SINTECHCOM: Science, Technology and Communication Journal
Vol. 7, No. 1, October 2026, pp. 97-112, DOI: 10.59190/stc.v7i1.420

AI-based feature selection for craniometric sex classification: A nested cross-validation study on the Howells dataset

Iis Afrianty, Liza Afriyanti*
Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

Sex estimation is an important part of forensic biological-profile reconstruction. However, machine-learning results can be overly optimistic when feature selection and hyperparameter tuning are performed without separating them from model evaluation. This study compared five feature-selection methods, namely Information Gain, ReliefF, LASSO, SHAP-based ranking, and support vector machine recursive feature elimination (SVM-RFE), using five classifiers: SVM, random forest, logistic regression, decision tree, and multilayer perceptron. The study used 2,524 individuals from a Howells-derived craniometric dataset, consisting of 1,156 females and 1,368 males. A stratified hold-out set containing 20% of the data was reserved before model development. After missingness filtering based on the development data, 71 predictors remained. Model development used 5-fold outer and 5-fold inner cross-validation, with 3, 5, 7, 10, and 15 candidate features. SVM-RFE combined with SVM obtained a nested-cross-validation accuracy of 0.8767 ± 0.0098, with an F1-score of 0.8852 and ROC-AUC of 0.9481. After retuning using the full development set, the final 15-feature model achieved 0.8970 accuracy, 0.9055 F1-score, 0.9511 ROC-AUC, 0.9088 sensitivity, and 0.8831 specificity on 505 internal hold-out observations. The AI-oriented selectors had a mean accuracy only 0.0074 higher than the filter selectors, which was below the predefined 0.05 improvement threshold. Classifier choice had a significant effect on accuracy, while the feature-selection method and its interaction with the classifier were not significant. SHAP analysis identified ZYB, AUB, and MDH as the three highest-ranked features. The results show that a relatively small number of craniometric features can provide good classification performance under the evaluation procedure used in this study. However, the differences between feature-selection approaches were relatively small, and further validation on independent populations and data collected at different time points is needed before the model can be considered for forensic use.

ARTICLE INFO

Article history:

Received Aug 25, 2026
Revised Oct 1, 2026
Accepted Oct 8, 2026
Published Oct 31, 2026

Keywords:

Feature Selection
Forensic Anthropology
Nested Cross-Validation
Sex Estimation
SHAP

Copyright (c) 2026:

Iis Afrianty, Liza Afriyanti

License:

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

***Corresponding Author**
E-mail address: liza.afriyanti@uin-suska.ac.id

1. INTRODUCTION

Sex estimation is a core component of biological profile reconstruction because it supports subsequent estimation of other attributes and narrows the pool of possible identities in forensic casework. Recent reviews describe forensic anthropology as an important source of biological evidence when remains are decomposed, skeletonized, incomplete, or otherwise unsuitable for conventional soft-tissue identification procedures [1,2]. In humanitarian, mass-disaster, and unidentified-remains contexts, forensic practitioners may need to work with incomplete skeletal evidence; consequently, combining physical skeletal assessment with complementary scientific approaches can strengthen bio-

logical-profile reconstruction [3]. The application of machine learning in forensic anthropology has expanded across biological-profile estimation tasks, although recent literature also emphasizes the need for transparent reporting, interpretability, and further validation before model outputs are relied upon in forensic practice [1]. Metric approaches translate skeletal morphology into reproducible quantitative variables, while machine learning can model multivariate patterns that may be difficult to express through fixed univariate thresholds. Machine-learning applications in forensic anthropology have expanded across skeletal element and biological profile tasks, but transparency, interpretability, and independent validation remain recurring methodological needs [1]. Consequently, predictive performance alone is insufficient: a defensible model must also disclose its data boundaries, selection procedure, calibration, failure modes, and explanation limits.

Studies of gender recognition using facial images and hand radiographs further illustrate that machine-learning and deep-learning models can extract sex-related patterns from biomedical or biometric representations; nevertheless, these image-based settings are methodologically distinct from classification based on numerical cranial measurements [4,5]. Cranial measurements are attractive because they provide standardized numeric representations of size and shape, yet they create several technical challenges. Measurements are correlated, sexual dimorphism varies among populations, incomplete observations are common, and a model may inadvertently learn population structure rather than sex-associated morphology. Broad reviews of skull-based sex estimation emphasize that population specificity can restrict transportability, even when models perform well within the development population [6]. These concerns are particularly important for globally assembled reference collections such as the Howells dataset.

Previous craniometric machine-learning studies have shown promising predictive performance, although their findings vary considerably due to differences in sample characteristics, selected features, learning algorithms, and evaluation strategies. For example, SexEst developed classifiers using cranial and postcranial measurements collected from diverse populations worldwide, achieving cranial classification accuracies ranging from 81.2% to 87.7%. In another study, a population-inclusive neural-network approach based on 10 cranial measurements achieved a cross-validation accuracy of 86.7% and maintained an accuracy of 84.3% when evaluated on an independent, unrelated test dataset [7,8]. A Bulgarian CT-based study reported more than 95% accuracy, but its sample, landmarks, derived variables, and repeated cross-validation protocol differ from those of Howells-based analyses; therefore, the numerical values do not establish direct superiority of one approach over another [9]. In classification problems based on numerous correlated measurements, predictive performance may depend not only on the classifier but also on whether the selected variables retain complementary information rather than redundant signals. This issue is particularly relevant to craniometric data, in which many linear measurements may encode related aspects of cranial size and shape.

Feature selection may improve parsimony, reduce redundancy, and facilitate interpretation, but the estimated benefit depends on its placement within the validation process. Evaluating a tuned model on the same cross-validation process used for optimization can bias estimated error; nested cross-validation substantially reduces this bias by separating inner-loop selection from outer-loop assessment [10]. Feature selection must therefore be repeated inside training folds rather than performed once on the complete dataset. Stability is also important because small changes in training data can yield different selected subsets, weakening claims about biological relevance even if predictions remain accurate [10,11].

No single selector paradigm is necessarily optimal for craniometric data. Information Gain ranks marginal label association, ReliefF incorporates neighborhood-based contrasts, LASSO embeds sparsity within a penalized linear model, SHAP ranks model-attributed contributions, and SVM-RFE removes variables recursively according to SVM-derived weights. SHAP provides a unified additive feature-attribution framework with local-accuracy and consistency properties, but its attributions describe model behavior rather than causal anatomy [12]. SVM-RFE can account for multivariable discrimination while producing a compact subset, although the selected features may vary with folds and hyperparameters [12-14].

Machine-learning applications in forensic anthropology have been reported for multiple components of the biological profile, including sex estimation. Faisal and Rogers [1] reviewed the growing use of machine learning across forensic-anthropological applications and stressed that methodological transparency and validation remain central requirements. In a related forensic anthropology applica-

tion, a gender determination model was developed using hybrid classifiers, demonstrating continued interest in computational approaches to sex-related inference. Hairuddin et al. [15] also examined gender classification in forensic anthropology using particle swarm optimization-based feature selection and an optimized backpropagation neural network. Collectively, these studies demonstrate the potential of algorithmic classification in forensic anthropology. However, differences in dataset composition, feature representation, and validation procedures make direct numerical comparisons with the present cranial-measurement experiment inappropriate.

This study addresses the identified gap by systematically evaluating five feature-selection methods across five classifiers using a fully nested model-development framework, followed by assessment on a fixed internal hold-out set. The analysis examines whether AI-based feature selectors provide meaningful advantages over conventional filtering methods, whether the choice of selector affects model stability, which selector–classifier pairing achieves the strongest internal generalization, and which craniometric variables consistently contribute to model predictions. To provide a more comprehensive evaluation, the study also reports model calibration, classification errors, SHAP-based interpretations, ablation results, sensitivity to sample size, and hold-out accuracy across population groups.

The study makes four main contributions. First, it systematically evaluates 25 selector–classifier combinations within a 5 × 5 nested validation framework. Second, 20% of the data are kept completely separate from model development and used only for the final internal evaluation. Third, the evaluation goes beyond accuracy by jointly considering discrimination, calibration, feature stability, consensus across selection methods, and model explainability via SHAP. Fourth, it explicitly reports unsupported hypotheses, including the absence of a 5-percentage-point selector-family gain and the non-significant feature-method main effect. The article proceeds from literature synthesis to reproducible methods, empirical results, critical discussion, and bounded conclusions.

2. RESEARCH METHODS

2.1. Research Design and Data

A quantitative experimental design evaluated five feature selectors crossed with five classifiers. The source dataset included 2,524 individuals described by 86 columns, comprising sex, three metadata variables, and 82 cranial measurements. The sample consisted of 1,368 males and 1,156 females distributed across 30 population labels. No duplicate identifiers or explicitly missing values were found. However, an audit of zero-valued entries revealed systematic zero patterns in several measurements. In addition, four population groups were represented exclusively by male individuals.

The source-file hash was retained for provenance. Sex was normalized to M and F and encoded as female = 0 and male = 1. ID, PopNum, and Population were excluded from predictors to reduce direct leakage through identifiers or population labels. The executed configuration used all records rather than a balanced 100-per-sex subsample. This choice increased statistical power but retained sex imbalance and population-composition heterogeneity.

2.2. Preprocessing

Zero values were treated as missing in GLS, BRR, LAR, OSR, BAR, RFA, RPA, ROA, BSA, SBA, SLA, and TBA. Using a stratified 80:20 split with a random seed of 42, the dataset was divided into 2,019 observations for model development and 505 observations for the final hold-out evaluation. The missingness filter was derived solely from the development set to avoid information leakage. Eleven variables BRR, LAR, OSR, BAR, RFA, RPA, ROA, BSA, SBA, SLA, and TBA, each showed a missing rate of 26.45% and were therefore excluded for exceeding the predefined 10% threshold. After this filtering step, 71 predictor variables remained for subsequent analysis.

Imputation, scaling, selection, and classifier fitting were implemented as training-fold operations. The hold-out set did not influence missingness filtering, subset selection, hyperparameter tuning, or configuration choice. This nesting is methodologically important because feature selection performed before splitting would leak label-dependent information into the validation set. Nested cross-validation is an established remedy for bias when the same data would otherwise support both tuning and assessment [10].

2.3. Feature-Selection Methods

Information Gain represented a univariate filter that ranks features according to class-relevant information. ReliefF was used as a neighborhood-based filter from the Relief family, assigning feature weights by examining differences between nearby observations from the same class and those from different classes, commonly referred to as near hits and near misses [16]. LASSO served as an embedded linear feature-selection method, in which less informative variables were progressively reduced through coefficient shrinkage. For SHAP-based selection, features were ranked according to their absolute SHAP values derived from the model. SHAP provides a unified framework for additive feature attribution, allowing local explanations of individual predictions to be aggregated into a global assessment of feature importance [12].

SVM-RFE served as the wrapper-style selector. It iteratively removes variables associated with smaller SVM weights and has been used to reduce redundancy while retaining discriminative subsets [13]. Candidate subset sizes were 3, 5, 7, 10, and 15. All ranking and selection decisions were learned within the appropriate training partition.

2.4. Classifiers and Optimization

The classifiers were SVM, random forest, logistic regression, decision tree, and multilayer perceptron. Hyperparameters and the number of selected features were optimized within the inner cross-validation loop, with accuracy used as the selection criterion. The outer loop was then used to evaluate each configuration on data not used in the tuning process, yielding a less biased estimate of model performance. After the best-performing model family was identified, it was retuned using the entire development set and subsequently evaluated once on the untouched hold-out set.

The winning final configuration was SVM-RFE with an SVM classifier. Inner optimization selected 15 variables, selector parameter 0.1, SVM penalty parameter C = 100, and gamma = 0.001. The selected variables were ZYB, MDH, AUB, FRS, FRA, SOS, FMB, EKB, ZMB, FOL, XCB, MLS, MDB, PAF, and XFB.

2.5. Validation Protocol

Five-fold outer and five-fold inner stratified cross-validation were used on the development set. The outer-fold results supported model-family ranking and inferential comparisons; the fixed hold-out was not consulted during this process. The hold-out was stratified by sex at the individual level but not separated by population, so it estimates internal generalization to unseen individuals drawn from the same pooled dataset rather than transfer to unseen populations.

The primary performance measures were accuracy, precision, recall/sensitivity, specificity, F1-score, ROC-AUC, and average precision. Calibration measures included the Brier score, expected calibration error, maximum calibration error, and Brier skill score. Confusion-matrix counts provided direct error analysis. Training-test gaps were monitored as an indicator of overfitting.

2.6. Statistical Analysis

AI-oriented selectors were defined as LASSO, SHAP, and SVM-RFE; filters were Information Gain and ReliefF. For each outer-fold and classifier combination, accuracy scores were first averaged within each selector family and then compared using a paired t-test. Paired Cohen's d was used to quantify the standardized mean difference, while Levene's test was applied to examine differences in variance. A priori, an absolute improvement in mean accuracy of at least 0.05 was defined as the threshold for substantive improvement.

A two-way ANOVA was used to examine whether outer-fold accuracy varied according to the feature-selection method, the classifier, or the interaction between the two. The consistency of feature selection across folds was assessed using the mean pairwise Jaccard similarity calculated from the feature subsets selected in each fold. Hypergeometric tests assessed top-10 feature overlap between selector pairs. Because multiple overlap comparisons were not explicitly corrected in experiment, their p values are exploratory.

2.7. Explainability and Ablation

SHAP was applied to the final model to estimate global mean absolute feature attributions and generate a local explanation for one hold-out observation. SHAP values were interpreted as contribu-

tions to the model's output under the fitted model and selected background, not as causal effects or universal anatomical importance [12]. Ablation evaluated the final model with increasing numbers of top-ranked variables. Univariate Mann-Whitney tests and sex-specific means supplied complementary evidence of dimorphism.

2.8. Implementation Environment

The Experiment used Python 3.13.15, NumPy 2.1.3, pandas 2.2.3, scikit-learn 1.6.1, SciPy 1.16.3, SHAP 0.52.0, statsmodels 0.15.0, and the random seed was 42.

3. RESULTS AND DISCUSSIONS

3.1. Dataset Audit and Dimorphism

The analysis retained 2,524 individuals. The 80:20 partition produced 2,019 development cases and 505 hold-out cases, each with 71 retained measurements. Among the strongest univariate differences, male means exceeded female means for ZYB (135.294 vs. 125.575), JUB (118.702 vs. 111.179), MDH (29.355 vs. 25.085), GOL (183.375 vs. 174.405), and FMB (99.201 vs. 94.571). All five had Mann-Whitney p values below 0.001. Figure 1 visualizes the correlation among leading univariate variables and underscores the redundancy that motivates multivariable selection.

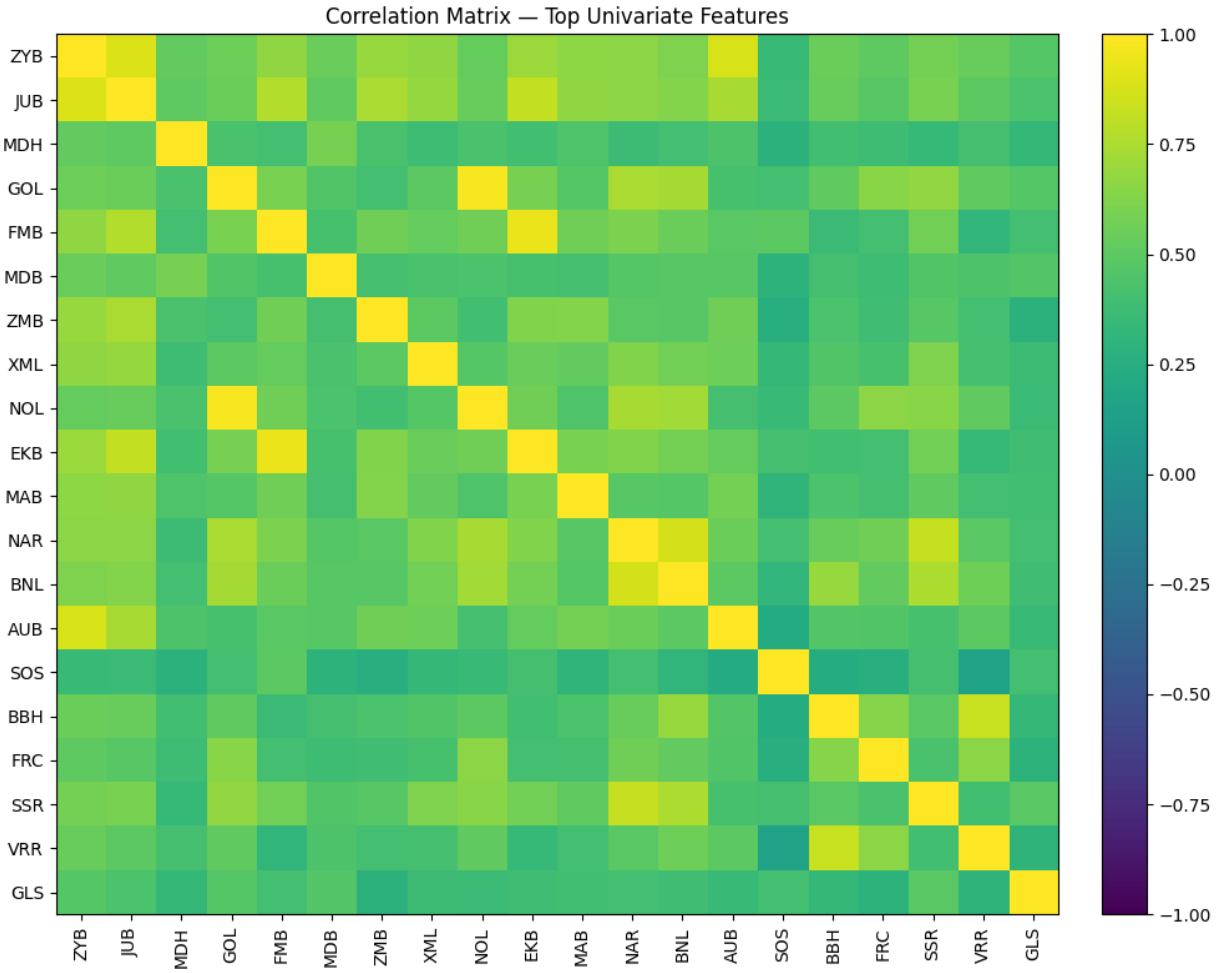


Figure 1. Correlation matrix.

3.2. Nested-Cross-Validation Performance

SVM-RFE with SVM ranked first among the 25 combinations, with mean accuracy 0.876671 (SD 0.009813), mean F1-score 0.885221, and mean ROC-AUC 0.948123. LASSO-SVM ranked second in accuracy at 0.868746, followed by SVM-RFE-LR at 0.867259, SHAP-SVM at 0.864780, and

SHAP-LR at 0.858842. The lowest mean accuracy was observed for SVM-RFE-DT at 0.802384. Table 1 shows that SVM and logistic regression generally occupied the top tier, while decision trees formed the lowest tier across selectors.

Table 1. Top 10 nested-cross-validation configurations.

Rank	Selector	Classifier	Accuracy, mean ± SD	F1	ROC-AUC
1	SVM-RFE	SVM	0.876671 ± 0.009813	0.885221	0.948123
2	LASSO	SVM	0.868746 ± 0.004994	0.878581	0.946077
3	SVM-RFE	LR	0.867259 ± 0.010902	0.877720	0.946372
4	SHAP	SVM	0.864780 ± 0.009111	0.873086	0.942082
5	SHAP	LR	0.858842 ± 0.015544	0.869065	0.937690
6	LASSO	LR	0.858348 ± 0.015369	0.868586	0.938288
7	Information gain	SVM	0.857861 ± 0.018489	0.867064	0.933584
8	ReliefF	LR	0.857854 ± 0.016802	0.868130	0.934753
9	LASSO	MLP	0.856373 ± 0.012391	0.868195	0.932716
10	SHAP	RF	0.856366 ± 0.012726	0.865569	0.930698

3.3. Selector-Family and ANOVA Findings

AI-oriented selectors had mean accuracy 0.849498 compared with 0.842105 for filter selectors, an absolute difference of 0.007393. The paired test yielded t = 2.959085, p = 0.006838, and paired Cohen's d = 0.591817. However, the 0.007393 increase was far below the prespecified 0.05 threshold; therefore, the substantive improvement hypothesis was not supported. Variances were 0.000524 and 0.000360, and Levene's p = 0.778428 did not indicate a variance difference.

The two-way ANOVA identified a strong classifier effect, F(4,100) = 49.281292, p = 7.70e-23. The feature-selection-method effect was not significant, F(4,100) = 2.229020, p = 0.071184, and the method-by-classifier interaction was not significant, F(16,100) = 1.248592, p = 0.245967. Thus, classifier choice accounted for clearer performance differences than selector choice in this experiment.

3.4. Feature Stability and Consensus

Mean Jaccard stability varied by method-classifier combination. Information Gain paired with RF or SVM reached 0.876471, the highest observed value. For the winning SVM-RFE/SVM pair, mean Jaccard stability was 0.513461, indicating moderate rather than high subset reproducibility. LASSO-SVM reached 0.662023, SHAP-SVM 0.609679, and ReliefF-SVM 0.490008.

Top-10 consensus exceeded or equalled 60% for Information Gain–SHAP (7 features), LASSO–SVM-RFE (7), ReliefF–SHAP (7), and SHAP–SVM-RFE (6). MDH and ZYB appeared with 100% selection frequency under Information Gain, LASSO, ReliefF, and SHAP and 96% – 100% under SVM-RFE configurations. Figure 2 shows the leading frequency profiles for all selectors.

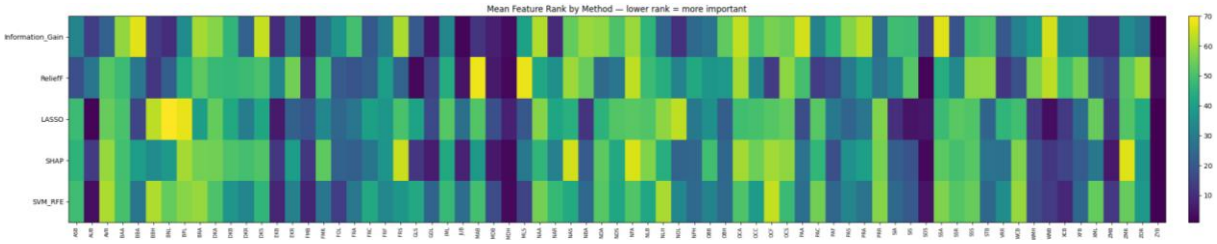


Figure 2. Mean feature rank by method.

3.5. Final Hold-Out Performance

The analysis dataset comprised 2,524 crania (1,156 female and 1,368 male). A stratified 80:20 split reserved 505 specimens for an internal independent hold-out evaluation, while the remaining 2,019 specimens were used for model development and nested cross-validation. Accordingly, the final hold-out confusion matrix contains 505 predictions rather than 2,524 observations.

The final SVM-RFE/SVM model selected 15 variables and achieved 0.897030 accuracy, 0.902174 precision, 0.908759 sensitivity, 0.883117 specificity, and 0.905455 F1-score. ROC-AUC was 0.951101 and average precision was 0.963287. The confusion matrix contained 204 true negatives, 27 false positives, 25 false negatives, and 249 true positives (Figure 3), while Figure 4 presents the ROC curve.

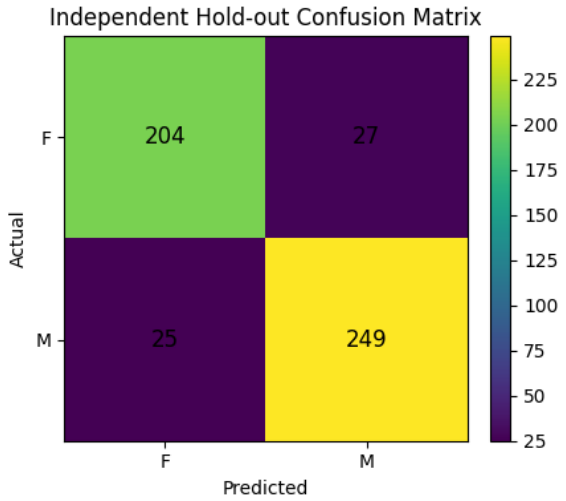


Figure 3. Independent hold-out confusion matrix.

The Brier score was 0.080417, expected calibration error 0.028816, maximum calibration error 0.237037, and Brier skill score 0.675983. The prespecified composite calibration hypothesis required Brier < 0.08, ECE < 0.08, and BSS > 0.65. ECE and BSS passed, but Brier exceeded its strict threshold by 0.000417; therefore, the composite hypothesis failed. Figure 5 presents the calibration curve.

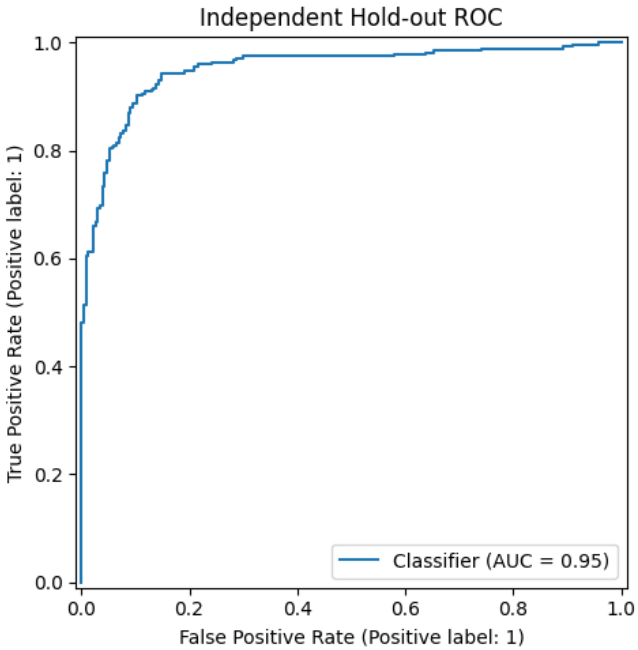


Figure 4. Independent hold-out confusion matrix.

Population-stratified hold-out accuracy ranged from 0.5556 for Andaman (n = 18) to 1.0000 in several small groups. These values are descriptive only because group sizes varied substantially and the model-selection protocol did not use population-disjoint validation.

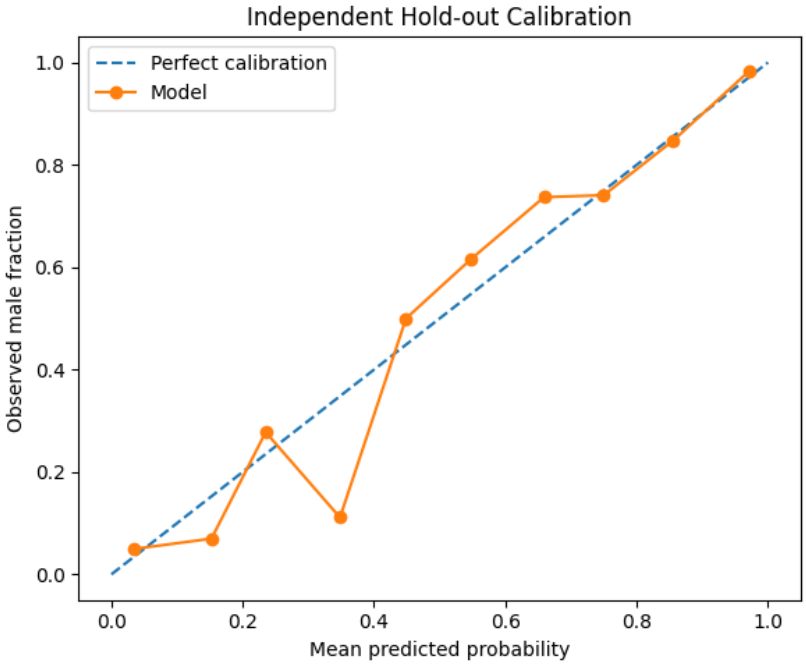


Figure 5. Independent hold-out calibration.

3.6. SHAP and feature interpretation

The final selected variables were ZYB, MDH, AUB, FRS, FRA, SOS, FMB, EKB, ZMB, FOL, XCB, MLS, MDB, PAF, and XFB. Global mean absolute SHAP values ranked ZYB first (0.238662), followed by AUB (0.110110), MDH (0.080476), EKB (0.079840), SOS (0.069707), FRA (0.066186), FMB (0.057592), and FRS (0.056094). Figure 6 shows both magnitude and directional distribution across observations.

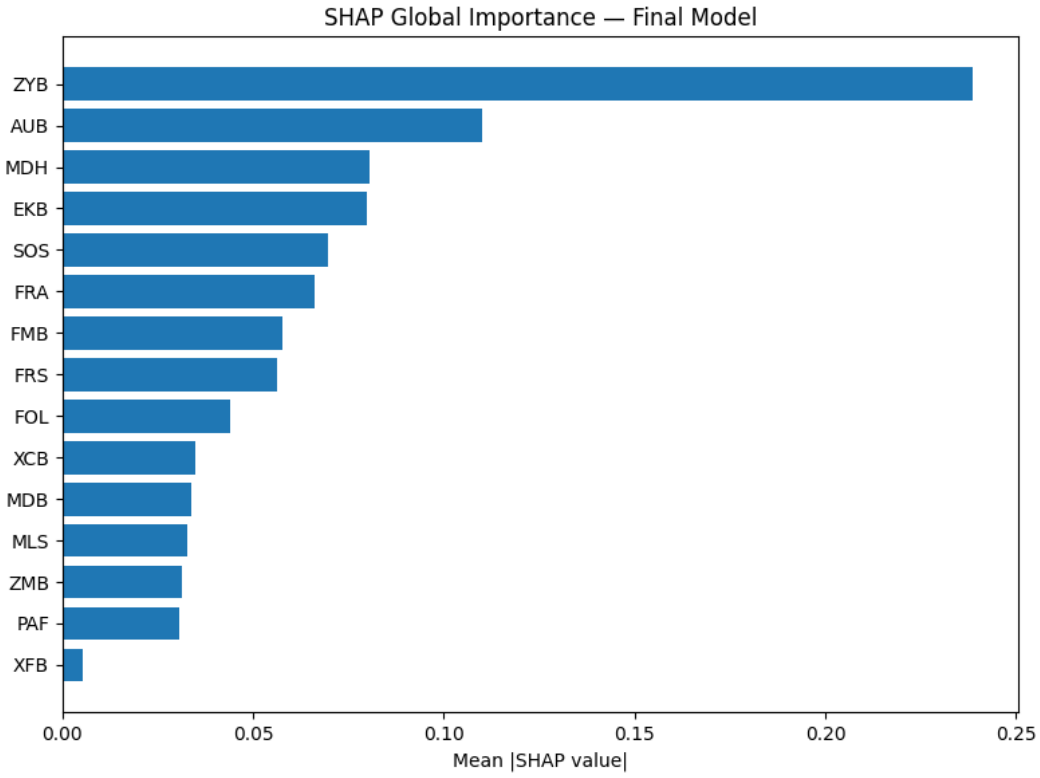


Figure 6. SHAP global importance.

For one hold-out observation, FRA contributed 0.248404 in the positive model direction, while ZYB, FRS, and SOS contributed -0.185133, -0.173432, and -0.143124, respectively. This case-level explanation demonstrates how opposing contributions combine within one prediction; it does not establish that changing an anatomical measurement would causally change sex. Figure 7 presents the local waterfall.

SHAP-based ranking produced an interpretable ordering of features, but the final best hold-out configuration was obtained with SVM-RFE rather than the SHAP-based selector. This distinction is important: feature-attribution outputs can assist in examining which variables influence a fitted model, but they do not automatically identify the subset that maximizes predictive performance. The present study shares a conceptual similarity with the skull-CT work of Çiftçi et al. [17], which applied deep feature selection and feature fusion to human sex estimation. However, the two studies are not directly comparable because Çiftçi et al. [17] relied on deep features extracted from CT images, whereas the present analysis was based on manually defined numerical cranial measurements. Therefore, the resulting feature rankings should be viewed as model-dependent indicators of association with classification outcomes rather than as evidence that any particular cranial measurement is a biological cause of sex differences.

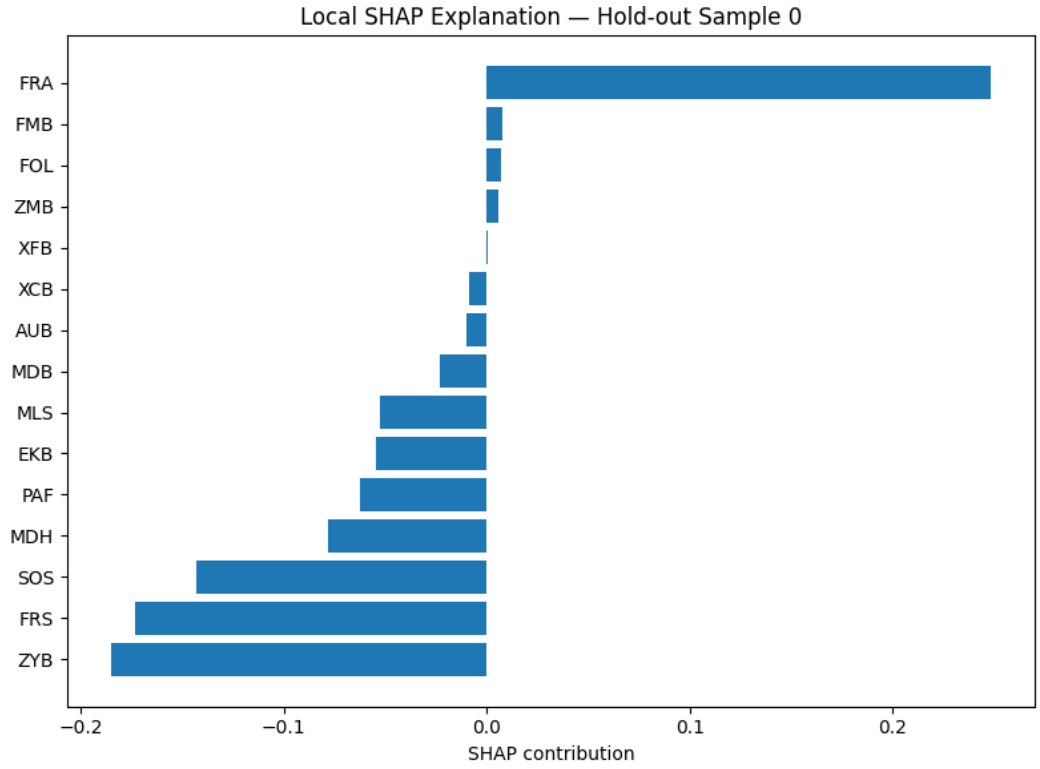


Figure 7. Local SHAP explanation.

3.7. Ablation and Sensitivity

The ablation curve evaluated performance as progressively more ranked variables were included (Figure 8). It should be interpreted as a compactness analysis of the fitted pipeline rather than an independent validation experiment. The sample-size sensitivity audit used matched subsamples of 75, 100, and 150 observations per sex. Across selectors, observed accuracies ranged from 0.665789 to 0.806505, with rankings changing by sample size. SHAP ranked highest at 100 and 150 per sex, whereas Information Gain ranked highest at 75 per sex. This variability cautions against treating the full-data ranking as invariant to sample composition.

3.8. Principal Findings and Comparison with Previous Research

The principal result was that SVM-RFE paired with SVM produced the highest nested-CV accuracy and subsequently achieved 89.70% accuracy on the 505-case hold-out. The performance is

within the broad range of useful cranial sex-estimation results reported in the literature, but its meaning depends on protocol. SexEst reported cranial accuracies between 81.2% and 87.7% using Howells data, while Del Bove and Veneziano reported 86.7% cross-validation accuracy and 84.3% on unrelated test material [7,8]. The present hold-out value is numerically higher, but no superiority claim is justified because SexEst used different classifier families and the neural-network study used a distinct 10-measurement protocol and an unrelated test set.onlinelibrary.

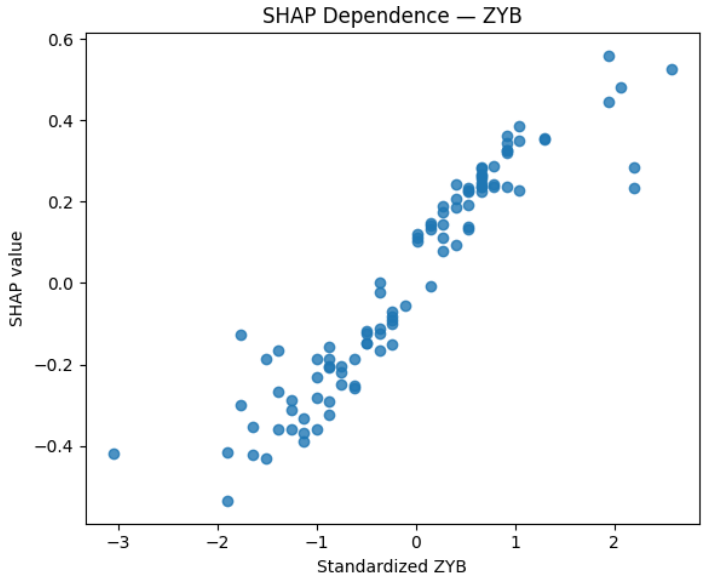


Figure 8. SHAP dependence.

The performance observed in the present study was lower than the 96.1% SVM accuracy reported for a Bulgarian CT-based sample. However, that study involved 393 adults, 47 cranial landmarks, multiple derived measurement sets, and repeated 10 × 10-fold validation, making direct comparison difficult. Differences in population structure, imaging technology, construction of candidate variables, and resampling strategy may reasonably contribute to the observed difference in accuracy. In contrast, the globally diverse sample used in the present study is likely to capture greater morphological variability than a single-population dataset and to provide a larger overall sample size.

This broader context is consistent with the systematic review by Faisal & Rogers [1], which shows that machine-learning applications in forensic anthropology now encompass a wide range of skeletal elements and biological profile estimation tasks, while also highlighting the need for clearer methodological reporting and more rigorous validation. The present study addresses these concerns by documenting nested hyperparameter tuning, independent hold-out evaluation, negative hypothesis testing, calibration analysis, and feature-selection stability.

Importantly, the strongest conclusion is not that SVM-RFE is universally superior. Given the non-significant feature-selection main effect, the first-place combination must be understood as the optimal configuration solely under this specific dataset and protocol. Its value is the combined evidence from ranking, discrimination, calibration, and parsimony, not proof of selector dominance.

3.9. Interpretation of Model Performance

The SVM-RFE/SVM pairing was selected because its selector and classifier optimize related linear-margin representations. Recursive feature elimination can discard measurements that contribute little additional information to the classification margin while preserving a subset of variables that work collectively to distinguish between classes. Guyon et al. [13] introduced SVM-RFE as a strategy for reducing feature redundancy and identifying compact sets of discriminative variables, which is consistent with the 15-variable subset obtained in the present study. The relatively low standard deviation of outer-fold accuracy (0.009813) further suggests that predictive performance remained consistent across folds, even though the specific feature subsets selected from one fold to another showed only moderate stability.

Classifier selection had a greater influence on performance than the choice of feature-selection method. SVM and logistic regression appeared most frequently among the highest-performing configurations, whereas combinations involving decision trees generally produced weaker results. This pattern may indicate that the main male–female separation captured by the standardized cranial measurements can be modeled effectively using relatively smooth decision boundaries, while a single decision tree may be more susceptible to local partitioning effects and variation across samples. Consistent with this interpretation, the classifier main effect was highly significant, whereas neither the feature-selector main effect nor the selector–classifier interaction was statistically significant. Consequently, effort devoted to leakage-controlled classifier selection may yield greater benefit than searching for increasingly elaborate selectors.

Discrimination and calibration conveyed different messages. ROC-AUC of 0.9511 and average precision of 0.9633 indicate a strong ranking of male and female observations, while the Brier score of 0.080417 narrowly missed the prespecified 0.08 threshold. The ECE of 0.028816 and positive Brier skill score show generally useful probability behavior, but the MCE of 0.237037 indicates that at least one calibration bin deviated more substantially. The strict composite calibration hypothesis appropriately remained unsupported. For forensic decision support, threshold-independent AUC should therefore not replace examination of probabilities, false-positive and false-negative consequences, or population-specific uncertainty.

The error counts were relatively balanced: 27 females were predicted male, and 25 males were predicted female. Because the sample contained more males, sensitivity exceeded specificity by approximately 2.56 percentage points. This asymmetry is modest, but its practical interpretation depends on case context and cannot be reduced to overall accuracy.

3.10. Explainable AI and Feature-Level Interpretation

ZYB was the dominant global SHAP feature, and MDH appeared among the leading features in univariate testing, selection frequency, and SHAP attribution. AUB and EKB also emerged as prominent variables in the final model interpretation. Their consistent appearance across multiple analytical perspectives strengthens the evidence that they are important within the fitted model, as their relevance is not determined by a single ranking approach. However, these measures of importance capture different aspects of the data and model. The Mann–Whitney p-value reflects marginal differences between groups, selection frequency indicates how consistently a feature is retained across resampled datasets, and SHAP values quantify the contribution assigned to a feature by the fitted model in the presence of correlated predictors.

SHAP is especially useful for distinguishing global and local explanations. The global distribution summarizes how strongly each variable contributes across hold-out observations, whereas the local waterfall illustrates how one prediction results from opposing contributions. Lundberg and Lee established SHAP as an additive feature-attribution framework with consistency and local-accuracy properties. These properties support a faithful description of the fitted prediction function, but they do not transform association into biological causation. Correlated craniometric variables can share or redistribute attribution, and the direction of this relationship can vary across individuals.

The local explanation further highlights the need for cautious interpretation. In this case, a strong positive contribution from FRA was offset by negative contributions from ZYB, FRS, and SOS. Focusing solely on the most influential global feature would therefore overlook how multiple measurements interact to shape an individual prediction. In forensic applications, case-level explanations can help analysts examine whether a prediction is supported by plausible measurement patterns or influenced by an unusual profile. Such interpretations, however, should always be considered together with data-quality assessment and an explicit acknowledgment of uncertainty. The experiment generated one local explanation; a stronger validation would examine the consistency of explanations across correct predictions, false positives, false negatives, populations, and missing-data patterns.

Comparison with earlier cranial work is necessarily indirect because many studies report accuracy but not feature-attribution distributions. The field-wide review's call for interpretability and comprehensive reporting supports inclusion of SHAP, but the present analysis should be viewed as a methodological demonstration rather than anatomical validation [1]. Replication in independent contemporary samples and consultation with standardized measurement definitions are needed before assigning broad biological meaning.

It is crucial to emphasize that these attributions explain the mathematical behavior of the model, not biological causation. While ZYB and MDH emerged consistently as strong model-associated predictors, this does not establish them as the singular anatomical drivers of sexual dimorphism in human crania.

3.11. Explainable Feature Selection and Ablation

The feature-selection results show a productive tension between predictive performance and subset stability. SVM-RFE/SVM ranked first in accuracy but had a mean Jaccard stability of 0.513461. This suggests that in highly correlated craniometric spaces, alternative features can effectively substitute for one another across training folds Information Gain with SVM or RF achieved 0.876471 stability, but lower predictive accuracy. Additionally, because multiple pairwise comparisons were performed without a formal adjustment for multiple testing, these p-values must be strictly regarded as exploratory. This indicates that the exact optimal subset was more sensitive than the resulting predictions. In correlated measurement spaces, alternative features may substitute for one another, so model accuracy can remain stable while subset membership changes.

Nogueira et al. define feature-selection stability as robustness to training-sample and stochastic perturbations and provide a chance-aware statistical framework. The Jaccard measure used in this study provides an intuitive indication of feature-set overlap, but it does not fully account for chance agreement or differences in subset size. For this reason, the results should be interpreted only as observed pairwise overlap between selected feature sets. Cross-method agreement nevertheless offers additional qualitative support. LASSO and SVM-RFE shared seven of their top 10 features, while SHAP and SVM-RFE shared six, and several pairs of methods showed hypergeometric overlap p-values below 0.001. However, because multiple pairwise comparisons were performed without a documented adjustment for multiple testing, these p-values should be regarded as exploratory rather than confirmatory.

The recurring appearance of ZYB and MDH across selectors suggests that they are robust candidates for compact measurement panels. Yet the winning final set contained 15 variables, and ablation should not be interpreted as proving that every selected measurement is indispensable. SVM-RFE can retain a group whose joint margin contribution is useful even when individual SHAP values are small. XFB, for example, had the smallest global mean absolute SHAP value among the selected variables, but its removal effect must be assessed within the complete ablation trajectory rather than from its isolated attribution.

Compared with conventional filter-based selection, AI-oriented feature-selection methods achieved a statistically detectable improvement ($p = 0.0068$). However, this 0.74-percentage-point gain was well below the prespecified 0.05 threshold, demonstrating statistical detectability but lacking practical significance. This finding suggests that greater algorithmic complexity does not necessarily translate into a substantial operational advantage.

3.12. Practical and Scientific Implications

The final model retained 15 features from the 71 candidate measurements, resulting in a more manageable measurement panel. This reduction may help lessen data-acquisition effort, streamline quality-control procedures, and make model auditing more focused. This reduction offers significant practical value in forensic anthropology by minimizing the manual data-acquisition burden and reducing the likelihood of missing variables when analyzing fragmented remains. In the hold-out evaluation, both sensitivity and specificity were above 0.88, with false-positive and false-negative counts remaining broadly comparable. These results support the model's use as an experimental decision-support tool within a similar data-generating setting, rather than as a stand-alone system for forensic identification.

An important scientific implication of these findings is the clear distinction between the effects of classifier choice and feature-selection strategy. Most of the variation in predictive performance was associated with the classifier, while differences among selector families were comparatively limited. This indicates that future craniometric machine-learning studies should evaluate new feature-selection approaches against strong classifier baselines and within a properly nested tuning framework. It also highlights that feature selection may still be valuable for producing more compact and interpretable models, even when improvements in predictive accuracy are relatively small. The broad-

er ML-forensic literature increasingly emphasizes transparent and validated decision support, consistent with this multi-criterion framing [1].

The observed feature consensus may guide future measurement protocols, but it cannot define a universal sexually dimorphic panel. Population composition, preservation, measurement technique, and secular change can alter distributions. SexEst's worldwide training strategy and the population-inclusive neural-network protocol demonstrate the practical interest in broadly applicable models, yet both lines of work still require evaluation in target contexts [7,8]. The present pooled individual-level hold-out does not resolve this transfer problem [7,8].

The practical relevance of the findings lies in their potential to support, rather than replace, forensic anthropological judgment. Pilli et al. [2] and Pilli et al. [3] emphasize that identification of human remains benefits from combining complementary forms of evidence. In this context, a craniometric classifier may provide an additional quantitative input when skeletal material is available. However, the present model was evaluated on an internal hold-out drawn from the same overall dataset, not on an independently collected external dataset. Its performance should therefore be interpreted as evidence of internal validation only. This limitation is especially important in forensic applications, where transparency, reproducibility, robustness, and cautious generalization are necessary for responsible deployment [1,18].

3.13. Limitations and Generalizability

he most important limitation of this study remains the lack of external validation. Because the internal hold-out set was not population-disjoint, the observed population-specific accuracies. As a result, similarities in collection history, measurement protocols, and population structure may have made the hold-out evaluation less challenging than applying the model to data from a different laboratory, historical period, or population indicate that the model may inadvertently learn population structure. The population-specific accuracy analysis further highlights this concern, as performance varied considerably across groups, and several populations were represented by only a small number of hold-out cases.

A second limitation is the absence of an all-feature baseline without feature selection. The current design allows direct comparison among feature-selection methods, but it cannot determine whether selecting features actually improves predictive performance relative to using all 71 retained measurements. Third, the aggregate comparison between AI-oriented and filter-based selectors was based on observations that were correlated across folds and classifiers. Although the paired t-test provides a useful summary of these differences, it does not fully account for the dependence created by repeatedly evaluating models on the same dataset. Similarly, the two-way ANOVA applies a conventional factorial analysis to fold-level performance estimates and should therefore not be interpreted as a definitive population-level causal assessment. Finally, feature-selection stability was quantified using mean Jaccard similarity rather than the chance-corrected stability framework proposed by Nogueira et al. [11].

4. CONCLUSION

This study developed and evaluated a leakage-controlled machine-learning framework for craniometric sex classification by comparing five feature-selection methods Information Gain, ReliefF, LASSO, SHAP-based selection, and SVM-RFE across five classifiers. Using 5-fold outer and 5-fold inner cross-validation, SVM-RFE combined with SVM produced the highest mean validation accuracy of 0.8767. After optimization on the complete development set, the resulting 15-feature model achieved an accuracy of 0.8970, an F1-score of 0.9055, a sensitivity of 0.9088, a specificity of 0.8831, and a ROC-AUC of 0.9511 on the internal hold-out set, which estimates internal generalization rather than external transportability.

The findings show that a relatively compact set of craniometric measurements can maintain strong discriminatory performance. ZYB, AUB, and MDH emerged as the most influential variables in the global SHAP analysis, while ZYB and MDH were also selected consistently by several feature-selection methods. However, the results do not provide evidence that AI-oriented feature-selection approaches offer a substantial advantage over conventional filter methods. Their average accuracy improvement was only 0.0074, well below the prespecified threshold of 0.05 for a meaningful gain. Con-

sistent with this result, the two-way ANOVA showed a significant effect of classifier choice, but no significant effect of feature-selection method and no significant interaction between classifier and selection method.

The principal contribution of this study is therefore a transparent and internally validated analytical framework that integrates nested model selection, independent hold-out evaluation, feature-stability assessment, calibration analysis, and explainable AI. The resulting model should not yet be regarded as a universally applicable forensic instrument because the hold-out observations originated from the same underlying dataset, and an all-feature baseline was not evaluated. Future research should incorporate population-disjoint and independent external validation, explicit comparison with a no-selection baseline, observer-error assessment, chance-corrected feature-stability analysis, and prospective calibration. These steps are essential for determining whether the identified feature panel can provide reliable and defensible decision support across contemporary forensic populations and operational settings.

AUTHOR CONTRIBUTIONS

Iis Afrianty: Conceptualization, methodology, research planning, supervision, validation, interpretation of findings, and writing—review and editing. Liza Afriyanti: Data curation, preprocessing, software implementation, feature selection experiments, nested cross-validation, model training and evaluation, statistical analysis, SHAP interpretation, visualization, and writing—original draft. Both authors contributed to discussing the findings and limitations, revising the manuscript, and approving the final version. Both authors agreed to be accountable for the accuracy and integrity of the work.

FUNDING STATEMENT

This research received no external funding.

DECLARATION OF AI USE

The authors declare that no AI tools were used in preparing this manuscript.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All data supporting this study are included in the article. No separate datasets are available.

ETHICAL APPROVAL

Ethical approval was unnecessary, as no humans or animals were involved.

ACKNOWLEDGMENTS

The author acknowledges the funder for its financial support of this research in 2026. Appreciation is expressed to Universitas Islam Negeri Sultan Syarif Kasim Riau for its support and for providing the laboratory facilities used in this research.

REFERENCES

[1] Faisal, E., & Rogers, T. L. (2025). A review of the literature on the applications of machine learning in forensic anthropology. *Forensic Science International*, **376**, 112579. DOI: https://doi.org/10.1016/j.forsciint.2025.112579

[2] Pilli, E., Palamenghi, A., & Cattaneo, C. (2025). Forensic skeletal and molecular anthropology face to face: Combining expertise for identification of human remains. *Annals of the New York Academy of Sciences*, **1550**(1), 77-107. DOI: https://doi.org/10.1111/nyas.15398

[3] Pilli, E., Palamenghi, A., Morelli, S., Mazzarelli, D., De Angelis, D., Jantz, R. L., & Cattaneo, C. (2023). How physical and molecular anthropology interplay in the creation of biological profiles of unidentified migrants. *Genes*, **14**(3), 706.
DOI: https://doi.org/10.3390/genes14030706

[4] Gupta, S. K., Yesuf, S. H., & Nain, N. (2022). Real-time gender recognition for juvenile and adult faces. *Computational Intelligence and Neuroscience*, **2022**, 1-15.
DOI: https://doi.org/10.1155/2022/1503188

[5] Ulubaba, H. E., Atik, İ., Çiftçi, R., Eken, Ö., & Aldhahi, M. I. (2025). Deep learning for gender estimation using hand radiographs: A comparative evaluation of CNN models. *BMC Medical Imaging*, **25**(1), 260.
DOI: https://doi.org/10.1186/s12880-025-01809-8

[6] Wang, X., Liu, G., Wu, Q., Zheng, Y., Song, F., & Li, Y. (2024). Sex estimation techniques based on skulls in forensic anthropology: A scoping review. *PLOS ONE*, **19**(12), e0311762.
DOI: https://doi.org/10.1371/journal.pone.0311762

[7] Constantinou, C., & Nikita, E. (2022). SexEst: An open access web application for metric skeletal sex estimation. *International Journal of Osteoarchaeology*, **32**(4), 832-844.
DOI: https://doi.org/10.1002/oa.3109

[8] Del Bove, A., & Veneziano, A. (2022). A generalised neural network model to estimate sex from cranial metric traits: a robust training and testing approach. *Applied Sciences*, **12**(18), 9285.
DOI: https://doi.org/10.3390/app12189285

[9] Toneva, D., Nikolova, S., Agre, G., Zlatareva, D., Hadjidekov, V., & Lazarov, N. (2021). Machine learning approaches for sex estimation using cranial measurements. *International Journal of Legal Medicine*, **135**(3), 951-966.
DOI: https://doi.org/10.1007/s00414-020-02460-4

[10] arma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**(1), 91.
DOI: https://doi.org/10.1186/1471-2105-7-91

[11] Nogueira, S., Sechidis, K., & Brown, G. (2018). On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research*, **18**(174), 1-54.
URL: https://www.jmlr.org/papers/v18/17-514.html

[12] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, **30**, 4768-4777.
URL: https://proceedings.neurips.cc

[13] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1-3), 389-422.
DOI: https://doi.org/10.1023/A:1012487302797

[14] Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, **19**(1), 432.
DOI: https://doi.org/10.1186/s12859-018-2451-4

[15] Hairuddin, N. L., Yusuf, L. M., Othman, M. S., & Nasien, D. (2021). Gender classification using a PSO-based feature selection and optimised BPNN in forensic anthropology. *International Journal of Computer Aided Engineering and Technology*, **15**(2/3), 232.
DOI: https://doi.org/10.1504/IJCAET.2021.117133

[16] Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *F. Bergadano & L. Raedt (Eds.), Machine Learning: ECML-94*, **784**, 171–-82.
DOI: https://doi.org/10.1007/3-540-57868-4_57

[17] Çiftçi, R., Dönmez, E., Kurtoğlu, A., Eken, Ö., Samee, N. A., & Alkanhel, R. I. (2024). Human gender estimation from CT images of skull using deep feature selection and feature fusion. *Scientific Reports*, **14**(1), 16879.
DOI: https://doi.org/10.1038/s41598-024-65521-3

[18] Greene, T., Shmueli, G., Fell, J., Lin, C.-F., & Liu, H.-W. (2022). Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **185**(Supplement_2), S692-S723.
DOI: https://doi.org/10.1111/rssa.12966