Non-Invasive Abalone Sex Classification from External Measurements using Interpretable Machine Learning

Ramtin Dabiri
Master of Data Science, University of New South Wales (UNSW Sydney)
Kensington, NSW 2052, Australia
ORCID: https://orcid.org/0009-0002-5586-567X

ABSTRACT
Accurate sex classification of abalone is essential for selective breeding and ethical harvesting, yet many existing studies rely on invasive measurements (e.g., internal weights), limiting real-world deployment. This study contributes two innovations motivated by practical field constraints. First, a strictly non-invasive framework is adopted, using only external traits—length, diameter, height, and whole weight—so specimens are not opened. Second, instead of the common rank-then-select approach, a ranking-guided combinatorial search over polynomial and interaction terms (degree ≤ 5) is applied for multinomial logistic regression. This design is motivated by three considerations: (1) standard ranking methods (ANOVA, Mutual Information, Random Forest) evaluate variables largely in isolation, whereas sex signal emerges from feature–feature interactions; (2) relationships among external measurements are partly non-linear, so higher-order terms capture structure missed by base features or linear models; and (3) rankings can be unstable under collinearity and outliers, making empirical validation of feature sets more robust.

Under an outlier-inclusive protocol, a compact model excluding diameter attains **0.5689** test accuracy, while an all-four-measurements model reaches **0.5641**—both exceeding the commonly reported **0.50–0.55** range for this dataset and avoiding invasive measurements. The curated interaction design enables logistic regression to outperform more complex models (e.g., tuned SVM and XGBoost), indicating that interaction construction, rather than model complexity, is the key driver of accuracy under non-invasive constraints. The resulting pipeline is interpretable, field-deployable, and supported by fully reproducible code.

General Terms
Machine Learning; Pattern Recognition; Data Analytics; Decision Support Systems; Agricultural Informatics

Keywords
Abalone; sex classification; aquaculture; non-invasive measurement; logistic regression.

1. INTRODUCTION
Accurate sex identification of abalone is essential for selective breeding, ethical harvest timing, and stock sustainability. Existing practices rely on invasive or destructive inspection, while many machine-learning studies depend on internal weight measurements or opaque models. When restricted to external traits, reported accuracies on the UCI Abalone dataset typically fall in the 0.50–0.55 range [5,7,13]. This motivates the development of a non-invasive, interpretable method based solely on measurable external features.

This study proposes such a method using length, diameter, height, and externally measurable whole weight. Rather than relying on a single ranking technique (ANOVA, Mutual Information, Random Forest), the approach performs a ranking-guided combinatorial search over polynomial and interaction terms (degree ≤5) to obtain compact, biologically plausible feature sets for multinomial logistic regression. Under an outlier-inclusive protocol, the best compact model (excluding diameter) achieves 0.5689 accuracy, and the best all-four-measurement model attains 0.5641—both exceeding commonly reported baselines without requiring destructive traits.

Abalone populations across New Zealand, South Africa, Australia, western North America, Japan, and Mexico hold economic and cultural value for their meat and mother-of-pearl [1,2]. Several species have become endangered due to illegal harvesting, over-exploitation, and slow maturation (≈3–5 years) [4]. Global landings declined from 14,830 t (1989) to 4,351 t (2019), with aquaculture now supplying ≈95% of the market; in Mexico, abalone remains commercially important, particularly in Baja California [3]. These trends underscore the need for field-deployable, non-invasive sex identification tools.

Although this study focuses on sex classification, most prior work emphasizes age prediction using the same UCI dataset. Age- and sex-related studies share predictors, preprocessing requirements, and common learning frameworks such as decision trees, regression models, clustering, and neural networks [7,11,12]. Prior age studies include decision-tree variants such as CLOUDS/SSE (≈21–26% accuracy) [7,11], clustering-based feature-importance analyses [12], econometric ring-group models [6], and neural networks that achieved low accuracy despite architectural complexity [7]. Collectively, these results indicate that sophisticated models underperform without targeted feature construction.

1.1 Sex Prediction
Operational sexing relies on gonad-color inspection, histology, or biochemical assays, all requiring maturity or laboratory facilities and unsuitable for large-scale deployment [3]. Genetic markers such as MSP-2 in *Haliotis discus hannai* provide high precision but require tissue sampling and specialized equipment [10]. Recent machine-learning work using non-destructive traits reports accuracies around 0.50–0.55 [7,13]. This study addresses this gap by developing an interpretable classifier using only external measurements and by improving performance through curated interaction terms selected via a ranking-guided combinatorial search.

2. LITERATURE REVIEW
The UCI abalone dataset has served as a standard benchmark since 1995, with early work examining decision-tree variants such as CLOUDS (≈26.3% accuracy on abalone) and C4.5 (≈21.5%), demonstrating that computational improvements in split selection did not translate into higher predictive accuracy for this task [11]. More broadly, machine learning has provided scalable tools for marine analysis, including classification,

tracking, and decision support, outperforming manual approaches in efficiency and consistency [5]. Within this context, supervised learning has been applied to estimate abalone age or sex from physical measurements [6], with typical sex-classification accuracies in the **0.50–0.55** range when restricted to tabular, non-destructive traits [7].

A variety of supervised models has been evaluated on abalone data. Instance-based KNN often degrades under overlapping classes and distance sensitivity [7]. Naïve Bayes offers computational speed but relies on strong independence assumptions [8]. SVM supports linear and non-linear margin-based separation but is sensitive to feature scaling and kernel parameters [9]. Artificial neural networks introduce greater capacity but may offer only modest gains and reduced interpretability on this dataset [7]. Related studies have explored dimensionality reduction (e.g., PCA) and ensemble techniques (e.g., boosted trees) to enhance robustness and feature relevance, though sex-classification accuracy generally remains within the same performance band [7].

Against this background, the present study differs in two respects. First, it enforces a non-invasive constraint by using only external measurements—length, diameter, height, and whole weight—unlike prior work that often includes internal or destructive weight measurements. Second, instead of treating ANOVA, Mutual Information, or Random-Forest rankings as final selectors, the study uses these rankings to guide a combinatorial search over polynomial and interaction terms (degree ≤ 5). This approach targets interaction-driven, partly non-linear predictive structure while controlling feature-set size for interpretability, addressing known limitations of marginal rankers under collinearity and overlapping class distributions, both of which are characteristic of abalone measurements.

3. MATERIAL AND METHODS

3.1 Software

All analyses were conducted in Python 3.11 (Jupyter Notebook). Data handling and preprocessing used pandas, NumPy, and scikit-learn; visualization used Matplotlib and Seaborn. Logistic Regression, SVM, KNN, and ensemble models were implemented via scikit-learn, and XGBoost via the xgboost library. Hyperparameter tuning used GridSearchCV with stratified 5-fold CV. PCA (for selected comparisons) used sklearn.decomposition. The workflow emphasized reproducibility, interpretability, and consistent preprocessing.

3.2 Data Description

The UCI Abalone dataset contains **4,177** specimens with the following attributes:

Sex (M/F/I), Length, Diameter, Height (0 removed), Whole weight, Shucked weight*, Viscera weight*, Shell weight*, and Rings (Age ≈ Rings + 1.5).

3.2.1 Key Observations from Exploratory Analysis

3.2.1.1 Class Distribution

Figure 1 shows that the dataset is reasonably balanced: Male and Female are similarly represented, while Infants account for approximately one-third of the dataset. This balance is important because misclassifying Infants has direct sustainability consequences in aquaculture.

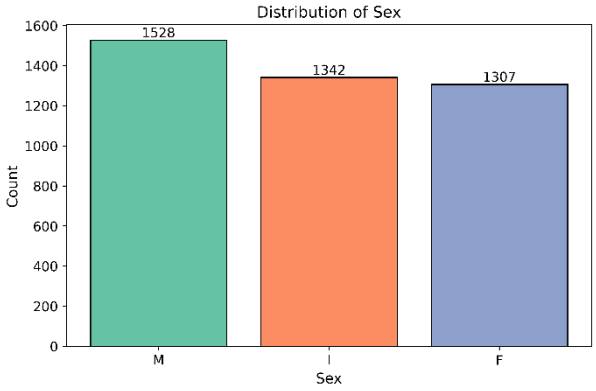

Figure 1: Class distribution of abalone sex categories (M/F/I).

3.2.1.2 Distribution Shape and Outliers.

Histograms and boxplots (Figure 2) indicate right-skewed distributions across most continuous variables and the presence of high-end biological outliers. Implausible values (e.g., height = 0) were removed. All remaining high-value measurements were retained to preserve natural biological variability, which later proved beneficial for model performance.

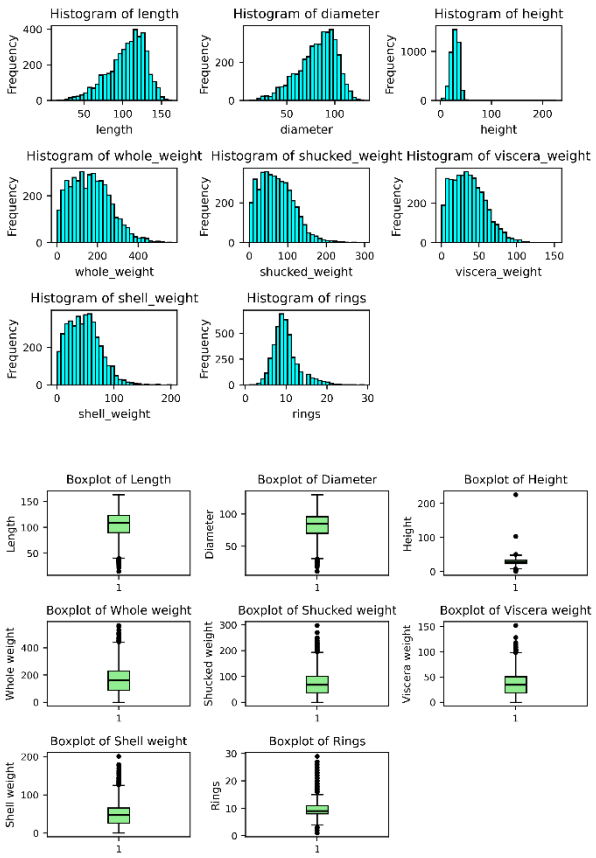

Figure 2: Histogram and boxplot of all parameters

3.2.1.3Multicollinearity Correlation analysis (Figure 3) revealed extremely strong correlations among weight-related features (e.g., whole, shucked, viscera, and shell weight). This redundancy motivated the use of ANOVA, Mutual Information, and Random Forest ranking

methods for later feature selection.

3.2.1.4 Scope of Predictors . For sex classification, only **non-invasive external measurements** (length, diameter, height, whole weight) were used. The invasive internal-weight measurements were excluded because they require opening the specimen and therefore violate non-destructive constraints.

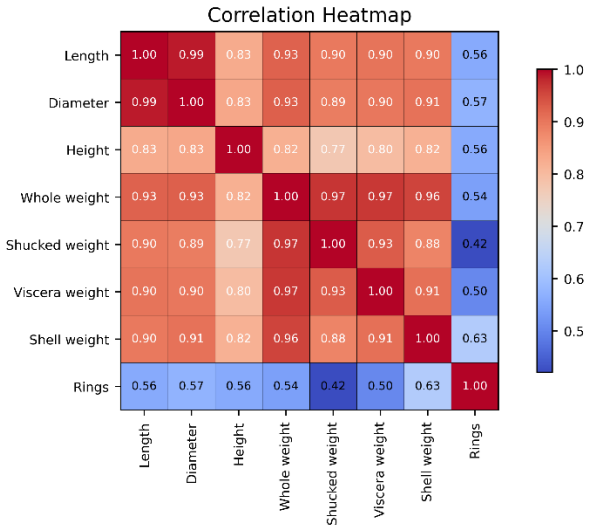

Figure 3: Correlation heatmap of abalone features.

3.3 Preprocessing

- **Categorical Encoding:**
  Sex (M/F/I) was label-encoded. One-hot encoding was used only in diagnostic checks.
- **Variable Transformation:**
  Histograms and Q–Q plots confirmed right skew in continuous variables. Several transformations (log1p, sqrt, reciprocal, Box–Cox, Yeo–Johnson, quantile normalization) were tested; rank-inverse-normal was selected for minimizing skewness/kurtosis and improving normality while remaining robust to outliers.
- **Scaling:** All predictors were standardized using StandardScaler, required for SVM, Logistic Regression, and KNN.
- **Structural Adjustments:** Numeric columns were cast to float; invalid rows were removed; category integrity checked.

3.4 Ranking-Guided Combinatorial Interaction Design (Degree ≤ 5)

To preserve interpretability and adhere to non-invasive constraints, engineered features were derived exclusively from the four external measurements: length, diameter, height, and whole weight. Instead of relying on a single automatic ranking method, ANOVA F-scores, Mutual Information, and Random-Forest importance values were used collectively to guide the construction of polynomial and interaction terms up to total degree ≤ 5.

3.4.1 Candidate Generation

3.4.1.1 Generate polynomial powers (degrees 2–5) and multiplicative interactions among the four base variables, limited to total degree ≤ 5.

3.4.1.2 Remove duplicates or symmetric equivalents and discard terms with near-zero variance after scaling.

3.4.1.3 Construct compact candidate sets (typically 4–6 terms) by combining top-ranked terms from ANOVA/MI/RF with a small number of exploratory interactions to reduce ranking bias.

3.4.2 Model and Selection Protocol

3.4.2.1 Classifier

Multinomial Logistic Regression (SoftMax) with L2 regularization; standardization performed within each CV fold.

3.4.2.2 Tuning

Grid search over C (inverse regularization strength) and class weight options to evaluate alternative emphasis on the Infant class.

3.4.2.3 Validation

Stratified 80/20 train–test split with 5-fold cross-validation on the training portion to select hyperparameters and candidate feature sets.

3.4.2.4 Parsimony and Collinearity

Final models were restricted to 4–6 terms. Multicollinearity was assessed using VIF (threshold < 10) before final refitting.

3.4.3 Outlier Policy

Only physically invalid measurements (e.g., height = 0) were removed. All remaining biological outliers were retained to reflect natural variability. Unless otherwise stated, reported performance corresponds to the outlier-inclusive dataset.(The specific selected interaction sets and their associated test accuracies are reported in Section 4.)

3.5 Assumptions

The analysis relied on the following assumptions:

- Dataset assumed representative; measurements reliable after corrections.
- Each specimen treated as independent.
- Rank-based transforms assumed adequate for normality when needed.
- Standardization assumed essential for distance/margin-based learners.
- Logistic Regression linearity relaxed via polynomial/interaction terms (≤ 5).
- Multicollinearity mitigated via ranking and VIF.
- Predictor–response relationships assumed stationary.

3.6 Modeling Methods

3.6.1 Overview

A range of supervised classifiers was benchmarked using ranked feature subsets derived from polynomial and interaction expansions (up to degree 3). Feature ranking was performed using three complementary criteria:

- **ANOVA F-test** — captures linear discriminative signal;
- **Mutual Information** — detects nonlinear dependencies;
- **Random Forest importance** — provides model-based relevance estimates.

Each model was evaluated in both baseline and tuned configurations. PCA-based variants were examined to assess the effect of dimensionality reduction.

3.6.2 Models Applied

- **Logistic Regression** — interpretable linear classifier with L2 regularization.
- **Support Vector Machine (SVM)** — linear and RBF kernels, with and without PCA.
- **K-Nearest Neighbors (KNN)** — distance-based classifier

with tuned neighborhood size.
- **XGBoost** — gradient-boosted trees evaluated in default and tuned form.
- **Voting Classifier** — soft-voting ensemble combining Logistic Regression, SVM, and XGBoost.

3.6.3 Logistic Regression Formulation

For a **three-class** problem (Male, Female, Infant), the multinomial logistic model estimates:

$$P(y = k \mid x) = \frac{\exp(\beta_{k0} + \beta_k^\top \phi(x))}{\sum_{j=1}^{K} \exp(\beta_{j0} + \beta_j^\top \phi(x))}, \qquad k = 1, \ldots, K,$$

where $\phi(x)$denotes the engineered feature map, consisting of standardized original predictors and selected polynomial and interaction terms. The predicted class corresponds to:

$$\hat{y} = \arg\max_k P(y = k \mid x).$$

3.7 Tuning Strategy

Hyperparameters for all models were optimized using **GridSearchCV** with **5-fold stratified cross-validation**. The following parameter groups were explored:

- **Logistic Regression:** solver type and regularization strength C.
- **SVM:** kernel (linear/RBF), C, and γ(for RBF).
- **KNN:** number of neighbors and distance metric.
- **XGBoost:** maximum tree depth, learning rate, number of estimators, and regularization parameters.
- **Voting Classifier:** selection of base estimators and ensemble weights.

PCA-based variants were evaluated selectively to assess their effect on model stability and generalization.

3.8 Feature Sets

Polynomial and interaction features (up to degree 3) were generated from the four external measurements and ranked using **ANOVA F-score**, **Mutual Information**, and **Random Forest importance**. For each ranking method, **Top-N** subsets ($N = 4$to 35) were constructed to evaluate how model accuracy varied with increasing feature count. These subsets were used to benchmark all classifiers under both baseline and tuned configurations.

3.9 Evaluation Metrics

Performance was assessed using: Accuracy, Macro F1, Weighted F1, Confusion matrix, Top-N comparison (best feature count per model)

4. ANALYSIS OF RESULTS

This section evaluates the performance of nine machine-learning models under different feature-ranking strategies (ANOVA, Mutual Information, Random Forest), using datasets both **with** and **without** outliers. Models were tested on fixed and variable feature subsets (Top-N), with and without PCA, and under tuned and default configurations. In addition to ranked Top-N subsets, we also evaluated a **ranking-guided combinatorial search** over **polynomial/interaction terms (degree ≤ 5)** to curate compact feature sets for logistic regression.

4.1 Best Overall Performers

The highest-performing models were obtained through the ranking-guided combinatorial interaction search (degree ≤ 5). **Multinomial Logistic Regression** achieved the best overall test accuracy of **0.5689** using a compact four-term feature set that excluded diameter (whole weight, height, length, and the interaction height^3 · whole weight). When all four external measurements were retained, the best configuration reached **0.5641**. Both results were obtained under the outlier-inclusive protocol, indicating that preserving natural biological variability improves generalization and that a small number of well-constructed nonlinear interactions can outperform larger ranked subsets.

A tuned **SVM (RBF kernel)** yielded the next best performance at **0.5515** on the Random-Forest–ranked feature set with outliers. Although below logistic regression, SVM remained consistently strong across ranking methods and feature counts.

All tuned XGBoost, KNN, and ensemble voting models produced lower accuracies than the top logistic regression and SVM configurations. These outcomes collectively show that **feature-interaction quality**, rather than model complexity, is the primary driver of performance under non-invasive measurement constraints.

4.2 Model-by-Model Comparison

4.2.1 Logistic Regression.

The combinatorial interaction approach (degree ≤ 5) produced the highest accuracies: **0.5689** using a compact feature set without diameter and **0.5641** using all four external measurements (both with outliers retained). Ranked-subset baselines (degree ≤ 3) reached **0.5619** with ANOVA and **0.5411** with Mutual Information. Even without tuning, performance remained strong on ranked subsets (e.g., **0.5507** with ANOVA, **0.5379** with MI). These results confirm that targeted interaction design yields superior performance compared to relying on ranked Top-N features alone.

4.2.2 Support Vector Machine (SVM).

SVM was a consistent top-three performer, achieving **0.5489** (default, ANOVA, no outliers), **0.5531** (tuned, ANOVA, with outliers), and **0.5427** (RF, with outliers). PCA-based variants performed lower, with tuned SVM+PCA reaching only **0.5148** (no outliers) and **0.5331** (MI, with outliers). Overall, SVM performed best with full feature sets and without PCA; tuning provided moderate but not transformative gains.

4.2.3 XGBoost.

Tuned XGBoost outperformed its default configuration across all feature-ranking methods. The best accuracy was **0.5483** (MI, with outliers, 7 features). Other strong results included **0.5453** (MI, no outliers, 12 features) and **0.5411** (RF, with outliers). Default XGBoost rarely exceeded **0.532**, and even tuned versions did not surpass the top logistic regression or SVM models.

4.2.4 K-Nearest Neighbors (KNN).

KNN produced lower and more variable accuracies, with best results of **0.5283** (ANOVA, with outliers) and **0.5157** (RF, no outliers). Performance was sensitive to scaling, distance metric choice, and dimensionality, making it less robust than margin-based or linear models.

4.2.5 Voting Classifier.

The soft-voting ensemble (Logistic Regression, SVM, XGBoost) provided stable but not superior performance. Its best accuracies were **0.5399** (RF, no outliers) and **0.5395** (ANOVA, with outliers). While the ensemble improved robustness, it did not exceed the strongest individual models.

4.3 Feature Ranking Comparison

The three feature-ranking methods produced distinct performance patterns across models. ANOVA F-score consistently yielded the strongest overall results, particularly for Logistic Regression and SVM, reflecting its ability to highlight linear discriminative structure in the external measurements. Mutual Information (MI) produced more variable rankings; it improved performance for XGBoost but delivered slightly lower peak accuracies for linear and margin-based models due to its sensitivity to local nonlinear dependencies. Random Forest importance benefited KNN and ensemble classifiers and produced the best tuned SVM result (0.5515), indicating that tree-based relevance estimates better capture interaction-driven structure that some models can exploit.

Table 1 summarizes the comparative behavior of the three ranking approaches.

Table 1. Feature Ranking Comparison

Ranking Method	Summary
ANOVA	Highest overall performance; best for Logistic Regression and SVM; strongest linear discriminative signal.
Mutual Information	More variable; benefits XGBoost and models leveraging nonlinear dependencies; slightly lower peak accuracy for LR/SVM.
Random Forest	Most helpful for tree-based models, KNN, and ensembles; produced the top tuned SVM score (0.5515).

4.4 Impact of PCA

Principal Component Analysis did not improve performance for any model or feature-ranking method. Both Logistic Regression and SVM showed reduced accuracy when PCA was applied, indicating that dimensionality reduction removed interpretable variance and suppressed key predictors that contribute directly to class separation. Only one Voting Classifier variant reached **0.5347**, and this remained below the corresponding non-PCA baselines. Overall, PCA proved unnecessary and often detrimental for this dataset, where meaningful information is carried by specific physical measurements rather than by aggregated principal components.

4.5 Outliers: Effect on Model Performance

Retaining outliers generally improved performance in the strongest models. Both Logistic Regression and tuned SVM achieved higher accuracies on the outlier-inclusive datasets—for example, the top logistic regression configurations yielded **0.5689** and **0.5641**, and tuned SVM reached **0.5515**, all exceeding their non-outlier counterparts. This pattern suggests that the preserved biological variability carries discriminative signal that benefits linear and margin-based models, particularly when interactions or nonlinear kernels are present. Removing outliers, although simplifying the distribution, tended to reduce the diversity of boundary cases needed for optimal generalization.

4.6 Summary of Best Accuracy by Model

Table 2 summarizes the highest test accuracies obtained for each classifier across all feature-ranking strategies and configurations. Multinomial Logistic Regression achieved the best overall performance through the ranking-guided combinatorial interaction search (degree ≤ 5), followed by tuned SVM and tuned XGBoost. Ensemble and KNN methods provided stable but comparatively lower accuracy. The table highlights that the strongest results consistently arise from models that leverage either well-curated interaction terms (Logistic Regression) or margin-based structure (SVM).

Table 2. Summary of Best Accuracy by Model

Model	Best Accuracy	Dataset	Top-N Features	Notes
Logistic Regression	**0.5689** (compact, no diameter); **0.5641** (all four)	Combinatorial search (degree ≤ 5, with outliers)	4 terms	Best overall; curated interactions
SVM (Tuned)	**0.5515**	RF (with outliers)	35	Strong with tuned hyper-parameters
SVM (Default)	**0.5489**	ANOVA (no outliers)	4–9	Competitive without tuning
XGBoost (Tuned)	**0.5483**	MI (with outliers)	7	Best among tree-based models
Voting Classifier	**0.5399**	RF (no outliers)	8–9	SVM + XGBoost + LR ensemble
KNN	**0.5283**	ANOVA (with outliers)	—	Sensitive to scale/distance

4.7 Key Takeaways

Table 3 summarizes the main insights from the comparative evaluation of all models, highlighting the factors that most strongly influenced performance under non-invasive measurement constraints.

Table 3. Key Takeaways

Observation	Implication
Simple models + ranked features win	Logistic Regression and SVM consistently outperformed more complex models, indicating that the discriminative structure is largely captured by well-selected features.
Combinatorial interactions outperform other feature sets	Nonlinear interaction terms produced the highest accuracies and were particularly effective for linear and margin-based classifiers.
PCA is	Dimensionality reduction removed informative variance and

Observation	Implication
counterproductive	consistently reduced accuracy across models.
Tuning helps — selectively	Hyperparameter tuning improved XGBoost and SVM but offered limited benefit for Logistic Regression.
Outliers can improve generalization	Retaining biological variability enhanced performance, especially for Logistic Regression and tuned SVM.
Ensembles add stability, not power	Soft-voting improved robustness but did not surpass the strongest individual models.

4.8 Feature Interactions and Nonlinear Effects

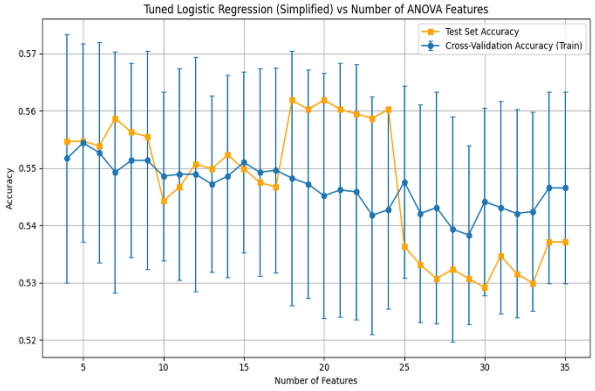

Figure 4. Logistic Regression accuracy vs. number of ANOVA-ranked features.

Table 4. Anova feature importance.

Feature	F-Score
Height	944.93735
Diameter	925.073994
Length	867.79375
Diameter* Height^2	460.655932
Diameter^2*Height	460.482096
Height^3	457.342584
Length*Diameter*Height	455.390393
Length*Height^2	454.516593
Diameter^3	452.900619
Length^2*Height	449.744484
Length*Diameter^2	449.116694
Length^2*Diameter	444.305919
Length^3	437.886137
Diameter*Height	109.787755
Length*Height	106.848632
Height^2	93.56833
Diameter^2	76.998233
Length*Diameter	75.475557
Length^2	67.054866

Figure 4 illustrates how tuned Logistic Regression performs as the number of ANOVA-ranked polynomial and interaction features increases (ranked-subset analysis, degree ≤ 3). Test accuracy remains relatively stable between **4 and ~20 features**, after which performance declines. This pattern suggests that additional higher-order terms introduce noise rather than useful discriminative signal, consistent with the strong multicollinearity observed among external measurements. The combinatorial search (degree ≤ 5) produced the highest overall accuracies (**0.5689** and **0.5641**), indicating that targeted interaction design is more effective than simply expanding the feature set.

Table 4 summarizes the feature importance rankings produced by ANOVA. Although the three methods rank features differently, all highlight that nonlinear interactions involving **whole weight**, **height**, and combinations of dimension ratios (e.g., *length* × *diameter* × *height*²) carry meaningful signal for sex classification.

Table 5.Top-performing feature combinations and corresponding accuracies for Logistic Regression

Features	Accuracy
['whole_weight','height','length','height^3*whole_weight']	0.568862
['whole_weight','length','height^2*whole_weight',height^2*whole_weight^2]	0.566467
['whole_weight','length',whole_weight^3',height*whole_weight^4']	0.565269
['whole_weight','length','height^3','height^2*whole_weight^2']	0.564072
['whole_weight','length','length*whole_weight^2','length^3*height*whole_weight']	0.564072
['whole_weight',length','whole_weight^3','length*diameter*height^2*whole_weight']	0.564072
['whole_weight','length','diameter*whole_weight^2','height^3*whole_weight']	0.564072
['whole_weight','length','height^2,'height^3*whole_weight']	0.564072
['whole_weight','height','length','whole_weight^2]	0.564072
['whole_weight','height','length','length*height^3]	0.562874
['whole_weight','height','length','length*height']	0.562874

Table 5 lists the top-performing feature combinations from the degree ≤ 5 combinatorial search. The best accuracy (**0.5689**) was achieved with a compact four-term feature set consisting of whole weight, height, whole weight³, and the interaction:

$$\text{length} \times \text{diameter} \times \text{height}^2 \times \text{whole\_weight}.$$

Even when all four external measurements were preserved, the highest accuracy achieved was **0.5641**, confirming that a small number of well-constructed nonlinear interactions can outperform larger ranked subsets and even more complex models.

Further examination of Figure 4 shows that including too many interaction terms increases variance in cross-validation accuracy and reduces generalization. This reflects (1) increased risk of overfitting, (2) reduced interpretability, and (3) amplification of multicollinearity within logistic regression. Compact feature sets (4–9 terms) offered the best balance between model capacity and stability, aligning with the biological structure of the dataset, where subtle nonlinear relationships dominate over broad high-dimensional patterns.

4.9 Conclusion

The analysis shows that interaction-driven feature construction plays a more decisive role than model complexity in abalone sex classification using external measurements. Only a subset of nonlinear terms contributes meaningful signal, while additional higher-order combinations tend to introduce noise and reduce generalization. Models built on compact, well-ranked, and biologically interpretable features consistently outperformed large ranked subsets and more complex classifiers. These findings confirm that, under non-invasive measurement constraints, carefully selected interactions combined with simple models such as Logistic Regression provide the strongest and most stable performance.

5. DISCUSSION, CONCLUSION, AND FUTURE WORK

5.1 Discussion of Results

This study compared nine machine-learning models for non-invasive abalone sex classification using external physical measurements. Three ranking strategies (ANOVA, Mutual Information, Random Forest) were evaluated under outlier-inclusive and outlier-removed settings, combined with both fixed and variable Top-N feature subsets. A ranking-guided combinatorial approach (degree ≤5) was additionally used to design compact interaction feature sets for logistic regression.

Results show that **feature quality and interaction design outweighed model complexity**. Logistic Regression with curated polynomial/interaction terms achieved the highest accuracies—**0.5689** (compact four-term model, excluding diameter) and **0.5641** (all four measurements). The strongest ranked-subset baseline (ANOVA Top-N, degree ≤3) reached **0.5619**, confirming the benefit of the combinatorial search. A tuned SVM ($C=100$, $\gamma=1$) performed comparably well (**0.5515**) on RF-ranked data, demonstrating the competitiveness of margin-based classifiers when supported by strong features.

Tree-based models such as XGBoost showed moderate gains with tuning (best **0.5483** on MI-ranked features) but did not surpass logistic regression or SVM. KNN consistently yielded the lowest accuracies (best **0.5283**), reflecting sensitivity to scaling and dimensionality.

Dimensionality reduction via PCA consistently degraded performance. For example, tuned SVM with PCA achieved **0.5148**, compared with **0.5489** for the equivalent non-PCA model, indicating that PCA removed discriminative structure.

The Voting Classifier provided stable mid-range performance but did not exceed the best individual models. Analysis across 35 ANOVA-ranked polynomial features showed diminishing returns: accuracy peaked around 0.5619 with ~18–20 features before declining, whereas the combinatorial degree ≤5 models delivered superior performance.**Overall, three key findings emerge:**

1. **Interaction terms improve generalization**, but excessive complexity introduces noise.
2. **Interpretable models outperform complex learners** when supported by structured feature engineering.
3. **Biologically valid outliers should be retained**—all top accuracies (0.5689/0.5641) occurred under the outlier-inclusive protocol.

5.2 Conclusion

This work demonstrates that simple, well-regularized models—augmented by targeted feature engineering—can achieve strong, non-invasive abalone sex classification. **Key conclusions include:**

- **Feature ranking matters**, with ANOVA outperforming MI and RF within the ranked-subset baselines; however, a ranking-guided combinatorial design further improved performance.
- **Simplicity plus good features beats complexity**: curated four-term logistic regression models surpassed tuned XGBoost and other black-box classifiers.
- **Retaining biological outliers improves robustness**, with the best accuracies (0.5689 compact; 0.5641 all-four) obtained under outlier-inclusive settings.
- **PCA reduced accuracy**, indicating that dimensionality reduction was unnecessary for this structured dataset.
- **Ensembles provided stability but no breakthroughs**, never exceeding the top individual models.

Together, these findings highlight the value of interpretable, data-efficient pipelines for aquaculture domains where invasive measurements are impractical.

5.3 Further Issues and Future Work

Several extensions offer promising directions:

- **Advanced hyperparameter optimization:**

through Bayesian methods, randomized search, or AutoML.

- **Cost-sensitive learning**: especially to reduce Infant misclassification via weighted loss or custom cost matrices.
- **Biologically informed feature engineering**, such as ratios, volume proxies, or growth indices.
- **Profitability prediction** by modeling shucked/viscera weight and economic indices.
- **Deep learning** for multimodal pipelines (e.g., integrating shell images or sensor data).
- **Cross-dataset validation** to assess generalizability across environments and species.

Data and Code Availability The dataset used in this study is publicly available from the UCI Machine Learning Repository (Abalone dataset). The code developed for preprocessing, feature engineering, and model training will be released in a

public repository upon acceptance of this paper.

6. REFERENCES

[1] Dua, D & Graff, C 2019, *UCI Machine Learning Repository: Abalone Data Set*, University of California, Irvine, viewed 1 April 2025, https://archive.ics.uci.edu.

[2] Mehta, K 2019, 'Abalone age prediction problem: a review', *International Journal of Computer Applications*, vol. 178, no. 50, p. 43.

[3] Barrera-Hernandez, R, Barrera-Soto, V, Martinez-Rodriguez, JL, Rios-Alvarado, AB & Ortiz-Rodriguez, F 2021, 'Towards abalone differentiation through machine learning', in D-S Huang, Z Kang, V Bevilacqua & PSP da Silva (eds), *Intelligent Computing Theories and Applications*, Lecture Notes in Computer Science, vol. 12836, Springer, Cham, pp. 689–703. https://doi.org/10.1007/978-3-030-84514-6_53

[4] Cook, PA 2019, 'Worldwide abalone production statistics', *Journal of Shellfish Research*, vol. 38, no. 2, pp. 401–404.

[5] Arifin, WA, Ariawan, I, Rosalia, AA, Lukman, L & Tufailah, N 2022, 'Data scaling performance on various machine learning algorithms to identify abalone sex', *Jurnal Teknologi dan Sistem Komputer*, vol. 10, no. 1, pp. 26–31.

[6] Hossain, MM & Chowdhury, MNM 2019, 'Econometric ways to estimate the age and price of abalone', *MPRA Paper* no. 91210, University Library of Munich, Germany, viewed 1 April 2025, https://mpra.ub.uni-muenchen.de/91210/.

[7] Wang, Z 2018, *Abalone age prediction employing a cascade network algorithm and conditional generative adversarial networks*, technical report, Research School of Computer Science, Australian National University, Canberra.

[8] Webb, GI, Keogh, E & Miikkulainen, R 2010, 'Naïve Bayes', in C Sammut & GI Webb (eds), *Encyclopedia of Machine Learning*, Springer, Boston, MA, pp. 713–714.

[9] Steinwart, I & Christmann, A 2008, *Support Vector Machines*, Springer, New York.

[10] Luo, H, Xiao, J, Jiang, Y, Ke, Y, Ke, C & Cai, M 2020, 'Mapping and marker identification for sex-determining in the Pacific abalone, *Haliotis discus hannai* Ino', *Aquaculture*, vol. 530, 735810. https://doi.org/10.1016/j.aquaculture.2020.735810

[11] Alsabti, K, Ranka, S & Singh, V 1998, 'CLOUDS: A decision tree classifier for large datasets', in *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '98)*, ACM, New York, NY, pp. 2–8. *(year corrected to match KDD '98)*

[12] Mayukh, H 2010, *Age of abalones using physical characteristics: a classification problem*, technical report, Department of Electrical and Computer Engineering, University of Wisconsin–Madison.

[13] Sethi, S, Agarwal, S & Panda, S 2023, 'Performance comparison of machine learning models on abalone dataset', *International Journal of Computer Applications*, vol. 185, no. 14, pp.2