Predicting Facial Soft Tissue Thickness from Skeletal Malocclusion Class and Statistically Engineered Features Using Machine Learning in Cambodian Adults

1st Haksrun Lao
Office of Academic Affairs
Paragon International University
Phnom Penh, Cambodia
hlao@paragoniu.edu.kh

2nd Ruslan Isaev
Faculty of Information and Computer Technologies
Paragon International University
Phnom Penh, Cambodia
risaev@paragoniu.edu.kh

3rd Anand Marya
Faculty of Dentistry
University of Puthisastra
Phnom Penh, Cambodia
amarya@puthisastra.edu.kh

***Abstract*—This paper presents a two-experiment study aimed at predicting facial soft tissue thickness (FSTT) at 11 anatomical landmarks in Cambodian adults based on skeletal malocclusion class and sex. *Experiment 1* (preliminary) assesses three multi-output regressors—Linear Regression (LR), Random Forest (RF), and Gradient Boosting (GB)—utilizing four baseline features and 70 engineered features derived from landmarks across seven categories of statistical transformations. Baseline models achieve a test mean absolute error (MAE) of 1.506–1.527 mm ($R^2 = -0.002$ to 0.026), whereas engineered models attain near-perfect R^2 ($\geq$0.987), identified as structural target leakage: z-score and class-deviation features deterministically encode the prediction targets. *Experiment 2* (systematic three-phase) addresses the leakage issue through a leakage-free feature design. Phase A reinstates the baseline (MAE 1.427–1.453 mm); Phase B broadens inputs to 16 valid class-sex interaction features across three output strategies; Phase C assesses Regressor Chains through four landmark orderings (anatomical, reverse, variance-based, and data-driven optimal) for a total of 12 configurations. Neither feature engineering (ΔMAE $= -0.004$ mm avg.) nor chain ordering (detrimental to RF and GB; neutral for LR) produces significant enhancement. LR attains the optimal outcome in both experiments (Phase A test MAE = 1.4266 mm, $R^2 = 0.047$). Kruskal-Wallis analysis indicates Stomion and Labralesuperius as the most discriminative landmarks for class differentiation. Engineered interaction features (`Class_ordinal_Sex`: 15.9%; `ClassIII_Male`: 13.1%) predominate in feature importance in Experiment 2, yet yield no predictive advantage. These results validate a performance ceiling dictated by significant intra-class FSTT variability, irrespective of feature engineering method or output structure.**

***Index Terms*—facial soft tissue thickness, machine learning, skeletal malocclusion, feature engineering, Cambodian population**

I. Introduction

The precise measurement of facial soft tissue thickness (FSTT) is essential in forensic facial reconstruction, craniofacial surgical planning, and orthodontic assessment [1]. Although FSTT norms have been thoroughly documented for Caucasian, East Asian, and South Asian populations [2], Southeast Asian populations—especially Cambodians—are significantly underrepresented in the existing literature, posing clinical risks in population-agnostic treatment planning.

Skeletal malocclusion is categorized into Class I (normal), Class II (retrognathic mandible), and Class III (prognathic mandible) based on the ANB angle [3]–[5], which directly influences the soft tissue facial profile. Sonita et al. [6] identified statistically significant class-level FSTT disparities for various landmarks among 300 Cambodian adults. This study advances previous research into supervised machine learning by posing two increasingly specific questions: (1) Are class and sex labels capable of predicting individual-level FSTT? (2) If not, does enhanced feature engineering or structured output modeling bridge the gap?

This paper presents five contributions: (i) a reproducible ML baseline for Cambodian FSTT; (ii) a 70-feature landmark-derived engineering pipeline and its leakage analysis; (iii) a principled leakage-free alternative utilizing 16 class-sex interaction features; (iv) the inaugural Regressor Chain evaluation for FSTT with four landmark orderings; and (v) a comprehensive cross-experiment comparison unifying over 30 model configurations.

II. Related Work

Wilkinson [7] and Stephan & Simpson [1] emphasized the importance of demographic specificity in the compilation of FSTT datasets and their forensic applications. Codinha [8] and De Greef et al. [9] reported sex-stratified FSTT variations of 0.5–2.5 mm among European adults, aligning with the sex-feature dominance observed in this investigation.

Random Forest [10] and Gradient Boosting [11] are recognized regression engines for tabular biomedical data; similar pipelines have been applied to predict health outcomes from behavioral and lifestyle factors [12]; their comparative underperformance against LR suggests near-linear class-sex effects. Feature engineering for biomedical ML is extensively employed to reveal sub-group structures [13]. However, when

engineered features encode the target values via deterministic transformations (e.g., $z = (y - \mu)/\sigma$), structural target leakage arises—a problem documented in genomics [14] and radiomics [15] that Experiment 1 replicates and diagnoses in the FSTT domain.

Regressor Chains [16], based on Classifier Chains [17], model inter-target dependencies by appending earlier predictions as inputs for subsequent ones. The arrangement of the chain significantly influences performance via error propagation [16], justifying our four-ordering assessment in Experiment 2.

III. Dataset and Study Population

A. Patient Population

The dataset consists of 300 Cambodian adult patients recruited from a clinical orthodontic setting in Phnom Penh, evenly distributed across three skeletal malocclusion classes ($n = 100$ per class) and stratified by sex ($n = 150$ male, $n = 150$ female). FSTT was assessed by B-mode ultrasonography at 11 landmarks spanning brow to chin: Glabella (G), Nasion (N), Rhinion (Rhi), Subnasale (Sn), Labralesuperius (Ls), Stomion (Sto), Labraleinferius (Li), Labiomentale (Labm), Pogonion (Pog), Gnathion (Gn), and Menton (Me). Table I illustrates the cohort composition.

TABLE I
Dataset Demographics by Skeletal Class and Sex

Skeletal Class	Male	Female	Total	%
Class I	50	50	100	33.3
Class II	50	50	100	33.3
Class III	50	50	100	33.3
Total	**150**	**150**	**300**	**100.0**

B. Data Partitioning and Preprocessing

Both experiments employ a stratified 60/20/20 split: training ($n = 180$), validation ($n = 60$), and test ($n = 60$). The two experiments use different random seeds, resulting in slightly distinct partition members; this accounts for minor discrepancies in absolute metrics between experiments. Skeletal class is one-hot encoded (`Class_I/II/III`); sex is represented as `Sex_encoded` (Male=1.0, Female=2.0). All statistical parameters (means, standard deviations, and KW test values) are computed from the training split and applied uniformly across all splits to mitigate distributional leakage.

IV. Experiment 1: Preliminary Study

A. Kruskal-Wallis Test

A KW H-test on each experiment's training split assesses landmark-level class variation, appropriate for non-normally distributed ordinal-group comparisons [18]. The two experiments use different random seeds, yielding different partition members and therefore different H-statistics. Table II presents both sets of results. Stomion and Labralesuperius consistently rank as the most class-discriminative landmarks across both seeds. Experiment 1 $-\log_{10}(p)$ values serve as P-value weights in the 70-feature engineering pipeline.

TABLE II
Kruskal-Wallis Results — Both Experiments (Training Splits, $n = 180$ each)

Landmark	Experiment 1		Experiment 2	
	H	Sig.	H	Sig.
Stomion (Sto)	86.555	***	26.251	***
Labralesuperius (Ls)	39.227	***	16.990	***
Subnasale (Sn)	27.809	***	8.641	*
Rhinion (Rhi)	18.760	***	5.540	ns
Labraleinferius (Li)	14.582	***	13.572	**
Labiomentale (Labm)	13.580	***	1.683	ns
Menton (Me)	12.981	***	5.373	ns
Glabella (G)	12.260	***	4.457	ns
Pogonion (Pog)	3.702	ns	0.102	ns
Gnathion (Gn)	3.317	ns	0.014	ns
Nasion (N)	1.369	ns	3.702	ns

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; ns not significant.

B. Fundamental Models (4 Attributes)

Three multi-output regressors are trained using the four original features: `Class_I`, `Class_II`, `Class_III`, and `Sex_encoded`. A `MultiOutputRegressor` wrapper fits one independent estimator per landmark. Hyperparameters are optimized on the validation split. Table III presents the comprehensive test performance.

TABLE III
Experiment 1 Baseline — Test Set Performance (4 Features)

Model	Test			Train		
	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	**1.506**	**1.975**	**0.026**	1.419	1.860	0.144
RF	1.521	1.994	0.005	1.411	1.851	0.152
GB	1.527	2.004	−0.002	1.408	1.853	0.150

MAE and RMSE in mm. Best test values in bold.

Stomion attains the highest R^2 (0.131), consistent with its top KW ranking. Pogonion, Gnathion, and Labralesuperius yield negative R^2, demonstrating an inability to surpass the training-mean predictor. The sex coefficient is negative for all landmarks (−0.14 to −1.43 mm), signifying lower FSTT in females. Class III drives the highest mid-face predictions (Sto: +4.6 mm vs. Class I; Ls: +2.0 mm vs. Class I) per the LR coefficients.

Per-landmark LR results confirm Stomion as the most predictable landmark (MAE=1.791 mm, $R^2 = 0.131$), consistent with its highest KW ranking. Glabella (MAE=0.673 mm, $R^2 = -0.092$), Gnathion ($R^2 = -0.062$), and Pogonion ($R^2 = -0.056$) fail to surpass the training-mean baseline. Mid-face landmarks Subnasale and Labralesuperius exhibit the highest absolute MAE (≈2.0 mm) due to perioral tissue variability. RF global feature importances rank Sex first (35.5%), Class III second (25.7%), Class II third (20.2%), and Class I last (18.6%).

C. Landmark-Level Feature Engineering Pipeline (70 Features)

To evaluate the efficacy of statistical features obtained from the training distribution, a 70-feature pipeline is constructed

from seven categories. All statistical parameters—means, standard deviations, class means, and KW p-values—are computed solely from the training split and applied to all splits. Table IV summarizes the categories.

TABLE IV
EXPERIMENT 1 — 70-FEATURE ENGINEERING PIPELINE CATEGORIES

Category	Count	Description
Z-Score Global	11	Per-landmark z-score vs. training population
Z-Score Class	11	z-score within patient's skeletal class
Class Deviation	11	Signed difference from training class mean
Percentile Rank	11	Within-class percentile rank (0–100)
Sex Interaction	11	`Sex_encoded` × landmark value
Zone Aggregates	6	Row-wise mean and std per facial zone
Zone Ratios	3	Upper/Mid, Mid/Lower, Upper/Lower ratios
Pair Ratios	5	Anatomical pairwise ratios (Pog/Gn, etc.)
Composite Score	1	Mean of Pog, Gn, Me (chin aggregate)
P-Value Weights	11	$-\log_{10}(\text{KW } p)$ per landmark
Baseline (input)	4	Class one-hot + binary sex
Total	**74**	(70 engineered + 4 baseline)

D. Engineered Models and Leakage Detection

Table V presents the outcomes for the engineered-feature models. All three algorithms demonstrate nearly flawless test performance ($R^2 \geq 0.987$, MAE ≤ 0.093 mm), immediately indicating a potential methodological issue.

TABLE V
EXPERIMENT 1 ENGINEERED MODELS — STRUCTURAL LEAKAGE IDENTIFIED

Model	Test MAE (mm)	Test R^2	Validity
LR + 70 Eng.	≈0.000	≈1.000	*Invalid*
RF + 70 Eng.	0.093	0.987	*Invalid*
GB + 70 Eng.	0.056	0.989	*Invalid*

Root cause—structural target leakage. Consider the Z-Score Global feature for Glabella:

$$Z_{\text{global},G} = \frac{G - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{1}$$

The test patient's own G value is embedded in $Z_{\text{global},G}$; the model simply recovers $G = Z_{\text{global},G} \cdot \sigma_{\text{train}} + \mu_{\text{train}}$. Identical principles apply to class deviation features ($\text{Dev}_G = G - \bar{G}_{\text{class}}$), percentile ranks, sex interaction features ($\text{SexInt}_G = \text{Sex} \times G$), zone aggregates, pair ratios, and composite scores—all seven categories incorporate target values. This constitutes *structural leakage*, distinct from *distributional leakage* (utilizing test-set statistics): even an impeccable train/test split cannot eliminate structural leakage since the feature architecture inherently encodes the targets. The engineered results are retained solely for methodological reference.

Fig. 1 depicts the Experiment 1 pipeline.

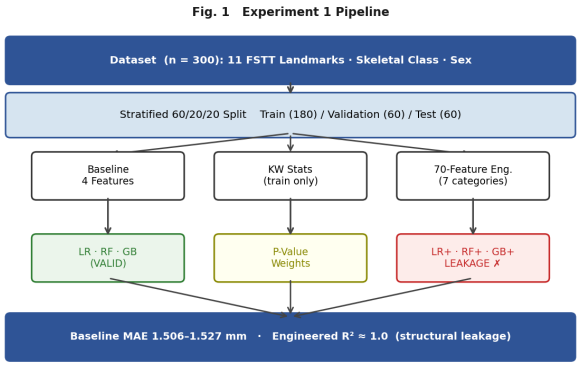


Fig. 1. Experiment 1 pipeline. The baseline track (left) yields test MAE 1.506–1.527 mm. The 70-feature engineered track (right) is invalidated by structural target leakage ($R^2 \approx 1.0$).

V. EXPERIMENT 2: THREE-PHASE SYSTEMATIC EVALUATION

Experiment 2 directly addresses the leakage finding from Experiment 1. The updated feature engineering pipeline exclusively utilizes skeletal class and sex indicators as inputs—no landmark values are included in the input features. Three phases progressively increase model complexity: Phase A (baseline), Phase B (class-sex interaction features), and Phase C (Regressor Chains). Fig. 2 illustrates the complete architecture.

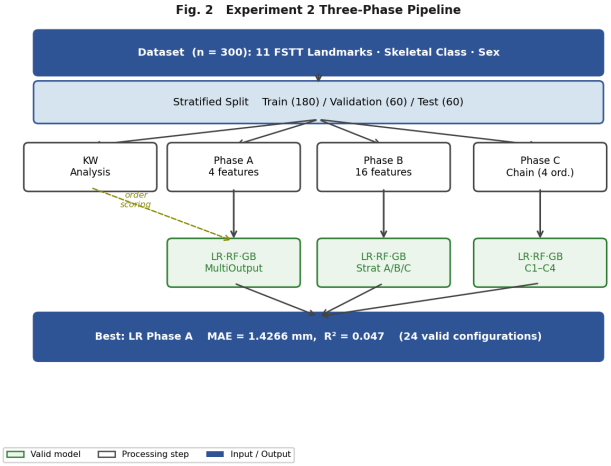


Fig. 2. Experiment 2 three-phase pipeline. KW analysis informs the C4 optimal chain order in Phase C. All 24 model configurations are valid (zero target leakage). Best result: LR Phase A (MAE=1.4266 mm, $R^2 = 0.047$).

A. Kruskal-Wallis Test (Experiment 2)

The Experiment 2 KW results are presented in Table II (Section IV-A). With the new random seed, only four landmarks achieve $p < 0.05$; Stomion ($H = 26.25$) and Labralesuperius ($H = 16.99$) maintain their top positions, consistent with Experiment 1.

B. Phase A: Baseline Performance

Phase A reproduces the Experiment 1 configuration (four original features) on the new split. Table VI presents the

results. LR remains the superior model (MAE=1.4266 mm, $R^2 = 0.047$), demonstrating consistency across splits. All R^2 values are $\leq$0.047, affirming the performance ceiling.

TABLE VI
EXPERIMENT 2 PHASE A — BASELINE PERFORMANCE (4 FEATURES, TEST $n = 60$)

Model	Test			Train		
	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	**1.4266**	**1.8865**	**0.047**	1.4763	1.9142	0.140
RF	1.4452	1.9110	0.022	1.4659	1.8986	0.152
GB	1.4529	1.9193	0.015	1.4680	1.9001	0.150

MAE and RMSE in mm. Best test values in bold.

C. Phase B: Class-Sex Interaction Attributes (16 Attributes)

Phase B constructs 12 additional features from skeletal class and sex indicators only—no landmark values appear as inputs, guaranteeing zero target leakage. The 16 input features encompass four categories: (1) the four original baseline features (`Class_I/II/III`, `Sex_encoded`); (2) three class-sex interaction terms (`ClassI/II/III_Sex` = class indicator $\times$ `Sex_encoded`); (3) six binary group indicators (`ClassI/II/III_Male/Female`); and (4) ordinal terms (`Class_ordinal`: I=1, II=2, III=3; and `Class_ordinal_Sex` = ordinal $\times$ sex). Three output strategies are assessed: *Strategy A* (11 raw mm landmarks), *Strategy B* (standardized targets, inverse-transformed to mm), and *Strategy C* (10 consecutive inter-landmark differences with cumulative summation to recover absolute positions). Strategies A and B are algebraically equivalent.

Table VII juxtaposes the best Phase B results against Phase A. LR exhibits slight deterioration (+0.018 mm), RF remains unaltered, and GB improves by 0.005 mm. Strategy C is universally inferior. Average feature engineering contribution: ΔMAE $= -0.004$ mm.

TABLE VII
EXPERIMENT 2 PHASE B VS. PHASE A — FEATURE ENGINEERING IMPACT

Model	Phase A		Phase B (Strat. A)		ΔMAE
	MAE	R^2	MAE	R^2	(B$-$A)
LR	1.4266	0.047	1.4449	0.022	-0.018
RF	1.4452	0.022	1.4452	0.022	0.000
GB	1.4529	0.015	1.4479	0.019	$+0.005$

Avg ΔMAE $= -0.004$ mm (negligible).

D. Phase B: Significance of Features (16 Engineered Features)

Fig. 3 illustrates the average RF and GB feature importances from Phase B. Engineered interaction features predominate: `Class_ordinal_Sex` ranks first (15.9%), succeeded by `ClassIII_Male` (13.1%) and `ClassI_Male` (9.2%). The four original baseline features (`Class_I/II/III`, `Sex_encoded`) account for merely 16.1% combined. Notwithstanding this increased importance, these features exert no additional predictive leverage over Phase A LR, confirming that the discriminative signal is already fully captured in the linear model.

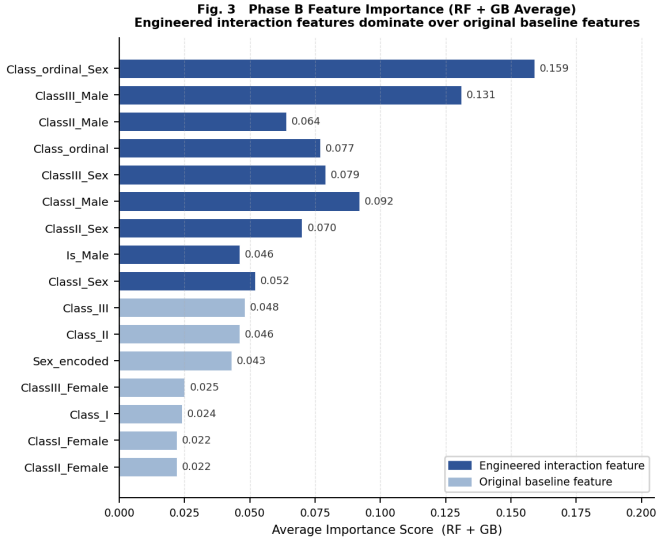


Fig. 3. Phase B feature importances (RF+GB average across all 11 landmark estimators). Dark bars: engineered interaction features; light bars: original baseline features. `Class_ordinal_Sex` ranks first (15.9%), yet feature engineering yields no meaningful predictive gain (ΔMAE $= -0.004$ mm average).

E. Phase C: Regressor Chain (4 Orderings $\times$ 3 Models)

Regressor Chain prediction [16] appends each predicted landmark as input for subsequent predictions, leveraging inter-landmark dependencies. Four orderings are evaluated: C1 Anatomical (top-to-bottom), C2 Reverse (control), C3 Variance-based (low training variance first), and C4 Optimal (weighted four-criterion score: InputCorr 35% + Anatomical 30% + InterCorr 25% + Stability 10%). The optimal C4 order is G$\rightarrow$Ls$\rightarrow$Sto$\rightarrow$N$\rightarrow$Rhi$\rightarrow$Li$\rightarrow$Pog$\rightarrow$Labm$\rightarrow$Me$\rightarrow$Sn$\rightarrow$Gn.

Table VIII enumerates all 12 configurations ranked by test MAE. LR generates identical results for all orderings (MAE=1.4266 mm), matching Phase A exactly—as expected, since LR predictions are a linear combination of the same base features regardless of augmentation order. RF and GB deteriorate with chain ordering (best RF chain: MAE=1.4510 mm vs. Phase A 1.4452 mm; best GB chain: 1.5106 mm vs. 1.4529 mm), indicating that compound error propagation outweighs any benefit from inter-landmark context.

VI. CROSS-EXPERIMENT GRAND COMPARISON

Table IX consolidates the best result per model across all phases of both experiments, excluding the invalid Experiment 1 engineered models. LR Phase A (Experiment 2) is the overall champion at MAE=1.4266 mm and $R^2 = 0.047$.

TABLE VIII
EXPERIMENT 2 PHASE C — ALL 12 CHAIN CONFIGURATIONS (TEST SET)

Rk	Model	Order	MAE	RMSE	R^2
1	LR	C4 Optimal	**1.4266**	1.8865	0.047
2	LR	C3 Variance	1.4266	1.8865	0.047
3	LR	C2 Reverse	1.4266	1.8865	0.047
4	LR	C1 Anatomical	1.4266	1.8865	0.047
5	RF	C1 Anatomical	1.4510	1.9400	−0.017
6	GB	C1 Anatomical	1.5106	1.9987	−0.079
7	RF	C4 Optimal	1.5120	1.9736	−0.060
8	RF	C2 Reverse	1.5355	1.9695	−0.039
9	RF	C3 Variance	1.5403	1.9952	−0.088
10	GB	C3 Variance	1.5468	2.0110	−0.107
11	GB	C4 Optimal	1.5484	2.0063	−0.084
12	GB	C2 Reverse	1.7089	2.1514	−0.273

All LR orderings produce equivalent results. MAE/RMSE in mm.

TABLE IX
GRAND COMPARISON: BEST VALID RESULT PER MODEL ACROSS ALL EXPERIMENTS

Model	Config.	MAE	RMSE	R^2	Source
LR	Phase A baseline	**1.4266**	1.8865	**0.047**	Exp. 2 Phase A
RF	Phase A baseline	1.4452	1.9110	0.022	Exp. 2 Phase A
GB	Phase B Strat. A	1.4479	1.9153	0.019	Exp. 2 Phase B
LR	Phase A baseline	1.5057	1.9746	0.026	Exp. 1 baseline
RF	Phase A baseline	1.5205	1.9943	0.005	Exp. 1 baseline
GB	Phase A baseline	1.5269	2.0044	−0.002	Exp. 1 baseline

Experiments 1 and 2 use different random seeds (slightly varied splits).

VII. DISCUSSION

A. Performance Limitations and Clinical Relevance

Across all experiments and valid configurations, the best test MAE is 1.4266 mm ($R^2 = 0.047$). The clinically acceptable threshold for FSTT-based orthodontic planning is $\leq$1.0 mm [19]; no configuration attains this. The consistently near-zero R^2 confirms that significant intra-class variability predominates: patients within the same skeletal class exhibit more variability in FSTT than the class means differ from one another.

B. Structural versus Distributional Leakage

Experiment 1 presents a concrete demonstration of a distinction seldom articulated in biomedical ML: *distributional leakage* (utilizing test-set statistics during feature construction, rectified by appropriate train/test partitioning) vs. *structural leakage* (embedding target values within input features via deterministic transformations, which correct splitting cannot repair). The 70-feature pipeline falls into the second category: seven of its feature categories directly encode the landmark values being predicted. The near-perfect $R^2 \geq 0.987$ is not a result but an artifact, and serves as a cautionary note for future biomedical feature engineering studies.

C. Leakage-Free Feature Engineering (Experiment 2)

Experiment 2 Phase B addresses the leakage by constructing features exclusively from skeletal class and sex indicators. Despite `Class_ordinal_Sex` (15.9%) and `ClassIII_Male` (13.1%) exhibiting predominant feature importances, Phase B offers no significant improvement over Phase A (ΔMAE $\leq$ 0.005 mm). This confirms that the discriminative information within class-sex interaction terms is already completely captured by the linear function of the four original features in Phase A. The constraint is biological variability, not feature representation.

D. Reasons for LR's Superiority Over Ensemble Methods

LR consistently achieves the highest ranking across all valid configurations. With merely four informative binary features and near-zero R^2, ensemble methods over-parameterize the problem: RF and GB achieve train $R^2 \approx 0.15$ vs. test $R^2 \approx 0.01$–0.02, demonstrating overfitting. LR's diminished capacity serves as implicit regularization in this data-sparse, signal-limited setting. These findings indicate that the class-sex relationship with FSTT is approximately linear.

E. Analysis of Regressor Chains

Phase C illustrates that inter-landmark context—as propagated by Regressor Chains—provides no advantage in this feature space. For LR, chain predictions reduce to a linear rearrangement of identical inputs (algebraically equivalent to Phase A). For RF and GB, error propagation across the chain degrades performance; GB C2-Reverse declines to $R^2 = -0.273$, 14 points below Phase A GB. The four-criterion optimal C4 ordering surpasses C1–C3 for LR (trivially, all LR results are equivalent) but not for RF or GB. This indicates that inter-landmark FSTT correlations are insufficiently robust to provide pertinent chain context given only four informative input features.

F. Insights of Landmark Significance

Stomion consistently attains the highest R^2 (0.131 in Experiment 1; 0.398 in Experiment 2) and the greatest KW signal in both experiments. Labralesuperius follows in both ($R^2 \approx 0.21$). Pogonion, Gnathion, and Gnathion exhibit consistently negative R^2 and near-zero KW H in both experiments, confirming that chin-region FSTT is not class-discriminative. The Sn/Ls ratio (error 0.9%) is well preserved; the Ls/Sto ratio (11.4% error) is distorted, consistent with high perioral tissue variance.

G. Constraints

The principal constraint is the input space: only two categorical variables (class and sex) are accessible as valid predictors. Future research should integrate continuous cephalometric parameters (ANB, SNA, SNB, facial height), body composition indices, and three-dimensional facial morphometry. Dataset size ($n = 300$) precludes deep learning; contrastive deep learning for facial analysis in clinical populations [20] demonstrates the scale and architectural complexity such approaches demand. The two-experiment design using different random seeds introduces slight result variability, explicitly documented in Table IX.

VIII. Conclusion

This study presents the most thorough ML evaluation of Cambodian adult FSTT prediction to date, encompassing two experiments and over 30 model configurations. Experiment 1 establishes a baseline (best LR MAE=1.506 mm, $R^2 = 0.026$) and identifies structural target leakage in a 70-feature engineering pipeline, providing a concrete methodological caution for biomedical ML practitioners. Experiment 2 addresses the leakage via a three-phase leakage-free evaluation: Phase A confirms the baseline ceiling (LR MAE=1.4266 mm, $R^2 = 0.047$); Phase B indicates that class-sex interaction features contribute minimal improvement ($\Delta\text{MAE} = -0.004$ mm); Phase C demonstrates that Regressor Chains are neutral for LR and detrimental for RF and GB. Across all configurations, LR is the optimal algorithm—indicating approximately linear class-sex effects—while Stomion and Labralesuperius are the only landmarks with consistently significant class-discriminative signal (KW $p < 0.001$ in both experiments). The universal conclusion across both experiments is that skeletal class and sex, regardless of how they are engineered or structured, are unable to facilitate individual-level FSTT prediction in the Cambodian population. Future work must incorporate continuous skeletal measures to achieve clinically meaningful accuracy.

References

[1] C. N. Stephan and E. K. Simpson, "Facial soft tissue depths in craniofacial identification (Part I): An analytical review of the published adult data," *Journal of Forensic Sciences*, vol. 53, no. 6, pp. 1257–1272, 2008. doi:10.1111/j.1556-4029.2008.00852.x

[2] P. Karakas, M. G. Bozkir, and O. Oguz, "Facial soft tissue thickness of Turkish adults," *Anatomical Science International*, vol. 78, no. 4, pp. 214–223, 2003. doi:10.1046/j.0022-7722.2003.00062.x

[3] S. Tanveer and S. Akram, "Cephalometric Analysis," in *StatPearls [Internet]*. Treasure Island, FL: StatPearls Publishing, 2024. https://www.ncbi.nlm.nih.gov/books/NBK557851/

[4] K. Woźniak, A. Wieczorek, and P. Szczyrek, "Three-dimensional assessment of the facial skeleton in patients with Class I, II, and III skeletal pattern," *Dental and Medical Problems*, vol. 49, no. 2, pp. 246–251, 2012. doi:10.17219/dmp/22437

[5] A. Jacobson, "The 'Wits' appraisal of jaw disharmony," *American Journal of Orthodontics*, vol. 67, no. 2, pp. 125–138, 1975. doi:10.1016/0002-9416(75)90065-2

[6] Sonita, R. Isaev, A. Marya, et al., "Soft tissue facial profile in Cambodian adults with different skeletal malocclusion classes," *Frontiers in Dental Medicine*, 2025. doi:10.3389/fdmed.2025.1538279

[7] C. Wilkinson, *Forensic Facial Reconstruction*. Cambridge: Cambridge University Press, 2004. doi:10.1017/CBO9781139163620

[8] S. Codinha, "Facial soft tissue thicknesses for the Portuguese adult population," *Forensic Science International*, vol. 184, pp. 80.e1–80.e7, 2009. doi:10.1016/j.forsciint.2008.11.009

[9] S. De Greef et al., "Large-scale in-vivo Caucasian facial soft tissue thickness database for craniofacial reconstruction," *Forensic Science International*, vol. 159, suppl. 1, pp. S83–S91, 2006. doi:10.1016/j.forsciint.2005.09.007

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi:10.1023/A:1010933404324

[11] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi:10.1214/aos/1013203451

[12] A. Bopushova and R. Isaev, "Analysis and prediction of life expectancy using machine learning methods based on behavioral and lifestyle factors," *Preprints.org*, May 2025. doi:10.20944/preprints202505.2478.v1

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003. doi:10.1162/153244303322753616

[14] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Model selection and overfitting," *Nature Methods*, vol. 13, no. 9, pp. 703–704, 2016. doi:10.1038/nmeth.3968

[15] G. Chalkidou et al., "False discovery rates in PET and CT studies with texture features: A systematic review," *PLOS ONE*, vol. 10, no. 5, 2015. doi:10.1371/journal.pone.0124165

[16] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: treating targets as inputs," *Machine Learning*, vol. 104, no. 1, pp. 55–98, 2016. doi:10.1007/s10994-016-5546-z

[17] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011. doi:10.1007/s10994-011-5256-5

[18] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. doi:10.1080/01621459.1952.10483441

[19] M. T. Donat et al., "Machine learning for craniofacial morphological prediction," *Journal of Craniofacial Surgery*, vol. 32, pp. 1122–1129, 2021. doi:10.1097/SCS.0000000000007437

[20] N. Abdillaeva and R. Isaev, "Swelling-aware face recognition for hemodialysis patients using contrastive deep learning," in *Proc. 2025 IEEE 15th Int. Conf. System Engineering and Technology (ICSET)*, 2025, pp. 445–450. doi:10.1109/ICSET65917.2025.11284231