Automated Age Prediction of Abalones Using Machine Learning Models

Authors: Xincheng Qiao* (z5327961)

**Xincheng Qiao uploaded the code on the ed platform.*

1. Abstract

Traditionally, estimating the age of abalone entails the labor - intensive and time - consuming process of manually counting shell rings. This study endeavors to automate the age classification process through the utilization of machine learning models based on physical measurements, with the aim of categorizing abalones into four age groups.

Several models were evaluated, namely Decision Trees, Random Forest, Gradient Boosting, XGBoost, and Neural Networks with Adam and SGD optimizers. Our analysis revealed that Random Forest achieved stable accuracy, attaining 64.0% with 300 trees. Pruning enhanced the interpretability of the Decision Tree model, with optimal performance at a maximum depth of 5 and a test accuracy of 60.3%. Neural Networks exhibited the best performance with the Adam optimizer, achieving 65.4% accuracy with a 200 unit hidden layer.

Our results demonstrate the efficacy of Random Forest for consistent performance and the value of pruning techniques in balancing accuracy and model complexity. This automated classification method provides a reliable alternative to manual age estimation, thereby augmenting research efficiency in marine biology.

2. Introduction

2.1 Background and Significance

Machine learning has emerged as an indispensable tool for automating complex predictive tasks across diverse fields, with the biological sciences being no exception [1]. Techniques such as neural networks, decision trees, and ensemble methods have achieved significant success in a wide range of applications, spanning from image recognition to species classification [2]. Among these, ensemble methods, particularly Random Forest and Gradient Boosting, are highly regarded due to their outstanding accuracy and robustness [3].

This study focuses on the application of these methods within the biological domain, specifically in predicting the age of abalones based on physical measurements. Conventionally, estimating abalone age requires the manual counting of growth rings, a labor - intensive process that is also subject to variability induced by human error. Automating this process holds the potential to substantially streamline research activities in marine biology, thereby enhancing the overall efficiency and reliability of related studies.

2.2 Challenges in Machine Learning for Biological Data

Biological datasets often present unique challenges for machine learning, such as class imbalance, high feature correlations, and the need for interpretable models [4]. In this study, the class distribution for abalone age categories was found to be imbalanced, with most samples concentrated in Class 2 and few in Class 4 (**Figure 1**: Distribution of Age Classes).

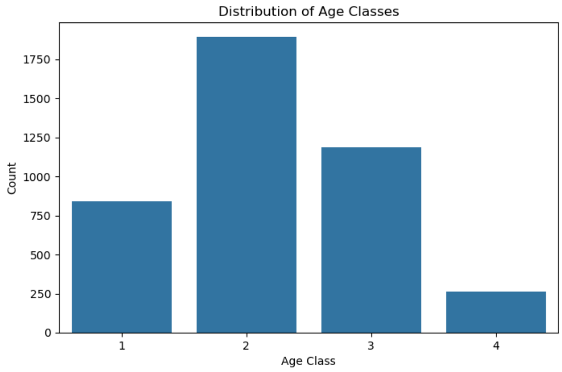

Figure 1: Distribution of Age Classes

Additionally, high correlations were observed between features like Length, Diameter, and Whole weight (**Figure 3**: Feature Correlation Heatmap). These correlations can introduce redundancy and impact model interpretability. Effective model selection and tuning are essential to mitigate these challenges and improve predictive performance [5].

2.3 Motivation and Study Objectives

This study is motivated by the need for an accurate and interpretable model for abalone age classification. While previous studies have explored machine learning in biological applications, there has been limited focus on developing an automated system specifically for abalone age prediction that balances accuracy with interpretability. The key objectives of this study are to:

- Compare the performance of Decision Trees, ensemble methods, and neural networks on this classification task.
- Investigate the effects of hyperparameter tuning and pruning on model accuracy and complexity.
- Provide interpretable insights through feature importance analysis and rule extraction.

2.4 Contribution and Paper Structure

In this study, we present a systematic evaluation of machine learning models for abalone age classification. Key contributions include identifying Random Forest as the most consistent performer, achieving 64.0% accuracy with 300 trees, and demonstrating that pruning can improve Decision Tree interpretability, with optimal accuracy at a depth of 5 (60.3% test accuracy). Neural networks with the Adam optimizer reached 65.4% accuracy, outperforming SGD.

The rest of the paper is organized as follows: Section 3 details the methodology, including data preprocessing and model training. Section 4 presents the results, with a comparative analysis of model performance. Finally, Section 5 concludes the paper with a summary of findings and suggestions for future work.

3. *Methodology*

3.1 Data Description and Preprocessing

The dataset used in this study is the Abalone dataset from the UCI Machine Learning Repository, containing 4,177 samples, each with eight physical measurements and one target attribute, Rings, which represents the age of abalones [6]. The attributes include Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight. For preprocessing, the Sex attribute was encoded as numeric values (0 for male, 1 for female, and 2 for infant) to make it compatible with machine learning algorithms. Additionally, since Rings is a continuous variable, it was transformed into categorical age classes for this study, defining four age classes: Class 1 (0-7 years), Class 2 (8-10 years), Class 3 (11-15 years), and Class 4 (>15 years). The resulting class distribution is significantly imbalanced, with Class 2 being the most prevalent and Class 4 the least frequent (see Figure 1: Distribution of Age Classes).

3.2 Exploratory Data Analysis (EDA)

EDA was performed to understand the distributions of each feature and examine potential correlations among them [7]. This step included generating histograms for feature distributions and a correlation heatmap to assess relationships among the physical measurements. The insights gained from EDA informed decisions on model selection and preprocessing techniques, such as handling correlated features in ensemble models.

3.3 Model Selection

Several models were selected to address the age classification task, each chosen for specific advantages. The Decision Tree model was selected for its interpretability, allowing rule extraction and clear decision-making pathways [8]. The Random Forest model was chosen to reduce variance and improve stability by averaging the predictions of multiple decision trees. Gradient Boosting and XGBoost, both ensemble methods, were included to enhance accuracy by sequentially improving on the errors of prior models. Lastly, Neural Networks were incorporated to explore complex feature interactions, with both Adam and SGD optimizers tested to evaluate training efficiency [9].

3.4 Hyperparameter Tuning and Pruning

Hyperparameter tuning and pruning were applied to each model to enhance performance. For the Decision Tree, pre-pruning involved restricting the maximum depth of the tree to prevent overfitting and reduce model complexity, while post-pruning used cost complexity pruning by adjusting ccp_alpha values, allowing nodes that contributed little to accuracy to be pruned, thereby balancing accuracy with interpretability. In the case of Random Forest, the number of trees was varied to identify a configuration that balanced accuracy and computational efficiency, recognizing that while Random Forest benefits from larger tree ensembles, we aimed to minimize redundancy. For Neural Networks, hidden layer sizes were adjusted to test the impact of different configurations on accuracy, and both Adam and SGD optimizers were compared to evaluate their convergence speeds and overall performance.

3.5 Experimental Setup

Experiments were conducted using Python with libraries such as scikit-learn, XGBoost, and Keras. To ensure reproducibility, data was split into training (80%) and test (20%) sets, and results were averaged over multiple runs. Our code can be found in the codebase corresponding to the ED platform, including scripts for preprocessing, model training, and evaluation.

4. Results

In this section, we present the results of each model, including Decision Tree, Random Forest, Gradient Boosting, XGBoost, and Neural Networks. Each model's performance was evaluated in terms of accuracy, with hyperparameter tuning and pruning applied to optimize results. Quantitative comparisons and visualizations illustrate the effectiveness of different approaches.

4.1 Feature Distribution

Each feature in the dataset represents a physical measurement of abalones, such as Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight. Figure 2: Distribution of Features shows histograms for each feature, revealing variations in their distributions:

- Height and Viscera weight are skewed, with most values concentrated near zero, indicating that the dataset contains a majority of smaller-sized abalones.
- Length, Diameter, and Whole weight show relatively normal distributions, suggesting these features have greater variability across samples.

These distributions provide insight into the range and scale of each feature, influencing decisions around feature scaling and potential transformations.

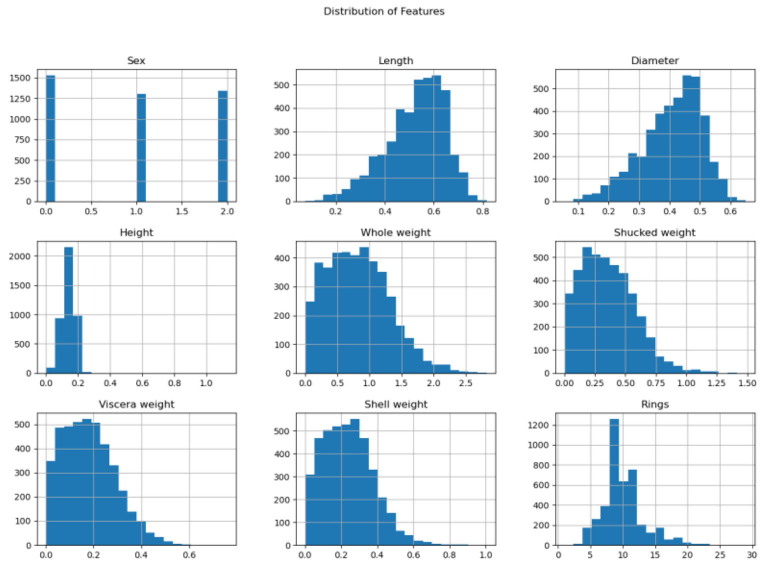

Figure 2: Distribution of Features

4.2 Feature Correlation Analysis

A correlation heatmap was generated to identify relationships between features (see **Figure 3: Feature Correlation Heatmap**). The heatmap highlights strong positive correlations among Length, Diameter, and Whole weight, with correlation coefficients close to 0.9. This high degree of correlation suggests that these features may carry redundant information, which could lead to multicollinearity in certain models, affecting interpretability and potentially increasing model complexity.

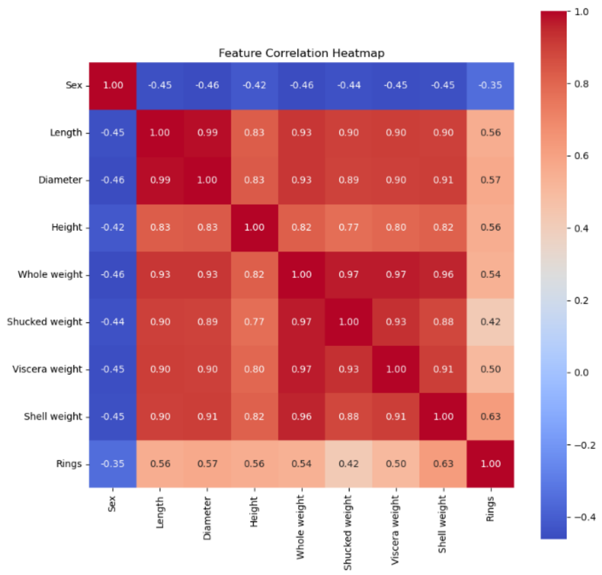

Figure 3: Feature Correlation Heatmap

4.3 Key Takeaways from Data Analysis

- **Class Imbalance**: The imbalance in age classes indicates the need for models that can handle or mitigate bias towards the majority classes.
- **Feature Variability**: The diverse distributions across features highlight the importance of selecting models that can capture both skewed and normally distributed data.
- **Correlated Features**: High correlations among features like Length, Diameter, and Whole weight suggest that ensemble methods, which are less sensitive to multicollinearity, may be effective choices.

These findings from EDA played a critical role in shaping our approach to model selection, tuning, and evaluation.

4.4 Decision Tree Performance

The Decision Tree model has furnished an interpretable baseline for classifying the age of abalone. We optimized this model through pre - pruning (by limiting the max_depth) and post - pruning (by adjusting the ccp_alpha). The optimal performance was obtained when the max_depth was set at 5, as this balanced the accuracy and simplicity. Consequently, the resulting Decision Tree model attained a test accuracy of 60.3%.

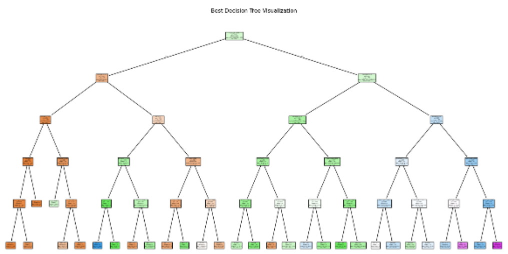

Figure 4: Best Decision Tree Visualization

To demonstrate the decision - making process of the optimized Decision Tree, Figure 4 (Best Decision Tree Visualization) presents the model's structure, which encompasses the decision points for different feature values. For example, nodes are split based on diameter, shell weight, and other physical measurements, thus offering clear decision paths. This visualization is helpful for interpreting the model. It enables us to derive rules like "If diameter ≤ 0.5 and shell weight > 0.2, then class = class 1." These rules are of great value to researchers who require transparent and interpretable models in biological studies.

To handle the complexity of the Decision Tree model and avoid overfitting, both pre - pruning and post - pruning methods were adopted [10]. Pre - pruning was carried out by restricting the maximum depth (max_depth) of the tree, as shown in Figure 5 (Pre - Pruning Accuracy vs Max Depth for Decision Trees). The accuracy on the test set rose as the depth increased, reaching its peak when the max_depth was 5. Beyond this value, further increases in depth caused a decline in test accuracy, suggesting overfitting. Setting max_depth to 5 achieved an optimal balance between accuracy and model simplicity, leading to a manageable tree structure that is conducive to interpretability.

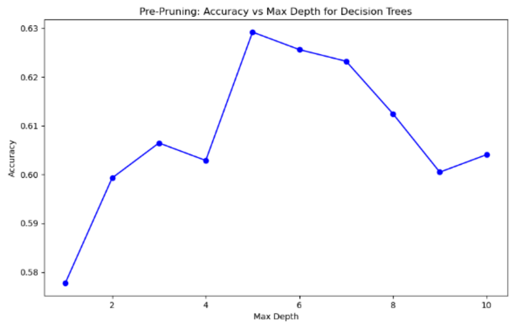

Figure 5: Pre-Pruning Accuracy vs Max Depth for Decision Trees

Additionally, cost complexity pruning was applied as a post-pruning method, adjusting the ccp_alpha parameter to prune nodes with minimal contribution to predictive power. As shown in Figure 6: Post-Pruning Accuracy vs Alpha for Pruned Decision Trees, test accuracy initially improved with slight increases in ccp_alpha, stabilizing at certain values before declining when over-pruned. The resulting pruned Decision Tree, visualized in the "Best Decision Tree Visualization," demonstrates how post-pruning can streamline the model, yielding a simplified, interpretable structure with competitive accuracy.

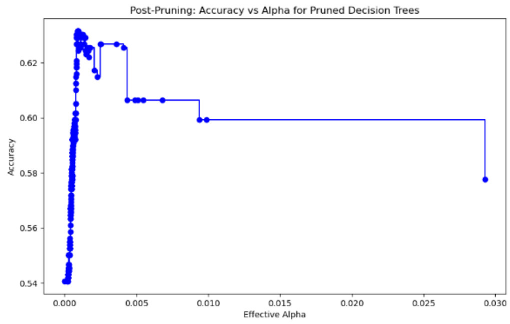

Figure 6: Post-Pruning Accuracy vs Alpha for Pruned Decision Trees

These results demonstrate that pruning is effective in simplifying the Decision Tree model while maintaining competitive accuracy, making it a suitable choice for applications requiring model interpretability.

4.5 Random Forest and Ensemble Model Results

Ensemble models were evaluated to leverage their robustness and ability to handle high feature correlations:

- **Random Forest**: The Random Forest model showed stable accuracy with an increasing number of trees. Performance plateaued around 300 trees, achieving a maximum accuracy of 64.0% (see Figure 7: Random Forest Accuracy vs Number of Trees). This result indicates that Random Forest can achieve consistent performance with a relatively small increase in model complexity.

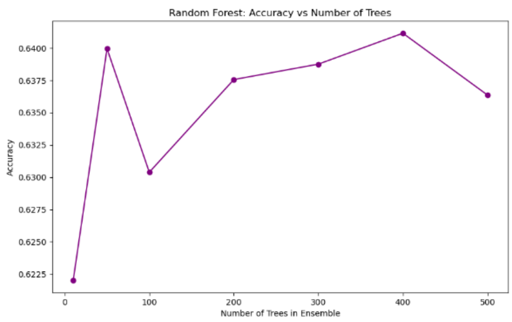

Figure 7: Random Forest Accuracy vs Number of Trees

- **Gradient Boosting and XGBoost**: Both Gradient Boosting and XGBoost were tested with varying tree counts. XGBoost initially outperformed other models, achieving an accuracy of 64.5% with 10 trees but declining as tree count increased. In contrast, Gradient Boosting achieved stable performance but did not exceed Random Forest (see Figure 8: Accuracy Comparison of Random Forest, Gradient Boosting, and XGBoost).

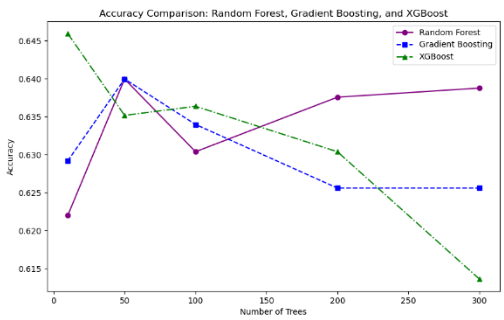

Figure 8: Accuracy Comparison of Random Forest, Gradient Boosting, and XGBoost

Overall, Random Forest proved to be the most consistent and efficient model in terms of balancing accuracy and computational cost, while XGBoost showed initial gains that diminished with complexity.

4.6 Neural Network Performance

Neural networks were employed to explore complex interactions between features. The models were trained with both Adam and SGD optimizers to evaluate their impact on accuracy and convergence.

- **Optimizer Comparison**: The Adam optimizer consistently outperformed SGD across all tested configurations. With a 200-unit hidden layer, the Adam-based neural network achieved the highest accuracy of 65.4%, while the best SGD configuration reached only 61.1% accuracy (see Figure 9: Accuracy Comparison for Neural Networks with Adam vs SGD).

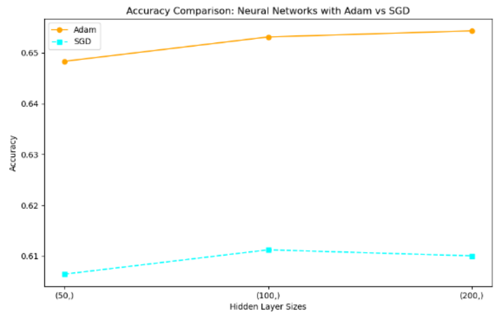

Figure 9: Accuracy Comparison for Neural Networks with Adam vs SGD

- **Layer Size Impact**: Increasing the hidden layer size improved accuracy, with the optimal configuration at 200 units. However, further increases could lead to overfitting, especially in models without sufficient regularization.

These results confirm that Adam is a more effective optimizer for this task, providing faster convergence and higher accuracy. Neural networks with Adam achieved the best overall performance, although at the cost of interpretability compared to tree-based models.

4.7 Feature Importance and Interpretability

To gain insights into the features contributing most significantly to predictions, feature importance was analyzed in the Random Forest model. The top features identified were Diameter, Shell weight, and Whole weight, aligning with the high correlations observed in the exploratory data analysis phase (Figure 3: Feature Correlation Heatmap). The Decision Tree model facilitated the extraction of interpretable "If-Then" rules, offering insights into how specific feature values influence certain age classifications, making Decision Trees a suitable choice for applications that require transparent decision-making processes.

4.8 Summary of Results

The results from each model are summarized as follows: the Decision Tree achieved 60.3% accuracy with optimal pruning at a max_depth of 5, where pruning enhanced model interpretability without compromising accuracy. The Random Forest reached 64.0% accuracy with 300 trees, providing stable and consistent performance across different configurations. For Gradient Boosting and XGBoost, XGBoost achieved the highest initial accuracy of 64.5% with 10 trees, though it showed diminishing returns with increased complexity, while Gradient Boosting remained stable but did not exceed the accuracy of the Random Forest. Finally, the Neural Network with the Adam optimizer attained the highest accuracy of 65.4% when it had a 200 - unit hidden layer. It outperformed the SGD and other models. However, it was lacking in interpretability. The Random Forest model strikes the best balance between performance and complexity. The Decision Tree model, on the other hand, provides interpretability. And the neural network with the Adam optimizer offers the highest accuracy. These results imply that each model has its own distinct advantages, depending on the desired trade - off among accuracy, interpretability, and computational efficiency.

5. *Conclusion*

This study aimed to develop a machine learning model to automate abalone age classification based on physical measurements, eliminating the need for manual ring counting. By evaluating multiple models, including Decision Trees, Random Forest, Gradient Boosting, XGBoost, and Neural Networks, we identified effective approaches for achieving a balance between model accuracy and interpretability.

Our findings indicate that the Random Forest model provided the most consistent performance, achieving 64.0% accuracy with an optimal configuration of 300 trees. Random Forest demonstrated robust handling of class imbalance and feature correlations, making it suitable for large datasets with overlapping feature information. The Decision Tree model, while less accurate, was significantly enhanced by pruning techniques that improved interpretability. Optimal pruning (at a max depth of 5) achieved 60.3% test accuracy and produced interpretable rules that could be valuable in biological studies requiring transparency in decision-making. Among the neural network models, those optimized with the Adam optimizer achieved the highest accuracy of 65.4%, though at the cost of interpretability.

5.1 Limitations

This study is limited by the class imbalance in the abalone dataset, which may bias model predictions towards the more frequent classes. Additionally, feature redundancy due to high correlations among physical measurements may have introduced complexity in certain models, potentially impacting interpretability.

5.2 Future Work

Future work can tackle class imbalance by means of techniques like synthetic sampling or class - weight adjustments. This will improve the predictions for classes that are underrepresented. Moreover, further hyperparameter tuning in neural network architectures, for example, adjusting the depth and width of hidden layers, has the potential to boost model performance. Exploring alternative evaluation metrics, such as the F1 - score, can provide a more well - balanced assessment of model accuracy across all classes [11].

In conclusion, this study shows that it is feasible to use machine learning to automate the estimation of abalone age. This has promising applications in marine biology research. The Random Forest and pruned Decision Tree models present valuable solutions that combine accuracy and interpretability, thus facilitating efficient and scalable age prediction.

6. *Reference*

[1] S. Guney, I. Kilinc, A. A. Hameed, and A. Jamil, “Abalone Age Prediction Using Machine Learning,” *Pattern Recognition and Artificial Intelligence*, pp. 329–338, 2022, doi: https://doi.org/10.1007/978-3-031-04112-9_25.

[2] M. Reichstein *et al.*, “Deep learning and process understanding for data-driven Earth system science,” *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019, doi: https://doi.org/10.1038/s41586-019-0912-1.

[3] T. Kavzoglu and A. Teke, “Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost),” *Arabian Journal for Science and Engineering*, Jan. 2022, doi: https://doi.org/10.1007/s13369-022-06560-8.

[4] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, “Interpretable machine learning: Fundamental principles and 10 grand challenges,” *Statistics Surveys*, vol. 16, no. none, Jan. 2022, doi: https://doi.org/10.1214/21-ss133.
[5] G. C. Cawley and N. L. C. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation - UEA Digital Repository,” *Journal of Machine Learning Research*, vol. 11, no. 70, pp. 2079–2107, Jul. 2010, Accessed: Oct. 15, 2024. [Online]. Available: http://jmlr.csail.mit.edu/papers/v11/cawley10a.html.

[6] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. "Abalone," UCI Machine Learning Repository, 1994. [Online]. Available: https://doi.org/10.24432/C55C7W.

[7] Li-C. Wang, “Experience of Data Analytics in EDA and Test—Principles, Promises, and Challenges,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 6, pp. 885–898, Jun. 2017, doi: https://doi.org/10.1109/tcad.2016.2621883.

[8] G. Vandewiele, K. Lannoye, O. Janssens, Femke Ongenae, F. D. Turck, and Sofie Van Hoecke, “A Genetic Algorithm for Interpretable Model Extraction from Decision Tree Ensembles,” *Lecture notes in computer science*, pp. 104–115, Jan. 2017, doi: https://doi.org/10.1007/978-3-319-67274-8_10/

[9] X. Wen and M. Zhou, “Evolution and Role of Optimizers in Training Deep Learning Models,” *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 10, pp. 2039–2042, Sep. 2024, doi: https://doi.org/10.1109/jas.2024.124806.

[10] F. M. J. Mehedi Shamrat, S. Chakraborty, Md. M. Billah, P. Das, J. N. Muna, and R. Ranjan, “A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm,” *IEEE Xplore*, Jun. 01, 2021. https://ieeexplore.ieee.org/abstract/document/9452898/

[11] P. Thölke *et al.*, “Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data,” *NeuroImage*, vol. 277, p. 120253, Aug. 2023, doi: https://doi.org/10.1016/j.neuroimage.2023.120253.