ABALONE

Small-scale literature review & Implementation

Wadiqa Baig
& others

Table of Contents

1. Introduction 1
2. Small-scale literature review 1
3. Experimental design 3
4. Experimental results and discussion 4
   4.1 Learning Curve 4
   4.2 XML with SHAP on Logistic Regression 5
   4.3 XML with LIME on Random Forest 7
5. Conclusion 9
6. References 10

1. Introduction

Machine Learning (ML) is utilised rapidly across numerous areas of application, particularly health, cybersecurity, and the financial sector. However, the "black-box" approach often hinders its adoption, where inner decision-making mechanisms are unrecognised. Such ambiguity is a severe issue, as it is not the fact that a model's outcome is correct that is significant, but rather the way it arrives there [1]. Therefore, there is a momentum for ML models to become explainable. Throughout this literature review, the terms Explainable Machine Learning (XML) and Explainable Artificial Intelligence (XAI) were interchanged since both refer to the broad goal of ensuring that ML models are understandable and interpretable to humans [2].

Decision trees are some of the easily understood models in XML, and they also feature SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) for explaining predictions. These methods try to explain the basis for a model's predicted outcomes. In this review, five studies from different application areas involving Support Vector Machines (SVM), XGBoost (XGB), Logistic Regression, and Random Forests are explored.

Moreover, we provide the outcomes of our research by analysing the use of SHAP with Logistic Regression and LIME with Random Forest on the Abalone dataset, considering how well the models predict outcomes and how easy they make it to explain their results.

2. Small-scale literature review

According to the study by J. Zhen and colleagues [3], SVC was applied to identify patients who would suffer from impaired quality of life (QOL) due to inflammatory bowel disease (IBD). SVC was the best of eight ML algorithms with 0.71 accuracy and 0.80 AUC results, exceeding Logistic Regression and Random Forest. The authors used local explanations given by LIME, a model-agnostic explanation technique, to address the problem of understanding how models work [3]. As a result of using LIME, doctors could see which elements affected the model's conclusions, making it more useful for medical work. Still, since LIME is interpreted within its context, it cannot give a broad understanding of the model, according to J. L. Imbwaga et al. [4], who also applied LIME in another setting.

J. L. Imbwaga et al. [4] used a Random Forest Classifier to detect hate speech in English and Kiswahili YouTube video comments. Their model achieved strong predictive performance with accuracy scores of 0.98 for English and 0.90 for Kiswahili. LIME was employed to identify which words most influenced classification decisions, with the word "people" having the highest positive weight (0.6%) and "that" reducing the prediction probability (0.02%). As in J. Zhen et al. [3], LIME's interpretability improved the model's usability for non-technical stakeholders, but its inability to capture overall model behaviour remained a limitation. In contrast to these local explanations, other studies in this review applied SHAP to achieve more comprehensive interpretability.

R. Ahmed et al. [5] utilised Logistic Regression with Recursive Feature Elimination (RFE) and SHAP to predict dementia in a dataset with 1,000 patients. The model showed a high level of accuracy, reaching 0.995, and it helps predict and practice medicine. SHAP explained the results on a global and local scale, showing the effect of each feature on each prediction in the data. In contrast to LIME, it looks at the bigger picture and helps fill the transparency gap in both J. Zhen et

al. [3] and J. L. Imbwaga et al. [4]. Similarly, RFE improved how the model was interpreted by removing unimportant features, a practice not used in other research. Even though SHAP explained the data well and was accurate, it was a challenge because it took too much processing time. The presence of this trade-off stands out more in important activities like cybersecurity.

S. A. Wali and I. Khan [6] built an intrusion detection system (IDS) by combining a Random Forest Classifier with SHAP to detect adversarial network attacks. They got perfect accuracy (1.0) and a CAM (Credibility Assessment Module) score of 96.5%, which is better than KNN, SVM, and the RFC without SHAP. SHAP values enabled users to identify the features that changed the system's decision and hence increased trust in the system. R. Ahmed et al. [5] note that, as shown by S. A. Wali and I. Khan [6], SHAP is important wherever trust in predictions is necessary. Nevertheless, reaching the limits of the computer, it was necessary to split the dataset into ten parts to function. In contrast, LIME-based models were made for small datasets with few constraints.

Lastly, M. Begum et al. [7] looked at thirteen ML models on four NASA datasets to spot software faults and found that XGBoost Regression (XGBR) was the most reliable performer. Though Decision Trees performed the best for specific metrics on CM1, XGBR did best in every case. The model interpretation was made using SHAP and LIME. LIME noted that "NOSI" (40%) and "CBO" (20%) were important, and SHAP selected "NOSI" (36%) and "CBO" (15%), which are both similar. Though LIME was transparent, users liked SHAP for its valuable insights about the model. Past studies show that SHAP helps uncover the details, while LIME focuses on the key factors.

The five studies discuss how different XML methods excel in various fields. LIME helped a lot in explaining human behaviours in situations related to the clinical and linguistic fields. Still, SHAP made accessing and understanding information from global and specific models simpler for all users. The performance varies between these methods because it is based on the system's complexity and what is needed regarding trust. To get the most out of XML, one should design the process to match the model and the business's needs, since it is not the exact solution for everyone.

Table. 1. Summary of Articles with XAI Methods, Models, and Performance Metrics

NO	Articles	XML	ML Model	Accuracy	Precision	F1-Score
1	[3]	LIME	SVM	0.71	0.59	0.68
2	[4]	LIME	Random Forest	English: 0.98 Kiswahili: 0.90	English: 0.99 Kiswahili: 0.93	English: 0.98 Kiswahili: 0.94
3	[5]	SHAP	Logistic Regression	0.995	1.0	0.995
4	[6]	SHAP	Random Forest	0.99	0.92	0.9
5	[7]	LIME	XGBoost regression	PC1: 0.4886 Baseline: 0.5088 CM1: 0.4407 JM1: 0.4472	-	-

3. Experimental design

For this study, the researcher used the Abalone data from the UCI Machine Learning Repository [8] to investigate using XML for classification. The number of cases in the dataset is 4,177, each with eight features showing abalone physical characteristics. The target variable 'Rings' tries to estimate the age of an abalone (age = Rings + 1.5).

Attributes:

1. **Sex** (nominal): M (male), F (female), I (infant)
2. **Length, Diameter, Height** (continuous, in mm)
3. **Whole, Shucked, Viscera, Shell Weights** (continuous, in grams)
4. **Rings** (integer): Target for classification (young < nine rings, old ≥ nine rings)

The dataset was chosen for its structured format, balanced features, and suitability for classification and regression, making it ideal for comparing explainability in linear and non-linear models.

After preprocessing the data by correcting the missing values, encoding the 'Sex' variable, and normalising it, the data was divided 80-10-10 for training, validation, and testing.

Two XML models were implemented:

- **Logistic Regression with SHAP**
- **Random Forest with LIME**

4. Experimental results and discussion

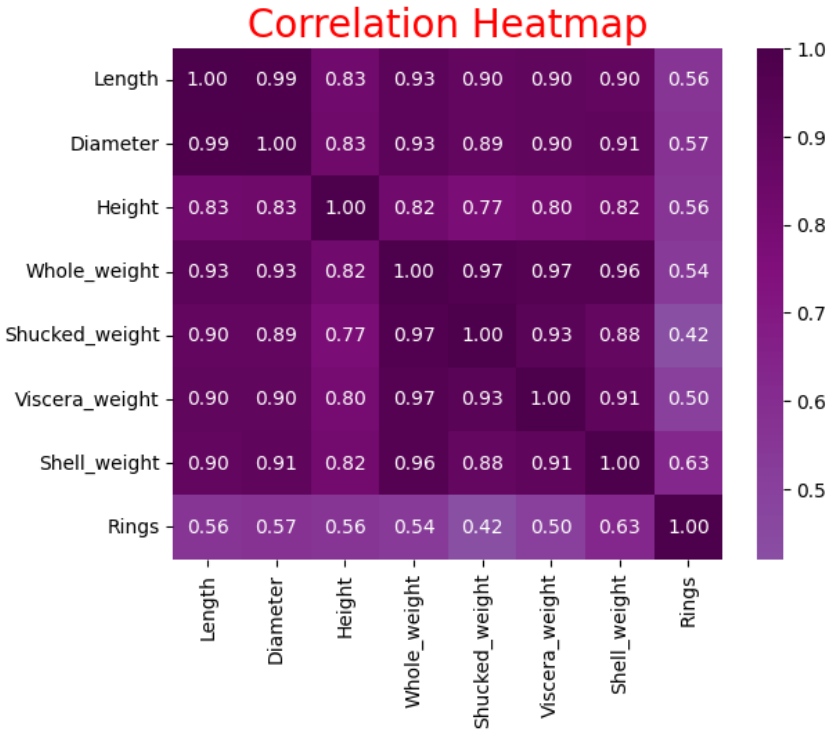

Fig. 1. Attribute Correlation Heatmap

Figure 1 shows a perfect correlation between Length and Diameter and strong correlations between weight and other weight-related features (Shucked, Viscera, and Shell weights). Length can be dropped in favour of Diameter to reduce multicollinearity, and Whole_weight can be removed due to its redundancy with other weight attributes.

4.1 Learning Curve

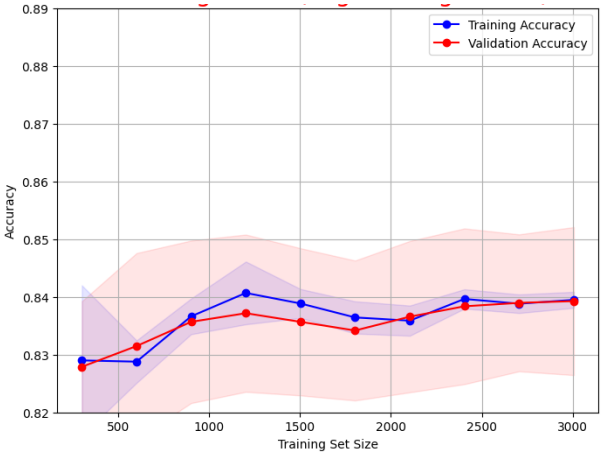

Fig. 3. Learning Curve of Logistic Regression

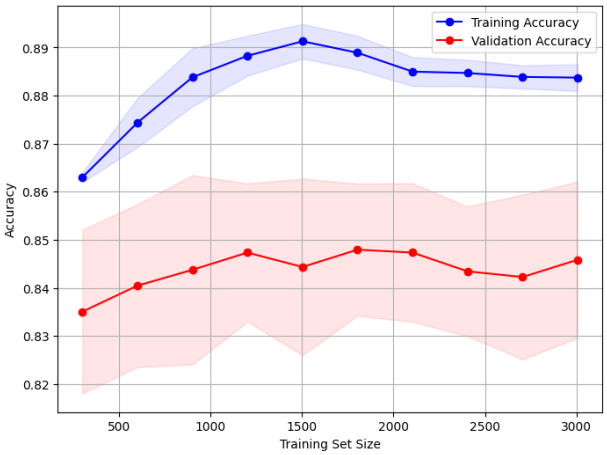

Fig. 2. Learning Curve of Random Forest

The logistic regression model showed stable but limited accuracy (83–84%) across sample sizes, flattening early and indicating underfitting (Fig. 2). In comparison, Random Forest gave better results with 88–89% accuracy on training and 84–85% on validation. At the same time, its

improvement on the learning curve slowed and finally stabilised around 1,500 samples (Fig. 3). Based on this comparison, it became clear that logistic regression could not adjust to differences, leading to underfitting. The gap in accuracy between what Random Forest learned from the training dataset and what it showed on the validation data was small. The two results are nearly identical, but Random Forest showed more ability to identify patterns.

4.2 XML with SHAP on Logistic Regression

SHAP is one of the tools for understanding and interpreting the model's decision-making process. SHAP gives the importance of each feature for a given prediction and lets users explain the model's behaviour globally and for each result.

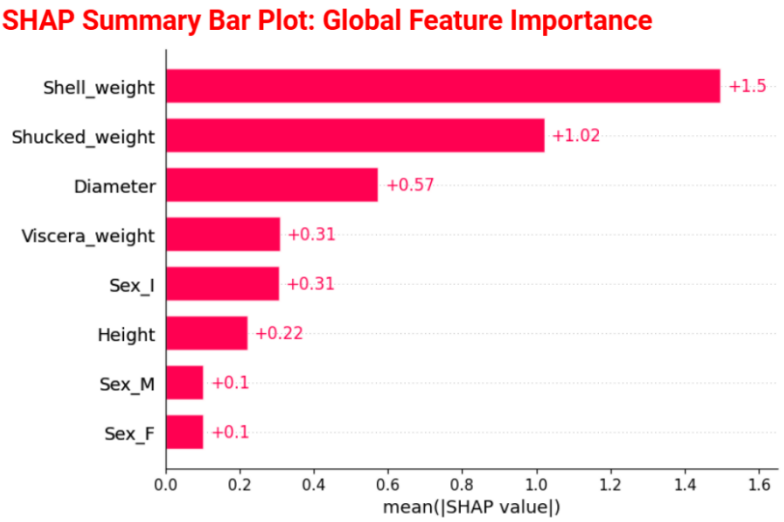

Fig. 4. Bar Plot

We began by producing a SHAP summary bar plot (Fig. 4) to understand how important each feature is in the test dataset as a whole. Shell_weight and Shucked_weight were the features that made the biggest difference, and Diameter and Viscera_weight also mattered. On the other hand, Sex and Height did very little to affect the results. The ranking aligns with biology since weight-related traits generally relate to a person's age. Because of SHAP analysis, the model's choices became more believable because they focused on important features.

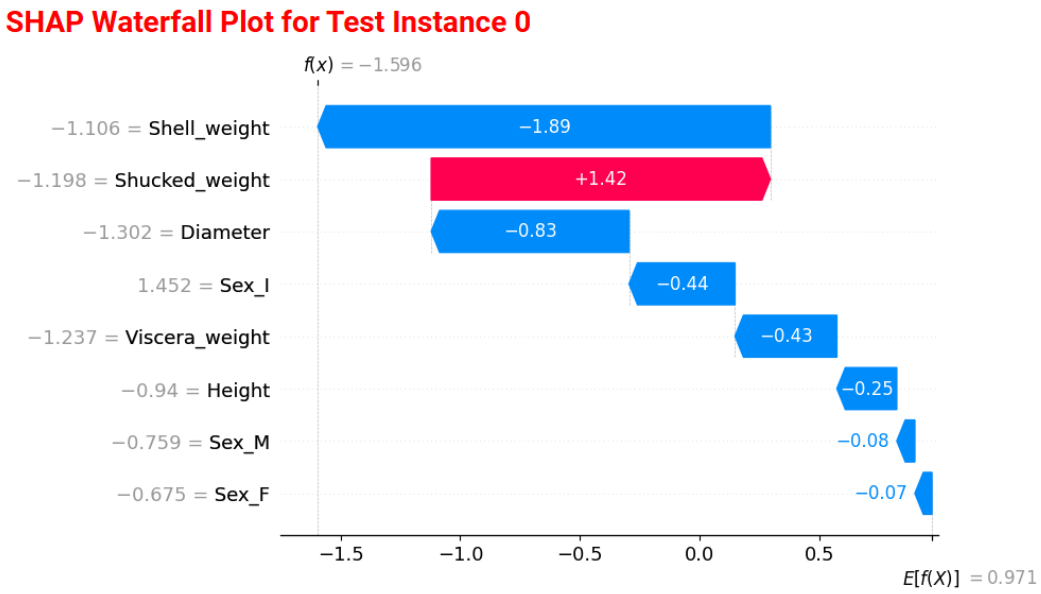

Fig. 5. Waterfall Plot

We relied on each prediction's SHAP waterfall plot (Fig. 5) to understand how each feature changed the model's output from the base value (≈0.971) to the final prediction. Based on the selected instance, the model predicted log-odds of −1.596, implying that it was most likely a "young" observation. Shell_weight and Diameter were the top factors lowering the age, outdoing the good effect from Shucked_weight. The local explanation points out that SHAP helps determine how each feature affects the prediction, which matters greatly for important models.

Fig. 6. Force Plot

In Figure 6, the SHAP force plot visualises how the influence of every feature changes when making a prediction. Shucked_weight is an example of a red feature that helps the model predict that the sample belongs to the "Old" class, whereas blue features make the opposite happen. The result of −1.60 appears once the impact of all these forces on the average prediction is added to reach the final prediction. It matches the waterfall plot (Fig. 5) by outlining the importance of different features and how they affect the outcome in a particular case.

With SHAP, it was possible to see the overall reason for predictions and determine how much the input affected the outcome and how uncertain the model was for close values. Showing all the details is very important for keeping non-technical stakeholders informed about the findings and decisions made by the model in biological or ecological settings.

4.3 XML with LIME on Random Forest

Since it is hard to explain Random Forests, we used LIME to create a straightforward model for each prediction and see how it works locally. The purpose was to examine the model's decisions by exploring each case separately and collectively through LIME.

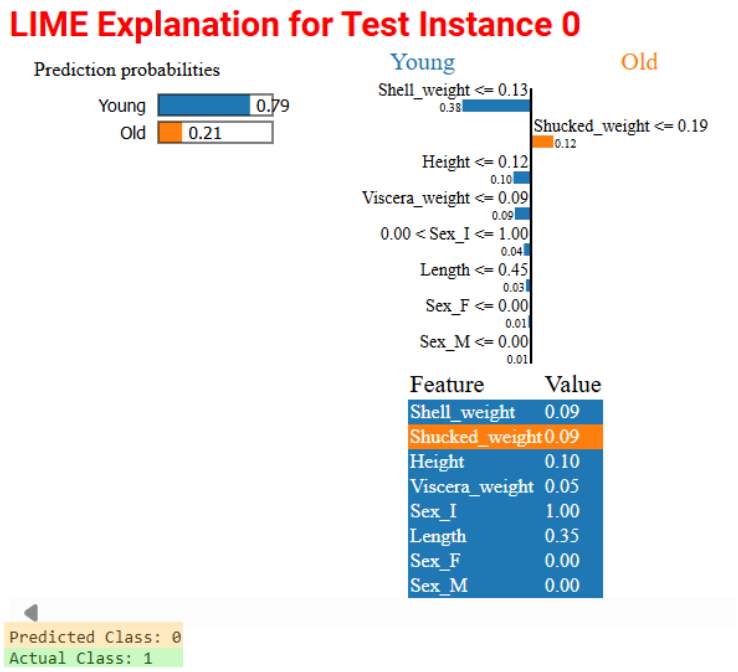

Fig. 7. LIME Agnostic Explanations Prediction

To use LIME, features from the input dataset are modified, and the changes in the predictions are noted so that a nearby feature can be identified. The model made one mistake in classifying the shell as 'Young', but it turned out to be 'Old'. This mistake was closely related to the low shell weight and height. LIME revealed that Shell_weight played a bigger role in 100 cases than categorical ones like Sex.

At one point, the model thought the subject was 'Young' with a confidence of 79% (Fig. 7), but the label said the Abalone was 'Old'. Because Shell_weight and height were the main factors, the model may not work well regarding significant outliers.

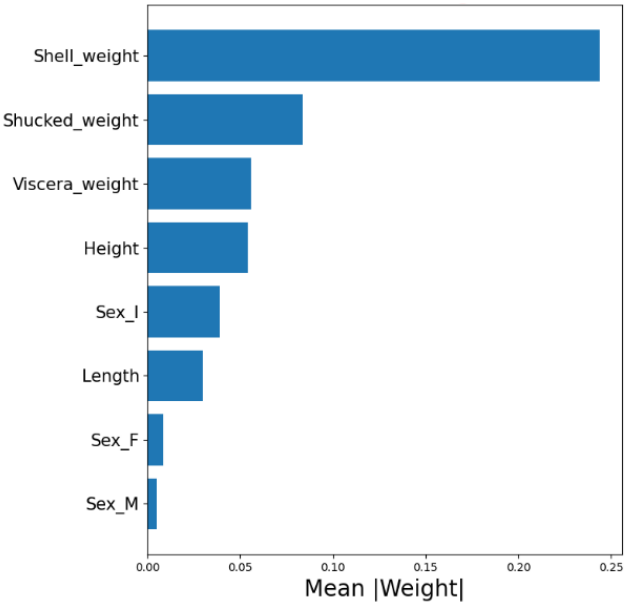

Fig. 8. Mean Absolute LIME Feature Weights (Classification)

To gain a global perspective, we computed Mean Absolute LIME Feature Weights across 100 test instances (Fig. 8). Shell_weight was the most important feature among Shucked_weight and Viscera_weight. Meanwhile, Sex_F and Sex_M had a small effect, so gender played only a minor role in the model's decisions.

As in local explanations (Fig. 7), Shell_weight played a significant role since it was stronger than Shucked_weight and pushed the prediction toward being "Old". When we study all the instances together, it becomes clear that Shell_weight is very important.

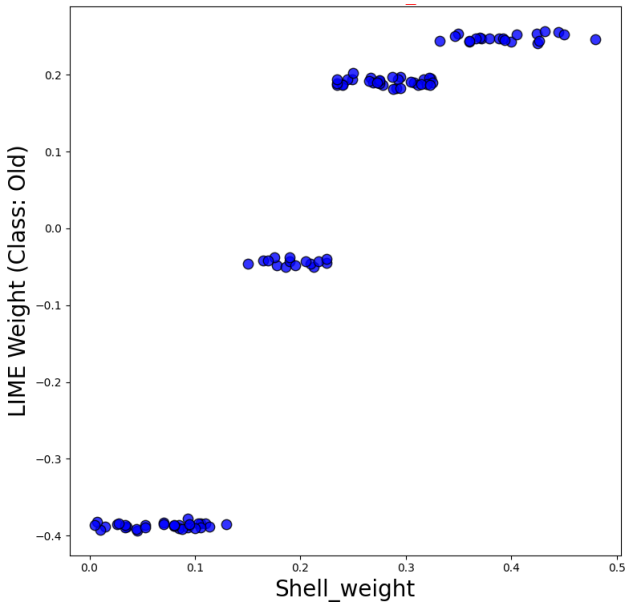

Fig. 9. LIME Trend for "Shell_weight"

We generated a LIME feature trend to know how Shell_weight affects choices (Fig. 9). As the Shell_weight gets higher (past approximately 0.25), its importance with LIME greatly increases and supports the "Old" class. Below 0.15, it has a negative influence and is a sign of "Young" players. The change is seen between 0.15 and 0.25, indicating that age in the model is indicated by whether Shell_weight is over or under a given threshold.

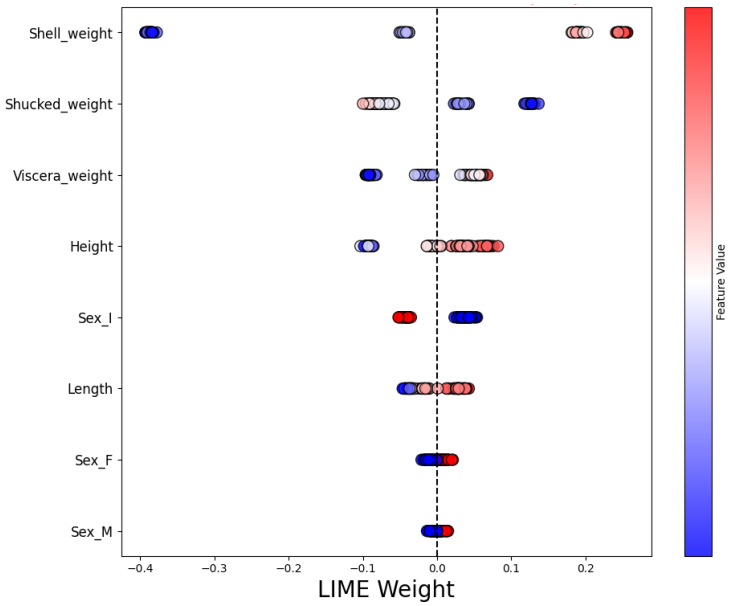

Fig. 10. LIME Beeswarm Plot

Lastly, Figure 10 shows another way to observe the relationship between features and the predictions. Low Shell_weight values help older predictions, whereas large values push toward younger predictions. Also, some, while closer, trends occur with Viscera_weight and Shucked_weight. Most features, including Sex, have values close to zero, proving that they hardly affect the result. The dots' colours and position highlight that weight, and several other characteristics determine the predictions, but gender does not add much value to this task. The Beeswarm plot proves our previous findings and emphasises that Shell_weight is an important threshold feature.

Overall, LIME helped change the Random Forest model into a transparent system by confirming feature importance trends and revealing flaws in the behaviour of edge cases.

5. Conclusion

Our research looked into five papers that combined XML techniques, LIME, and SHAP to interpret models clearly in various areas. The first two articles use SVM and XGBoost with LIME or SHAP to show that transparency helps improve people's trust in model decisions. The next utilised Logistic Regression with SHAP and Random Forest with SHAP, showing how feature attribution helps clinical and security applications. The final study applied Random Forest with LIME, proving that local explanations allow the model to make fair decisions.

In our Abalone dataset experiment, the accuracy of Random Forest with LIME was 88-89%, more than the 83-84% attained by Logistic Regression with SHAP. Although SHAP explained global features, the local aspects were explained by LIME, which made the model's work easier to detect. This explains how XAI techniques help make irregular systems clear and reliable by showing how they work.

6. References

[1] U. Kamath and J. Liu, *Explainable artificial intelligence: an introduction to interpretable machine learning*. Cham: Springer, 2021.

[2] R. Zhou and T. Hu, "Evolutionary Approaches to Explainable Machine Learning," *in Handbook of Evolutionary Machine Learning*, pp. 487–506, Nov. 2023, doi: https://doi.org/10.1007/978-981-99-3814-8_16.

[3] J. Zhen *et al.*, "Evaluating Inflammatory Bowel Disease-Related Quality of Life Using an Interpretable Machine Learning Approach: A Multicenter Study in China," *Journal of Inflammation Research*, vol. Volume 17, pp. 5271–5283, Aug. 2024, doi: https://doi.org/10.2147/jir.s470197.

[4] J. L. Imbwaga, N. B. Chittaragi, and S. G. Koolagudi, "Explainable hate speech detection using LIME," *International Journal of Speech Technology*, vol. 27, no. 3, pp. 793–815, Aug. 2024, doi: https://doi.org/10.1007/s10772-024-10135-3.

[5] R. Ahmed *et al.*, "A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction," *Healthcare Analytics*, vol. 6, pp. 100362–100362, Dec. 2024, doi: https://doi.org/10.1016/j.health.2024.100362.

[6] S. A. Wali and I. Khan, "Explainable AI and Random Forest Based Reliable Intrusion Detection system," Dec. 2021, doi: https://doi.org/10.36227/techrxiv.17169080.v1.

[7] M. Begum, M. H. Shuvo, I. Ashraf, A. A. Mamun, J. Uddin, and M. A. Samad, "Software Defects Identification: Results Using Machine Learning and Explainable Artificial Intelligence Techniques," *IEEE Access*, vol. 11, pp. 132750–132765, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3329051.

[8] 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.