}$	97.50 ± 1.40	0.339 ± 0.059	$\mathbf{0.339 \pm 0.059}$	99.20 ± 0.70	0.446 ± 0.077	0.598 ± 0.196
	MCD	79.40 ± 0.03	$\mathbf{0.127 \pm 0.012}$	51.30 ± 62.30	84.00 ± 0.03	$\mathbf{0.151 \pm 0.014}$	69.80 ± 73.50	89.80 ± 0.02	$\mathbf{0.199 \pm 0.019}$	27.40 ± 22.84
	Vanilla-CP	89.40 ± 0.60	0.140 ± 0.002	0.314 ± 0.100	94.30 ± 0.50	0.174 ± 0.003	0.429 ± 0.055	98.80 ± 0.20	0.256 ± 0.007	$\mathbf{0.517 \pm 0.091}$
	kNN-CP	89.30 ± 0.90	0.161 ± 0.005	0.374 ± 0.152	94.70 ± 0.60	0.216 ± 0.010	0.429 ± 0.154	$\mathbf{99.00 \pm 0.40}$	0.411 ± 0.050	0.680 ± 0.242
	RF-CP	$\mathbf{89.60 \pm 0.70}$	0.162 ± 0.007	0.325 ± 0.113	$\mathbf{94.90 \pm 0.70}$	0.209 ± 0.014	0.353 ± 0.152	98.90 ± 0.20	0.337 ± 0.030	0.520 ± 0.192
	CQR	89.30 ± 0.80	0.159 ± 0.005	0.390 ± 0.103	94.50 ± 0.40	0.211 ± 0.012	0.482 ± 0.046	98.70 ± 0.20	0.342 ± 0.025	0.710 ± 0.138
C	BNN	93.00 ± 3.70	0.271 ± 0.039	0.605 ± 1.037	95.70 ± 2.10	0.323 ± 0.047	0.573 ± 0.427	98.20 ± 0.90	0.425 ± 0.061	0.993 ± 0.292
	MCD	83.30 ± 0.03	0.163 ± 0.012	16.30 ± 33.01	87.50 ± 0.03	$\mathbf{0.195 \pm 0.014}$	21.81 ± 38.44	92.20 ± 0.02	$\mathbf{0.256 \pm 0.019}$	12.14 ± 13.94
	Vanilla-CP	89.50 ± 0.80	$\mathbf{0.160 \pm 0.008}$	0.343 ± 0.153	$\mathbf{95.00 \pm 0.40}$	0.199 ± 0.010	0.315 ± 0.123	$\mathbf{98.90 \pm 0.20}$	0.290 ± 0.018	0.504 ± 0.139
	kNN-CP	89.70 ± 0.70	0.180 ± 0.010	0.313 ± 0.152	95.20 ± 0.50	0.246 ± 0.018	$\mathbf{0.289 \pm 0.141}$	99.10 ± 0.30	0.494 ± 0.091	0.594 ± 0.181
	RF-CP	$\mathbf{90.10 \pm 0.60}$	0.195 ± 0.011	$\mathbf{0.273 \pm 0.126}$	95.20 ± 0.40	0.254 ± 0.015	0.344 ± 0.149	99.20 ± 0.20	0.447 ± 0.043	$\mathbf{0.485 \pm 0.102}$
	CQR	89.90 ± 0.80	0.190 ± 0.006	0.340 ± 0.131	94.40 ± 0.50	0.252 ± 0.025	0.582 ± 0.160	98.90 ± 0.30	0.509 ± 0.085	0.852 ± 0.300

regardless of backbone. For instance, it achieves 79.4–83.3% at δ=90%, 84.0–87.5% at δ=95%, 89.8–92.2% at δ=99%. Although it produces the narrowest intervals, with PINAW dropping 0.127 for EfficientNet-B3, under-coverage is heavily penalized. As a result, its CWC scores reach up to 69.8 ± 73.5 for EfficientNet-B3 at δ=95%. On the other hand, BNN mostly covers all three targets and for all backbones. For example, its empirical coverage reaches 95.7–97.5% at the δ=95% target, and rises to 98.2–99.2% at δ=99%. However, its intervals are 1.5–2$\times$ wider than the conformal methods at matched coverage. Additionally, it has a high PICP standard deviation, going up to $\pm$4.80 for ConvNeXtV2-Tiny. This indicates that some seeds are still under-covering. The three conformal variants track the desired target closely. As a direct result, they achieve the lowest CWC across nearly every backbone and coverage combination. RF-CP provides the best coverage-width trade-off at the δ=90%. Meanwhile, among the conformal methods, CP generates the narrowest intervals. kNN-CP is generally the least efficient of the three conformal variants, particularly at δ=99%. CQR also reaches near-desired coverage. However, it is the only calibrated method whose efficiency decreases sharply as target coverage increases. At δ=90% and

δ=95%, its PINAW and CWC metrics are comparable to the conformal baselines. In contrast, at δ=99%, its interval width roughly increases. For example, its width increases from 0.207 to 0.749 for ConvNeXtV2-Tiny, and from 0.190 to 0.509 for ViT-B/16. Consequently, this makes CQR the least efficient calibrated method at high coverage. Across the different backbones, EfficientNet-B3 consistently gives the tightest intervals for nearly every method. Conversely, ConvNeXtV2-Tiny shows the widest and most variable intervals at the δ=99% target.

Figure 2 illustrates these trends qualitatively, showing the mean 95% prediction intervals generated by each of the six methods over 10 different random seeds on the same 500 randomly selected test samples, using EfficientNet-B3 as a representative backbone. The BNN bands are the widest overall, consistent with its relatively large PINAW, whereas MCD produces the visually tightest intervals but frequently fails to enclose the ground-truth value, directly illustrating the systematic under-coverage reported above. In contrast, kNN-CP and RF-CP exhibit stronger variation in interval width because their normalized conformal intervals depend on local difficulty estimates rather than a single global residual quantile. The narrower prediction intervals produced by kNN-CP and RF-CP in certain age ranges (e.g., 12.5–15 years) suggest lower estimated local uncertainty, demonstrating how these methods adapt interval width to local difficulty.

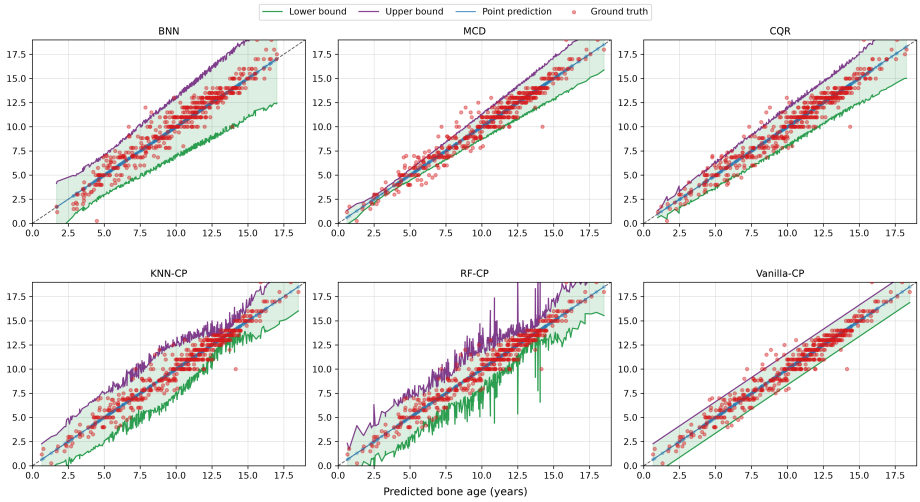

Fig. 2. PIs generated by the six UQ methods using the EfficientNet-B3 backbone at 95% desired coverage. Test samples are ordered by increasing predicted bone age.

4 Conclusion and Future Work

We benchmarked six UQ methods for bone age estimation from hand radiographs across three backbone architectures. Our results identify RF-CP as the most effective UQ method, providing the best trade-off between coverage and interval width through adaptive PIs. Yet, this performance comes at the cost of training

an additional RF model to estimate sample difficulty. kNN-CP also produces adaptive PIs but requires careful tuning of k, while CQR requires training the predictor to estimate conditional quantiles. In contrast, Vanilla CP is simple, computationally inexpensive, and parameter-free, but produces constant-width PIs that cannot adapt to sample difficulty. On the other hand, MCD tends to underestimate uncertainty, whereas BNNs yield overly conservative intervals. Thus, neither approach produces efficient PIs for medical image regression.

The findings of this study suggest that CP is particularly valuable for uncertainty-aware medical image regression. Future work should evaluate whether the same trends hold across other anatomical sites and imaging modalities, and systematically assess the sensitivity of kNN-CP and RF-CP to their key hyperparameters.

Acknowledgments. This paper has been produced benefiting from the 2510 Program of The Scientific and Technological Research Council of Türkiye (project no: 124N419). The paper also benefited from Istanbul Technical University Scientific Research Projects (ITU BAP) funds, grant number 47653.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Basora, L., Viens, A., Chao, M.A., Olive, X.: A benchmark on uncertainty quantification for deep learning prognostics. Reliability Engineering & System Safety **253**, 110513 (2025)
2. Bench, C., Pfeffer, O., Desai, V., Moulaeifard, M., Coquelin, L., Charlton, P.H., Strodthoff, N., Hegemann, N., Aston, P.J., Thompson, A.: A systematic evaluation of uncertainty quantification techniques in deep learning: a case study in photoplethysmography signal analysis. Machine Learning: Health **2**(1), 015011 (2026)
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning. pp. 1613–1622. PMLR (2015)
4. Boström, H., Linusson, H., Löfström, T., Johansson, U.: Accelerating difficulty estimation for conformal regression forests. Annals of Mathematics and Artificial Intelligence **81**, 125–144 (2017)
5. Bram, J.T., Pareek, A., Beber, S.A., Jones, R.H., Shariatnia, M.M., Daliliyazdi, A., Tracey, O.C., Green, D.W., Fabricant, P.D.: Determination of skeletal age from hand radiographs using deep learning. The American Journal of Sports Medicine **53**(11), 2715–2725 (2025)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. pp. 1050–1059. PMLR (2016)
9. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., et al.: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019)
10. Huang, J., Song, J., Zhou, X., Jing, B., Wei, H.: Torchcp: A python library for conformal prediction. Journal of Machine Learning Research **26**(266), 1–25 (2025)
11. Iglovikov, V.I., Rakhlin, A., Kalinin, A.A., Shvets, A.A.: Paediatric bone age assessment using deep convolutional neural networks. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 300–308. Springer (2018)
12. Johansson, U., Boström, H., Löfström, T., Linusson, H.: Regression conformal prediction with random forests. Machine Learning **97**, 155–176 (2014)
13. Khosravi, A., Nahavandi, S., Creighton, D.: Construction of optimal prediction intervals for load forecasting problems. IEEE Transactions on Power Systems **25**(3), 1496–1503 (2010)
14. Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.F.: Lower upper bound estimation method for construction of neural network-based prediction intervals. IEEE Transactions on Neural Networks **22**(3), 337–346 (2011)
15. Kim, J.K., Park, D., Chang, M.C.: Assessment of bone age based on hand radiographs using regression-based multi-modal deep learning. Life **14**(6), 774 (2024)
16. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)

17. Krishnan, R., Esposito, P., Subedar, M.: Bayesian-torch: Bayesian neural network layers for uncertainty estimation (Jan 2022)
18. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018)
19. Msaddi, S., Kumbasar, T.: An alliance for uncertainty quantification: Type-2 fuzzy logic systems with conformal prediction. IEEE Transactions on Emerging Topics in Computational Intelligence **10**(2), 2022–2032 (2026)
20. Papadopoulos, H., Gammerman, A., Vovk, V.: Normalized nonconformity measures for regression conformal prediction. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008). pp. 64–69 (2008)
21. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: European conference on machine learning. pp. 345–356. Springer (2002)
22. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. Journal of Artificial Intelligence Research **40**, 815–840 (2011)
23. Prokop-Piotrkowska, M., Marszałek-Dziuba, K., Moszczyńska, E., Szalecki, M., Jurkiewicz, E.: Traditional and new methods of bone age assessment-an overview. Journal of Clinical Research in Pediatric Endocrinology **13**, 251–262 (08 2021)
24. Romano, Y., Patterson, E., Candès, E.J.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
25. Sutha, J., Dhanasekaran, S., Murugan, B., Manjunath, T.: An enhanced approach for bone-age assessment using efficientnet-b3 architecture. In: 2025 International Conference on Data Science and Business Systems (ICDSBS). pp. 1–6. IEEE (2025)
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PmLR (2019)
27. Upalananda, W., Chairat, S., Prapayasatok, S., Nalampang, S., Chaichulee, S.: Conformalized deep regression for age estimation from panoramic radiographs: An uncertainty-aware approach. Forensic Sciences **6**(2), 50 (2026)
28. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer (2005)
29. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142. Ieee (2023)
30. Yang, Z., Cong, C., Pagnucco, M., Song, Y.: Multi-scale multi-reception attention network for bone age assessment in x-ray images. Neural Networks **158**, 249–257 (2023)
31. Yuan, W., Fan, P., Zhang, L., Pan, W., Zhang, L.: Bone age assessment using various medical imaging techniques enhanced by artificial intelligence. Diagnostics **15**(3) (2025)
32. Zhang, H., Wu, J., Zhou, L., Li, J.: Uncertainty-aware multi-stream deep learning for pediatric orthopedic bone age assessment from hand radiographs. Journal of King Saud University Computer and Information Sciences (05 2026)
33. Zhang, J., Chen, W., Joshi, T., Zhang, X., Loh, P.L., Jog, V., Bruce, R.J., Garrett, J.W., McMillan, A.B.: Bae-vit: an efficient multimodal vision transformer for bone age estimation. Tomography **10**(12), 2058–2072 (2024)