Trustworthy Biometrics: Adaptive Uncertainty Quantification in Age Estimation

Team ‘Honolulu’:
Mohammadreza Heibati
Sandro Khizanishvili
Emre Yesil
Juan Iñaki Larrea

Advanced Machine Learning
Sapienza University of Rome
Dec 2025

Introduction

Problem

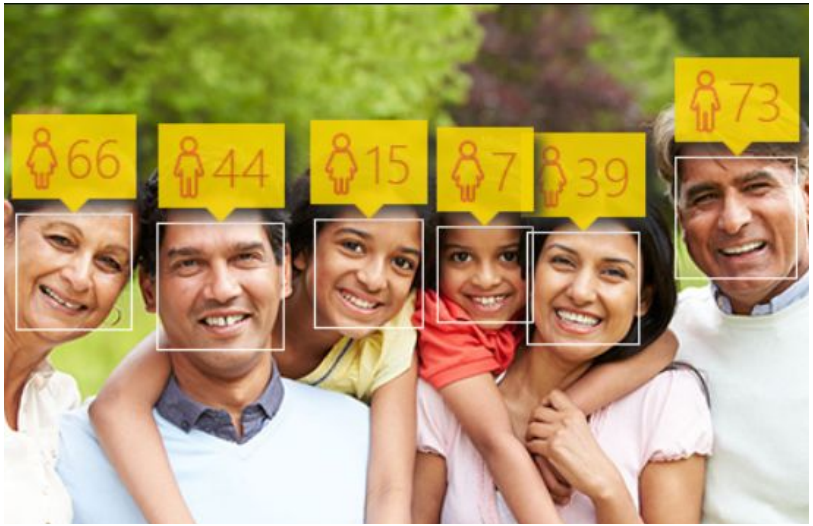
- Standard Age Estimation systems output point estimates (e.g, Age: 25).
- Issue: Models could be over/under confident but mostly they are wrong.

Goal

- Develop a robust framework for Uncertainty Quantification in age estimation.
- Output calibrated prediction intervals (e.g., [17.5, 20.2]) instead of unsafe point estimates.

Target

- Achieve statistical validity with a **95% (1-α; α=5%)** marginal coverage guarantee via Conformal Prediction.



Method & Dataset

Dataset

- **UTKFace**: more than 20k aligned images.
  - Full age range 0-116
  - Data is split into three unique sets:
    - Training - 60%
    - Validation/Calibration - 20%
    - Test - 20%

ResNet50: [1]

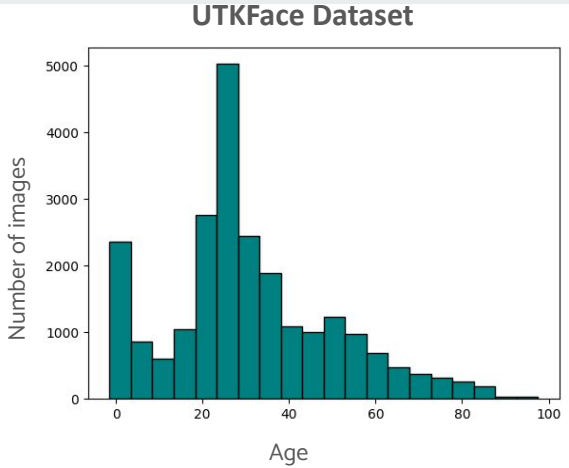
- Standard CNN. Focuses on local features.
- Provides a robust, established baseline for comparison.

Swin Transformer V2: [2]

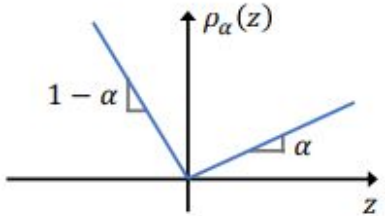
- Captures global facial geometry.
- See if a state of the art model advanced global feature learning improves prediction and CQR performance.

Output Layer:

- Modified heads to predict ***q*_0.025**, ***q*_0.5**, ***q*_0.975** via **Pinball Loss.**



$$L_\alpha(y, \hat{y}) = \begin{cases} \alpha \cdot (y - \hat{y}) & \text{if } y \geq \hat{y} \\ (\alpha - 1) \cdot (y - \hat{y}) & \text{if } y < \hat{y} \end{cases}$$



where $z = y - \hat{y}$.

Uncertainty Quantification (CQR)

Quantile Regression

- Estimate a given quantile of Y conditional on X.
- The uncertainty in the prediction of Y is reflected in the length of the interval.

Conformal Prediction

- Construct prediction intervals that attain valid coverage in finite samples, without making distributional assumptions.
- Can be conservative: they form intervals of constant or weakly varying length.

Conformalized Quantile Regression (CQR) [3]

Algorithm 1: Split Conformal Quantile Regression.

Input:

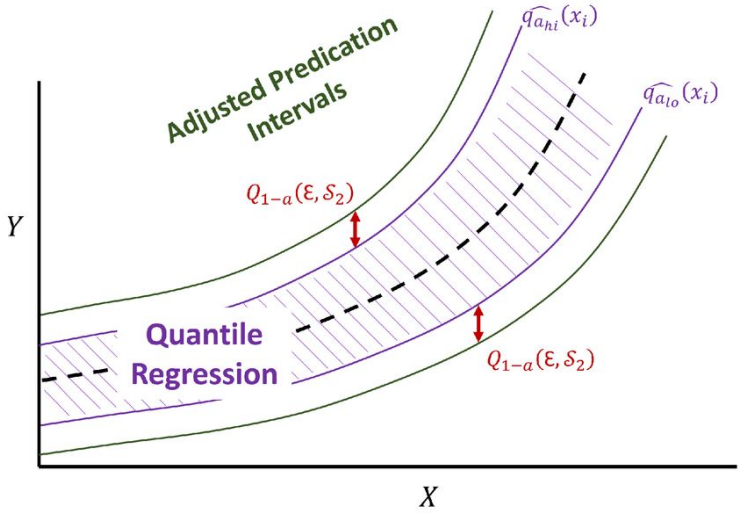
Data $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R},\ 1 \le i \le n$.
Miscoverage level $\alpha \in (0, 1)$.
Quantile regression algorithm $\mathcal{A}$.

Process:

Randomly split $\{1, \ldots, n\}$ into two disjoint sets $\mathcal{I}_1$ and $\mathcal{I}_2$.
Fit two conditional quantile functions: $\{\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}\} \leftarrow \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$.
Compute E_i for each $i \in \mathcal{I}_2$, as in equation (9).
Compute $Q_{1-\alpha}(E, \mathcal{I}_2)$, the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th empirical quantile of $\{E_i : i \in \mathcal{I}_2\}$.

Output:

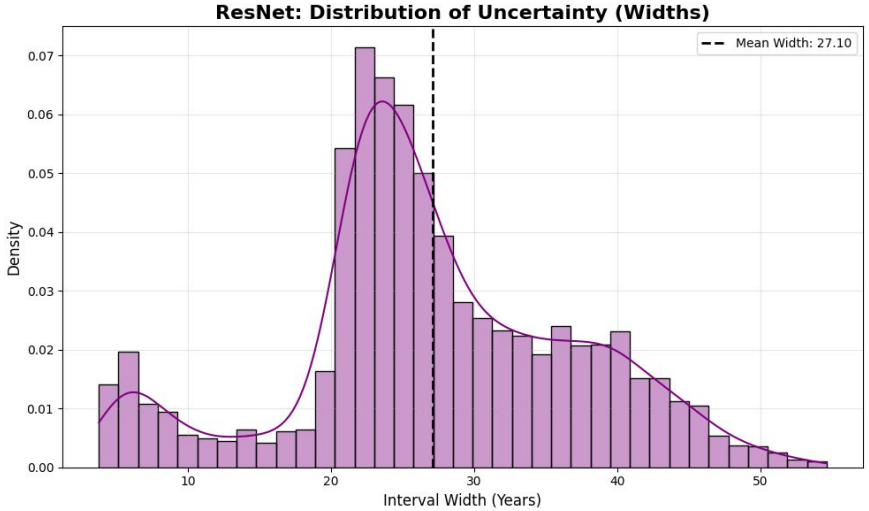
Prediction interval $C(x) = [\hat{q}_{\alpha_{\text{lo}}}(x) - Q_{1-\alpha}(E, \mathcal{I}_2),\ \hat{q}_{\alpha_{\text{hi}}}(x) + Q_{1-\alpha}(E, \mathcal{I}_2)]$ for unseen input $X_{n+1} = x$.

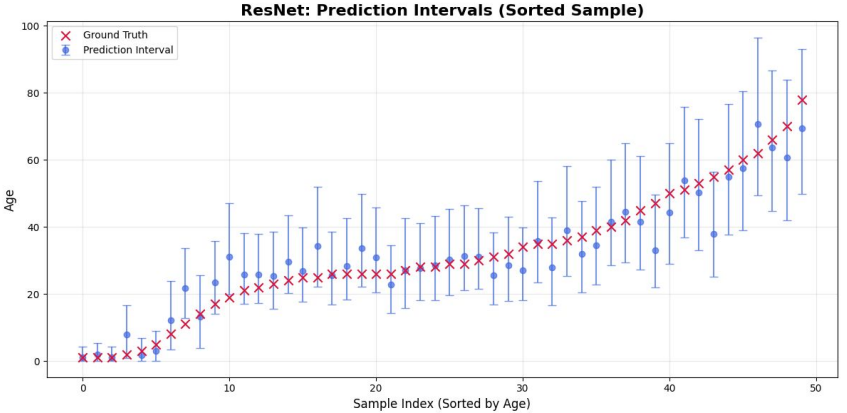


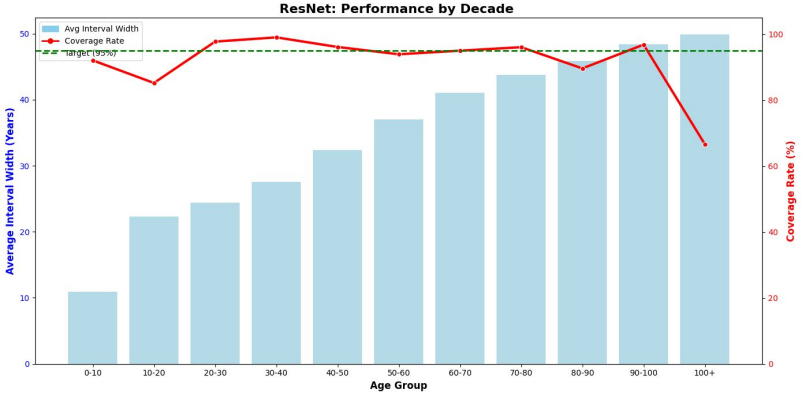
$$E_i := \max\{\hat{q}_{\alpha_{\text{lo}}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{\text{hi}}}(X_i)\} \quad (9)$$

ResNet

- Coverage Rate: **95.5% (before cal. 93%)**
- Average Interval Width: **27.2 years (before cal. 25.0)**
- Mean Absolute Error (MAE): **5.2 years**

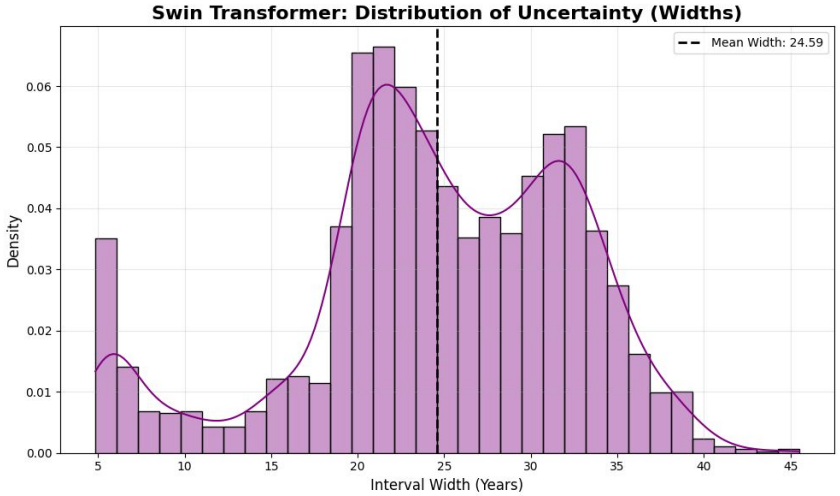


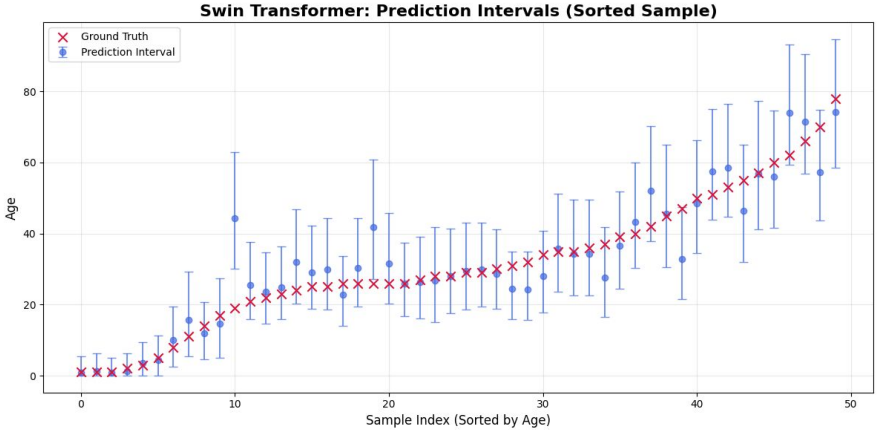


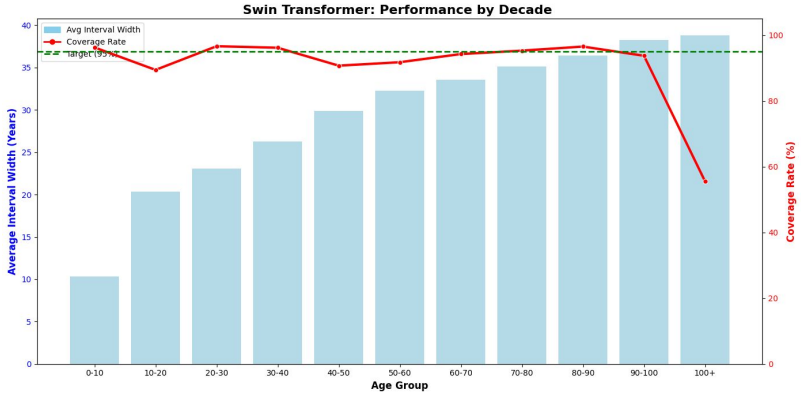


Swin

- Coverage Rate: **94.7% (before cal. 86%)**
- Average Interval Width: **24.8 years (before cal. 17.8)**
- Mean Absolute Error (MAE): **4.8 years**







Conclusions & Next steps

- **High Target Coverage (95%):** We prioritized reliability. Capturing 95% of cases requires a "wide net" to include outliers.
- **Inherent Ambiguity:** Visual age does not map perfectly to biological age. The model captures this natural variance.

"Better Data" over "More Data"

- Fairness via JPD Approach: Using a wide age range and global demographic balanced dataset that includes more selfies from real people (as JPD applied this approach with more than 100,000 selfies from their customers) to ensure fairness.[4]

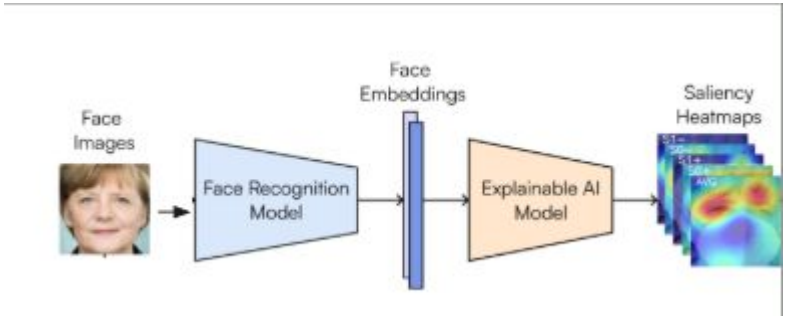
Model Explainability (XAI)

- Verify that the model focuses on relevant facial features (e.g., wrinkles, skin texture) rather than background noise to estimate age.

Table 1. IMDB-WIKI dataset and its partitions sizes in number of images.

IMDB-WIKI	IMDB	Wikipedia	IMDB-WIKI used for CNN training
524,230	461,871	62,359	260,282 images

IMDB&WIKI could be misleading at this point with celebrity images



References

1. Kaiming He et al. (2015) “Deep Residual Learning for Image Recognition” - https://arxiv.org/pdf/1512.03385
2. Ze Liu et al. (2022) “Swin Transformer V2: Scaling Up Capacity and Resolution” - https://arxiv.org/pdf/2111.09883
3. Romano et al. (NeurIPS 2019) “Conformalized Quantile Regression (CQR).“ - https://arxiv.org/pdf/1905.03222
4. François D. et al. “JAM: A Comprehensive Model for Age Estimation, Verification, and Comparability1 - https://arxiv.org/html/2410.04012v2

Thank you for the attention!

Dec 2025

EXTRA: OUR TRIES

Different architectures

- ResNet50, VGG, Swin Transformer

Different approaches

- Different hyperparameters
- Transfer learning, only train the last layer (6147 parameters for ResNet50)
- Train the whole neural network (23514179 parameters for ResNet50)

More data

- Concatenate UTKFaces and Morph 2 (73719 images)
- **Morph-2:** ~50,000 images (mostly mugshots).
    - In contrast to UTKFace, have consistent lighting, neutral expressions, and plain backgrounds.

Filtered and balance data

- Have same size of data in intervals (20,29), (30,39), etc.
- Keep only data from 20 to 40, 30 to 60, etc.