Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Demographic Baseline for Pediatric Forearm Fracture Localization: A Machine- Learning Benchmark Using the PediURF Dataset

Volkan Kızılkaya

dr_volkankizilkaya@hotmail.com

MetroLife Hospital

Murat Aslanyürek

Kırklareli University

Canan Aslanyürek

Kırklareli University

Ali Levent

MetroLife Hospital

Research Article

Keywords: pediatric forearm fracture, machine learning, demographic baseline, class imbalance, calibration, multimodal learning, PediURF

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11195635/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Image-based artificial intelligence models for pediatric fracture localization may exploit demographic signals that are correlated with fracture distribution. Quantifying the discrimination available from demographic metadata alone provides a reference for interpreting image-only and multimodal models.

Methods

We performed a secondary analysis of the open PediURF dataset, preserving its original patient-level training (n = 4,212) and test (n = 1,053) split. Only age and sex were used as predictors. Logistic regression (LR), Gaussian naive Bayes, radial-basis-function support vector machine, random forest, and gradient boosting were tuned exclusively in the training set. Standard training, inverse-frequency cost-sensitive training, and a post hoc equal-prior decision rule applied to standard LR were compared. The primary discrimination measure was macro one-vs-rest ROC-AUC. Test uncertainty was quantified with 2,000 bootstrap resamples. Paired ΔAUC comparisons were Holm-adjusted. Calibration, class-specific recall, and a training-estimated saturated age-by-sex reference were also evaluated. A test-label-derived saturated model was reported only as an optimistic empirical oracle reference, not as a deployable or theoretical upper bound.

Results

Among 5,265 children with confirmed fractures, 64.1% were distal, 25.1% midshaft, and 10.9% proximal. Standard LR achieved a test macro AUC of 0.704 (95% CI 0.674–0.734), comparable with the training-estimated saturated reference (0.698). Random forest (0.701) and gradient boosting (0.699) were within the prespecified analytic equivalence framework relative to LR, whereas SVM was lower (0.648; ΔAUC −0.056; Holm-adjusted $p < 0.004$). Standard models had poor proximal-class recall. Cost-sensitive training increased proximal recall to 0.450 for most algorithms but worsened probability calibration. Modifying the LR decision rule using equal-prior rescaling reproduced the balanced classification profile (balanced accuracy 0.528; proximal recall 0.450) while retaining the original LR probabilities for probability-based discrimination and calibration assessment.

Conclusions

Age and sex contain moderate information about fracture location in PediURF, but flexible machine-learning algorithms did not materially improve on a parsimonious LR reference. The principal effect of imbalance handling was on class decisions rather than on the amount of demographic information

available. This demographic benchmark should be interpreted as a non-imaging reference for future image-only and multimodal studies, not as a diagnostic substitute for radiography.

1. Background

Pediatric forearm fractures are among the most frequent skeletal injuries in children and adolescents, but their distribution is not uniform across development. Both fracture incidence and anatomical pattern vary with age and sex, reflecting changes in skeletal maturation and injury exposure during childhood [1, 2]. This demographic structure is relevant to artificial-intelligence research because age- and sex-related information may contribute to fracture localization even before radiographic features are considered.

Artificial-intelligence methods have shown strong performance in pediatric fracture assessment, and systematic reviews indicate growing evidence for radiograph-based AI in children [3, 4]. Open radiographic datasets have also accelerated model development [5]. The recently released PediURF dataset contains paired anteroposterior and lateral radiographs of pediatric ulna and radius fractures categorized as proximal, midshaft, or distal, together with age and sex metadata [6]. Its original dual-view URFNet experiments demonstrated high image-based classification performance [6]. However, the amount of localization information available from the accompanying demographic metadata alone has not been quantified.

This distinction matters for both benchmarking and multimodal model development. A demographic-only model is not intended to replace radiography. Rather, it provides a non-imaging baseline against which the incremental contribution of image features can be judged, helps identify the extent to which demographic correlates could contribute to shortcut learning, and may provide prior probabilities for future image-plus-metadata models [7, 8]. Because PediURF includes only children with confirmed fractures, all probabilities in the present analysis are conditional on fracture presence and must not be interpreted as probabilities of having a fracture.

A second methodological issue is class imbalance. Proximal fractures constitute only about one tenth of PediURF [6]. In prediction modeling, class weighting can improve minority-class classification at a chosen decision threshold while altering predicted probabilities and degrading calibration [9]. Discrimination, classification performance, and calibration therefore answer different questions and should be evaluated separately [9, 10].

Accordingly, this study aimed to (1) quantify the fracture-localization information contained in age and sex alone; (2) determine whether flexible machine-learning algorithms extract materially more demographic signal than a parsimonious multinomial logistic-regression reference; (3) characterize a training-estimated saturated age-by-sex benchmark and a clearly labeled, non-deployable test-label oracle reference; (4) compare cost-sensitive training with a decision-level equal-prior adjustment; and (5) provide a reproducible demographic baseline for future image-only and multimodal studies using PediURF.

2. Methods

2.1 Study design and data source

This was a retrospective secondary analysis of the publicly available PediURF metadata [6]. The dataset was collected at Shenzhen Children's Hospital between 2013 and 2024, de-identified, and released for research. Only children with confirmed fractures are included. Each patient is assigned to one localization class (proximal, midshaft, or distal) based on expert radiologist annotation [6]. We preserved the original patient-level stratified training set (n = 4,212) and independent test set (n = 1,053); patient identifiers did not overlap between the two sets. No radiographic pixels were used in the present analysis.

2.2 Data preprocessing

Age was encoded as years (Y1–Y17) for most patients and as months (M5–M12) for 36 infants. Month-coded ages were divided by 12 for continuous-model analyses. Sex was coded as a binary predictor. Patient folder identifiers were not used as predictors. No missing values or duplicate patient identifiers were identified. For saturated age-by-sex cell analyses, the source age code was retained so that distinct infant-month categories were not collapsed; a sensitivity interpretation is provided because this representation differs from the continuous-year encoding used by the fitted classifiers.

2.3 Descriptive and exploratory analyses

Age distributions across localization classes were compared using the Kruskal–Wallis test with epsilon-squared effect size; sex and localization were assessed with the chi-square test and Cramér's V. Training and test sets were compared descriptively for age, sex, and class distribution. A multinomial logistic model with natural-spline age terms and an age-by-sex interaction was used to visualize conditional localization probabilities. These analyses were exploratory and were not used to select the final test-set model.

2.4 Demographic reference models and empirical oracle

Two saturated age-by-sex references were constructed. First, a deployable training-estimated saturated reference used class frequencies within training-set age-by-sex cells, shrunk toward the overall training class proportions with a total of 1.5 Dirichlet pseudo-observations. Unseen test cells received the training class proportions. Because sparse cells introduce sampling noise, this is an empirical reference rather than a strict upper bound.

Second, an oracle reference used class proportions calculated from the test set itself. This deliberately uses test labels and is therefore non-deployable and optimistically biased. It is reported only to describe the discrimination obtainable from the observed age-by-sex cells in this particular test sample. We avoid interpreting it as a theoretical population upper bound.

2.5 Candidate algorithms and model selection

Five classifier families were evaluated: multinomial logistic regression (LR), Gaussian naive Bayes (GNB), radial-basis-function support vector machine (SVM), random forest (RF), and gradient boosting (GB). LR and SVM predictors were standardized. Hyperparameters were selected exclusively within the training set using five-fold stratified cross-validation with macro one-vs-rest ROC-AUC as the optimization criterion. The independent test set was not used for hyperparameter selection or model selection.

The tuning grids were: LR C={0.01,0.1,1,10,100}; GNB var_smoothing={10^-9,10^-6,10^-3}; SVM C={0.1,1,10} with gamma={scale,0.1,1}; RF maximum depth={3,5,unlimited} and minimum leaf size={1,10,50} with 300 trees; and GB n_estimators={100,300}, learning rate={0.03,0.1}, and maximum depth={2,3}. After tuning, each configuration was refitted on the complete training set and evaluated once on the official test set. Repeated 5×5-fold stratified cross-validation in the training set was used to describe internal variability.

LR was designated the parsimonious reference a priori because it is simple, probabilistic, and clinically interpretable. This choice also reflects evidence that more complex machine-learning methods do not necessarily outperform logistic regression in clinical prediction settings [11]. Comparisons with flexible models therefore tested whether additional algorithmic complexity extracted materially more information from the same two predictors; they were not a post hoc competition to select the numerically highest test AUC.

2.6 Class-imbalance strategies

Three strategies were compared. Standard training assigned equal weight to all observations. Cost-sensitive training used inverse-frequency class weighting (class_weight='balanced' for LR, SVM, and RF; balanced sample weights for GB; equal class priors for GNB). Because imbalance corrections can change probability calibration even when they improve minority-class classification [9], calibration was evaluated separately. Finally, standard-LR probabilities were transformed only for class assignment by rescaling them to equal class priors, $p'_k \propto p_k/\pi_k$, followed by argmax classification, following the general prior-adjustment principle described by Saerens et al. [12]. The underlying standard-LR probabilities were retained for probability-based AUC, Brier, and calibration analyses; therefore, no claim is made that the transformed p' probabilities themselves are calibrated.

2.7 Outcomes and statistical analysis

The primary discrimination metric was macro one-vs-rest ROC-AUC. Secondary metrics were accuracy, balanced accuracy, macro-F1, Matthews correlation coefficient (MCC) [13], multiclass Brier score, and class-specific recall. Test-set 95% confidence intervals were estimated using 2,000 bootstrap resamples. Within each training strategy, algorithms were compared with the LR reference using paired bootstrap differences in macro AUC; four p values were adjusted with the Holm procedure. Class-specific AUC comparisons used DeLong's method [14].

Equivalence was treated as a supportive, exploratory analysis because the margin was not preregistered. The main margin was ± 0.02 in macro AUC, with equivalence defined when the entire 95% CI for ΔAUC

lay within the margin. To show dependence on the post hoc margin, we additionally interpreted the same confidence intervals against ± 0.01 and ± 0.03 sensitivity margins. No new test-set fitting was performed for this sensitivity analysis.

Calibration was evaluated one-vs-rest for each class using calibration curves, calibration slope, calibration-in-the-large (intercept with slope fixed at 1), and expected calibration error (ECE) calculated over deciles of predicted probability. Calibration was interpreted as a distinct property from discrimination, consistent with contemporary prediction-model guidance [10]. Reporting was informed by TRIPOD + AI [15]. Analyses used Python 3.12 with pandas, NumPy [16], SciPy [17], statsmodels, scikit-learn, and Matplotlib.

3. Results

3.1 Cohort and demographic gradients

The analysis included 5,265 patients: 3,374 (64.1%) distal, 1,319 (25.1%) midshaft, and 572 (10.9%) proximal fractures; 74.3% were male. Training and test sets were similar in age (Mann–Whitney U p = 0.649) and class distribution (p = 0.931); the sex-distribution difference was borderline (p = 0.063) (Table 1). [Insert Table 1 here]

Table 1
Characteristics of the study population by fracture location (N = 5,265)

Location	N (%)	Age, mean ± SD	Age, median (IQR)	Female, n (%)	Male, n (%)
Distal	3374 (64.1)	8.50 ± 3.26	9 (6–11)	706 (20.9)	2668 (79.1)
Midshaft	1319 (25.1)	5.79 ± 3.41	5 (3–8)	389 (29.5)	930 (70.5)
Proximal	572 (10.9)	7.27 ± 2.84	7 (5–9)	257 (44.9)	315 (55.1)
Total	5265 (100.0)	7.69 ± 3.45	8 (5–10)	1352 (25.7)	3913 (74.3)

Age is shown in years; 36 infants encoded in months in the source metadata were converted to years for continuous analyses.

Localization varied with age (Kruskal–Wallis H = 578.5, $p < 0.001$; $\varepsilon^2=0.110$) and sex ($\chi^2=161.1$, $p < 0.001$; Cramér's V = 0.175). The spline model showed a nonlinear age effect (likelihood-ratio $\chi^2=63.2$, $p < 0.001$) and an age-by-sex interaction (p = 0.007). Estimated proximal localization probability ranged from approximately 2% in 1-year-old boys to 22–23% in girls aged 5–8 years.

3.2 Feature ablation and empirical references

Sex alone produced macro AUC 0.564 and did not change majority-class decisions. Age alone produced AUC 0.672. Adding sex increased macro AUC to 0.702 (Δ = 0.030, 95% CI 0.009–0.051; p = 0.004), driven

mainly by the proximal-vs-rest AUC (0.591 to 0.676; DeLong p = 0.004). A spline age specification did not materially increase overall discrimination (AUC 0.703).

The training-estimated saturated age-by-sex reference yielded macro AUC 0.698 (95% CI 0.668–0.726). Standard LR yielded 0.704, a difference of + 0.006 (95% CI − 0.004 to + 0.016; p = 0.232), consistent with the fact that the saturated training estimate is noisy in sparse cells. The non-deployable oracle reference was 0.732 (95% CI 0.705–0.758). Because it was calculated from test labels, this value is interpreted only as an optimistic empirical reference for this test sample (Table 2). [Insert Table 2 here]

Table 2
Feature ablation and demographic reference performance on the independent test set (n = 1,053)

Feature/model	Accuracy	Balanced acc.	Macro-F1	MCC	Macro AUC	Brier	Proximal recall
Majority reference	0.643	0.333	0.261	0.000	0.500	0.512	0.000
Sex	0.643	0.333	0.261	0.000	0.564	0.507	0.000
Age	0.680	0.409	0.394	0.253	0.672	0.460	0.000
Age + Sex	0.680	0.409	0.394	0.253	0.702	0.457	0.000
Spline age + Sex	0.680	0.409	0.394	0.253	0.703	0.455	0.000
Saturated, training-estimated	0.678	0.407	0.391	0.247	0.698	0.458	0.000
Oracle reference†	0.681	0.409	0.394	0.256	0.732	0.442	0.000

†Uses test-set class proportions within age×sex cells and is intentionally non-deployable.

3.3 Algorithm benchmark and imbalance handling

Under standard training, LR achieved macro AUC 0.704 (95% CI 0.674–0.734) and training repeated-CV AUC 0.681 ± 0.014. RF (0.701) and GB (0.699) produced the same balanced accuracy and macro-F1 as LR. GNB was statistically lower than LR (ΔAUC − 0.010; Holm-adjusted p = 0.006), although the 95% CI remained within the ± 0.02 exploratory equivalence margin. SVM was lower both statistically and practically (AUC 0.648; ΔAUC − 0.056; Holm-adjusted p < 0.004). Overall performance metrics for the standard and cost-sensitive training strategies are summarized in Table 3.

Table 3
Benchmark under standard and cost-sensitive training (Age + Sex; test n = 1,053)

Algorithm / strategy	Train CV AUC	Test AUC	Balanced acc.	Macro-F1	Proximal recall	Brier
LR, standard	0.681 ± 0.014	0.704	0.409	0.394	0.000	0.457
LR, standard + equal-prior decision	0.681 ± 0.014	0.704*	0.528	0.487	0.450	0.457*
GNB, standard	0.678 ± 0.014	0.693	0.396	0.378	0.000	0.465
SVM, standard	0.618 ± 0.025	0.648	0.383	0.359	0.000	0.475
RF, standard	0.680 ± 0.013	0.701	0.409	0.394	0.000	0.455
GB, standard	0.680 ± 0.013	0.699	0.409	0.394	0.000	0.455
LR, cost-sensitive	0.666 ± 0.013	0.682	0.528	0.487	0.450	0.599
GNB, cost-sensitive	0.680 ± 0.015	0.694	0.529	0.488	0.450	0.593
SVM, cost-sensitive	0.668 ± 0.016	0.682	0.413	0.398	0.000	0.460
RF, cost-sensitive	0.678 ± 0.015	0.699	0.520	0.491	0.450	0.598
GB, cost-sensitive	0.675 ± 0.015	0.698	0.529	0.488	0.450	0.598

***AUC and Brier use the original standard-LR probabilities; only the class-decision rule is changed.**

[Insert Table 3 **here]**

Cost-sensitive training increased proximal recall from 0 to 0.450 for LR, GNB, RF, and GB and increased balanced accuracy to approximately 0.52–0.53, but overall accuracy fell to approximately 0.56 and Brier scores worsened to approximately 0.59–0.60. The equal-prior LR decision rule reproduced the cost-sensitive LR class decisions (balanced accuracy 0.528; proximal recall 0.450) without retraining. Because AUC, Brier score, and calibration were calculated from the retained original LR probabilities, these probability-based metrics remained those of standard LR; no claim is made that the rescaled p' probabilities are themselves calibrated. Paired macro-AUC differences relative to LR, Holm-adjusted p values, and equivalence assessments are presented in Table 4.

Table 4
Macro-AUC differences relative to LR within each training strategy

Training	Algorithm	Selected hyperparameters	ΔAUC vs LR (95% CI)	Holm p	± 0.02 interpretation
Standard	GNB	var_smoothing = 0.001	−0.010 (−0.017, −0.003)	0.006	Equivalent; statistically lower
Standard	SVM	C = 1; gamma=scale	−0.056 (−0.079, −0.033)	< 0.004	Not equivalent
Standard	RF	depth = 3; leaf = 1	−0.003 (−0.008, + 0.002)	0.418	Equivalent
Standard	GB	lr = 0.03; depth = 2; n = 100	−0.004 (−0.012, + 0.002)	0.418	Equivalent
Cost-sensitive	GNB	var_smoothing = 0.001	+ 0.012 (−0.002, + 0.027)	0.214	Inconclusive
Cost-sensitive	SVM	C = 1; gamma = 0.1	+ 0.000 (−0.007, + 0.008)	0.969	Equivalent
Cost-sensitive	RF	depth = 3; leaf = 50	+ 0.017 (+ 0.001, + 0.033)	0.136	Inconclusive
Cost-sensitive	GB	lr = 0.03; depth = 2; n = 100	+ 0.017 (+ 0.002, + 0.032)	0.136	Inconclusive

Sensitivity to the post hoc margin: under standard training RF remains equivalent at ± 0.01, whereas GB does not because its CI extends to − 0.012; both RF and GB are equivalent at ± 0.03. This analysis reinterprets the same paired-bootstrap CIs and does not refit models.

[Insert Table 4 **here]**

3.4 Error analysis and calibration

Standard LR correctly classified 95% of distal fractures but only 28% of midshaft fractures and assigned 102 of 111 proximal cases to the distal class. Cost-sensitive GB produced more balanced recall (58% distal, 55% midshaft, 45% proximal), at the cost of labeling approximately 22% of distal cases as proximal.

Standard LR was reasonably calibrated: class-specific calibration intercepts were near zero and ECE ranged from 0.025 to 0.049. Calibration slopes were 1.12–1.24, consistent with mildly conservative probability spread. The spline age-by-sex model had similar discrimination but wider proximal probability

range and lower ECE for midshaft and proximal classes. In contrast, cost-sensitive LR and GB showed substantial calibration shifts, with distal probabilities systematically underestimated and proximal probabilities overestimated (Table 5). [Insert Table 5 here]

Table 5
One-vs-rest calibration on the independent test set

Model	Class	Slope	Intercept	ECE	Predicted probability range
LR, standard	Distal	1.15	0.01	0.030	0.270–0.910
LR, standard	Midshaft	1.12	0.03	0.049	0.032–0.599
LR, standard	Proximal	1.24	−0.08	0.025	0.058–0.198
Spline age×sex	Distal	1.13	0.01	0.032	0.226–0.875
Spline age×sex	Midshaft	1.14	0.04	0.034	0.091–0.762
Spline age×sex	Proximal	1.15	−0.09	0.021	0.012–0.232
Saturated, training	Distal	1.13	0.01	0.037	0.274–0.878
Saturated, training	Midshaft	1.11	0.04	0.041	0.050–0.713
Saturated, training	Proximal	0.95	−0.09	0.034	0.002–0.297
LR, cost-sensitive	Distal	1.18	1.30	0.287	0.096–0.694
LR, cost-sensitive	Midshaft	1.21	−0.33	0.066	0.056–0.684
LR, cost-sensitive	Proximal	1.20	−1.48	0.226	0.179–0.521
GB, cost-sensitive	Distal	1.25	1.29	0.287	0.096–0.661
GB, cost-sensitive	Midshaft	1.25	−0.35	0.063	0.158–0.866
GB, cost-sensitive	Proximal	1.24	−1.48	0.224	0.039–0.546

Calibration intercept is calibration-in-the-large with slope fixed at 1. ECE was calculated over deciles of predicted probability.

4. Discussion

4.1 Principal findings

This study provides a demographic-only benchmark for fracture localization in the PediURF cohort and shows that age and sex contain a measurable, but clearly limited, amount of information about whether a confirmed pediatric forearm fracture is proximal, midshaft, or distal. Standard multinomial logistic regression achieved a macro AUC of 0.704 on the untouched official test set. Age accounted for most of this signal, whereas sex provided a smaller but statistically supported incremental contribution,

particularly for discrimination of proximal fractures. More flexible algorithms did not produce a meaningful improvement over logistic regression under standard training. A second important finding was that strategies aimed at class imbalance changed the pattern of class assignment much more than they changed the underlying discriminatory information. Proximal recall increased from 0 to 0.450 after cost-sensitive training, but this was accompanied by poorer probability calibration. A simple equal-prior decision adjustment reproduced the more balanced LR classification pattern without retraining the probability model. Taken together, these results suggest that the main contribution of demographic information in PediURF is not to provide a stand-alone diagnostic classifier, but to define the non-imaging signal against which image-only and multimodal models can be interpreted.

4.2 Demographic pattern of pediatric forearm fractures

The demographic signal observed in our analysis is clinically plausible and consistent with the established epidemiology of pediatric forearm fractures. Large epidemiologic studies have shown that forearm-fracture occurrence is strongly age dependent and that the age distribution differs between boys and girls [1, 2]. Hedström et al. reported marked variation in childhood fracture incidence according to age and sex, while Husum et al. demonstrated that the anatomical distribution of pediatric forearm fractures is not uniform throughout childhood [1, 2]. These observations are important for interpreting our model: age and sex are not arbitrary metadata fields but proxies for developmental and exposure-related factors that influence fracture patterns.

Our results extend this epidemiologic observation from fracture occurrence to fracture localization within a cohort in which every child already had a confirmed forearm fracture. Distal fractures represented 64.1% of the PediURF cohort, whereas proximal fractures accounted for only 10.9%. Localization nevertheless varied substantially with age (ε^2=0.110) and, to a lesser extent, sex (Cramér's V = 0.175). Age alone yielded a macro AUC of 0.672, and the addition of sex increased the AUC to 0.702. The increase was driven mainly by the proximal-versus-rest comparison, for which AUC rose from 0.591 to 0.676. Thus, sex contributed relatively little to overall classification but contained useful information for the least frequent anatomical category. This distinction would be obscured if interpretation were based only on overall accuracy.

The age effect should not be interpreted as implying that demographic variables can identify fracture location with sufficient certainty for clinical use. Rather, it demonstrates that the class distribution itself changes across developmental strata. In children, skeletal maturation, changing activity patterns, mechanisms of injury, and age-related differences in bone geometry may all contribute to these gradients. The present analysis cannot determine which of these mechanisms is responsible because PediURF provides only age and sex as structured demographic predictors. Nevertheless, the finding is relevant to AI studies because a model trained on radiographs may indirectly recover some of the same age-related information from growth plates, ossification patterns, and skeletal morphology. Consequently, part of the apparent information available to an image model may overlap with information already encoded in patient age.

4.3 Comparison with image-based pediatric fracture AI

The magnitude of discrimination obtained from demographics alone is substantially below that reported for radiograph-based pediatric fracture AI. Systematic reviews have described strong diagnostic performance for AI systems applied to pediatric fracture radiographs [3, 4]. More recently, Ximenes et al. pooled 16 diagnostic-accuracy studies including more than 10,000 pediatric patients and reported sensitivity of 93%, specificity of 91%, and an AUC of 0.96; performance remained high in studies using external datasets [18]. Individual pediatric studies using convolutional neural networks have likewise demonstrated that radiographic information can support high-performance fracture detection [19, 20]. These studies address fracture detection rather than the exact three-class localization task examined here, so their performance estimates are not directly comparable with our macro AUC. The contrast is nevertheless informative: demographic metadata contain a genuine signal, but radiographic information remains the dominant source of clinically useful fracture information.

This distinction is central to the purpose of the present study. Tang et al. introduced PediURF as a paired-view radiographic dataset for pediatric ulna and radius fracture analysis and reported strong performance for their image-based URFNet framework [6]. Our analysis asks a different question: how much of the localization structure can be recovered without looking at a single pixel? The answer appears to be a moderate amount, but far from enough for diagnostic substitution. This demographic benchmark therefore provides a useful reference point for future PediURF experiments. An image-only model should substantially exceed this baseline, while a multimodal image-plus-demographic model should demonstrate incremental value beyond both the image-only model and the demographic-only model.

4.4 Implications for multimodal modeling and shortcut learning

The findings also have implications for multimodal model design. Huang et al. emphasized that combining medical images with structured clinical information can improve prediction, but that the contribution of each modality should be evaluated explicitly rather than inferred from the performance of the final fused model [7]. In PediURF, age and sex already provide a macro AUC around 0.70. A future multimodal model that includes these variables should therefore report ablation experiments showing the performance of demographics alone, images alone, and their combination. Without such comparisons, an apparent multimodal improvement may be difficult to interpret.

Demographic information may also interact with shortcut learning and dataset-specific correlations. Zech et al. demonstrated in chest-radiograph AI that performance can change substantially across institutions when models exploit site- or population-associated signals that do not transport reliably [8]. Our study does not demonstrate shortcut learning in PediURF because no radiographic pixels were analyzed. However, it identifies a concrete demographic structure that an image encoder could potentially capture indirectly. If fracture localization varies with age and sex, and if age-related skeletal

appearance is visible on radiographs, an image model may partially exploit this correlation. Future PediURF studies could test this possibility by reporting performance across age and sex strata, assessing calibration within subgroups, and comparing image-only predictions with explicit demographic models.

4.5 Why complex machine-learning models did not outperform logistic regression

A notable finding was the absence of a meaningful advantage from greater algorithmic complexity. Standard LR achieved a macro AUC of 0.704, compared with 0.701 for random forest and 0.699 for gradient boosting; SVM performed worse at 0.648. This result is consistent with the systematic review by Christodoulou et al., who examined 71 studies comparing logistic regression with machine-learning methods for clinical prediction and found no evidence of superior ML performance in comparisons at low risk of bias [11]. Their review also highlighted that apparent advantages of complex models were often associated with methodological limitations in validation. Our use of the original independent PediURF test set and training-only hyperparameter selection was intended to minimize this source of optimism.

In our setting, the result is also intuitively understandable. Only two predictors—age and sex—were available. Flexible algorithms are most useful when they can exploit nonlinearities, interactions, or high-dimensional structure that simpler models cannot adequately represent. Here, the amount of information contained in the predictors appears to be the limiting factor rather than model capacity. The spline model produced almost the same discrimination as standard LR, and the training-estimated saturated age-by-sex reference achieved a macro AUC of 0.698. Even the deliberately optimistic test-label-derived oracle reached only 0.732. These converging results suggest that little additional localization information can be extracted from age and sex alone in this dataset, regardless of algorithmic complexity.

This finding should not be interpreted as a general argument against machine learning. PediURF is fundamentally an imaging dataset, and high-capacity models are appropriate when the input contains thousands or millions of potentially informative image features. Instead, our results illustrate a narrower principle: model complexity cannot compensate for limited predictor information. For demographic benchmarking, the simpler probabilistic model may therefore be preferable because its performance is comparable, its behavior is easier to inspect, and its predicted probabilities can be evaluated directly for calibration.

4.6 Class imbalance, proximal fractures, and calibration

The proximal class provides the clearest example of why discrimination, classification, and calibration should be reported separately. Under standard LR, macro AUC was 0.704, yet proximal recall at the default argmax decision rule was zero. The model assigned 102 of 111 proximal test cases to the distal class. This apparently contradictory result occurs because AUC evaluates ranking across probability

thresholds, whereas class recall at argmax depends on which class has the highest predicted probability. With proximal fractures representing only about one tenth of the cohort, a model may rank proximal cases better than non-proximal cases without making proximal the most probable class often enough to win the final argmax decision.

Cost-sensitive training changed this behavior substantially. Proximal recall increased to 0.450 for LR, GNB, RF, and GB, and balanced accuracy increased to approximately 0.52–0.53. However, these gains came with lower overall accuracy and markedly worse Brier scores. This pattern closely parallels the findings of van den Goorbergh et al., who showed that commonly used class-imbalance corrections can improve the balance between sensitivity and specificity without improving discrimination, while producing substantial probability miscalibration [9]. They further showed that a similar classification effect can often be obtained by changing the decision threshold rather than altering model training [9]. Our equal-prior LR experiment produced the same practical message: the balanced class-decision profile was reproduced without retraining, while the original LR probabilities were retained for AUC and calibration assessment.

This distinction has practical importance. If the purpose of a model is to output clinically interpretable probabilities, altering the training distribution to force more minority-class predictions may be undesirable. Calibration is essential when predicted probabilities are intended to support decisions, because a nominal probability should correspond reasonably to the observed frequency of the outcome [10]. Conversely, if the immediate objective is a classification rule with greater sensitivity for a rare class, threshold or decision-rule adjustment may be more transparent. The present findings therefore argue against describing a class-weighted model as simply “better” because its proximal recall is higher; the preferred strategy depends on whether the intended output is ranking, probability estimation, or categorical assignment.

4.7 Strengths and limitations

Several features strengthen the present analysis. First, the original patient-level PediURF training and test split was preserved, and the independent test set was not used for hyperparameter selection. Second, multiple aspects of model performance were evaluated rather than relying on accuracy alone: discrimination, class-specific recall, balanced accuracy, MCC, Brier score, and calibration were considered separately. Third, the analysis included both a training-estimated saturated age-by-sex reference and a deliberately labeled test-label oracle, allowing the observed performance of fitted models to be interpreted in relation to the amount of demographic structure present in the dataset. Finally, the comparison between cost-sensitive training and a decision-level equal-prior adjustment helps separate probability modeling from the choice of a classification rule.

The study also has important limitations. PediURF contains only children with confirmed fractures; consequently, none of the reported probabilities should be interpreted as the probability that a child has a fracture. The data were obtained from a single tertiary pediatric center, and the age- and sex-specific localization patterns may differ in other populations. Only age and sex were available as demographic

predictors. Potentially relevant variables such as mechanism of injury, laterality, activity, fracture morphology, displacement, and specific injury patterns were unavailable. External-center validation was not possible, and the analysis relied on a single official test split. Some age-by-sex cells were sparse, which limits the stability of saturated cell estimates. The ± 0.02 equivalence margin was not preregistered and should therefore be viewed as an exploratory aid to interpretation rather than formal proof of equivalence. Finally, because we did not analyze radiographs, the study cannot determine whether an image model actually uses demographic shortcuts; it only quantifies the demographic signal that is available to be learned directly or indirectly.

4.8 Clinical and research implications

The immediate clinical implication is deliberately modest: age and sex should not be used to replace radiographic assessment of a suspected pediatric forearm fracture. Their value in this study is methodological. For researchers using PediURF, the demographic-only results provide a reproducible baseline against which image and multimodal models can be compared. Future work should report demographic-only, image-only, and combined models; assess calibration as well as discrimination; evaluate performance across age and sex strata; and, where possible, validate findings in an external pediatric population. Recent pooled evidence on pediatric fracture AI has likewise emphasized the importance of external validation and subgroup assessment before broad clinical generalization [18].

More broadly, our results illustrate why benchmark construction matters in medical AI. A high-performing model should not only outperform a majority-class reference; it should also be compared with simple, clinically interpretable baselines that capture known structure in the data. In PediURF, age and sex alone explain a non-trivial portion of localization discrimination, yet increasingly flexible tabular algorithms extract little additional information from those variables. Demonstrating improvement beyond this demographic baseline would therefore provide stronger evidence that a future imaging or multimodal model is learning clinically relevant information that is not already available from basic patient characteristics.

5. Conclusions

In the PediURF cohort, age and sex contained moderate information about the anatomical localization of confirmed pediatric forearm fractures, but this signal was insufficient for stand-alone clinical classification. A parsimonious multinomial logistic-regression model captured essentially the same demographic discrimination as more flexible standard machine-learning algorithms. The most important effect of class-imbalance handling was not an increase in discriminatory information, but a change in threshold-dependent class decisions: proximal recall improved at the expense of probability calibration. These findings establish a reproducible non-imaging benchmark for PediURF and support future evaluation of image-only and multimodal models against simple demographic baselines, with discrimination, calibration, and class-specific performance reported as distinct properties.

Declarations

Ethics approval and consent to participate:
This study was a secondary analysis of the publicly available and de-identified PediURF dataset and involved no direct interaction with human participants and no collection of new patient data. The original PediURF study was approved by the Ethics Committee of Shenzhen Children's Hospital (No. 202505902), as reported by Tang et al. The present study was limited to secondary analysis of anonymized, publicly accessible data, with no participant contact, intervention, or collection of new patient-level information. Accordingly, no additional informed consent was obtained for the present secondary analysis.

Consent for publication:
Not applicable.

Availability of data and materials:
The data analyzed in this study are publicly available as part of the PediURF dataset described by Tang et al. No new patient-level data were collected or generated for the present study. The analyses reported in this manuscript were performed using the publicly available demographic and fracture-localization data from the PediURF dataset.

Competing interests:
The authors declare that they have no competing interests.

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions:
V.K. contributed to the conception and clinical interpretation of the study, interpretation of the results, and drafting and critical revision of the manuscript. M.A. contributed to the computational methodology, data analysis, machine-learning implementation, statistical analysis, interpretation of the results, and revision of the manuscript. C.A. contributed to interpretation of the findings, manuscript review, and language editing. A.L. contributed to clinical interpretation, manuscript review, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements:
Not applicable.

References

1. Hedström EM, Svensson O, Bergström U, Michno P. Epidemiology of fractures in children and adolescents. Acta Orthop. 2010;81(1):148–53. 10.3109/17453671003628780.

2. Husum HC, Kold S, Rahbek O. The distribution of paediatric forearm fractures: a five-year retrospective cohort study of 4546 forearm fractures in children. Child (Basel). 2025;12(6):711. 10.3390/children12060711.
3. Shelmerdine SC, White RD, Liu H, Arthurs OJ, Sebire NJ. Artificial intelligence for radiological paediatric fracture assessment: a systematic review. Insights Imaging. 2022;13(1):94. 10.1186/s13244-022-01234-3.
4. Ashworth E, Allan E, Pauling C, et al. Artificial intelligence (AI) in radiological paediatric fracture assessment: an updated systematic review. Eur Radiol. 2025;35(9):5264–86. 10.1007/s00330-025-11449-9.
5. Nagy E, Janisch M, Hržić F, Sorantin E, Tschauner S. A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning. Sci Data. 2022;9(1):222. 10.1038/s41597-022-01328-z.
6. Tang S, Ou L, Li W, et al. A comprehensive X-ray dataset for pediatric ulna and radius fractures analysis. Sci Data. 2026;13(1):308. 10.1038/s41597-026-06666-w.
7. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med. 2020;3:136. 10.1038/s41746-020-00341-z.
8. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. 10.1371/journal.pmed.1002683.
9. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inf Assoc. 2022;29(9):1525–34. 10.1093/jamia/ocac093.
10. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. 10.1186/s12916-019-1466-7.
11. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
12. Saerens M, Latinne P, Decaestecker C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Comput. 2002;14(1):21–41. 10.1162/089976602753284446.
13. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405(2):442–51. 10.1016/0005-2795(75)90109-9.
14. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
15. Collins GS, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.

16. Harris CR, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62. 10.1038/s41586-020-2649-2.
17. Virtanen P, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. 10.1038/s41592-019-0686-2.
18. Ximenes GF, Costa ÁL, Leite LL, et al. Are artificial intelligence models reliable for clinical application in pediatric fracture detection on radiographs? A systematic review and meta-analysis. Clin Orthop Relat Res. 2026;484(2):371–85. 10.1097/CORR.0000000000003660.
19. Till T, Tschauner S, Singer G, Lichtenegger K, Till H. Development and optimization of AI algorithms for wrist fracture detection in children using a freely available dataset. Front Pediatr. 2023;11:1291804. 10.3389/fped.2023.1291804.
20. Zech JR, Jaramillo D, Altosaar J, Popkin CA, Wong TT. Artificial intelligence to identify fractures on pediatric and young adult upper extremity radiographs. Pediatr Radiol. 2023;53(12):2386–97. 10.1007/s00247-023-05754-y.