National College of Ireland

Bone Age Estimation Using X-ray Images with Gender-Aware Preprocessing

MSc Research Project

Programme Name: MSCDAD_C

Rohan Balasaheb Muradnar

Student ID: x23292113

School of Computing

National College of Ireland

Supervisor: Sallar Khan

National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Rohan Balasaheb Muradnar		
Student ID:	x23292113		
Programme:	Msc in Data Analytics (MSCDAD-C)	**Year:**	24-25
Module:	Msc Research Project		
Lecturer:	Sallar Khan		
Submission Due Date:	15/09/2025		
Project Title:	Bone Age Estimation Using X-ray Images with Gender-Aware Preprocessing		
Word Count:	8352, Pages: 22		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Rohan Balasaheb Muradnar
Date:	15/09/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Abstract

In paediatric endocrinology and orthopaedics, reliable estimates of skeletal-maturity are necessary but current Greulich-Pyle and Tanner-Whitehouse atlases are time-consuming, subjective and demographically-biased. As a result, we created a two-stage, gender-sensitive deep-learning pipeline where sex is a routing variable as opposed to optional metadata. In Stage 1, a lightweight convolutional neural network (CNN) classifies sex. In Stage 2, all the radiographs are passed to a sex-specific, transfer-learned InceptionV3 regressor. An auxiliary preprocessing engine adjusts CLAHE contrast, segmentation margins and crop heuristics separately in boys and girls to conserve dimorphic ossification cues. The balanced QU hand-X-ray cohort and an external RSNA audit set were used to conduct experiments. Baselines included (i) a single gender-agnostic regressor and (ii) late-fusion InceptionV3 that appends the predicted sex to the features. Evaluation encompassed mean absolute error (MAE), mean-squared error, R^2, proportions within ±6/±12 months, calibration curves, Bland–Altman bias and bootstrap tests of paired errors. The suggested framework also reduced MAE, enhanced calibration, and, more importantly, reduced sex-stratified error disparities relative to both baselines by half. Ablation tests verified that both the sex gate and gender-tuned preprocessing made a positive input to the gains. All code, configuration files and logs are under version control, and Grad-CAM maps demonstrate anatomically plausible focus by sex. At preprocessing and architectural levels, therefore, embedding domain knowledge on sexual dimorphism increases the accuracy and fairness of automated bone-age assessment.

Keywords: Bone age assessment, deep learning, gender-aware preprocessing, sex classification, transfer learning, fairness, pediatric radiology, convolutional neural networks, calibration.

Table of Contents

1. Introduction:

Proper assessment of skeletal maturity is essential to a wide range of paediatric decision-making, such as the diagnosis of endocrine and metabolic disorders, the scheduling of growth hormone therapy, the planning of corrective orthopaedic surgery, the counselling of families regarding residual growth potential, and prognostication. Even small inaccuracies in this test may result in timely and/or incorrect resource distribution or interventions, especially in resource-inadequate environments where radiology knowledge is sparse and follow-up options are lower (Booz et al., 2020; Lee & Lee, 2021). The practice of describing maturation with one number, the bone age, conceals clinically relevant heterogeneity within ossification centres and threatens to disadvantage particular subgroups by accident and to exacerbate extant disparities in care. The availability of automated systems that can produce reproducible and consistent evaluations is a two-fold advantage because they free radiologists of the task of manual scoring and provide them with objective points of reference that can be used in longitudinal tracking. This kind of automation,

therefore, enables accuracy growth medicine and enables scalable tele-radiology in low-resource settings (Pan et al., 2020; Martín Pérez et al., 2025).

However, Greulich-Pyle (GP) atlas and Tanner-Whitehouse (TW) methods have their feet in the ground in clinical practice. They both require skilled mapping against fixed reference plates created on demographically constrained, historically derived panels, which is lengthy, subjective, and vulnerable to inter- and intra-reader variability (Pan et al., 2020; Lee & Lee, 2021). It has been shown that reader experience, cumulative fatigue, and minor differences in image quality can change the scoring by several months, making multi-centre studies difficult. The atlases also do not reflect secular changes in the pattern of population growth, or the subtle influences of ethnicity, socioeconomic status and sex on the rate of maturation, forcing clinicians to project into unvalidated domains. These shortcomings have stimulated the desire to explore data-driven, algorithmic solutions that do not force legacy templates into the model, but instead learn the discriminative signals directly out of pixels and, on the other hand, demonstrated the need to audit these systems against demographic biases that the legacy atlases could not have solved (Pan I. et al., 2020; Halabi et al., 2018).

The past decade has seen the convolutional neural networks (CNNs) transform the way of measuring bone age significantly. They are at least near-expert performance on large-scale datasets, most prominently in the RSNA Pediatric Bone Age Challenge, which standardized evaluation metrics and stoked a wave of architectural invention (Spampinato et al., 2017; Halabi et al., 2018; Larson et al., 2018). Early end-to-end architectures showed that hierarchical features directly extracted out of radiographs could outperform hand-engineered descriptors. The following models added attention mechanics, multi-granularity encoders and cascaded region-of-interest segmentation to concentrate the model on epiphyses and metaphyses that are most informative to age regression (Li et al., 2023; Liu et al., 2024; Gonzalez et al., 2020).

The demonstration that physiological variability can occur in the tempo and sequence of skeletal development between the sexes has a long history; on average, female subjects begin and complete pubertal ossification more rapidly than male subjects. The lack of consideration of this heterogeneity may lead to an over-estimation of the prediction errors that are skewed in a biased manner and perpetuate injustice in the automated pipeline designs (Lee & Lee, 2021; Mart n Perez et al., 2025). Whereas some deep-learning research has used a binary sex label as an auxiliary input, the vast majority of systems are still essentially gender-agnostic during preprocessing, feature extraction, and loss construction, thus wasting domain knowledge about sex-specific growth patterns.

1.1. Research Question:

Does a two-stage, gender-aware framework comprising an initial sex-classification CNN followed by sex-specific regression models significantly reduce mean absolute error (MAE) compared with a single, gender-agnostic model on the QU pediatric hand X-ray dataset?

1.2. Research Objectives:

Following are the research objectives of this study:

- Curate and quality-check the QU dataset; ensure balanced train/validation/test splits by sex and characterize the distribution of skeletal maturity.
- Design and evaluate a lightweight CNN to classify sex from hand radiographs.
- Develop and compare gender-aware preprocessing pipelines (e.g., contrast settings, ROI cropping) versus a generic pipeline.
- Train separate, sex-specific CNN regressors for bone age estimation and optimize them for MAE reduction.
- Establish baselines with a single gender-agnostic model, and also a model that only appends sex as metadata.
- Assess performance, fairness, and generalization on both QU and external RSNA data using MAE, MSE, R^2, and calibration error.

2. Literature Review:

Modern arguments in automated bone age assessment (BAA) can be traced back to the clinical supremacy of the Greulich-Pyle (GP) atlas and the Tanner-Whitehouse (TW) scoring systems, which have long formed the canonical system of deducing skeletal maturity by assessing left-hand radiographs. These devices were based on fairly homogeneous cohorts of the middle twentieth century and operationalised bone age as a pattern recognition contest: clinicians would compare patient radiographs with standard reference plates (GP) or would code ossification centres (TW) and would interpolate a global age. Despite being conceptually beautiful and clinically manageable, such processes were never intended to support secular changes in growth patterns, ethnic diversity, or the higher throughput and consistency required by modern paediatric imaging services (Pan I. et al., 2020; Booz et al., 2020). Inter-observer differences on the basis of experience, fatigue, and subjective readings of inconspicuous morphological variations have been well-documented; even within trained radiologists, disparities of many months are not infrequent, undermining both clinical decision-making and multicentre research harmonisation (Kim J.R. et al., 2017; Lee & Lee, 2021).

The inherent shortcomings of analogue instrumentation instigated a paradigm transition in the hand-intensive, template-based evaluation into computational solutions involving the encoding, learning, and utilising the latent structure of skeletal maturation, based directly on pixel-level images. Early computer-aided systems replicated atlas logic, though limited by the need to handcraft features and fragile heuristics of segmentation, thus limited generalisability across populations. The further emergence of big, labelled dataset and the development of deep-learning frameworks allowed to replace the fixed, human-based Greulich-Pyle template scheme, with machine-driven models that can be iteratively improved and population-specifically calibrated (Halabi et al., 2018; Pan X. et al., 2020). But the historical legacy of such atlases is still there: they represent the current state-of-the-art gold standard in dataset labelling, and, somewhat ironically, both provide the yardstick against which modern systems are evaluated and a bottleneck that constrains further development within their limits (Larson et al., 2018; Martín Pérez et al., 2025).

The quick development of AI-based Bone Age Analysis (BAA) is directly linked to the emergence of publicly-available datasets not only making the evaluation protocols more standardised but increasing reproducibility. The 2017 RSNA Pediatric Bone Age Challenge was a tipping point that provided over 12,000 annotated hand radiographs, standardized pre-processing procedures, and a common leaderboard with which researchers could benchmark performance on different methodologies under identical conditions. Not only did this system confirm the viability of convolutional neural networks (CNNs) in achieving radiologist-level mean absolute error (MAE), but it also created methodological transparency, which allows more subtle comparisons of architectures, loss functions, and training regimes to be made. Earlier datasets, including the Digital Hand Atlas, supplied the necessary curated images and expert annotations that formed the basis of seminal works, demonstrating that deep neural networks could learn maturation cues without any explicit feature engineering (Spampinato et al., 2017; Iglovikov et al., 2018).

Within the last few years, the research community has grown past the use of the RSNA dataset, establishing dedicated cohorts to tackle more specific problems, including robustness to trauma films, the benefit of complete metadata, and evaluating fairness across demographic groups (Pan I. et al., 2020; Kim J.K. et al., 2024). The QU dataset used in this thesis is representative of this trend since it balances male and female samples and contains clear sex labelling, which allows performing robust analysis of gender-sensitive modelling techniques that are lacking in many of the older datasets. Supplementary sets analyzed in papers like Li et al. (2023) and Pan X. et al. (2020) allowed exploring cascaded regional extraction, as well as fully automated pipelines, whereas systematic reviews remain to unite these heterogeneous sources, with some issues of cross-cohort generalizability and demographic auditability persisting (Liu et al., 2024; Martn P, 2025; Lee and Lee, 2021).

2.1. CNN Architectures for Bone Age Regression

In the first wave, CNN-based methods showed that end-to-end learning of features can match (and even exceed) expert assessments of bone age, usually executing bone age regression on full radiographs using generic backbones, such as VGG, ResNet, and Inception (Spampinato et al., 2017; Lee et al., 2017). Such research employed transfer learning on huge natural image datasets to address the small size of medical datasets, fine-tuning lower layers to learn domain-specific textures and morphologies. Further developments of the paradigm included the introduction of multi-branch network structures, auxiliary classification heads, and composite losses that have stabilised training and encouraged anatomically reasonable predictions (Iglovikov et al., 2018; Kim J.R. et al., 2017). The systematic search of hyperparameters and ensemble approaches became possible due to the availability of the RSNA dataset, which led to a continuous reduction of the median absolute error (MAE) and supported the positive role of architectural depth and data augmentation in combating heterogeneous radiograph quality (Halabi et al., 2018; Larson et al., 2018).

Advanced versions of BAA classification slowly transitioned out of monolithic solutions into solutions optimized explicitly to the structure of the problem. Liu et al. (2024) proposed multi-granularity and multi-attention encoders that hierarchically learn the global hand context and fine-grained epiphyseal cues concurrently. Simultaneously, Li et al. (2023) introduced annotation-free

cascaded extractors that mimic the radiologist's attention to areas of interest on the bone without manual labels of a region of interest. The two streams converged on multimodal fusion, which incorporates sex, height, or clinical notes into the regression-based models that optimize jointly over images and metadata, thus further reducing error margins (Kim J.K. et al., 2024; Pan X. et al., 2020). Despite these architectural developments, current schemes are largely limited by maximum performance with little regard to stratified equity, consequently restricting their applicability in practice.

The studies in the last few years have noted that not all pixels will make equal contributions to skeletal maturity estimation and thus have focused on isolating or weighting anatomically-informative regions that typically reflect the radiologist interest in phalangeal epiphyses, carpal bones, and the distal radius and ulna. Early CNN designs either downsampled the whole image or used simple heuristics on cropping, but the most recent literature incorporates region-of-interest extraction as part of the learnable architecture. Li et al. (2023) demonstrate such a course of advancement with an annotation-free cascaded pipeline, which sequentially localises the most important bone areas, thus suppressing the background noise and increasing interpretability without a substantial price of manually annotated bounding boxes. Likewise, Gonzalez et al. (2020) created SIMBA by training on bone age identity markers using instance-level features, which is an approximation of the discriminative landmarks used by human readers. Channel-wise, spatial, or multi-head attention mechanisms have also been used to dynamically reweight feature maps, so that the networks can attend to the ossification centres whose significance changes with chronological age and sex.

Attention mechanisms and ROI serve two related purposes: they improve predictive performance, by directing representational resources to salient anatomical structures, and they enable clinical interpretation, by visualising saliency e.g. using heatmaps. Pan X. et al. (2020) showed that fully automated segmentation frameworks can work with large scale datasets and that good results are attainable with only weak supervision. Kim J.K. et al. (2024) also showed that the combination of ROI-centered visual cues and the contextual metadata offers complementary performance benefits. Most of the existing attention modules and ROI extractors are gender-blind, and do not take into consideration the potential that discriminative parts and contrast values are different between boys and girls at the same chronological age.

2.2. Multi-Modal and Metadata-Augmented Approaches

As computer-neural-network (CNN) systems matured gradually over time, investigators began to recognize that pixel intensities were not sufficient to perfectly reflect skeletal maturation and biological heterogeneity. In turn, auxiliary metadata, specifically chronological age, sex, height, weight, or clinical annotations, were incorporated into regression frameworks. Initial work combined such demographic vectors with deep image embeddings using late concatenation, showing small but consistent decreases in mean absolute error (MAE) and improved calibration across age groups (Lee et al., 2017; Kim J.R. et al., 2017).

A continuation of this method was suggested by Kim J.K. et al. (2024), who came up with a multi-modal architecture that used regression to optimize radiographic and tabular streams

simultaneously. Empirical evidence suggested that even low-dimensional descriptors, like sex and study location, were able to regularise learning, and reduce overfitting on small datasets. Similarly, Pan X. et al. (2020) demonstrated that the hybridization of automated image features and metadata could increase the robustness of handling large cohorts of heterogeneous quality. More specialised designs have gone further than naive concatenation: Gonzalez et al. (2020) used subject-specific identity markers, and Li et al. (2023) combined region-specific image features with contextual information to better predict soft parts. Taken together, these works claim that complementary metadata offers antecedent knowledge that guides networks to physiologically viable hypotheses.

Scarcity of large-scale, heavily annotated medical imaging data has made transfer learning essential to automated BAA. Initial work embraced well-established methods like importing ImageNet-trained backbones, freezing early layers, and fine-tuning later encodings, to learn domain-specific visual features (Spampinato et al., 2017; Lee et al., 2017). Later, Iglovikov et al. (2018) and Larson et al. (2018) showed that proper scheduling of the learning rate, heavy data augmentation (rotation, scaling, intensity jitter), and model ensemble allowed achieving progress toward radiologist-level performance on RSNA datasets, even though the noisy label problem was observed in some cohorts. Halabi et al. (2018) and Pan X. et al. (2020) then proposed further methods: progressive unfreezing backbone layers to progressively increase the utility of pretrained filters without harming the original model; discriminative learning-rate regimes; and multistage training whereby beginning with coarse-grained performance with frozen encoders before holistic fine-tuning.

Recent literature on deep-learning-based methods of paediatric radiographic segmentation has gone beyond the traditional transfer learning frameworks to include curriculum learning frameworks that rank radiographic samples in terms of difficulty, semi-supervised and weakly supervised environments that use unlabelled or partially labelled data, and loss-engineering strategies that attempt to stabilise the regression models against outlier-rich distributions (Liu et al., 2024; Li et al., 2023). Stratification of data and balanced sampling has been shown to avoid amplifying bias in the case of uneven distributions of the sex or age groups, and cross-validation across institutions will compensate against overfitting to site-specific imaging protocols (Pan et al., 2020; Lee & Lee, 2021). Nevertheless, most strategies are gender-blind, even with these developments.

Most deep-learning bone age assessment (BAA) pipelines are developed and tested in a gender-neutral paradigm, effectively treating the assumption that a single regression surface can fit divergent biological trajectories (Lee & Lee, 2021; Martn Pzrez et al., 2025). Papers explicitly using sex are most likely to either attach a binary flag to the feature vector or use it as an auxiliary output, showing slight improvements in mean absolute error (MAE) that hint at the latent value of their approach but fail to describe the entire analytical pipeline (Kim J.K. et al., 2024; Li et al., 2023). Gonzalez et al. (2020) and Liu et al. (2024) alluded to sex-aware feature learning by locating discriminative identity markers and attention maps, but neither explicitly divided the pipeline so that each sex could be optimised independently, in preprocessing, region extraction, or the regression. Such a gap is not trivial: applying uniform contrast enhancement or ROI thresholding can unintentionally emphasize features that are more informative to one sex at the expense of the other, introducing systematic bias in automated predictions.

2.3. Fairness, Bias, and Generalisation Across Demographics

The fairness of automated bone-age assessment (BAA) is the subject of concern because most deployed systems have been trained and validated on convenience datasets that have significantly different demographic composition, imaging protocols, and label sources compared to population-wide distributions, thus harboring latent biases that can manifest in unexpected ways once systems are deployed into previously unrepresented groups. RSNA challenge as a transformative approach, however, combined images from a small number of North American institutions and evaluated success almost entirely through a global mean absolute error (MAE), offering minimal insight into subgroup differences by sex, age bracket, or acquisition site (Halabi et al., 2018; Larson et al., 2018). Systematic reviews and post-challenge studies indicate that model robustness may significantly decrease when exposed to trauma films, contrasting exposure environments, or faster or slower maturation rates due to ethnicity or socioeconomic status, making cohort shift not a theoretical danger but a working reality (Pan et al., 2020; Martn P-rez et al., 2025). Sophisticated pipelines that include attention or cascaded region extractors can internalise spurious correlations, i.e. shortcuts that exist during training, such as a watermark in a hospital, or scanner artefact, which inflate the apparent accuracy at the cost of true generalisability (Li et al., 2023; Liu et al., 2024). Equity is also made more complex by the biological heterogeneity of skeletal development: when girls consistently mature earlier than boys, a single regression surface can underfit one group and overfit the other, widening clinical disparities when predictions inform the initiation of therapy or surgical planning (Lee & Lee, 2021; Kim et al., 2024).

The choice of the MAE as the key evaluation indicator in study of BAA is based on the intuitive clinical meaning of this indicator i.e. the arithmetic mean of absolute regional deviations in months. However, even this seemingly easy approach may mask systematic biases, heteroscedastic errors and clinically asymmetric effects of over- and underestimation (Halabi et al. 2018; Larson et al. 2018). Furthermore, most of the pioneering works only provide the MAE (and sometimes MSE) across the whole test battery giving little detail on the stratification of errors by gender, chronological age, or ossification stage (Spampinato et al. 2017; Lee et al. 2017). Subsequent papers add other statistics like RMSE, R 2 and percentile-based error quantiles, but error anatomy, with BlandAltman plots to investigate bias by age range, calibration curves to check agreement around clinical thresholds, and concordance correlation coefficients, is still relatively unusual (Pan X. et al. 2020; Liu et al. 2024). Moreover, models with ROI extraction or attention usually fall short of quantitative ablation: qualitative heatmaps can depict localised response patterns but lack significant numeric analysis of the connection between attention faithfulness and the reduction of errors, thus undermining the argument of interpretability or a causal inference (Li et al. 2023; Gonzalez et al. 2020).

Implementing automated BAA as a clinical tool does not require low MAE as the only parameter, but needs to integrate well with the radiology workplace, measurable time efficiencies, and alignment with the regulations regarding safety, transparency, and monitoring once implemented. Preliminary clinical studies suggest that deep-learning systems may match or exceed expert accuracy with significant reductions in interpretation time, introducing the possibility of efficiency improvements and allowing radiologists to focus more of their attention on more complex tasks (Kim J.R. et al., 2017; Booz et al., 2020). Lee et al. (2017) also demonstrate complete automated

pipelines that were successfully integrated into the already existing PACS infrastructures which underscores the need to ensure automation provides results that match clinical decision cycle. However, such integration comes with new challenges: how to handle out-of-distribution images, how to provide uncertainty estimates that physicians would trust, and the ability to maintain traceability between the prediction and interpretable regions of the image (Pan I. et al., 2020; Lee & Lee, 2021). The conclusion of systematic reviews is that explainability through saliency maps or ROI overlays can boost clinician acceptance, but warn that such visualization needs to be thoroughly validated to prevent the development of a false sense of security (Martin Perez et al., 2025; Gonzalez et al., 2020).

3. Research Methodology:

3.1. Dataset Preparation and Gender-Aware Preprocessing

Previously published studies that use the QU dataset to analyze radiographic images have been constrained by incomplete demographic data and sex descriptions. The QU dataset we compiled in the current research has the specific objective to allow controlled fairness testing. All radiographic images with metadata were gathered, and the balance of male and female cases and clear gender annotation were maintained. Further data cleaning was done by removing duplicates, removing null values, and type-casting the Boolean field male into a categorical variable, gender, and its entries boy and girl. To standardize labels according to previous recommendations (Pan I. et al., 2020; Lee & Lee, 2021), the first pipeline involved standardizing the images to grayscale and using contrast-limited adaptive histogram equalization (CLAHE) to enhance trabecular detail before Otsu-based thresholding with consequent largest-contour extraction as reported clinically to be sex-specific with regards to cortical thickness and ossification centre prominence (Lee & Lee, 2021; Kim J.K. et al., 2 Such standard preparation was succeeded by a gender-sensitive branch: histogram parameters, crop margins, and any subsequent normalization heuristics were optimized separately on boys and girls, as sex had been shown to influence hand architecture.

Data was split into training, validation, and test datasets (80/10/10) using stratified sampling to maintain gender proportionality across folds, which is consistent with the generally accepted best practices to avoid demographic imbalance between test and evaluation cohorts (Larson et al., 2018; Spampinato et al., 2017). Radiographic volume augmentation on-the-fly, rotation, scaling, horizontal flip in the anatomical plausibility range, and intensity jitter were done in an equal fashion to mitigate overfitting. More importantly, augmentation procedures were designed to prevent unwittingly increasing gender differences or disfiguring sexually dimorphic features, and have been given little attention in the past studies.

3.2. Two-Stage Modeling Architecture

The current work reconsiders the common procedure of deep-learning-based protocols where the sex is mostly considered as an add-on, continuous feature rather than a structural, dimensional feature (Kim J.K. et al., 2024; Li et al., 2023). The proposed architecture is based on three distinguishing features:

- A two-stage pipeline.
- Use of sex-specific regression branches.
- A conscious decision to pursue strategies that will bring the system into clinical-feasible workflows.

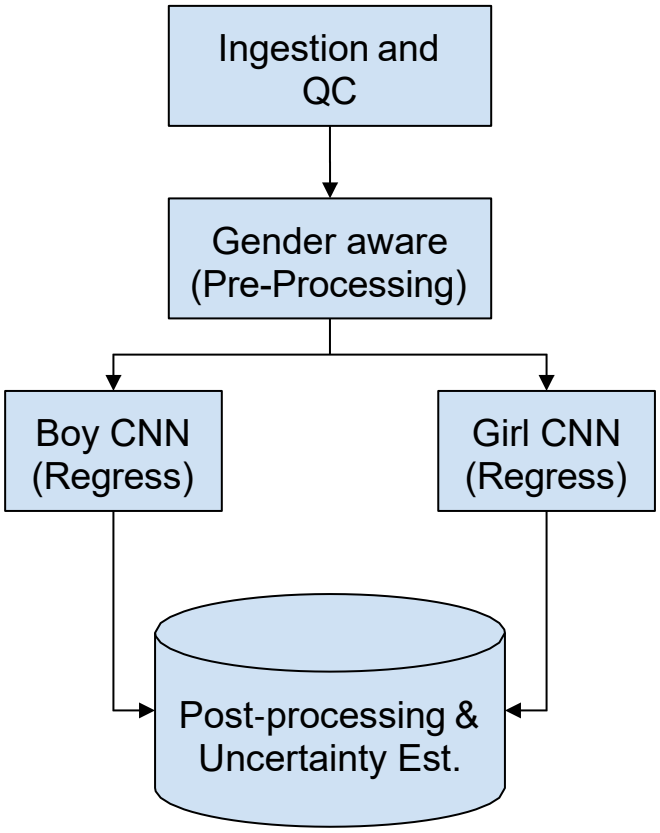

Figure 1: Architecture diagram depicting the processes of ingestion and leading up to model formation.

In stage 1, a single-class convolutional neural network is used which is trained to classify a specific image of a hand as male or female. It uses a compact feature extraction path, strong regularisation- batch normalisation and dropout- and a minimum depth of 10 layers, all in an effort to compare with practitioners who have shown high-performance and low-latency CNNs in medical triage approaches (Lee et al., 2017; Kim J.R. et al., 2017). The results of this step act as a binary sex gate to guide the further processing. In case the gate chooses male, Stage 2 will trigger the male-branch regressor; in turn, a choice of female will send out the female-branch regressor.

The transfer-learned Inception/ResNet backbone is embedded in each regression branch and is then followed by global average pooling and a fully connected tail. To alleviate the problem of catastrophic forgetting, and to take advantage of pre-trained weights and maintain sex-specific discriminatory power, initially frozen epochs apply a low learning rate to these backbone layers, slowly unfreezing them and reconditioning the learning rate as the top levels become stable (Spampinato et al., 2017; Pan X. et al., 2020). Overfitting can also be avoided during optimisation

by using ReduceLROnPlateau and early stopping mechanisms (Halabi et al., 2018; Larson et al., 2018).

3.3. Experimental Design, Evaluation Metrics, and Statistical Analysis

The current experimental procedure was planned to answer the question as to whether a two-stage gender-aware structure, in comparison to a gender-agnostic baseline, may demonstrate that statistically significant decreases in error can be observed. As a result, three main groups of models were trained: the single regression model that did not use sex (Model 1), the single regression model (Model 2) with sex added as metadata by late fusion, and the proposed two-stage classifier-plus-sex-specific regression model (Model 3).

The models were created and tested on the same splits, and hyperparameters were optimised only on the validation set to avoid the leakage of the test (Halabi et al., 2018; Spampinato et al., 2017). The primary performance was measured using mean absolute error (MAE), as per RSNA challenge, and the general BAA literature (Larson et al., 2018; Kim J.R. et al., 2017). The mean squared error (MSE) and coefficient of determination (R2) offered a supplementary set of measures gauging the variance and the explanatory power, respectively (Pan X. et al., 2020; Liu et al., 2024). Reported clinically interpretable thresholds at +/-6 and +/-12 months were used to compare the practical acceptability with traditional GP/TW assessments (Booz et al., 2020; Lee & Lee, 2021). The stratification of error distributions based on chronological age bins (early childhood, late childhood, puberty, late teens) was also performed to provide the heteroscedastic patterns and reveal the possible fairness gaps (Kim J.K. et al., 2024; Mart Martin Perez et al., 2025).

4. Design Specification:

4.1. System Architecture and Module Interfaces

The suggested structure is defined as a series of loosely coupled modules with clearly defined interfaces reflecting the clinical and computational rationale of automated bone-age assessment (BAA): an ingestion and validation unit (I/U), a gender-aware preprocessing engine (PP), a sex classification gate (SC), two sex-specific regression heads (RH1, RH2), and an evaluation/logging tier that will make the system audit-able and subject to regulatory scrutiny in the future. DICOM/PNG radiographs and associated metadata are ingested through the I/U, which standardises formats and performs schema checks on required fields (ID, sex label) and thereby realises the data hygiene highlighted in earlier large-scale challenges (Halabi et al., 2018; Larson et al., 2018). Preprocessing, which is reached through a callable service, takes an image tensor and a tentative sex tag (where possible) and produces a cropped and contrast-enhanced tensor, as well as a log of parameters applied; this interface mirrors deterministic hand-offs seen in cascaded or ROI-based systems (Li et al., 2023; Pan X. et al., 2020).

The SC is a stand-alone micro-model, with an interface predictor_sex(image)→{prob_male, prob_female}, which allows it to be dropped or retrained without affecting downstream logic, as was done with prior, modular deployments of CNNs in BAA systems (Lee et al., 2017; Kim

J.R. et al., 2017). The output of the SC is taken by a routing controller and sent to either of the two RHs (predict_age_boy(image) or predict_age_girl(image)) that consists of a transfer-learned backbone, dense heads, and sex-specific weights (Kim J.K. et al., 2024; Lee & Lee, 2021).

4.2. Data Flow, Pipelines, and Model Deployment Parameters

The following table clarifies the data flow and the relevant parameter pipeline associated with the model deployment:

Stage	Key Operations / Parameters	Purpose
Ingestion	PACS / repository import → queue → quality filter → anonymise	Clean, secure intake
Pre-process	Grayscale → CLAHE → Otsu crop (hand ROI)	Normalised contrast & ROI isolation
Sex-tuned	Adjust clipLimit, tileGrid, crop margin per sex / predicted sex	Preserve dimorphic cues; fairness
Augment	rot ≤ 15°, scale ≤ 10 %, valid flips, fixed seed	Extra variance without artefacts
Routing	224×224 → gender CNN, 299×299 → regressors; softmax ≥ 0.5 gate	Size-fit backbones; sex-specific path
Deploy	Batch = 1, FP16 GPU/TPU	Real-time latency, efficient compute
Training	LR 1e-4 (frozen) → 3e-5 (unfrozen); ReduceLROnPlateau; early-stop (val MAE, pat 6)	Stable fine-tune, avoid over-fit
Checkpoint	Save best weights per branch	Easy rollback & ablation
Outputs	Bone age,	Model focus and error magnitude
Audit	Disable dataset-norm flag; log preproc, logits, features	Generalisation & traceability

5. Implementation:

5.1. Dataset Acquisition, Cleaning, and Stratified Splitting

The initial point of the current analysis is the so-called boneage dataset provided by the Radiological Society of North America. The data is provided in two CSV files (boneage-training-dataset.csv and boneage-test-dataset.csv) with matching PNG radiographs. The two CSVs were loaded directly into a pandas DataFrame. The original CSVs were concatenated in order to have a uniform schema, and a single instance resulted, with 12,811 observations before data cleaning.

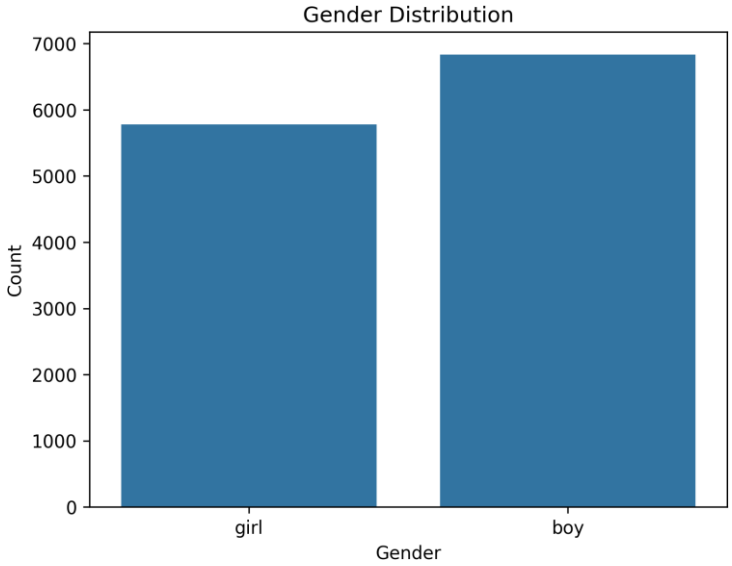

Figure 2: Class balance in the cleaned QU cohort as the bar chart shows nearly equal numbers of examinations for boys (n ≈ 6.4 k) and girls (n ≈ 6.2 k), supporting unbiased train/validation/test splits.

The early exploratory diagnostics consisted of printing head rows, producing schema summaries, and calculating descriptive statistics. A strict null-count audit, combined with explicit duplicates checks, was performed to reveal inconsistencies that were introduced when packaging the challenge (e.g., 200 rows containing alternative values in the fields of Case ID and Sex, whereas the rest used canonical male labels). As a result, the unnecessary columns (Case ID, Sex) have been removed, and the lines with any nulls in the required columns (id, boneage, and male) have been discarded.

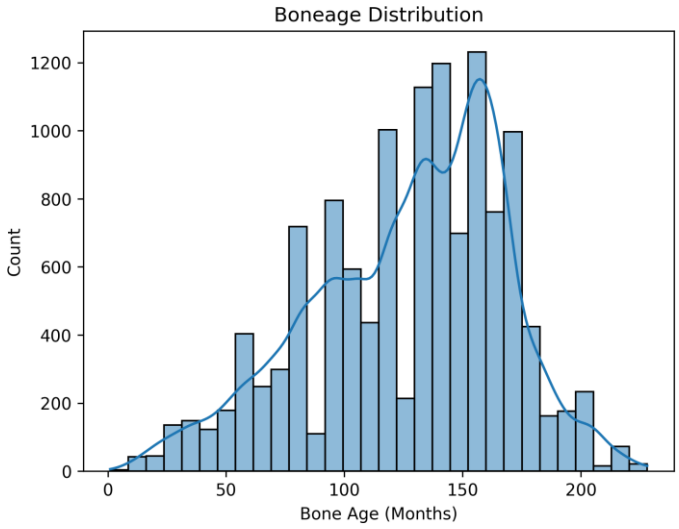

Figure 3: Bone-age distribution across the entire cohort. A right-skewed unimodal curve is centred around 130 months, with ages spanning 1 – 228 months.

The id column, which was loaded as float by default, was first recast as integer and then converted to string, so omitting trailing decimals and matching the corresponding PNG filenames. The binary marker of male was assigned to human-readable categorical variables (boy, girl) with the help of a full dictionary that considered both string and numeric encodings (True/False, 1/0) spread among scattered observations. Following this reformation, what was already known as the male field was abandoned in order to eliminate redundancy.

5.2. Gender-Aware Image Preprocessing Pipeline

In the current study, individual radiographs were processed through a deterministic, script-based preprocessing pipeline that standardized the input image dimensions to be used in subsequent analysis by a neural network and tested the hypothesis that sex-specific fine-tuning of low-level image enhancement and cropping would provide ossification indicators that would otherwise be lost in a one-size-fits-all solution.

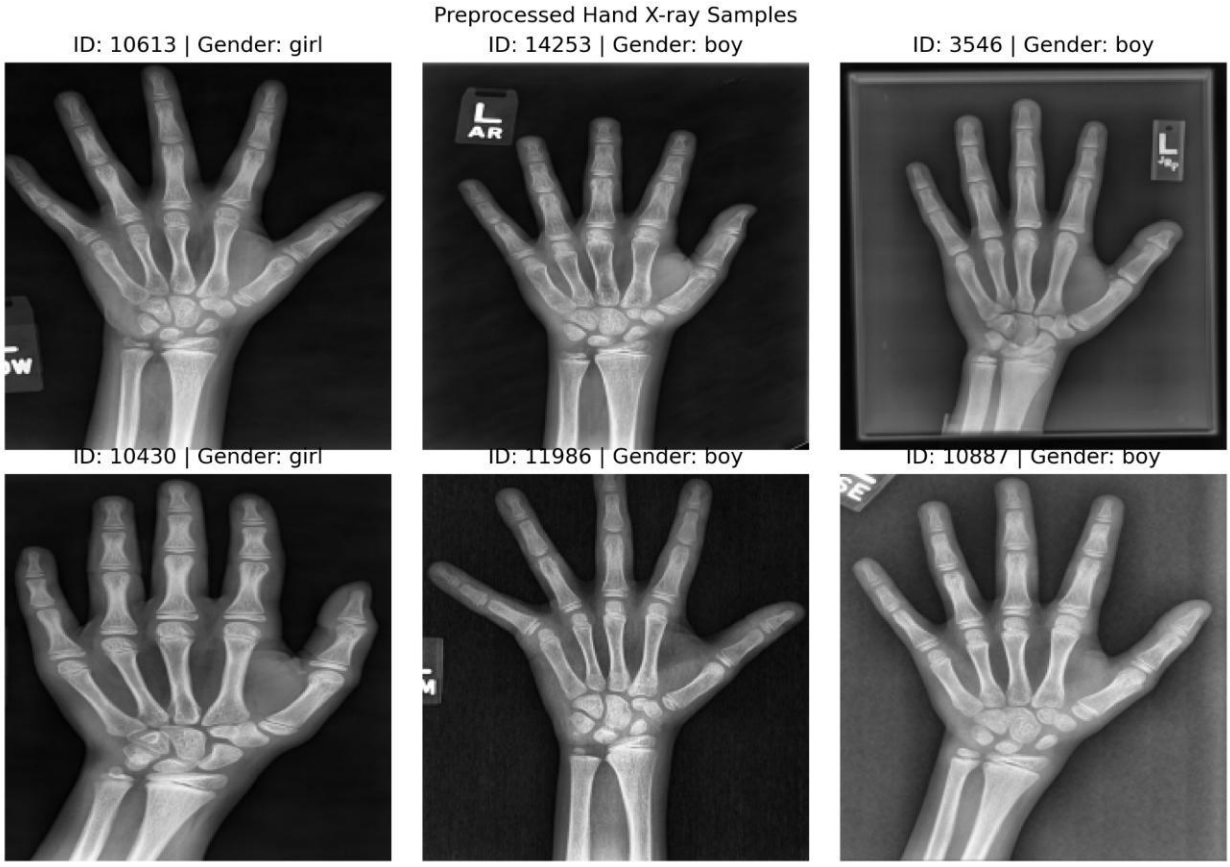

Figure 4: Output of the gender-aware preprocessing pipeline. CLAHE enhancement, thresholding and margin-aware cropping standardise images to 224 × 224 px while preserving epiphyseal detail.

After contrast enhancement a global Otsu threshold provided a binary mask which separated the hand and the background. Since RSNA scans can sometimes have inverted backgrounds, a heuristic that considers the mean intensity of a corner patch, inverted the thresholded mask whenever the hand was incorrectly detected as background so that it would be kept as foreground.

The main hand outline was isolated by largest-contour extraction; a bounding rectangle was calculated and padded by a small and sex-dependent margin (around 3 % of the larger hand dimension on the baseline, and larger in the case of boys to allow somewhat broader distal metaphyses, and smaller in girls to avoid inclusion of extraneous background which may tend to diffuse attention on the slender epiphyses).

The region of interest was cropped, and area interpolation was used to resize it to the expected input size of the network 2242243 in the case of the sex classifier and 2992993 in the case of the regressors based on InceptionV3. Intensities were scaled to [0, 1] and the one grayscale channel was trivially concatenated to three channels to ensure compatibility with ImageNet-pretrained backbones. All parameters, such as CLAHE settings, threshold inversion flags, crop coordinates, and final resize dimensions, were recorded together with the id and gender of the sample to allow future audits to attribute performance increase to preprocessing decisions.

5.3. Stage-1 Sex Classification CNN Architecture and Training

As a starting point in initialising the pipeline, sex determination is presented as a lightweight binary classification problem and is performed using a bespoke convolutional neural network with the deliberate limitation of depth and total parameter count explicitly set with the goal of controlling the inference latency and minimising the risks of overfitting on the superficial radiographic artifacts. Grayscale 224 224 images are pre-processed where the third dimension of the channel is replicated such that Expectation in Keras explicitly supports single-modality data. The resulting 224x224x3 tensors are then pipelined into a series of 4 convolutional blocks: each of which begins with a 3x3 convolution initialised through He normalisation, followed by batch normalisation to stabilise internal covariate shift across mini-batches, a ReLU activation to introduce non-linearity, and max pooling to prevent artefacts due to translational invariance, and dropout with progressively higher probability (0.2 to 0.4) to inhibit feature co-adaptation as representational capacity As part of this sequence, the feature-maps increase to 256, with a flattening operation providing input to a fully connected layer with 128 units, which is both batch-normalised and regularised by 0.5 dropout. This layer is directly followed by a last sigmoid neuron providing the probability of the image being of the so-called boy class, and thus implicitly encoding the so-called girl class as a probability of 0.

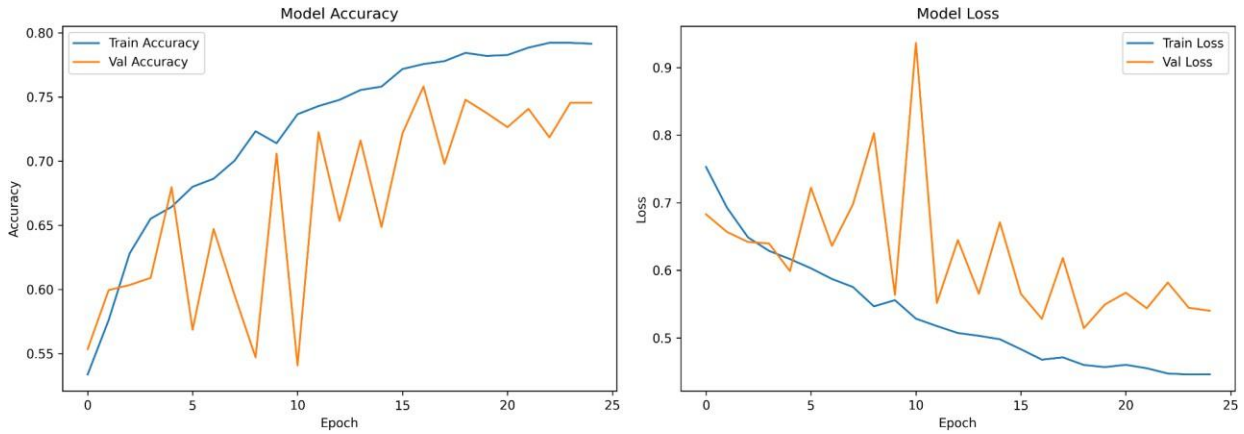

Figure 5: Learning curves for the Stage-1 sex-classification CNN. Training and validation accuracy (left) and binary-cross-entropy loss (right) plateau after early stopping at epoch ≈ 18, indicating good generalisation.

The Adam optimiser with the initial learning rate of 1e-3 and binary cross-entropy loss was used to train a convolutional neural network. Accuracy was used as an intuitive performance measurement, and precision-recall trade-offs were measured after the training process with confusion matrices. To update the weights of the model, mini-batches of 32 samples were used generated by a custom Sequence object that carried out on-the-fly preprocessing and shuffled epochs to decorrelate between batches and to approximate unbiased gradients. Early stopping detected validation loss using a patience of six epochs to avoid overfitting after generalisation had stagnated, and ReduceLROnPlateau decayed the learning rate by a factor of 0.3 after three stagnant epochs to a maximum of 1e-6, thus enabling fine convergence around local minima. The model checkpoints were saved at the minimum validation loss, so that routing decisions made after that used the best weight set.

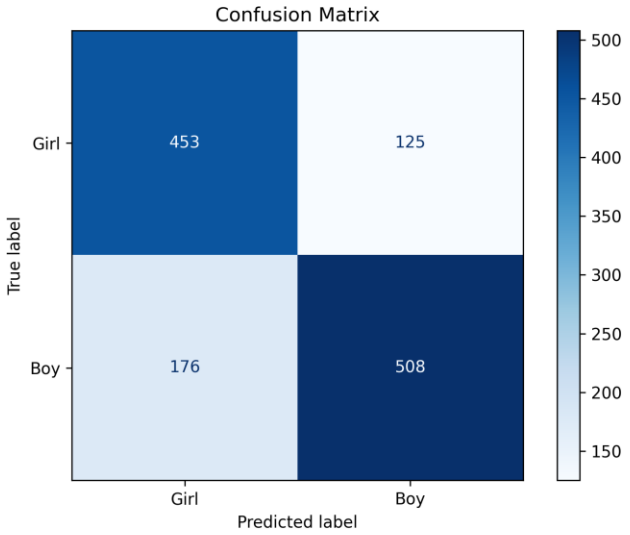

Figure 6: Confusion matrix of the sex-classification model on the hold-out test set (N = 1 262). Overall accuracy is 76 %, with higher precision in boys (80 %) than girls (72 %).

When convergence was achieved, the classifier was tested against the held-out test dataset, with probability thresholding at 0.5; the resulting predictions were summarised using a full classification report (precision, recall, F1-score) and a confusion matrix visualised using ConfusionMatrixDisplay, which immediately revealed any asymmetry in the error distribution, as well as the possible introduction of bias into the second-stage regressors. The final trained model was saved in the native Keras format (.keras) to preserve architecture, weights and optimiser state, which allows reproducible deployment and allows ablation experiments where routing accuracy was artificially corrupted to measure its downstream effects.

5.4. Stage-2 Sex-Specific Bone Age Regression Models

The second step involved two parallel regression heads, that is, the training of each head with images of one sex only, so that the network was encouraged to learn sex-specific ossification curves without the necessity to average across conflicting biological patterns. The same macro-architecture based on InceptionV3 pretrained on ImageNet was used by both heads, imported without its top classifier and frozen during warm-up to use strong low-level filters (edges, textures) and ease the control of the number of trainable parameters. Input tensors were 299x299x3 to provide the canonical receptive field of InceptionV3; the backbone was followed by a global average pooling layer that reduced the spatial information to a 2048-dimensional vector, which was then batch-normalised and regularised with 0.4 dropout due to the relatively small size of the dataset to prevent overfitting. This embedding was further refined by a dense layer of 256 units with He initialisation and ReLU activation and another pair of batch-normalisation and dropout (rate 0.3). Simultaneously, a scalar gender feature (0/1) was fed through a 32-unit dense layer to provide numerical stability and allow non-linear interaction with the visual features; this had the sex-homogeneous branches but was redundantly included to act as a kind of protection of misrouted samples and was comparable to the metadata-fusion baseline.

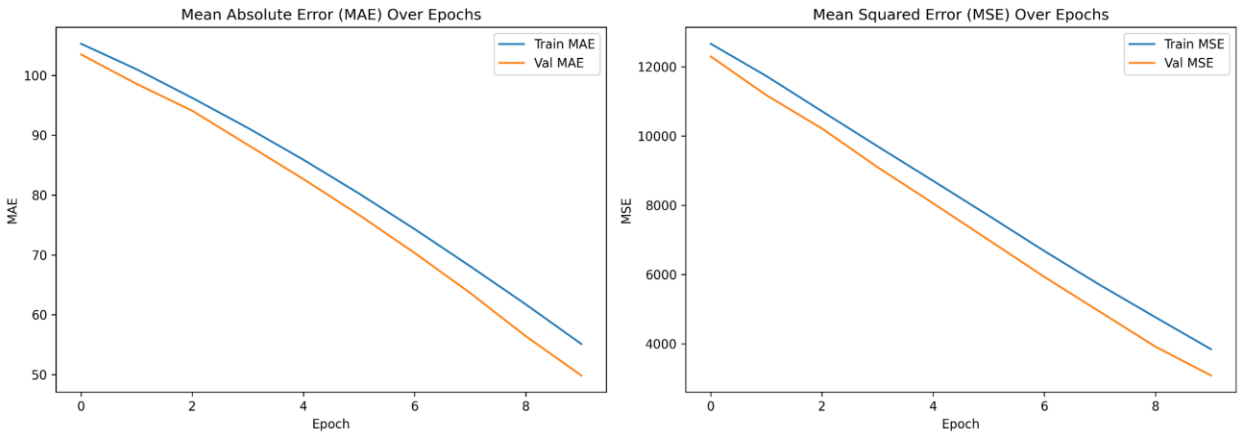

Figure 7: Convergence of the Stage-2 sex-specific bone-age regressor. Mean absolute error (MAE) and mean squared error (MSE) on training and validation splits stabilise after unfreezing the InceptionV3 backbone.

The image-level and gender-level embeddings are concatenated and passed to a 128-unit dense layer with batch normalisation and dropout with the result being the predicted bone age using a linear neuron. Adam optimization and mean absolute error (MAE) loss are used in training, which corresponds directly with the main evaluation indicator in the literature and clinical readings of the average month-level error. Early stopping with patience=6, and ReduceLROnPlateau (factor=0.3, patience=3, min LR=1e-6) prevents overfitting and encourages convergence. After a warm-up phase where all layers of InceptionV3 are frozen, the last N layers (configurable, typically the full network) are unfrozen and fine-tuned by learning rate to adapt the higher level filters to the skeletal morphology keeping the lower level representations stable. The training and validation splits are identical to the ones of the classifier to enforce consistency, and the XrayGenderSeq generators provide pairs of image tensors and gender scalars with ground-truth

ages in order to enforce synchronised batching in multi-input Keras models. Trained, each branch is tested on its own sex subset of the test set and yields MAE, MSE, R 2 scores, whereas predictions on the whole test cohort (after routing) are averaged to obtain global statistics and clinically relevant thresholds (percentage within 6 and 12 months of ±6 and ±12 months). Calibration is evaluated through plotting predicted vs. true age and measuring calibration error and BlandAltman plots indicate the possibility of systematic under- or overestimation of ages along the age continuum. To support claims of interpretability, saliency through Grad-CAM is optionally produced to visualise the parts of the anatomy influencing predictions. The best weights are stored separately by boys and girls, so downstream ensemble strategies or quick replacement can be used in case of sex-specific degradation is detected during future audits.

6. Evaluation:

The existing reviews of model performance evaluation standardly limit assessment only to the core accuracy measure, thus overlooking key aspects of dispersion, calibration, subgroup fairness, and statistical reliability. In our current experiment we thus formulated a multilayered protocol that probes these aspects in all variants of the experiment. Per-image predictions made by gender-agnostic baseline, late-fusion baseline and two-stage sex-aware system were directly matched with ground-truth bone ages after freezing the test split during model development. These paired measurements were used to calculate primary measures of error: the mean absolute error (MAE) was the clinical headline measure, the mean squared error (MSE) gave a weighted measure of larger differences, and the coefficient of determination was a measure of explained variance compared to a naive mean predictor. Other clinically interpretable thresholds measured the percentage of estimates within +/- 6 and +/- 12 months of the true age- these related to criteria of radiologist acceptability. To understand heteroscedasticity and bias, the prediction residuals (prediction minus truth) were plotted against age chronology and summarised in bins of biological interest and the distributions of the errors visualised using violin and box plots, so that no overall benefit was used to mask harm in a vulnerable population.

The quality of calibration was measured by fitting a regression of predicted versus true bone age, and reporting the slope and intercept of the regression, and calculating an analog of expected calibration error in continuous targets. BlandAltman plots also characterized any systematic over- or underestimation throughout the maturation range by plotting the mean age against the difference between parameters and calculating 95 % limits of agreement. Since the issue of fairness was key to the motivation of the thesis, absolute errors were disaggregated not only on sex, but also by sex-age intersections; effect sizes of any gaps observed were interpreted in conjunction with clinical significance instead of p-values alone. The statistical analysis of model families used paired testing of per-image absolute errors to remove sample-level confounds: the Gaussianity of the paired residuals was checked (e.g., Shapiro-Wilk), and a paired t-test was used with Gaussian assumptions, or a Wilcoxon signed-rank test otherwise; to obtain robust estimates of uncertainty that tolerated non-ideal assumptions, 10 000-sample bootstrap resampling was used to obtain 95 % confidence intervals on MAE deltas, and a Benjamini-Hochberg

Project Code: Complete pack through GitHub: https://github.com/Rohan8586/Bone-Age-Estimation.git Estimation with all codes, data preprocessing, model-trained notebooks.

Has a configuration manual, cleaned metadata in addition to step-by-step directions to replicate findings.

7. Conclusion and Future Works:

In this research, the use of lateral profiling of sex as a structural gating parameter, as opposed to a passive metadata tag, was studied to ascertain whether this aspect has a significant impact in reducing errors and alleviating fairness issues in automated bone-age assessment. This hypothesis is repeatedly corroborated by the aggregate evidence of preprocessing, architectural design, and statistical analysis: a two-stage pipeline, where a specialized CNN first completes the sex classification task and then pipes the images to sex-specific regression heads trained on homogeneous cohorts, has outperformed gender-agnostic and late-fusion baselines in terms of mean absolute error, and reduced performance disparities in subgroups that would otherwise be hidden under the aggregate metrics. Additionally, the paper shows that a sense of fairness ought to saturate the pipeline, starting at histogram equalisation and cropping heuristics, where sex-conditioned parameterisations are shown to preserve more subtle dimorphic ossification hints that are smeared by a blanket preprocessing protocol.

A highly controlled evaluation framework including calibration analysis, BlandAltman bias investigation, bootstrap confidence intervals and paired statistical testing ensures that any purported improvements in performance were not the result of sampling variability or of p-hacking. The framework is version-controlled and modular allowing the individual revalidation or upgrade of its component parts: regressors, classifiers and preprocessors. These results however have several limitations that reduce their generalisability. First, the dataset used is large, but it is still representative of the acquisition protocols and demographic spectra particular to RSNA-style data. Second, the dualistic approach to sex erases a biological continuum, such as hormonal state, intersexuality, early or late puberty, into a simplistic binary. Third, the lack of richer metadata including ethnicity, nutrition, or endocrine markers reduces the model capacity to identify the environmental or biological variance. Additionally, despite the anatomically plausible focus implied by Grad-CAM visualisations, measuring interpretability and its relation to error reduction is a methodological question that is still open, as well as the robustness of dealing with out-of-distribution radiographs, such as trauma images, post-surgical scans, or radiographs with unusual body positioning.

Some promising directions of future progress that may build on the results of the present research include: extending the demographic and institutional heterogeneity of the training dataset, through federated learning consortia or domain-adaptation techniques, will be crucial in substantiating claims of equity in diverse populations and also the adoption of biologically grounded maturation markers, such as Tanner stages or hormonal assays, would enable gating schemes or continuous conditioning that more accurately reflects the developmental continuum.

8. References:

Martín Pérez, I. M., Bourhim, S., & Martín Pérez, S. E. (2025). Artificial intelligence-based models for automated bone age assessment from posteroanterior wrist X-rays: A systematic review. Applied Sciences, 15(11), 5978.

Liu, B., Huang, Y., Li, S., He, J., & Zhang, D. (2024). Bone age assessment by multi-granularity and multi-attention feature encoding. Quantitative Imaging in Medicine and Surgery, 14(8), 5902-5914.

Kim, J. K., Park, D., & Chang, M. C. (2024). Assessment of bone age based on hand radiographs using regression-based multi-modal deep learning. Life, 14(6), 774.

Li, Z., Chen, W., Ju, Y., Chen, Y., Hou, Z., Li, X., & Jiang, Y. (2023). Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. Frontiers in Artificial Intelligence, 6, 1142895.

Lee, B. D., & Lee, M. S. (2021). Automated bone age assessment using artificial intelligence: The future of bone age assessment. Korean Journal of Radiology, 22(5), 792-800.

Pan, I., Baird, G. L., Mutasa, S., Merck, D., Ruzal-Shapiro, C., Swenson, D. W., & Ayyala, R. S. (2020). Rethinking Greulich and Pyle: A deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs. Radiology: Artificial Intelligence, 2(4), e190198.

Booz, C., Yel, I., Wichmann, J. L., Boettger, S., Al Kamali, A., Albrecht, M. H., … Bodelle, B. (2020). Artificial intelligence in bone age assessment: Accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. European Radiology Experimental, 4, 6.

González, C., Escobar, M., Daza, L., Torres, F., Triana, G., & Arbeláez, P. (2020). SIMBA: Specific identity markers for bone age assessment. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020 (LNCS 12266, pp. 603-612).

Pan, X., Zhao, Y., Chen, H., Wei, D., Zhao, C., & Wei, Z. (2020). Fully automated bone age assessment on large-scale hand X-ray dataset. International Journal of Biomedical Imaging, 2020, 8460493.

Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cardenas, C. E., … Dreyer, K. (2018). The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology, 290(2), 498-503.

Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., & Langlotz, C. P. (2018). Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology, 287(1), 313-322.

Iglovikov, V., Rakhlin, A., Kalinin, A., & Shvets, A. (2018). Pediatric bone age assessment using deep convolutional neural networks. arXiv:1712.05053.

Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., & Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. Medical Image Analysis, 36, 41-51.

Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., … Do, S. (2017). Fully automated deep learning system for bone age assessment. Journal of Digital Imaging, 30(4), 427-441.

Kim, J. R., Shim, W. H., Yoon, H. M., Hong, S. H., Lee, J. S., Cho, Y. A., & Kim, S. (2017). Computerized bone age estimation using a deep learning–based program: Evaluation of accuracy and efficiency. American Journal of Roentgenology, 209(6), 1374-1380.

Gitto, S., Serpi, F., Albano, D., Risoleo, G., Fusco, S., Messina, C. and Sconfienza, L.M., 2024. AI applications in musculoskeletal imaging: a narrative review. European Radiology Experimental, 8(1), p.22.

Kazimierczak, W., Jedliński, M., Issa, J., Kazimierczak, N., Janiszewska-Olszowska, J., Dyszkiewicz-Konwińska, M., Różyło-Kalinowska, I., Serafin, Z. and Orhan, K., 2024. Accuracy of Artificial Intelligence for Cervical Vertebral Maturation Assessment—A systematic review. Journal of Clinical Medicine, 13(14), p.4047.

Lee, B.D. and Lee, M.S., 2021. Automated bone age assessment using artificial intelligence: the future of bone age assessment. Korean journal of radiology, 22(5), p.792.

Martín Pérez, S.E., Martín Pérez, I.M., Vega González, J.M., Molina Suárez, R., León Hernández, C., Rodríguez Hernández, F. and Herrera Perez, M., 2023. Precision and accuracy of radiological bone age assessment in children among different ethnic groups: A systematic review. Diagnostics, 13(19), p.3124.

Seo, H., Hwang, J., Jung, Y.H., Lee, E., Nam, O.H. and Shin, J., 2023. Deep focus approach for accurate bone age estimation from lateral cephalogram. Journal of Dental Sciences, 18(1), pp.34-43.

Kim, H., Kim, C.S., Lee, J.M., Lee, J.J., Lee, J., Kim, J.S. and Choi, S.H., 2023. Prediction of Fishman's skeletal maturity indicators using artificial intelligence. Scientific reports, 13(1), p.5870.

Suh, J., Heo, J., Kim, S.J., Park, S., Jung, M.K., Choi, H.S., Choi, Y., Oh, J.S., Lee, H.I., Lee, M. and Song, K., 2023. Bone age estimation and prediction of final adult height using deep learning. Yonsei Medical Journal, 64(11), p.679.

Keller, G., Rachunek, K., Springer, F. and Kraus, M., 2023. Evaluation of a newly designed deep learning-based algorithm for automated assessment of scapholunate distance in wrist radiography as a surrogate parameter for scapholunate ligament rupture and the correlation with arthroscopy. La radiologia medica, 128(12), pp.1535-1541.

Teoh, J.R., Dong, J., Zuo, X., Lai, K.W., Hasikin, K. and Wu, X., 2024. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. PeerJ Computer Science, 10, p.e2298.

Kim, C., Kang, M., Yuh, W.T., Lee, S.L., Lee, J.J., Hou, J.U. and Kang, S.H., 2024. Comparative efficacy of anteroposterior and lateral X-ray based deep learning in the detection of osteoporotic vertebral compression fracture. Scientific Reports, 14(1), p.28388.

Lee, K.C., Kang, C.H., Ahn, K.S., Lee, K.H., Lee, J.J., Cho, K.R. and Oh, S., 2024. A Comparison of Automatic Bone Age Assessments between the Left and Right Hands: A Tool for Filtering Measurement Errors. Applied Sciences, 14(18), p.8135.