SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 327-338, DOI: 10.59190/stc.v6i3.390

Application of recursive feature elimination for sex classification of skull bones using random forest

Zam Afryan, Iwan Iskandar*, Iis Afrianty, Benny Sukma Negara, Fadhilah Syafria
Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

In forensic anthropology, sex estimation from the skull was a crucial initial step when visual identification of a body was not possible. Conventional methods relied on morphological assessment by experts, which was inherently subjective and dependent on the observer's experience. To address this limitation, this study implemented a computational approach using the random forest algorithm combined with the Recursive Feature Elimination feature-selection technique. The approach was evaluated using craniometric measurements from 2,524 individuals, comprising 1,368 males and 1,156 females, sourced from the Howell's craniometric dataset. The main challenge was the high dimensionality of the data, comprising 85 measurement features after non-biological attributes were removed. Using an excessive number of variables simultaneously introduced irrelevant information that lowered the model's ability to recognize true patterns, so the feature-selection technique was used to iteratively select the most informative measurements. The results showed that the model using all 82 features achieved an accuracy of 86.49 percent, while the optimized model using only 20 selected features achieved a higher accuracy of 86.85 percent. This indicated that by reducing the feature set by 75 percent, the model became lighter while remaining more accurate. The selection process further identified that cheekbone width and the height of the posterior ear protrusion were the most discriminative measurements between male and female crania, consistent with established biological evidence. In conclusion, the combination of random forest and recursive feature elimination produced an efficient and accurate sex-identification model, opening opportunities for its development as an objective forensic identification tool in the future.

ARTICLE INFO

Article history:

Received Jun 21, 2026
Revised Jun 22, 2026
Accepted Jun 23, 2026

Keywords:

Craniometry
Feature Selection
Forensic Anthropology
Machine Learning
Sexual Dimorphism

* **Corresponding Author**
E-mail address: iwan.iskandar@uin-suska.ac.id

1. INTRODUCTION

Forensic anthropology is a branch of science that applies biological methods to assist legal processes in establishing the identity of individuals from skeletal remains [1, 2]. In field practice, a forensic expert needs to build a biological profile that includes estimates of sex, age, ancestry, and stature, whether in criminal investigations, mass disasters, or archaeological research [1].

Among all components of the biological profile, sex determination is considered the earliest and most decisive step [3, 4]. This is because the methods used to estimate age, ancestry, and stature are generally influenced by biological differences between males and females, so if sex is unknown, the estimation of other parameters becomes less accurate [3, 5].

In sex identification from human skeletal remains, the pelvis is known as the most reliable indicator because its shape differs greatly between males and females [1]. However, in real field conditions, the pelvis is often unavailable in complete form due to damage, fragmentation, or simply not being found [2]. Under such conditions, the skull (cranium) becomes the most important

alternative because it has high resistance to post-mortem damage and still retains many features that distinguish males from females [2].

To date, forensic experts have determined sex from the skull by visually observing physical traits such as forehead shape, brow ridge prominence, and jaw size [6]. This approach relies entirely on the experience and expertise of each observer, so the results can differ from one researcher to another, a problem known as inter-observer variability [6]. This inconsistency has driven scientists to seek a more objective approach that can be repeated with consistent results [7].

Advances in computer technology have opened new opportunities through Machine Learning (ML), the ability of computers to automatically learn to recognize patterns from data without being manually programmed for every rule [7]. Previous studies have proven that ML algorithms can classify sex from skull measurement data with competitive accuracy and far greater consistency than visual observation [4]. Several algorithms that have been tested in forensic research include Support Vector Machine (SVM), which has proven effective in classifying sex from the sacrum bone with a highest accuracy of 85.56% [8], Artificial Neural Network, which is able to capture non-linear patterns in morphometric data [9], and discriminant analysis, which is still widely used as a comparison method, with a combination of SVM and logistic regression (LR) achieving 90% accuracy [10]. This is supported by earlier research [11] using Random Forest and XGBoost for sex determination through mandibular index analysis using lateral cephalograms. The results showed that the Random Forest model performed better, achieving 97.20% accuracy and 97.65% precision, while XGBoost achieved 96.26% accuracy and 95.40% precision. Based on the comparison of these ML methods, Random Forest demonstrates superior accuracy performance. Therefore, this study applies Random Forest (RF) because the algorithm works by building many decision trees simultaneously and combining their predictions through majority voting [12]. RF is also more stable and less easily affected by unrepresentative data, since the error of one tree can be corrected by the other trees [12].

Although these algorithms show promising results, a major challenge arises when using standard craniometric datasets such as the Howell's Craniometric Dataset, which contains up to 82 measurement features from various populations worldwide [4]. Using too many features at once without careful selection can make the model overly complex and fail to recognize true patterns in new data, a phenomenon known as overfitting [13]. In addition, many of these features are redundant or do not contribute meaningfully to distinguishing sex, so their presence can interfere with the model's learning process [13].

To optimize Random Forest performance on such high-dimensional data, one solution proven effective in various studies is the Recursive Feature Elimination (RFE) feature-selection technique [14]. RFE works by evaluating and discarding the least useful features step by step and repeatedly until only the most informative features remain [15]. This iterative approach is superior to ordinary feature-selection methods because it can capture inter-feature relationships that are not visible when features are evaluated one by one [16]. A number of previous studies have recommended the use of RFE on complex biological data, as it has been shown to improve model accuracy while reducing computation time, an important advantage in the forensic context [14]. Other research has also shown that RFE is effective in optimizing classification models in the health field, such as breast cancer detection [15], stunting classification [17], and environmental data analysis [18], all of which share high-dimensional data characteristics similar to craniometric data.

Based on the above, this study aims to apply a combination of the Random Forest and Recursive Feature Elimination methods to classify sex based on craniometric data from the Howell's Craniometric Dataset, while also identifying the craniometric features that are most influential in distinguishing male and female skulls.

2. RESEARCH METHODS

The research workflow was designed systematically to build an objective and efficient sex classification model. The research stages include data collection, data pre-processing, feature selection using Recursive Feature Elimination (RFE), modeling using Random Forest (RF), and model performance evaluation, as shown in Figure 1.

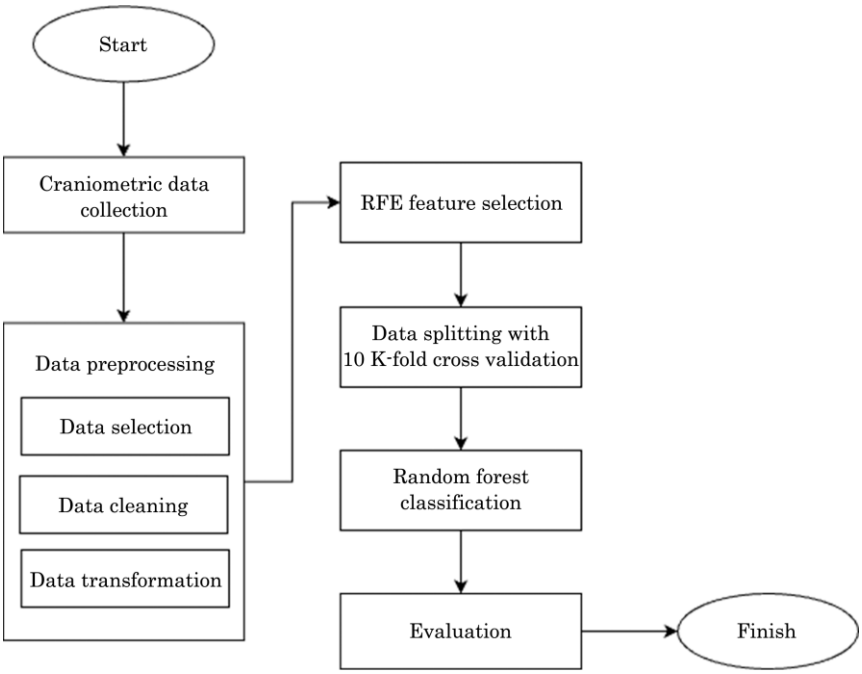

Figure 1. Research methodology.

2.1. Research Dataset

The data used in this study is sourced from the Howell's Craniometric Dataset (https://web.utk.edu/~auerbach/HOWL.htm), a standard anthropometric data repository covering skull measurements from various global populations. The dataset totals 2,524 samples, consisting of 1,368 males (Male) and 1,156 females (Female). Each sample includes linear craniometric measurement features (such as Glabello-Occipital Length and Maximum Frontal Breadth). The use of these linear cranial dimensions is consistent with prior research showing that metric data has high and stable diagnostic value for sex estimation across various populations [10]. which is further supported by [19] who emphasized that standardized craniometric measurements provide an objective and legally robust framework for sex determination.Performance analysis of sex classification using skeletal morphometric data is an important aspect of forensic anthropology in supporting accurate identification [8].

Table 1. Sample craniometric measurement variables in the Howell dataset.

CLASS	GOL	ZYB	AUB	RFA	OCA	…	TBA
M	189	133	119	0	117	…	0
M	182	137	125	0	119	…	0
M	191	134	125	0	111	…	0
…	…	…	…	…	…	…	…
F	160	117	112	60	131	…	156

2.2 Data Pre-processing

Before modeling, a pre-processing stage was carried out to ensure data quality and integrity:

1. Data selection: non-metric and categorical attributes such as “Population”, “PopNum”, and “ID” were removed from the dataset, as they only provide information about population names and counts. Removing these attributes reduced the number of features to 82, forcing the model to learn pure patterns of sexual dimorphism based on bone morphology (morphometrics) rather than memorizing demographic labels. making the model more applicable to forensic cases involving unidentified remains.
2. Data cleaning: several values of 0 were found in the dataset, representing missing values. These zero values were converted to NaN and imputed using the Mean Imputation method. To prevent

data leakage, this imputation process was strictly isolated within the Pipeline architecture so that the mean calculation was performed only on the training data during cross-validation. Mathematically, the calculation of the mean value used to fill missing data for a given measurement feature is formulated as:

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \quad (1)$$

where:

$\bar{x}_j$ = The mean value used to fill missing or zero-valued data;

m = The number of skull records that have an actual (non-missing) value in that column;

x_{ij} = The original measurement value of the skull record being calculated.

3. Data transformation: the target variable (Sex) was converted into numeric type, with the label 'M' (Male) converted to 1 and 'F' (Female) converted to 0.

2.3 Recursive Feature Elimination (RFE)

RFE was applied to reduce the dimensionality of the data from 82 features. This method works by training the model on all features, calculating feature importance, and iteratively eliminating the features with the lowest scores [13]. This stepwise elimination process has proven effective in filtering out the most important variables from complex biological data, resulting in a more stable predictive model [14]. The use of RFE in this study is supported by various studies showing the effectiveness of systematic feature reduction in finding an optimal point.

The RFE testing in this study refers to [15], which performed feature reduction consistently with an interval of 5 (starting from 30, 25, 20, 15, 10, down to 5 features) and found that 15 features were the most optimal number. Similarly, [14] used an interval of 10 to reduce 100 features down to 10 main features.

Based on this foundation, the feature subsets determined in this study were set at 80, 60, 40, and 20 features, performed systematically with a reduction interval of 20 features. This is supported by the approach in [16] regarding decision thresholds in Random Forest classification, which tested feature trimming at large percentage scales, namely 20%, 30%, 40%, up to 50% at once. This range was used to evaluate model stability at various levels of data dimensionality and to identify the optimal threshold at which dimensionality reduction does not compromise classification accuracy. This stepwise comparison approach is consistent with the strategies in [14, 15] balancing computational efficiency and predictive performance. By limiting the feature set to the best 20 features (a 75% reduction), the Random Forest model is shown to work more efficiently while still maintaining high diagnostic accuracy for sex identification [17].

The general stages of RFE can be explained as follows:

1. Assume a dataset with n features: $X = \{x_1, x_2, x_3, \dots, x_n\}$;
2. Build a classification model (e.g., Random Forest) using all features;
3. Calculate the feature importance from the training results;
4. Remove the feature or group of features with the lowest importance value;
5. Repeat steps 2 – 4 recursively until the desired number of features remains [5].

The Random Forest (RF) algorithm calculates feature importance using an impurity decrease measure. One of the most commonly used impurity measures is Gini Impurity (the Gini Index). This Gini-based metric is highly efficient at separating data patterns and has proven reliable when applied specifically to human skeletal morphometric data [20]. Gini Impurity measures the probability of misclassification if an element is randomly selected from a given node; a value of 0 means the node is perfectly "pure" (all samples belong to one class). Based on the mathematical calculation of the Gini index in Random Forest modeling [16], the formula is:

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \quad (2)$$

where:

C = The number of classes.

p_i = The proportion of data belonging to class i within a given node.

The importance value of a feature (referred to as Gini Importance or Mean Decrease Impurity) is calculated based on how much that feature reduces Gini Impurity when used as a node split. This reduction, called Gini Impurity Decrease, is calculated as the parent node's Gini Impurity minus the weighted average Gini Impurity of its two child nodes [16]:

$$\mathrm{VIM}_{\mathrm{node}}^{\mathrm{Gini}} = \mathrm{GI}_{\mathrm{parent}} - (\mathrm{w}_{\mathrm{left}} \times \mathrm{GI}_{\mathrm{left}}) - (\mathrm{w}_{\mathrm{right}} \times \mathrm{GI}_{\mathrm{right}}) \quad (3)$$

where:
$\mathrm{VIM}_{\mathrm{node}}^{\mathrm{Gini}}$ = The Gini Importance score (impurity decrease) at that node.
$\mathrm{GI}_{\mathrm{parent}}$ = The Gini Impurity of the parent node (before the split).
$\mathrm{GI}_{\mathrm{left}}$ = The Gini Impurity of the left/right child node.
$\mathrm{w}_{\mathrm{right}}$ = The proportion (weight) of samples going to the left/right child node.

This score is then summed across all nodes where the feature is used, and averaged across all trees in the Random Forest to obtain the total Gini Importance score for that feature.

2.4. Random Forest

Random Forest generally uses an impurity function to determine the best split at each node. As explained in Formula (2), the Gini index is the most commonly used metric to measure this impurity. Furthermore, as shown in Formula (3), Gini Impurity Decrease is calculated to determine how much a feature is able to reduce impurity, which becomes the basis for calculating feature importance [16].

This algorithm forms a collection of decision trees randomly through bootstrap aggregating and determines the final result through majority voting. The main advantage of RF is its ability to keep variance low (preventing overfitting) even when trained on data with a large number of features [20]. Mathematically, the classification prediction in Random Forest can be expressed as:

$$\hat{y} = \mathrm{mode}\{h_1(x), h_2(x), \ldots, h_N(x)\} \quad (4)$$

where, $h_i(x)$ is the prediction result of the i-th decision tree, N is the total number of trees in the Random Forest, and $\hat{y}$ is the final prediction result obtained through majority voting across all trees.

In this study, the 10-Fold Stratified Cross-Validation method was used to ensure fair evaluation. The entire skull dataset (2,524 samples) was split into 10 equally sized groups. In each round, 9 groups (around 2,272 samples) were combined and used as training data, while the remaining 1 group (around 252 samples) was used as test data. This process was repeated 10 times with the test group rotating each time, so that every data point was used as test data exactly once. This method ensures that the final accuracy obtained is truly representative and that the model does not "memorize" specific data patterns (overfitting).

2.5. Model Evaluation

The classification model's performance was evaluated using a Confusion Matrix through the 10-Fold Cross-Validation mechanism. The Confusion Matrix compares the model's predictions with the actual labels of the data. From this matrix, four basic values are extracted: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These four values are used to calculate the following evaluation metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

3. RESULTS AND DISCUSSIONS

3.1 Implementation

Implementation was carried out based on the results of the analysis and design of the sex classification model using the Random Forest method and the Recursive Feature Elimination (RFE) feature-selection technique. The implementation was developed using the Python programming language. The dataset used consists of 2,524 samples, namely 1,368 males (Male) and 1,156 females (Female), with performance evaluated using a cross-validation approach.

3.2. Data Pre-processing

The data pre-processing stage aims to transform raw data into a clean and optimal dataset through three main steps: data selection, data cleaning, and data transformation.

1. Data selection: in this phase, non-biological and administrative attributes such as ID, PopNum, and Population were eliminated, given that the entire database in this study purely consists of physical (metric) measurements of skull anatomy. These demographic attributes were excluded because they do not represent any dimension or morphological feature of the skull structure being measured. In real field conditions (such as the discovery of unidentified remains), the population origin of the victim is in fact never known beforehand [4]. This step is also consistent with the research methodology used by previous researchers [21].
2. Data cleaning: missing values and outliers were identified and handled. Data inspection revealed several zero values representing missing data. These missing values were handled using the Mean Imputation method, referring to Formula (1) described in Section 2.2, where missing values were automatically substituted with the mean value of the relevant measurement feature within the Pipeline to avoid data leakage. Next, abnormal data points were detected using the Interquartile Range method. As a result, 819 skull records were found to have measurements that were either too large or too small compared to typical human dimensions. The largest numbers of abnormal measurements were found in parietal arc length (PAS: 93 records), cranial base angle (BBA: 85 records), frontal arc fraction (FRF: 58 records), and maximum cranial breadth (XCB: 48 records). The distribution of these abnormal data points is shown in Figure 2.

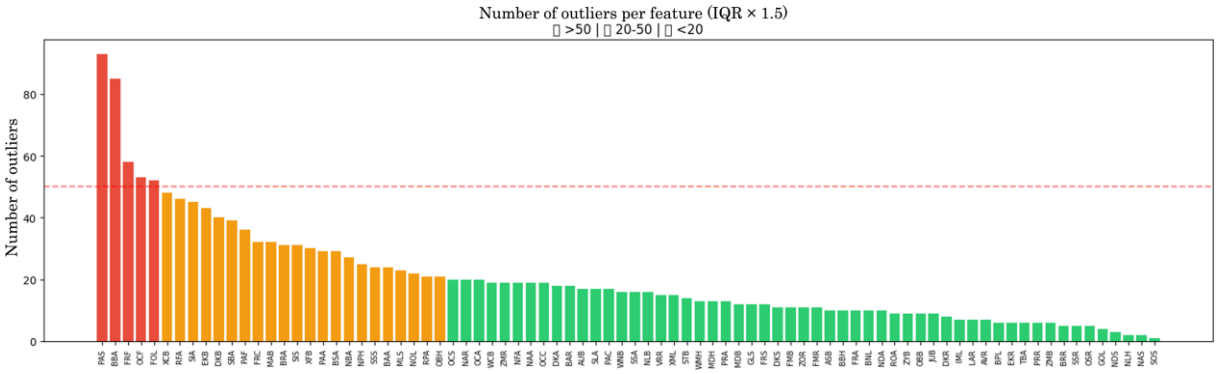

Figure 2. Outlier distribution analysis.

Although they have abnormal (outlier) measurements, all 819 of these rows were retained and not removed from the original 2,524 data points. From an anthropological perspective, such extreme measurements represent natural physical heterogeneity across the world's populations; removing this data was a concern, as it could cause a lack of data density that would reduce the model's ability to recognize global population variation, as highlighted by [22, 23] asserts that decision trees are classifiers with an inherent advantage of robustness to outliers. This trait is automatically inherited by ensemble models, where [12] explicitly states that the Random Forest structure is relatively robust to outliers and noise. The binary node splitting mechanism in this algorithm ensures that the presence of extreme values at the edges of the data distribution does not disrupt the model's main decision boundary. The ability of machine learning models to maintain reliable classification performance despite the inherent variability of skeletal measurements has also been

observed in other anatomical. Previous studies [24, 25] demonstrated successful sex estimation using ensemble and Random Forest-based approaches on piriform aperture and lumbar vertebral morphometric data, respectively, highlighting the suitability of these methods for complex skeletal datasets.

3. Data transformation: the raw data type was converted using label encoding on the target attribute Sex. The categorical string data 'M' (Male) and 'F' (Female) were converted into a discrete binary numeric format, namely 0 for female (F) and 1 for male (M). This numerical transformation is necessary because the internal calculation of Gini Impurity and the formation of split nodes in Random Forest require class proportions in numeric form.

3.3. RFE (Recursive Feature Elimination) Feature Selection Process

The selection of the most important skull measurements was carried out using the Recursive Feature Elimination (RFE) method. This method works by training the model on all measurements, calculating feature importance, and progressively eliminating the measurements with the lowest scores. The elimination was carried out consistently with a reduction interval of 20 measurements. The results of the 20 selected features are presented in Table 2.

Table 2. The 20 selected features based on Gini importance values (RFE-RF model).

Rank	Code	Measurement name	Importance value	Biological interpretation
1	ZYB	Bizygomatic breadth	0.145966	Facial / cheekbone width
2	MDH	Mastoid height	0.104613	Height of the posterior ear protrusion
3	JUB	Bijugal breadth	0.094402	External facial arch width
4	SOS	Supraorbital projection	0.063118	Brow ridge prominence
5	GOL	Glabello-occipital length	0.062699	Maximum skull length
6	ZMB	Zygomaxillary breadth	0.053157	Zygomaxillary region width
7	MDB	Mastoid breadth	0.049623	Mastoid width
8	FMB	Fronto-malar breadth	0.044045	Fronto-malar width
9	AUB	Auricular breadth	0.042528	Cranial width at the ear canal area
10	NOL	Naso-occipital length	0.037281	Naso-occipital length
11	XML	Zygomatic length	0.034389	Cheekbone length
12	PAF	Parietal arc fraction	0.033699	Fraction of parietal curvature on the skull roof
13	BBH	Basion-bregma height	0.032691	Skull height (vertical cranial dimension)
14	SIA	Simotic index angle	0.031762	Simotic index angle
15	SIS	Simotic subtense	0.031499	Nasal bridge depth
16	PAC	Parietal arc	0.029600	Parietal arc
17	XCB	Maximum cranial breadth	0.027834	Maximum cranial width
18	MAB	Mandibular breadth	0.027389	Lower jaw width
19	FRC	Frontal chord	0.027028	Straight-line distance between points on the forehead
20	WNB	White-nasale breadth	0.026676	Width at the nasal area

Based on Table 2, cheekbone width (ZYB) and the height of the posterior ear protrusion (MDH) are the two most important measurements for distinguishing sex. This is consistent with biological evidence, as the cheekbone and posterior ear protrusion naturally develop larger in males than in females.

Systematically, the importance value of each feature is calculated based on the impurity decrease measure using the Gini index formulated in Formula (2). This importance value is then calculated based on the feature's ability to reduce impurity through Gini Impurity Decrease, formulated in Formula (3).

3.4. Data Splitting

10-Fold Stratified Cross-Validation was used to test the fairness of the model and ensure that the system does not merely memorize the training data (overfitting). Through the Python program code, the total 2,524 skull records were automatically split into 10 balanced groups, with each group

maintaining a proportion of male and female samples consistent with the original dataset. This repeated testing process ran for 10 rounds with the following data rotation rules:

1. In the first round, data group number 1 was locked as the test set (252 samples), while groups 2 through 10 were combined as the training set (2,272 samples).
2. In the second round, data group number 2 was locked as the test set, while group 1 and groups 3 through 10 became the training set.
3. This rotation logic continued automatically until all 10 data groups had served as the test set.

At the end of each round, the system recorded the accuracy value obtained. After all 10 rounds were completed, all accuracy values were averaged to obtain the model's final performance value.

3.5. Random Forest Classification Process

After the Recursive Feature Elimination (RFE) algorithm successfully determined the 20 best skull measurements, the sex classification process was carried out using the Random Forest algorithm. Based on the results of parameter variation testing via GridSearchCV, the final model was locked using the most optimal configuration, namely 200 decision trees (n_estimators), unlimited tree depth (max_depth = None), and a minimum split threshold of 2 (min_samples_split). The data classification process runs through four main computational stages:

1. Bootstrapping: the system draws random samples with replacement from the 2,272 training records to build 200 independent decision trees. This randomization process is important so that each tree learns a different combination of data samples, making the model more stable and less prone to generalization errors.
2. Node splitting: at each branch of the decision tree, the algorithm selects the best measurement threshold based on the 20 skull measurements selected by RFE. The purity of the sex split is calculated based on the Gini Impurity Decrease value, formulated in Formula (3). The parameter min_samples_split = 2 means that a branch will continue to split into smaller branches as long as it still has at least 2 skull data samples; splitting stops naturally once all samples in the final branch belong to a single sex class.
3. Majority voting: when the test data (252 skull samples) is fed into the system, it is processed by all 200 decision trees simultaneously. Each tree casts one vote to determine whether the skull belongs to a male (value 1) or a female (value 0). The final sex decision is taken based on the majority vote across all trees. Through this Random Forest scheme, if there is a misclassification by one or a few decision trees, the error can be automatically corrected by the majority vote of the remaining trees.

3.6 Evaluation

Before the final classification, the Random Forest model was tested through a hyperparameter tuning process to compare the performance of default values against various parameter variations. The parameter search space tested is presented in Table 3.

Table 3. Comparison of random forest parameters.

Hyperparameter	Default value	Tested variations	Selected value
n_estimators	100	50, 100, 200, 500	200
max_depth	None	None, 10, 20	None
min_samples_split	2	2, 5, 10	2
min_samples_leaf	1	1, 2, 4	1
max_features	sqrt	sqrt	sqrt
bootstrap	True	True	True

Based on Table 3, through the grid search method, the best Random Forest hyperparameter configuration in this study was found at n_estimators 200, max_depth None (unlimited), and min_samples_split 2. Based on these findings, the RFE-Random Forest model was rebuilt using this configuration to obtain the most optimal classification performance. The resulting accuracy comparison is presented in Table 4 and Figure 3.

Table 4. Comparison of model accuracy across different numbers of features.

Scenario	Number of features	Accuracy	Change analysis
Random forest	82	86.49%	–
RFE – Stage 1	80	86.49%	(same)
RFE – Stage 2	60	86.13%	(down −0.36%)
RFE – Stage 3	40	86.73%	(up +0.60%)
RFE – Stage 4	20	86.85%	(highest, up +0.34% from no RFE)

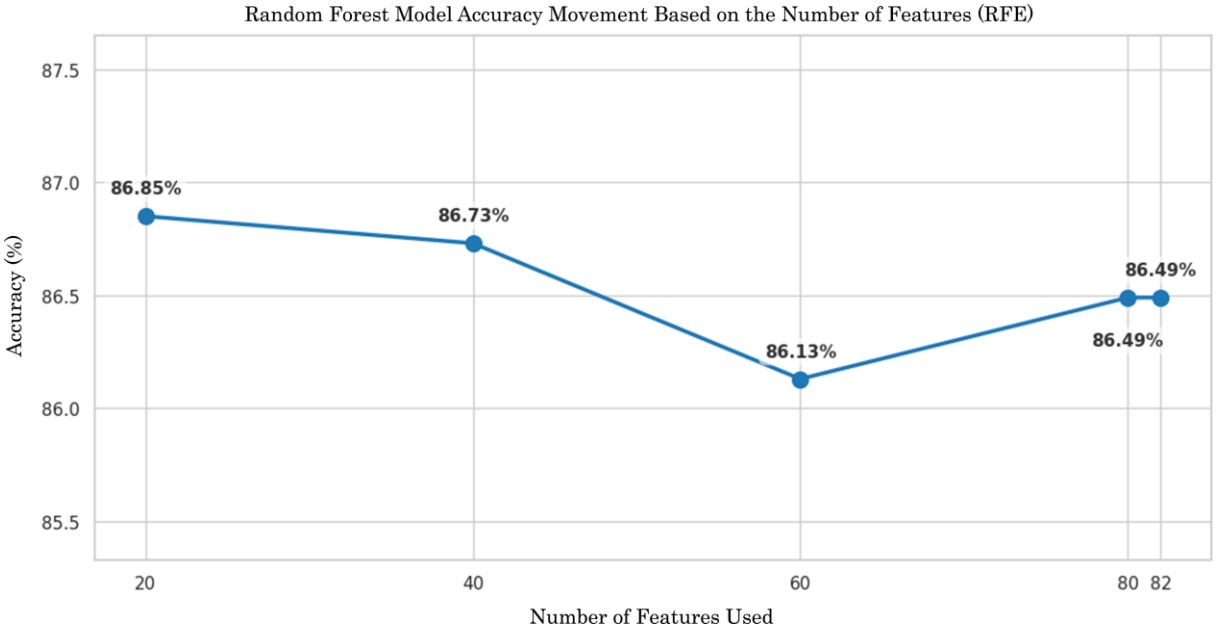

Figure 3. Increase in model accuracy based on the number of selected features.

Based on Table 4 and Figure 3, it can be clearly seen how the model reaches its optimal efficiency point. The initial reduction from 82 to 80 features did not decrease performance, indicating the removal of less relevant features without sacrificing important information. Although there was a slight drop in stability at 60 features, the model regained its best pattern and reached a peak accuracy of 86.85% when the data dimensionality was trimmed to 20 features.

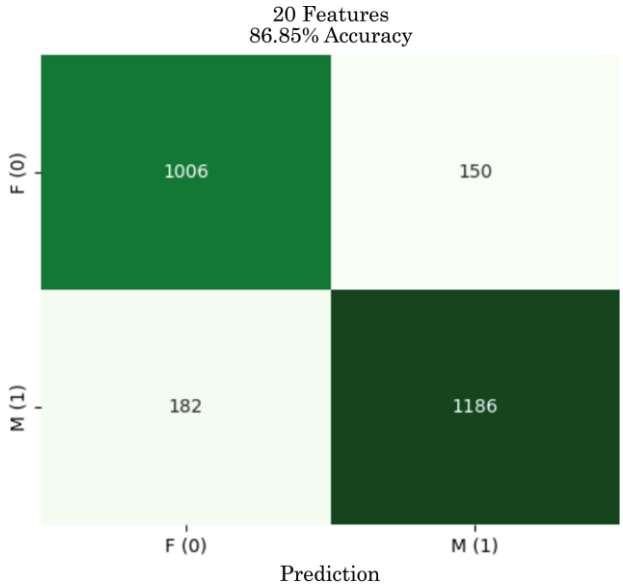

Figure 4. Confusion matrix for the best model scenario (20 features).

3.6.1. Confusion Matrix Evaluation

In addition to overall accuracy, the performance of the best model on the 20-feature subset was also evaluated in detail using a Confusion Matrix to observe the prediction distribution for each target class. Based on the test results (see Figure 4), the model showed balanced predictive ability and was not biased toward either class. This study split the dataset into training and test data through the 10-Fold Cross-Validation mechanism, in which the dataset is divided into ten equally sized folds. Each iteration uses nine folds as training data and one fold as test data, repeated ten times until every subset has served as test data once. This method helps reduce the variability of evaluation results that may arise from the selection of a particular subset as training or test data.

Out of a total of 1,368 actual male (M) samples, the model correctly classified 1,186 samples (True Positive), with a misclassification rate of 182 samples as female (False Negative). On the other hand, out of 1,156 actual female (F) samples, the model correctly predicted 1,006 samples (True Negative), with only 150 samples misclassified as male (False Positive).

Table 5. Evaluation metric results for the best model (20 features).

Metric	Performance value (%)
Accuracy	86.85
Precision	86.90
Recall / sensitivity	86.85
F1-score	86.86

Based on Table 5, it can be calculated that the model has a sensitivity (Recall) of 86.70% in recognizing male skulls and a specificity of 87.02% in recognizing female skulls. The nearly equal success rate between the two classes indicates that the proposed Random Forest model is fairly robust and is able to consistently distinguish patterns of sexual dimorphism based on the morphological features reduced by the RFE algorithm, rather than simply guessing the majority class.

4. CONCLUSION

This study successfully demonstrated the effectiveness of integrating the Random Forest (RF) algorithm with Recursive Feature Elimination (RFE) implemented within a Pipeline architecture for sex classification based on craniometric data. The use of the RFE method proved crucial in addressing the challenge of high data dimensionality in the Howell's Dataset. The baseline model using all 82 features produced an accuracy of 86.49%. After iterative feature reduction, the model's accuracy actually increased, reaching an optimal performance of 86.85% using only 20 features. This result confirms that feature selection successfully reduced data dimensionality by 75% by eliminating noisy attributes, resulting in a computationally more efficient model without sacrificing, and in fact improving, prediction quality. Furthermore, the feature importance analysis biologically validates the model's reliability, with the facial width dimension (Bizygomatic Breadth/ZYB), external facial arch width (Bijugal Breadth/JUB), and posterior ear protrusion height (Mastoid Height/MDH) identified as the main predictors with the highest degree of sexual dimorphism.

As a suggestion for future research, it is recommended to explore the use of other advanced ensemble algorithms such as XGBoost or LightGBM. In addition, more complex imputation methods such as the K-Nearest Neighbors Imputer could be added to handle missing data, along with testing the model's performance on local population datasets to verify the consistency of the main diagnostic features. The use of hyperparameter optimization techniques based on Bayesian Optimization is also recommended to explore the parameter space more efficiently in order to further improve model accuracy.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, for the facilities, data, and academic guidance provided during the completion of this research.

REFERENCES

[1] Christensen, A. M., Passalacqua, N. V., & Bartelink, E. J. (2019). *Forensic anthropology: current methods and practice*. Academic Press.

[2] Triantafyllou, G., Botis, G. G., Piagkou, M., Papanastasiou, K., Tsakotos, G., Paschopoulos, I., Matsopoulos, G. K., & Papadodima, S. (2024). Sex estimation through orbital measurements: A machine learning approach for forensic science. *Diagnostics*, **14**(24), 2773.

[3] 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.

[4] Jerković, I., Bašić, Ž., Krešić, E., Jerković, N., Dolić, K., Čavka, M., Bedalov, A., Anđelinović, Š., & Kružić, I. (2024). Developing a fully applicable machine learning (ML) based sex classification model using linear cranial dimensions. *Scientific Reports*, 14(1), 30969.

[5] Secgin, Y., Kaya, S., Harmandaoğlu, O., Öztürk, O., Senol, D., Önbaş, Ö., & Yılmaz, N. (2025). Sex estimation with parameters of the facial canal by computed tomography using machine learning algorithms and artificial neural networks. *BMC Medical Imaging*, **25**(1), 291.

[6] Del Bove, A., Menéndez, L., Manzi, G., Moggi-Cecchi, J., Lorenzo, C., & Profico, A. (2023). Mapping sexual dimorphism signal in the human cranium. *Scientific Reports*, **13**(1), 16847.

[7] Toy, S., Secgin, Y., Oner, Z., Turan, M. K., Oner, S., & Senol, D. (2022). A study on sex estimation by using machine learning algorithms with parameters obtained from computerized tomography images of the cranium. *Scientific Reports*, **12**(1), 4278.

[8] Afrianty, I., Nasien, D., & Haron, H. (2022). Performance Analysis of Support Vector Machine in Sex Classification of The Sacrum Bone in Forensic Anthropology. *Jurnal Teknik Informatika*, **15**(1), 63–72.

[9] Arthanari, A., Sureshbabu, S., Yadalam, P. K., Ravindran, V., & Raaj, S. (2025). Prediction of gender from radiographic condylar and coronoid measurements using elastic net and random forests. *Journal of Oral and Maxillofacial Pathology*, **29**(2), 309–317.

[10] Toneva, D. H., Nikolova, S. Y., Agre, G. P., Zlatareva, D. K., Hadjidekov, V. G., & Lazarov, N. E. (2020). Data mining for sex estimation based on cranial measurements. *Forensic Science International*, 315, 110441.

[11] Prabha, P. S., Ganesan, A., Lakshmi, K. C., & Murugan, A. J. (2025). Sex determination through analysis of mandibular indices using lateral cephalogram: An Artificial intelligence diagnostics. *Discover Artificial Intelligence*, **5**(1), 108.

[12] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[13] Xia, S. & Yang, Y. (2023). A model-free feature selection technique of feature screening and random forest-based recursive feature elimination. *International Journal of Intelligent Systems*, **2023**(1), 2400194.

[14] Han, Y., Huang, L., & Zhou, F. (2021). A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. *Bioinformatics*, **37**(15), 2183–2189.

[15] Sundari, H., Amrustian, M. A., & Wicaksono, A. D. P. (2024). Penerapan Recursive Feature Elimination pada Support Vector Machine untuk Klasifikasi Kanker Payudara. *LEDGER: Journal Informatic and Information Technology*, **3**(2), 60–65.

[16] Chen, C., Liang, J., Sun, W., Yang, G., & Meng, X. (2025). An automatically recursive feature elimination method based on threshold decision in random forest classification. *Geo-Spatial Information Science*, **28**(4), 1494–1519.

[17] Marpaung, S. H., Sinaga, F. M., Rambe, K. H., Simamora, F. P., & Kelvin, K. (2025). Random forest optimization using recursive feature elimination for stunting classification. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, **8**(1), 281–287.

[18] Demarchi, L., Kania, A., Ciężkowski, W., Piórkowski, H., Oświecimska-Piasko, Z., & Chormański, J. (2020). Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of poland based on airborne hyperspectral and LiDAR data fusion. *Remote Sensing*, **12**(11), 1842.

[19] Techataweewan, N., Hefner, J. T., Freas, L., Surachotmongkhon, N., Benchawattananon, R., & Tayles, N. (2021). Metric sexual dimorphism of the skull in Thailand. *Forensic Science International: Reports*, **4**, 100236.

[20] Torimitsu, S., Nakazawa, A., Flavel, A., Iwase, H., Makino, Y., Hisham, S., & Franklin, D. (2025). Estimation of population affinity using cranial measurements acquired in multidetector

computed tomography images of Japanese and Malay individuals. *International Journal of Legal Medicine*, **139**(2), 863–873.

[21] Rahayu, S. S. (2024). Klasifikasi Tulang Tengkorak Berdasarkan Jenis Kelamin Dalam Antropologi Forensik Menggunakan Metode Support Vector Machine. *Jurnal Inovtek Polbeng*, **9**(01), 243–256.

[22] Del Río, S., López, V., Benítez, J. M., & Herrera, F. (2014). On the use of mapreduce for imbalanced big data using random forest. *Information Sciences*, **285**, 112–137.

[23] Ghosh, A., Manwani, N., & Sastry, P. S. (2017). On the robustness of decision tree learning under label noise. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 685–697.

[24] Parlak, M. E., Etli, Y., Beyhan, M., Kanat, K., & Kızıloğlu, H. A. (2025). Sex estimation with ensemble learning: an analysis using anthropometric measurements of piriform aperture. *Egyptian Journal of Forensic Sciences*, **15**(1), 10.

[25] Diac, M. M., Toma, G. M., Damian, S. I., Fotache, M., Romanov, N., Tabian, D., Sechel, G., Scripcaru, A., Hancianu, M., & Iliescu, D. B. (2023). Machine Learning Models for Prediction of Sex Based on Lumbar Vertebral Morphometry. *Diagnostics*, **13**(24), 3630.