Application of Ensemble Learning in Abalone Rings Classification

Sishi Zhang
z5502647

***Abstract*—This research investigates the use of ensemble learning techniques to predict the age of abalones based on their physical characteristics. The task is formulated as a multiclass classification problem, with ages categorized into four distinct groups. Traditionally, the determination of abalone age involves tedious processes, such as counting shell rings under a microscope. To solve this problem, this study tris to streamline through the application of machine learning models. The open-access Abalone dataset is utilized for a comprehensive analysis of feature distributions, and the performance of models including Decision Trees, Random Forests, XGBoost, Gradient Boosting, and neural networks optimized using Adam and SGD is assessed. Advanced hyperparameter tuning, such as tree depth adjustments, pruning methods, dropout rates, and L2 regularization, is employed to enhance model performance. It is demonstrated that ensemble learning approaches significantly improve classification accuracy, with neural networks yielding promising results when fine-tuned. Detailed methodologies are presented, providing a scalable framework for automating age estimation in abalones and potentially other biological datasets.**

***Index Terms*—ensemble learning, classification task**

I. Introduction

Recently, neural networks and deep learning models have been highly favored by researchers for their success in solving complex problems in various fields such as image recognition, biometrics, and autonomous driving systems. [1] The field has seen key breakthroughs, such as the introduction of the backpropagation algorithm, which made it possible to train multilayer neural networks efficiently, and the development of convolutional neural networks, which transformed the way computers understand visual data. [2] At the same time, ensemble learning methods like Decision Trees, Random Forests, and Gradient Boosting have proven to be powerful tools for working with structured data, offering a good mix of accuracy and interpretability. As these core techniques have advanced over time, they have become essential for tackling complex problems in areas like multimedia analysis, healthcare, and biological data prediction. [3]

Although neural networks are highly powerful, they also face several significant challenges in practical applications. One common issue is overfitting, [4] where the model achieves high accuracy on training data but, due to its complexity and tendency to memorize the details and noise in the data, performs poorly on new data, making it difficult to generalize effectively. Additionally, neural networks often require substantial computational resources, especially when working with large datasets or time series data, which can place high demands on hardware performance. Another frequent challenge is hyperparameter optimization, [6] which involves fine-tuning parameters such as learning rates, regularization terms (e.g., dropout rates), and the number of hidden layers. This process typically requires extensive experimentation, debugging, and domain expertise. Addressing these issues is critical to the efficient application of neural networks.

In this research, multiple machine learning techniques, including integrated models (random forests, XGBoost, and gradient boosting), decision trees, and neural networks, were applied and compared to address the multi-classification problem of abalone age classification. To improve the performance of the models, hyperparameter adjustment was used to optimize their settings, and LabelEncoder was employed to process the target variables, converting classification labels into numerical forms recognizable by the models. Additionally, dropout regularization technology was introduced into the neural network to mitigate the effects of overfitting. [5] The combination of these methods not only led to improvements in the accuracy of the models but also enhanced their stability and robustness in processing complex biological data.

The motivation for this project lies in the business value it can bring, especially in industries that require efficient processing and analysis of biological data. For abalone classification, age is an important measure of price and fishing plan. However, the traditional method of manual counting shell rings is time-consuming and laborious, and easy to make mistakes, which affects the work efficiency and accuracy. By using machine learning models, we hope to make this process faster and more stable, not only reducing operating costs, but also helping fisheries better manage their resources and improve overall efficiency.

II. Exploratory Data Analysis

In this section, Data preprocessing was performed, including the transformation of the discrete feature Sex into numerical values, where 'M' was replaced by 1, 'F' by -1, and 'I' by 0 to facilitate model processing. The target variable Rings was classified into four groups based on predefined criteria for further analysis. A pie chart analysis was conducted to visualize the proportion of each class. Additionally, the eight features and the categorical variable were analyzed using frequency histograms, a heatmap for correlation analysis, and a chi-square test visualization to explore relationships and dependencies among the variables.

A. Abalone data

The dataset, which has already removed the missing value from the original data, can be downloaded from https://archive.ics.uci.edu/dataset/1/abalone. This dataset contains 4177 instances and 8 features, and our task is to predict the classification of rings according to the given 8 features. The classification of rings is shown in TABLEI.

TABLE I: Classification of Abalone's Rings

Classification Type	**Description**
Class 1	0 - 7 years
Class 2	8 - 10 years
Class 3	11 - 15 years
Class 4	Greater than 15 years

B. Rings distribution after classification

The abalone's rings distribution after classification is illustrated in the Fig1.

It is obvious to see that Class 2 accounts for the majority, about half, and Class 4 accounts for a very small part, only 6.2 percent, which may lead to imbalance problems.

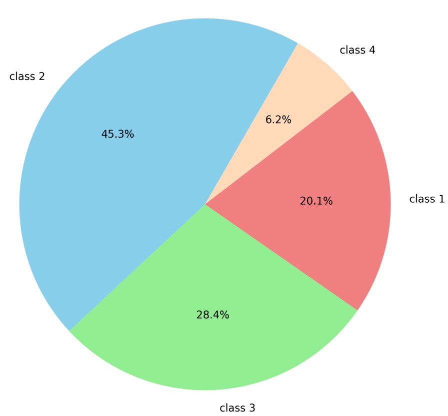


Fig. 1: The proportion of 4 classes.

C. Correlations between Features and classified Rings

The correlations between different features and classified Rings can be interpreted by heatmap, from which the positive correlation or negative correlation can be obtained. In most cases, heat maps are suitable for analyzing correlations between continuous variables.

The Fig2 shows the correlation between rings and other features, such as whole weight and length, which is almost close to 0. It indicates the linear relationship between rings and features is weak. In this context, Chi-square test is more suitable to analyze it. [7]

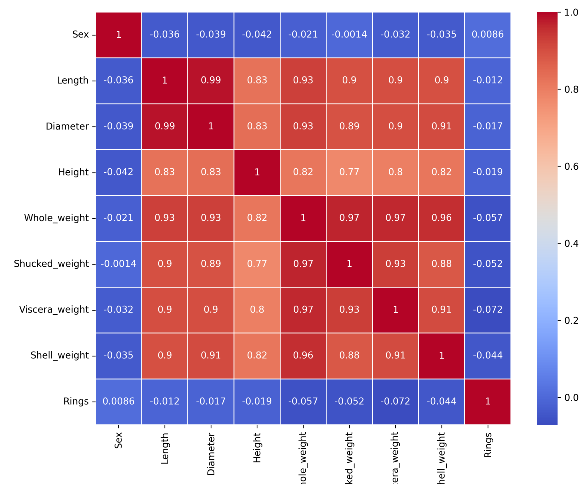


Fig. 2: The heatmap of different features and classified rings.

D. Correlations between Rings and Other Features

Correlations between target(classified variable) and other features(continuous variable) are commonly shown by chi-square bar.

From Fig3, it can be seen that the weight feature is the most critical feature to predict the ring number, the correlation of size feature is weak, and the influence of sex is the least.

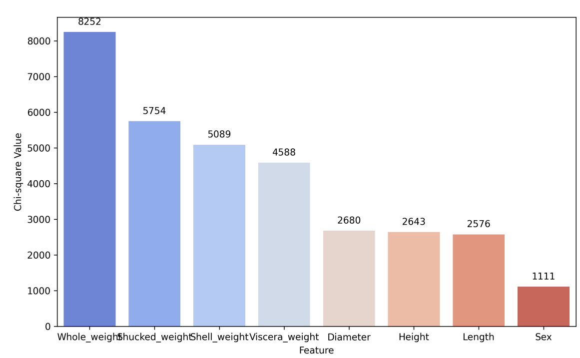


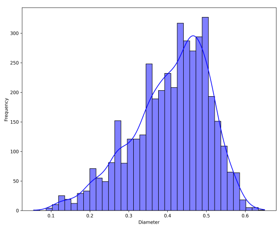
Fig. 3: The chi-square bar between classified rings and other features.

E. Distribution of Features

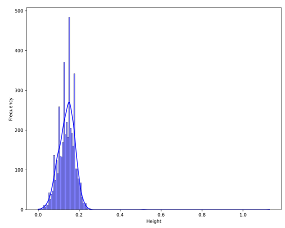
Distribution of abalone's features can be shown by the histograms.

From Fig4, it is vivid to see the frequency of each feature. The diameter and length characteristics are significantly skewed to the left and can be adjusted by taking the square root, while the variables related to weight are skewed to the right and can be adjusted by taking log.

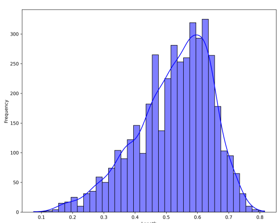
conclusion: The visualization results above indicates that the linear relationship between rings and these variables is weak. In this context, the Chi-square test is more appropriate for analyzing the relationship between the categorical variable rings and other features. Furthermore, non-linear models (e.g., decision trees, random forests) might be necessary to capture potentially complex relationships.



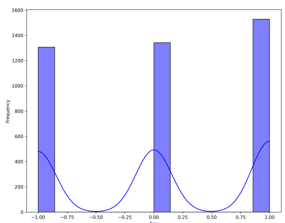
(a) Diameter of abalone.



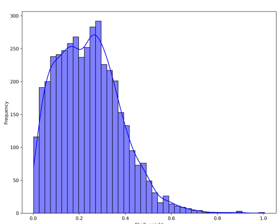
(b) Height of abalone.



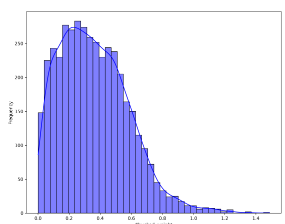
(c) Length of abalone.



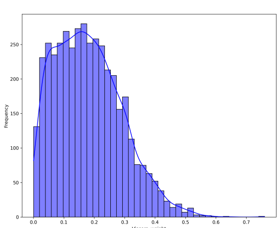
(d) Sex of abalone.



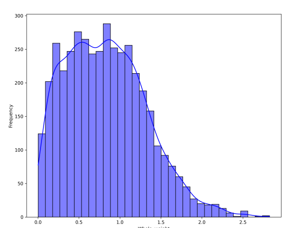
(e) Shell weight of abalone.



(f) Shucked weight of abalone.



(g) Viscera weight of abalone.



(h) Whole weight of abalone.

Fig. 4: Distribution of abalone's features

III. Methodology

In this section, Decision Trees, Random Forests, XGBoost, Gradient Boosting, and neural networks were applied to explore model accuracy.

A. Decision Tree

The optimal decision tree for the abalone multi-class dataset was obtained through 7 experiments, each using different random seeds and maximum depth parameters, and the fifth run was selected as it owned the highest accuracy according to TABLEII.

TABLE II: Accuracy of Different Runs by Random Forest.

Run Number	Train Accuracy	Test Accuracy
Run 1	0.607	0.601
Run 2	0629	0.569
Run 3	0.645	0.620
Run 4	0.664	0.600
Run 5	0.665	0.628
Run 6	0.710	0.596
Run 7	0.725	0.578

With the fifth run, which has the highest accuracy, the corresponding decision tree model is shown in Fig5.

Based on the best decision tree, it is evident that Class 1 and Class 2 dominate as the main categories, while Class 3 and Class 4 are relatively less represented. When the shell weight in the left branch of the tree is less than or equal to 0.032, Class 1 is relatively well-separated with high purity; however, some overfitting occurs at this stage. However, the splits here are less pure compared to the left branch, with Gini indices like 0.616 and 0.572 indicating significant impurity. [**?**]

When dealing with continuous variables gini is more efficiently than entropy, so the default value gini is used here.

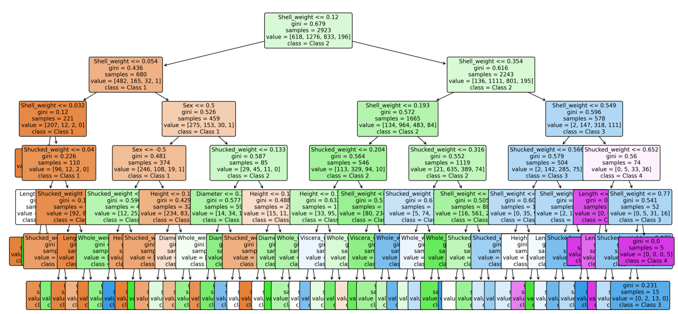


Fig. 5: The best decision tree.

Some of the left branches in the decision tree are ineffective. To enhance the performance of the decision tree, pruning is applied to simplify the model. As a result, the post-pruning technique is employed in this project. [8] The accuracy achieved after applying the post-pruning technique is shown in Fig. 6.

From this figure, it is evident that pruning initially leads to a significant improvement in accuracy. When alpha reaches 0.002, the accuracy peaks at 0.62. However, as alpha continues to increase, the accuracy begins to decline, indicating that excessive pruning adversely affects model performance.

B. Random Forest

Random Forest, as an ensemble learning method based on multiple decision trees, is also applied in this project. The accuracy score, especially test accuracy score, can describe the

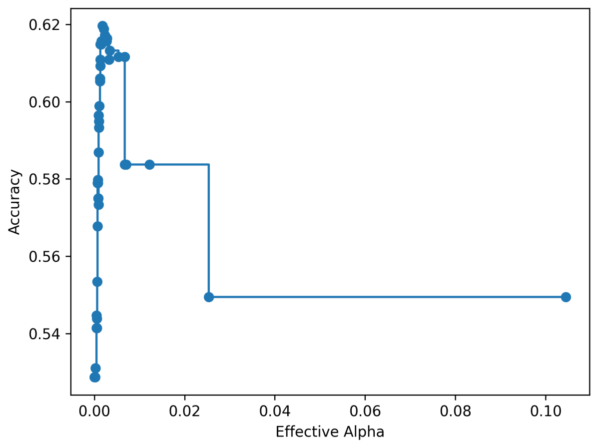


Fig. 6: Accuracy after post-pruning the decision tree.

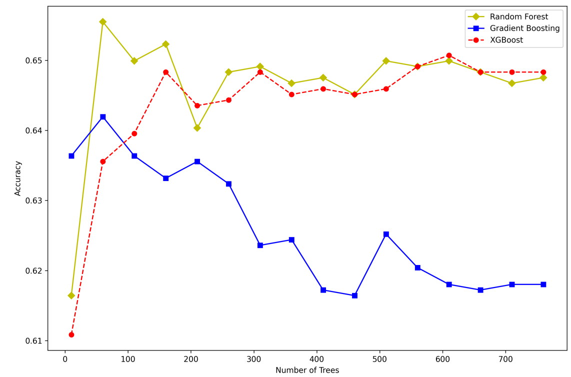


Fig. 8: Comparison of random forest, gradient boosting and XGBoosting.

performance of random forest, and accuracy score of random forest according to the number of trees is shown in Fig7.

Compared to decision trees, random forests incorporate data and feature randomization, training multiple independent trees and determining the final result through majority voting. This approach reduces the risk of overfitting and enhances model stability.

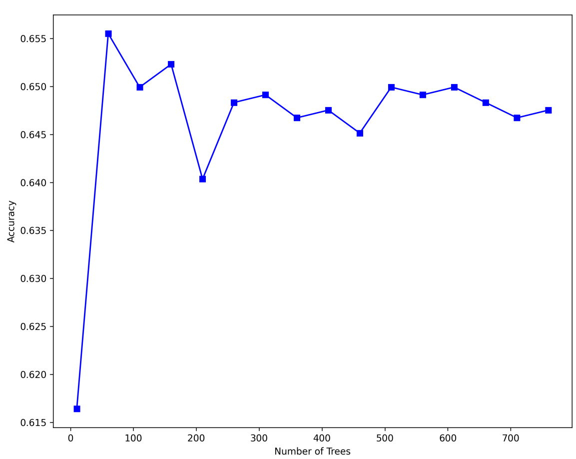


Fig. 7: The accuracy score of random forest.

To obtain the performance of different algorithms in this abalone rings classification task, the comparison result among random forest, gradient boosting and XGBoost is illustrated in Fig8.

It is easy to see from the graph that random forest and XGBoost perform better than gradient lift, which initially overfits after increasing the number of trees, while random forest and XGBoost tend to stabilize after 260 trees. [9]

Considering neural network, the comparison among random forest, gradient boosting, XGBoost, is demonstrated in Fig9. According to the figure, it can be seen that MLP uses different optimizers to define the upper and lower limits of performance, among which SGD, as a basic optimization algorithm, represents the lower limit of neural network, while Adam, as an improved algorithm, represents the upper limit of MLP. It is easy to see that integrated methods like random forest and gradient rise outperform Adam, which means that ensemble methods can also excel in solving data processing problems.

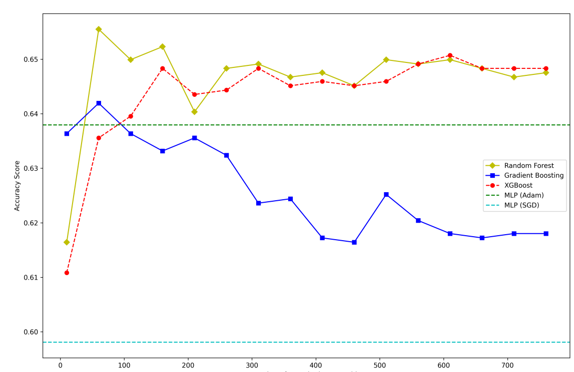


Fig. 9: Comparison of random forest, gradient boosting and XGBoosting.

C. L2 regularisation

These results in TABLEIII indicate that increasing the dropout rate while reducing the weight decay leads to improved model performance in this specific setup. This suggests that a higher dropout rate might better prevent overfitting, while a lower weight decay allows the model to learn more flexible weight parameters. The best accuracy is observed with the combination of Dropout Rate = 0.3 and Weight Decay = 0.0001.

In the regularization process, not only the hyperparameters such as Dropout Rate, Weight Decay affect the result, but other parameters such as the number of neuron in hidden layer, number of iterations and the size of batch also affect the final result. When the number of iterations is not enough, the accuracy will be lower than enough iterations.

TABLE III: Accuracy of Different hyper-parameters.

Hyper-parameter Combination		**Accuracy**
Dropout Rate	**Weight Decay**	
0.1	0.01	0.588
0.2	0.001	0.583
0.3	0.0001	0.594

IV. Result

Summary

In this study, Decision Trees, Random Forests, XGBoost, Gradient Boosting, and neural networks were applied to the abalone rings classification dataset to evaluate model accuracy. Decision Trees showed signs of overfitting, whereas Random Forests improved prediction accuracy and reduced overfitting by aggregating multiple decision trees, demonstrating greater stability. When comparing the results of XGBoost and Gradient Boosting with Random Forests, it was observed that XGBoost performed comparably to Random Forests, while Gradient Boosting exhibited overfitting prematurely.Further comparisons with neural networks optimized using Adam and SGD showed that both XGBoost and Random Forests achieved better classification accuracy than the Adam-optimized neural networks, proving the effectiveness of ensemble methods for the abalone rings classification task.

Importance of parameters

Model performance in this study was heavily influenced by the choice of parameters and settings. For Random Forests, a random seed of 42 consistently produced better and more reliable results compared to seeds like 17 or 47. Accuracy improved rapidly when the number of trees was relatively small (e.g., fewer than 100), but beyond 260 trees, the improvements plateaued, with little to no gains from adding more trees. Overall, Random Forests demonstrated a clear advantage over Decision Trees in this project. In neural networks, settings like dropout rate and weight decay played an important role in controlling overfitting and improving generalization. Other factors, such as the number of neurons in the hidden layer, the number of training iterations, and batch size, also had a significant impact on the final accuracy. For instance, when the hidden layer had 10 neurons, the training iterations were set to 100, and the batch size was 100, the accuracy was lower compared to a configuration with 40 neurons in the hidden layer, 1000 training iterations, and a batch size of 4. This example highlights the importance of tuning these parameters, as insufficient neurons, limited iterations, or inappropriate batch sizes can lead to underperformance.

V. Conclusion

Major contributions

We systematically compared the performance of Decision Trees, Random Forests, XGBoost, Gradient Boosting, and neural networks on the abalone age classification problem, providing a comprehensive evaluation of their strengths and weaknesses. The study showed that Random Forests and XGBoost worked very well, achieving high accuracy and reliable results, even outperforming neural networks in this classification task. Issues like overfitting, which were common in models like Decision Trees and Gradient Boosting, were better handled by Random Forests and XGBoost. We also used practical techniques like LabelEncoder() for feature processing and dropout regularization in neural networks to improve performance and reduce overfitting. These results provide useful guidance for applying machine learning to biological data classification and creating more efficient and scalable solutions.

Directions for future research

The study's approach could be extended to other biological datasets, such as fish or shellfish, by optimizing model parameters through hyperparameter tuning to enhance the performance of Random Forests and XGBoost, while also developing simple, user-friendly tools to make these models accessible to industries like seafood and aquaculture.

References

[1] Fujiyoshi, Hironobu, Tsubasa Hirakawa, and Takayoshi Yamashita. "Deep learning-based image recognition for autonomous driving." IATSS research 43.4 (2019): 244-252.

[2] Leonard, James, and Mark A. Kramer. "Improvement of the backpropagation algorithm for training neural networks." Computers & Chemical Engineering 14.3 (1990): 337-341.

[3] Samborska, Izabela A., et al. "Artificial neural networks and their application in biological and agricultural research." J. NanoPhotoBioSciences 2 (2014): 14-30.

[4] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.

[5] Smith, Leslie N. "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay." arXiv preprint arXiv:1803.09820 (2018).

[6] Feurer, Matthias, and Frank Hutter. "Hyperparameter optimization." Automated machine learning: Methods, systems, challenges (2019): 3-33.

[7] Sharpe, Donald. "Your chi-square test is statistically significant: now what?." Practical assessment, research & evaluation 20.8 (2015): n8.

[8] Esposito, Floriana, et al. "A comparative analysis of methods for pruning decision trees." IEEE transactions on pattern analysis and machine intelligence 19.5 (1997): 476-491.

[9] Sahin, Emrehan Kutlug. "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest." SN Applied Sciences 2.7 (2020): 1308.