A Hybrid Human-Centric approach to combining Rule-based and Attribute based Explanations*

Bahavathy Kathirgamanathan[1,2,*], Gennady Andrienko[1,2,3] and Natalia Andrienko[1,2,3]

[1] *Fraunhofer Institute IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany*
[2] *Lamarr Institute for Machine Learning and Artificial Intelligence, Schloss Birlinghoven, 53757 Sankt Augustin, Germany*
[3] *City St George's University of London, London EC1V 0HB, UK*

Abstract

As machine learning systems become increasingly integrated into critical decision-making processes, the need for clear and comprehensible model explanations has grown significantly. Traditional interpretability methods, such as rule-based models and feature attribution techniques, each offer unique strengths but also face significant limitations when used independently. Rule-based approaches provide intuitive, logic-driven explanations but can be difficult to scale. In contrast, feature attribution methods like SHAP can work well with large amounts of data and complex models, yet often lack clarity and alignment with human reasoning. In this paper, we propose a hybrid human-centric approach that combines rule-based and feature attribution techniques in a visual way to deliver more coherent and actionable explanations. By extracting and analyzing rules from a Random Forest classifier and integrating SHAP values to guide rule exploration, we offer a unified framework that enhances interpretability. We demonstrate the effectiveness of this approach through a case study on a vessel trajectory classification task, highlighting how the combination of rule distributions and feature attributions can provide deeper insight into model behavior and decision-making processes.

Keywords

Explainable AI, Visual Analytics, Rule-based, Attribute-based

1. Introduction

Interpretability in machine learning applications is crucial for trust, transparency, and accountability of the models. While machine learning models are often able to achieve high predictive performance, their complexity makes it difficult to understand. Two prominent approaches to interpretability are rule-based methods and feature attribution techniques. Rule-based interpretability involves expressing model decisions through logical rules, often in the form of if-then statements. Examples include decision trees, rule lists, and decision sets. Rule-based methods are advantageous as they are often intuitive and align with human reasoning. However, they do not scale well and large rule sets can become difficult to interpret. Feature attribution methods assess the contribution of individual features to a models predictions. Examples of this include SHAP (SHapeley Additive exPlanations), LIME (Local Interpretable Model agnostic Explanations), and gradient-based saliency maps. Feature attributions are good as they generally work with any black-box model and can be used even on high dimensional data. However, they lack the ability to explain individual predictions and are often not easy to align with human reasoning. Feature attribution techniques are also known to be instable and it is a challenge to provide consistent explanations.

While rule-based and feature attribution methods have distinct strengths and weaknesses, they are often used in isolation whereas their integration can lead to more interpretable and trustworthy machine learning models. By leveraging their complementary advantages, we can develop systems that are both transparent and effective in decision-making, ultimately increasing user trust and model accountability.

[*] Corresponding author.

bahavathy.kathirgamanathan@iais.fraunhofer.de (B. Kathirgamanathan); gennadiy.andriyenko@iais.fraunhofer.de (G. Andrienko); nathaliya.andriyenko@iais.fraunhofer.de (N. Andrienko)
0000-0002-2329-8909 (B. Kathirgamanathan); 0000-0002-8574-6295 (G. Andrienko); 0000-0003-3313-1560 (N. Andrienko)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

A further challenge lies in presenting these explanations. Rule-based techniques can provide a large number of rules that align with human reasoning, yet are not easily readable for the user. Furthermore, attribution techniques such as SHAP provide qualitative importance scores often presented as charts showing feature rankings, which do not necessarily explain how and under what conditions these features affect the prediction.

In this work, we develop a hybrid approach and couple it with visual analytics to provide human-centered explanations. We extract the rules generated by a rule-based model (in this case, random forest) and use the rule distributions from the ruleset to gain a visual understanding of how feature space impacts the classification. We couple this with a feature attribution strategy (in this case, SHAP) to help further guide the rule exploration. We present this approach using a case study on a vessel trajectory dataset.

2. Related Work

Human-centered Explainable Artificial Intelligence (XAI) is an area that focuses on understanding and evaluating explainability from the perspective of the users of AI systems. It acknowledges that explainability is not just a technical property but is inherently tied to how people perceive and comprehend explanations [1]. Initially, XAI evaluation focused on the objective quality of generated explanations, such as correctness and completeness, viewed from the system developer's perspective. However, there's been a growing recognition that the effectiveness and benefit of an explanation depend on the person receiving it [2].

When looking at tabular data, two of the most frequently used methods for explanation are feature importance scores and rule-based explanations. For feature importance, an algorithm calculates a vector containing a value for each feature indicating the importance of that feature for the model's decision. Rule-based methods display a set of premises that must be satisfied to meet the outcome of the rule [3]. Rule-based models are generally acknowledged to be interpretable and intuitive [4]. However, they do not scale well, and rules containing multiple logical conditions can be hard to read and understand. Feature importance is one of the most popular explanation strategies used. SHAP (SHapeley Additive exPlanations) is a widely used attribution based explanation technique which works by computing feature importance using Shapeley values, a concept used in cooperative game theory [5]. LIME (Local Interpretable Model-Agnostic Explanations) also generates feature importance scores by perturbing the input data and fitting a local surrogate model to approximate the behavior of the original model in the neighborhood of the instance to be explained [6]. Both SHAP and LIME offer certain insights into model behavior but can sometimes be challenging to interpret by a user as the explanations do not match well with the human mental model.

Recent advances in Visual Analytics (VA) have demonstrated the usefulness of employing VA to help bridge the gap between ML and humans [7]. VA has proven to be useful for enhancing the interpretability of decision tree ensembles, rule-based models, and even black-box classifiers [8]. There are rule-based explanation techniques that use visualization to further improve the comprehensibility of the explanations. RuleMatrix is one such technique that extracts rules, arranges them into a structured matrix format, and enables detailed interactive inspection of feature relationships and impacts on rule outcomes [9]. iForest allows the user to interactively explore random forest models and predictions by showcasing the decision paths [10]. Explainable matrix is another technique which visualises the rules in a matrix-like representation [11]

Overall, the growing body of work in human-centered XAI and visual analytics underscores the importance of not only generating accurate explanations but presenting them in ways that support human comprehension, reasoning, exploration, and decision-making. Our work builds on this foundation by proposing a hybrid approach combining rule-based representations, feature attribution techniques, and visual analytics methods (Fig. 1).

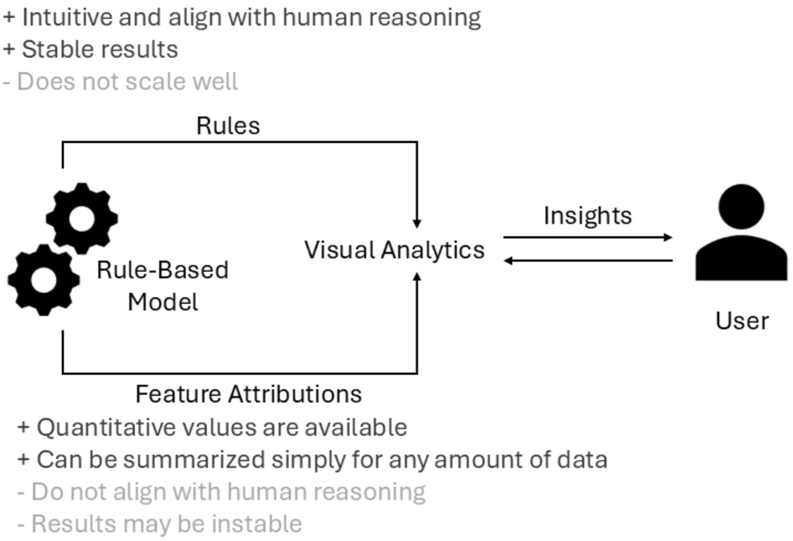

+ Intuitive and align with human reasoning
+ Stable results
- Does not scale well
Rules
Rule-Based Model
Visual Analytics
Insights
User
Feature Attributions
+ Quantitative values are available
+ Can be summarized simply for any amount of data
- Do not align with human reasoning
- Results may be instable

Figure 1: Combining rule-based and feature attribution techniques within an interactive visual interface enhances explanation quality by leveraging the strengths of both XAI approaches and the advantages of visual analytics in supporting human perception and cognition.

3. Hybrid approach for model understanding

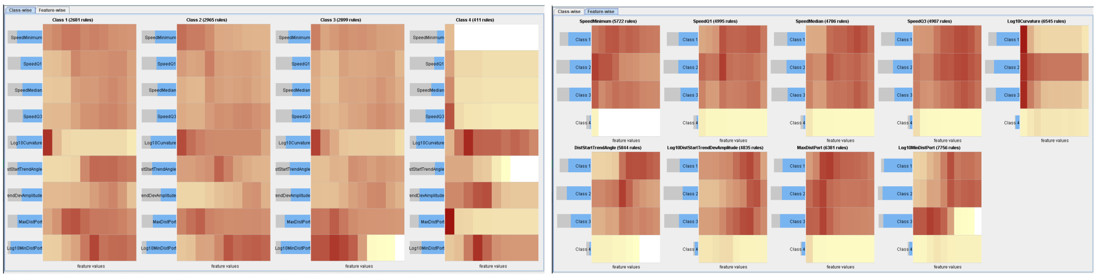
We propose a hybrid methodology that integrates rule-based model representations with feature attribution techniques to support and enrich model interpretability. While our study focuses on rule sets extracted from a random forest model, the approach is broadly applicable to rule sets derived from other models. The methodology operates in two complementary ways. First, feature attributions are used to **enhance understanding of individual rules**, clarifying which features contribute most to a rule's predictive power and how they influence the model's output. Second, attributions serve to **guide the exploration of the rule space** by highlighting the most influential features across the dataset, thus helping users prioritize which rules or feature combinations to investigate more deeply. Together, these two perspectives support a more focused and holistic analysis of the decision logic of the model.

To visualize and interact with the rules, we employ and further develop a prototype system called **RuleExplorer** [12], which integrates computational, visual, and interactive techniques to facilitate model understanding. The central visualization provides an overview of the distributions of features and their value intervals across the rule set, presented in two complementary layouts: grouped by predicted class (Fig.2a) or by feature (Fig.2b). Both layouts use heatmap matrices, where columns represent user-defined, equal-length intervals of feature values (10 intervals in the example shown in Fig. 2). Rows correspond to class-feature combinations, grouped by classes or features. Each heatmap row is accompanied by two bars: a blue bar indicating the number of rules using the feature to predict the class, and a gray bar representing the total number of rules predicting that class.

While RuleExplorer can be used without applying rules to a dataset, feature attribution techniques require data instances and their corresponding model predictions. In our study, we use SHAP to compute feature attributions, although the methodology is agnostic to the specific attribution method used. Calculated feature scores are aggregated across instances for each feature-class combination and integrated into the feature distribution visualization.

The typical user workflow under this approach is as follows:

1. **Visualize rules**: Extract rules from a trained model and load them into RuleExplorer to examine the distribution of feature values.
2. **Compute and visualize feature attribution scores**: Run a feature attribution method (e.g., SHAP) on a dataset with model predictions. Aggregate feature scores by class, and overlay the



(a) (b)

Figure 2: Visualization of the distributions of feature values for class-feature combinations grouped by predicted classes (a) and by features (b). Darker shades indicate higher frequencies of rules whose conditions include the corresponding intervals of the feature values.

results onto the feature distribution display (see Fig. 3).

3. **Interactively explore the rule set**: Use the attribution-enhanced visualizations to filter, sort, and explore rule subsets based on feature importance.
4. **Iterate:** Refine exploration by focusing on selected rules or features to build deeper understanding of model behavior.

This hybrid approach supports model validation, enhances trust, and helps uncover insights about both the data and the model's internal logic.

4. Case study: Vessel Activity Recognition

To demonstrate the proposed hybrid methodology in a real-world scenario, we apply it to a classification task aimed at recognizing types of fishing vessel activities based on their movement patterns.

4.1. Data and model

The dataset used in this study comprises trajectory segments (episodes) from 71 fishing vessels operating northwest of France between October 1, 2015, and March 31, 2016[1]. Each episode was represented by nine interval-based features derived from sequences of vessel positions, as described in [13]. These features capture speed characteristics (SpeedMinimum, SpeedQ1, SpeedMedian, SpeedQ3), trajectory shape (Log10Curvature, DistStartTrendAngle, Log10DistStartTrendDevAmplitude), and proximity to ports (MaxDistPort, Log10MinDistPort). To address skewed value distributions, logarithmic transformations were applied where appropriate. Episodes were labeled into four activity classes: Forward movement, Trawling, Port enter/exit, and Anchoring.

A random forest classifier was trained to distinguish between these movement types using a 75:25 train-test split (random state = 0). Prior to training, all numerical features were standardized using z-score normalization to ensure equal treatment of features regardless of magnitude. The resulting model demonstrated strong predictive performance, achieving a test accuracy of 0.9733. The trained model comprised 100 decision trees, each trained on a bootstrapped sample of the training data. Rules were generated by traversing each tree from root to leaf, resulting in one rule per path. Because the model was trained on standardized features, all threshold values were transformed back to the original scale for interpretability. Each extracted rule consisted of a predicted class and a set of feature constraints, expressed as value ranges. For interpretability, open-ended intervals were bounded using the minimum and maximum observed values for the respective features. This process produced an initial rule set of 9,939 unique rules.

However, due to the ensemble nature of the model and the inherent redundancy in decision forests, the rule set included inconsistencies and overlaps. To address this, we used RuleExplorer to clean the

[1] https://zenodo.org/records/1167595

rule set. We removed 113 automatically detected contradictory rules (where identical conditions led to different predicted classes) and 311 subsumed rules (fully contained within more general rules). After this automated cleaning step, 9,515 rules remained. We further refined the rule set using interactive exploration features of RuleExplorer to identify rules that were inconsistent with domain knowledge or human expectations. We detected and removed rules that omitted features critical for predicting certain classes, rules containing feature intervals beyond plausible limits for the predicted activity, and rules with predictive accuracy below 50% on the test data.

Following this expert-driven pruning, the final rule set consisted of 8,956 rules. Importantly, this refinement process did not reduce the model's overall accuracy but significantly improved the interpretability and logical consistency of the extracted rules from a human-centric perspective.

4.2. Combining rules with feature attributions

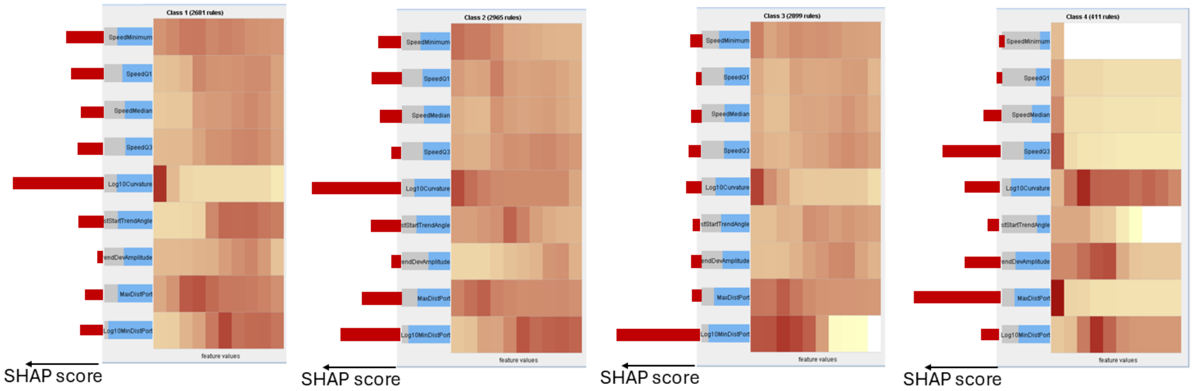


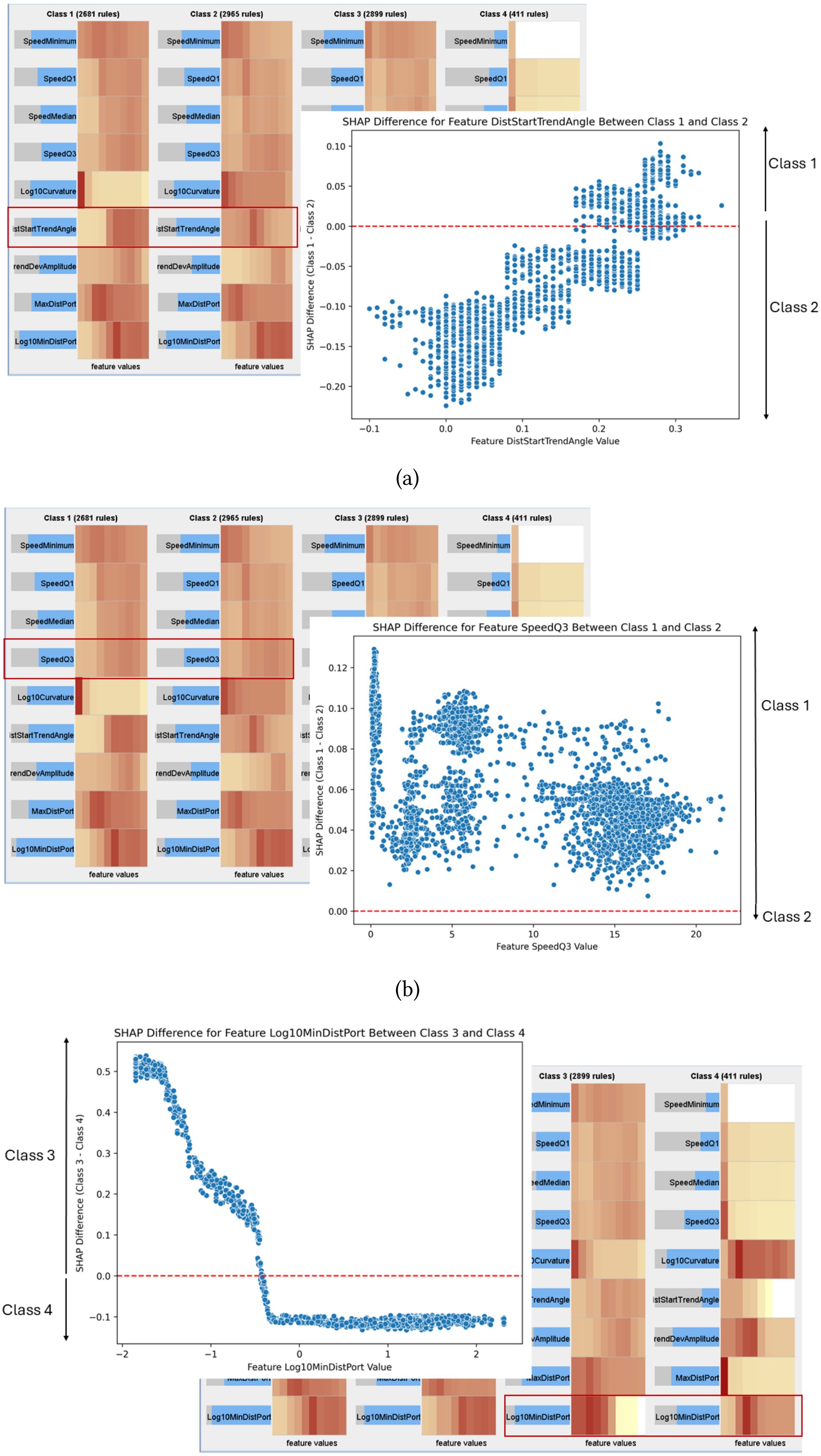
Figure 3: Visualization of the feature intervals distributions with integrated representation of aggregated SHAP values by proportional lengths of red bars.

We extend the functionality of RuleExplorer by integrating the computation and visualization of feature impact scores derived from feature attribution methods. This extension is designed to serve two key objectives:

- **Enhancing interpretability of model behavior**: Feature impact scores provide complementary information that is not directly accessible from rule-based explanations alone. By quantifying the contribution of individual features to model predictions, they support a more detailed understanding of the model's reasoning and foster greater transparency.
- **Supporting interactive exploration and alignment**: Awareness of feature importance enables users to more effectively navigate and interrogate the rule space, prioritizing rules that include or omit key features. This guided interaction helps users align the model's logic with their domain knowledge. Additionally, the comparison of rule-based and attribution-based explanations allows users to assess their consistency, offering opportunities to identify discrepancies and improve mutual trust in the system.

In this way, both explanation techniques contribute to a more robust and adaptive human-AI decision-making process.

Figure 3 presents a composite visual representation that integrates two layers of information: (1) the distribution of feature value intervals across predicted classes, and (2) feature importance scores aggregated (averaged) over all instances assigned to each class by the model. The importance scores are visualized as red horizontal bars with lengths proportional to the mean impact of each feature on the corresponding class prediction. These bars are positioned to the left of the heatmap rows that depict feature value distributions, enabling better understanding of how the model uses the features to derive its predictions.



(a)

(b)

(c)

Figure 4: Pairwise comparison of SHAP values between classes for individual features. Each scatterplot shows the difference in SHAP values between two classes plotted against the feature values. The red horizontal line indicates the point of no class preference (zero difference). (a) illustrates a clear transition in class preference as the feature value changes. (b) shows a case where no such trend is observed. (c) demonstrates a distinct class preference with relatively strong agreement across instances.

For example, in the class 1 part, the feature Log10Curvature has the longest red bar, indicating it is the most influential feature for predicting this class. While this importance could be partially inferred from the heatmap, where most rules associated with class 1 include the lowest interval of Log10Curvature, this pattern alone is not definitive. Notably, the heatmap for class 3 appears visually similar, with a comparable emphasis on low curvature values. However, the attribution bars clarify that Log10Curvature has minimal impact on predictions for class 3. At the same time, the feature shows high importance for class 2, even though the corresponding value distribution is more dispersed across intervals. In such cases, feature attribution scores help disambiguate overlapping rule patterns and provide a more direct and interpretable signal of feature relevance in class predictions.

An interesting observation emerges when comparing rule frequency with feature importance scores: some features appear in a large number of rules but exhibit low SHAP values, indicating that they may be frequently used due to how the trees are constructed, yet have limited influence on model predictions. This distinction highlights how the presence of features alone does not necessarily reflect predictive power, especially in ensemble models such as random forests, where features may be included in splits without significantly contributing to the final decision.

For instance, Log10MinDistPort appears frequently in rules predicting class 1 (Forward movement), as indicated by the blue bar in the visualization. However, its low SHAP value suggests that, while this feature helps narrow down the candidate instances, perhaps by filtering out vessels close to port, it does not play a central role in the classification itself. This interpretation is consistent with domain knowledge: vessels in forward movement are typically not near ports, but distance from port is not a defining characteristic of this movement type.

In contrast, features like Log10Curvature not only occur frequently in the rule set for class 1 but also exhibit high SHAP scores and distinct distribution patterns. This alignment across multiple signals, namely, frequency, importance, and value range, indicates that Log10Curvature is both structurally and functionally critical for predicting forward movement.

To gain deeper insight into the decision boundaries between classes, we explore how feature attributions vary across class predictions by comparing SHAP scores pairwise. Figure 4 presents scatterplots where the Y-axis encodes the difference in SHAP values between two classes for a given feature, while the X-axis shows the corresponding feature values. These visualizations help reveal how specific value ranges of a feature influence the model's preference toward one class over another.

For example, in Fig. 4a, we compare the SHAP scores of the feature DistStartTrendAngle for class 2 (Trawling) versus class 1 (Forward movement). The scatterplot reveals a clear pattern: lower values of the feature strongly favor predictions of class 2, while higher values support predictions of class 1. Specifically, values below approximately 0.2 consistently contribute to class 2 predictions, values above 0.3 favor class 1, and the intermediate range shows more ambiguous influence. While the feature distribution heatmaps for DistStartTrendAngle in classes 1 and 2 provide some indication of this trend, the scatterplot conveys the relationship much more clearly by directly mapping feature values to their impacts on class differentiation. This type of visualization enables a more interpretable and refined understanding of how feature values shift decision boundaries. Such an insight would otherwise be difficult to extract from rule coverage or attribution values alone.

Having access to both rule-based value distributions and SHAP-based impact scores may be particularly beneficial when the feature distributions alone do not reveal clear differences between classes. This is illustrated in Fig. 4b, which compares the role of the feature SpeedQ3 for the same classes 1 and 2. The heatmaps show that the usage frequency and value distributions of this feature are nearly identical for both classes, suggesting a limited role in distinguishing between them. This observation is reinforced by the scatterplot of SHAP value differences: unlike in the previous example with DistStartTrendAngle, there is no visible trend linking SpeedQ3 values to preference for one class over the other. This implies that while SpeedQ3 may appear frequently in rules for both classes, it does not contribute significantly to their differentiation. Notably, the SHAP scores indicate that it still tends to be slightly more influential for class 1 predictions overall.

One more example is shown in Fig. 4c, where the feature Log10MinDistPort is analyzed for class separation between classes 3 (Port enter/exit) and 4 (Anchoring). Here, the scatterplot reveals a much stronger trend: lower values of Log10MinDistPort clearly favor class 3, while higher values support class 4. This suggests that the feature is a decisive factor in differentiating between these two movement types. There is only a small interval approximately around value -0.4 where either of the two classes can be predicted, requiring other features to be used for the class differentiation. The contrast with DistStartTrendAngle in Fig. 4a is notable: while that feature exhibited some class-separating behavior, it lacked such a strong and direct trend, implying that interactions with other features were more critical in that case.

Overall, the combination of SHAP-based scatterplots and rule-based value distributions enhances the depth of exploration and facilitates gaining more meaningful insights. It allows users to probe not only whether a feature is important, but how and where in its value range it contributes to class separation. This dual representation supports more informed hypotheses, helps identify borderline or ambiguous regions in the feature space, and ultimately enhances the interpretability of complex model behavior.

5. Discussion

In this work, we have presented a hybrid, two-way methodology that combines rule-based representations with feature attribution scores (such as SHAP values) within an interactive visual interface. By first extracting and visualizing rules from a model, then overlaying class-aggregated SHAP scores, and finally enabling iterative, user-driven selection and inspection, our approach supports a human-centered exploration of model logic. We now discuss the benefits of this approach for both model developers (debugging and refinement) and end users (transparent explanations and trust), as well as its current limitations and directions for future work.

5.1. Model Developer Perspective

Diagnosing overrepresented features. Low importance scores of features that appear frequently in rules may suggest that the model overuses these features structurally (e.g., due to tree-splitting mechanisms), even though they contribute little to the predictive outcome. Through detailed investigation of the uses of such features, developers may find a way to restructure and simplify the model without compromising its accuracy.

Validating feature relevance. When domain knowledge suggests that a feature should (or should not) be important for a class, the composite visualization allows users to check whether those expectations are met based on empirical model behavior. Features having unexpectedly low or high SHAP scores may signal potential data quality or confounding issues.

Refining model logic. Developers can iteratively prune or refine rules by removing those that rely heavily on weak or misleading features, improving both interpretability and robustness without sacrificing model accuracy.

Supporting collaborative workflows. Domain experts and data scientists can jointly investigate why the model behaves as it does, what it pays attention to, and how its decisions align (or not) with human reasoning.

5.2. End User Explanations

Combining rule-based explanations with feature attribution scores promises significant benefits also for communicating model decisions to end users.

Contextualized rule explanations. Rules, by their nature, provide intuitive if-then statements that resemble human reasoning. They are particularly effective for conveying why a specific decision was made by identifying which conditions were met. Feature attribution complements this by quantifying how much each aspect contributed to a decision, which helps users focus on the most influential features relevant to a prediction.

Contrastive reasoning. Users can explore how changes in specific features might alter the prediction. For example, if a vessel was classified as Trawling, users can inspect which features pushed the prediction in that direction and which changes of these features might have led to a different classification (e.g., Forward movement).

Uncertainty and exception handling. In cases where feature distributions and attribution scores do not align, users gain insight into uncertainty or ambiguity in the model's reasoning. This prevents blind trust and encourages informed decision-making.

Trust through consistency. When rule-based logic, feature distributions, and SHAP impact align, users gain confidence in the model's internal coherence. This is highly valuable for high-stakes domains such as maritime monitoring, healthcare, or finance.

Overall, this hybrid explanation framework supports a more interactive, transparent, and user-centered approach to decision support.

5.3. Limitations and Future Work

SHAP stability and causality. SHAP values can be sensitive to small perturbations in data or model parameters, which may undermine explanation reliability. Moreover, SHAP quantifies correlational importance rather than true causality. By combining SHAP with aggregate rule distributions, we enable one to be validated based on the other. To further enhance robustness, techniques such as deletion–insertion sensitivity analysis [14] can be integrated to stabilize SHAP scores before visualization.

Refined visualization of feature attributions. Our current implementation displays only the mean SHAP value for each feature-class combination, which may hide important variations in how feature contributions differ across instances. A more granular representation, such as visualizing average SHAP scores across value intervals or using histograms or box plots to show score distributions, could support a deeper understanding of feature effects.

Generalization and validation. Our preliminary case study focuses on vessel activity recognition. Future work will apply the methodology across diverse domains and model types (e.g., gradient-boosted trees, neural rule extraction) to validate its generality and uncover domain-specific refinements.

Local explanations. While our current workflow aggregates SHAP scores by class, extending the interface to support per-instance rule and attribution views would enable detailed local explanations thereby facilitating use cases where understanding individual decisions is critical. Our vision of applying the hybrid approach to local explanations is outlined in Section 5.2.

The proposed combination of rule-based representations with feature attribution scores shows promise for empowering both developers and end users to interrogate, validate, and build trust in complex machine learning models, thereby advancing the design of hybrid human–AI systems.

Acknowledgments

This work was supported by Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (Lamarr22B), and by EU in project CrexData (grant no. 101092749).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for the following: Grammar and spelling check, rewording. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Kim, H. Maathuis, D. Sent, Human-centered evaluation of explainable ai applications: a systematic review, Frontiers in Artificial Intelligence 7 (2024) 1456486. doi:10.3389/frai.2024.1456486.

[2] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance, Frontiers in Computer Science 5 (2023) 1096257. doi:10.3389/fcomp.2023.1096257.

[3] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, Data Mining and Knowledge Discovery 37 (2023) 1719–1778. doi:10.1007/s10618-023-00933-9.

[4] R. Guidotti, S. Ruggieri, On the stability of interpretable models, in: 2019 international joint conference on neural networks (IJCNN), IEEE, 2019, pp. 1–8. doi:10.48550/arXiv.1810.09352.

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017). doi:10.5555/3295222.3295230.

[6] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[7] N. Andrienko, G. Andrienko, L. Adilova, S. Wrobel, Visual analytics for human-centered machine learning, IEEE Computer Graphics and Applications 42 (2022) 123–133. doi:10.1109/MCG.2021.3130314.

[8] C. Maçãs, J. R. Campos, N. Lourenço, P. Machado, Visualisation of random forest classification, Information Visualization 23 (2024) 312–327. doi:https://doi.org/10.1177/14738716241260745.

[9] Y. Ming, H. Qu, E. Bertini, Rulematrix: Visualizing and understanding classifiers with rules, IEEE transactions on visualization and computer graphics 25 (2018) 342–352. doi:10.1109/TVCG.2018.2864812.

[10] X. Zhao, Y. Wu, D. L. Lee, W. Cui, iforest: Interpreting random forests via visual analytics, IEEE transactions on visualization and computer graphics 25 (2018) 407–416. doi:10.1109/TVCG.2018.2864475.

[11] M. P. Neto, F. V. Paulovich, Explainable Matrix - Visualization for Global and Local Interpretability of Random Forest Classification Ensembles , IEEE Transactions on Visualization & Computer Graphics 27 (2021) 1427–1437. doi:10.1109/TVCG.2020.3030354.

[12] L. Adilova, M. Kamp, G. Andrienko, N. Andrienko, Re-interpreting rules interpretability, International Journal of Data Science and Analytics (2023). doi:10.1007/s41060-023-00398-5.

[13] N. Andrienko, G. Andrienko, A. Artikis, P. Mantenoglou, S. Rinzivillo, Human-in-the-loop: visual analytics for building models recognising behavioural patterns in time series, IEEE Computer Graphics and Applications (2024). doi:MCG.2024.3379851.

[14] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, arXiv preprint arXiv:1806.07421 (2018). doi:10.48550/arXiv.1806.07421.