EuroVis Workshop on Visual Analytics (2026)
D. Archambault and A. Vilanova (Editors)

Going Beyond Model-Centric XAI towards Human-Centric Explanations

B.Kathirgamanathan[†1,2], G.Andrienko[1,3] and N.Andrienko[1,3]

[1] Fraunhofer Institute IAIS, Sankt Augustin, Germany
[2] University of Cologne, Köln, Germany
[3] City St George's University of London, UK

Abstract

Model-Centric Explainable AI (XAI) techniques are most commonly used, although they typically do not focus on how useful, meaningful, or actionable the explanations are for humans. In this paper, we argue that this model-centric paradigm is insufficient for building trustworthy AI systems, particularly in high-stakes domains where human decision-makers must interpret and validate AI recommendations. We propose a shift toward human-centred XAI through an interactive visual analytics workflow and interface that integrate rule-based explanations with feature attribution methods, allowing users to dynamically explore and compare multiple explanation techniques side by side. By making agreements and contradictions between methods visible and explorable, the interface allows domain experts to identify inconsistencies and validate model behaviour against their own knowledge, and helps model developers assess which explanations to trust under which conditions. We illustrate this approach through a case study on fishing vessel movement classification, where the hybrid visualisation reveals discrepancies between the rule-based explanation, TreeSHAP, KernelSHAP, and LIME that would remain hidden when using any single method in isolation. This conceptually demonstrates the value of interactive visualisation as a bridge between model-centric explanations and human-centred understanding.

1. Introduction

Traditional explainable AI (XAI) methods are generally model-centric, where they seek to describe how a model arrives at its predictions regardless of whether the explanations are meaningful, trustworthy, or usable for humans. XAI methods are generally evaluated based on technical metrics such as faithfulness, robustness, and computational efficiency. While these properties are important, they are not sufficient. An explanation can be faithful to a flawed model, stable yet uninformative, and computationally efficient while still being of minimal utility to the end user [PSGH*21, LGM20]. We argue that this model-centric way of explaining is insufficient for building trustworthy AI systems, particularly in high-stakes domains where human decision-makers must understand and validate AI recommendations. Hence, there should be a shift toward human-centred XAI that prioritises alignment with human reasoning and domain knowledge over merely explaining model behaviour.

It is also known that explanation techniques rarely fully agree with each other, and each technique has own strengths and limitations. For example, feature attribution methods, such as SHAP (SHapeley Additive exPlanations) [LL17] or LIME (Local Interpretable Model-agnostic Explanations) [RSG16], are very popular due to their advantage of being model-agnostic and suitable for high-dimensional data. However, they may not align well with human logic and are known to be unstable, raising concerns about consistency and trust in the explanations. Rule-based methods, on the other hand, explain model decisions using logical if-then rules and are often perceived as more intuitive and better aligned with human reasoning. However, they can struggle to scale effectively and are hence not easy to use for most real-world applications. Hybrid approaches that bring together different explanation methods can highlight discrepancies between them and show where model behaviour diverges from domain expectations or where explanation methods fail to faithfully represent model logic. Furthermore, with recent trends showing visualisation as a popular means of improving explanations [CKK24], we aim to explore whether combining different XAI techniques can leverage their complementary benefits to provide explanations that are more meaningful and better aligned with human mental models.

This paper introduces a hybrid, human-centred workflow that integrates rule-based and attribution-based explanations within a visual analytics workflow. Starting from a rule-based machine learn-

† 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*

ing model (a random forest), we extract decision rules and compute feature attributions using multiple methods (TreeSHAP, KernelSHAP, and LIME). These explanations are then brought together in a rule–attribution visualisation that allows users to compare explanations and see where methods agree, where they contradict each other, and how these patterns relate to domain knowledge. We demonstrate the workflow on a real-world case study of fishing vessel movement classification and show how the integrated visualisation helps domain experts identify inconsistencies, assess which explanations are more trustworthy under which conditions, and refine both data and models accordingly. Ultimately, the goal is to support a continuous feedback loop in which human insight informs interpretation of model behaviour and guides iterative improvement of the models themselves.

2. Related Work

2.1. Limitations of Model-Centric XAI

Model-centric XAI typically reports which features influence a model's prediction, but it may not be clear to users how they can act on this information. There is now a growing recognition that the effectiveness and benefit of an explanation depend on the person receiving it [HMKL23].

For tabular data, two commonly used explanation approaches are feature attribution and rule-based explanations. Feature attribution methods compute importance scores expressing the contribution of each feature to the model's prediction. Rule-based approaches present a combination of logical conditions that must be satisfied for a specific outcome to occur [BGG*23]. Rule-based models are regarded as intuitive and interpretable [GR19]; however, they do not scale effectively, and rules with many conditions can be difficult to read and understand.

Feature attribution methods remain among the most widely adopted explanation strategies. SHAP (SHapley Additive exPlanations) is a popular method that estimates feature importance using Shapley values, a concept derived from cooperative game theory [LL17]. Similarly, LIME (Local Interpretable Model-Agnostic Explanations) computes feature importance by perturbing input data and fitting a local surrogate model to approximate the behaviour of the original model in the vicinity of the instance being explained [RSG16]. Although both SHAP and LIME provide valuable insights into model behaviour, their explanations can be difficult for users to interpret, as they may not align well with human mental models. Furthermore, they have been shown to be unstable: similar instances can produce very different explanations, motivating the need to cross-check and validate explanations from such techniques.

Several factors are important for building trust in explanations, including scalability, human-alignment, stability, and actionability. Scalability means that a method remains useful even when the data or model size becomes large. Human-alignment refers to the extent to which humans agree with the reasons models provide for their predictions [LWH*22]. It may be crucial for fostering trust and user acceptance of these systems [MFP*23]. Stability requires that similar inputs receive similar explanations [LZED25]. Actionability means that an explanation tells the user what concrete changes can be taken to achieve a different prediction or outcome. With all these aspects in mind, we primarily focus our work on human-alignment, analysing stability across explainers, and supporting feedback-driven refinement.

2.2. Visual Analytics for human-aligned explanations

The increasing complexity of ML models has motivated the development of visual analytics tools that support interpretation and exploration of model structure and behaviour [AAAW22, MCLM24, MJEP*21]. Most existing systems focus on helping users inspect models, understand their internal workings, and assess performance, rather than explicitly providing user-facing explanations. While visual analytics as such has inherent potential to mediate between human reasoning and model behaviour, its use for constructing and communicating explanations remains relatively underexplored.

Several rule-based explanation methods use visualisation to facilitate understanding of model behaviour. For instance, RuleMatrix extracts decision rules and organises them into a structured matrix format, enabling interactive exploration of feature relationships and their influence on rule outcomes [MQB18]. Similarly, iForest supports interactive analysis of random forest models by visualizing decision paths and predictions [ZWLC18]. Explainable Matrix represents extracted rules using a matrix-based visualisation framework [NP21], and VisRuler allows for decision rules to be explored interactively by inspecting the model using various views [CMK23].

Although many visualisation and explanation methods have been proposed, few focus on comparing different methods. Visagreement [SGB*25] provides an interactive visual interface to evaluate how explanations disagree with each other, but it focuses on local feature attribution methods alone. Other systems, such as ExplainExplore [CvW20] and GALE [XCD*22], also allow users to explore and compare explanations but again concentrate on local attribution-based exploration.

Current research in human-centred XAI and visual analytics highlights the need not only for accurate explanations but also for their presentation in forms that support human understanding, reasoning, exploration, and decision-making. Previous work has shown that incorporating domain knowledge into model development and explanation generation helps improve human alignment [CKR*25]. Building on this perspective, our work introduces a hybrid workflow and interface that integrate rule-based representations, feature attribution methods, and visual analytics to obtain useful insights for feedback-driven refinement of data and models.

3. Workflow

The proposed workflow is shown in Fig. 1. It is designed to support: (i) comparing multiple explanation methods, (ii) identifying regions of agreement and contradiction, and (iii) relating these patterns to domain knowledge to guide refinement. The workflow consists of six steps summarised below:

Step 1: Train an ML model with a rule representation. The analysis starts by training a predictive model whose decision logic

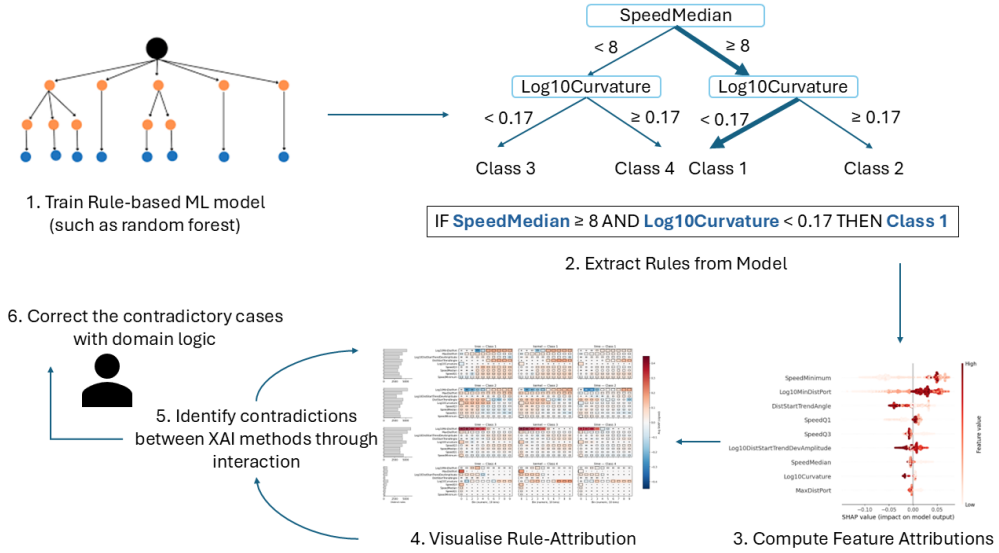

Figure 1: Workflow for exploration and cross-checking of different XAI methods.

can be expressed as a set of rules. This may be a model that is inherently tree- or rule-based (e.g., decision trees, random forests), or a surrogate model mimicking the behaviour of a more complex model. Here we focus on a random forest classifier, a widely used algorithm known to perform well in practice. Random forests are also particularly difficult to understand due to their ensemble nature, which makes them a suitable test case for our workflow, while their tree-based structure facilitates rule extraction and enables the use of TreeSHAP as a model-specific attribution method.

Step 2: Extract rules from the model: Once the model has been trained, rules are extracted by traversing every branch of each tree and recording the sequence of feature conditions and associated class predictions at each leaf. This produces a large initial set of rules that collectively describe the model's decision logic. Rules are pruned to remove redundancies and contradictions and aggregated into a concise representation of model logic.

Step 3: Compute feature attributions: In parallel to the rule extraction, feature attribution values are also computed. The interface allows for multiple methods to be used simultaneously. Although any attribution method can be used in principle, our case study involves TreeSHAP [LEC*20], which computes the exact SHAP values by exploiting the tree structure of the model, KernelSHAP [LL17], which works by estimating SHAP values model-agnostically, and LIME [RSG16], which approximates local model behaviour by fitting a linear surrogate to perturbed samples in the neighbourhood of each instance. Each of these methods produces a signed attribution score for every feature, which indicates the direction and magnitude of that feature's contribution to a given class prediction.

Step 4: Integrate and visualise rules and attributions: The rules and feature attributions are brought together in the rule–attribution visualisation. This hybrid interactive view allows users to compare attribution-based explanations with both the rule-based explanations and with other attribution-based methods. Key components of this visualisation, shown in Fig. 1, include the y-axis of each subplot, which lists the different features, and the x-axis, where feature values are discretised into a user-defined number of bins (10 in our case). This enables users to see how the impact of a feature on the class prediction changes as the feature value changes.

The colour scale of the heatmaps represents the weighted attribution scores. A diverging colour scale is used, where red indicates a positive contribution (the feature value in that bin pushes the model towards predicting that class) and blue indicates a negative contribution (the feature value pushes the model away from that class); darker shades indicate stronger effects. The size of the squares inside each bin represents the rule count extracted from the rule-based classifier. Larger squares indicate that more rules fall into that feature bin and thus that this feature value is more frequently used in rules predicting that class. Hence, where there is alignment between the rule-based and attribution-based methods, larger squares tend to correspond to strongly positive attributions (dark red bins). The grey bars on the side represent the total number of distinct rules present for a given feature and class. Up to three attribution techniques of the user's choice can be visualised side by side for direct comparison.

Step 5: Identify contradictions through interaction: Where rules and feature attributions align, the large squares (rules) coincide with dark red bins, indicating that many rules and high positive attributions support the same feature ranges and classes. Where the explanations conflict, this pattern breaks down, for example when dark red bins contain only small squares, or when large squares appear in blue (negatively attributed) bins. The user can toggle between different explanation methods as well as different bin divisions to systematically explore and locate such contradictions.

Step 6: Resolve contradictions with domain knowledge and generalise: The contradictions identified in the previous step can then be investigated using domain knowledge to judge which explanations are plausible and, where necessary, make decisions to

refine data, rules, or the choice of explainer. From these investigations, users can derive generalisations that flag the conditions under which the model explanations work well and the conditions under which they are unreliable, providing guidance for future analysis and model refinement.

4. Recognition of Vessel Movement Patterns

As a case study, we focus on identifying movement patterns of fishing vessels. The model takes as input numerical attributes that characterise segments of vessel trajectories. The training data used for model development were prepared as described in [AAA*24]. The model is designed to classify vessel behaviour into four movement categories: Forward movement (class 1), Trawling (class 2), Port entry/exit (class 3), and Anchoring (class 4). It relies on nine features describing vessel speed characteristics, trajectory shape, and proximity to ports. Features with highly skewed distributions in the training data were transformed using a logarithmic scale.

Speed statistics: `SpeedMinimum`, `SpeedQ1`, `SpeedMedian`, `SpeedQ3`. These features capture the distribution of vessel speed, ranging from the minimum observed speed to the quartile-based spread.

`Log10Curvature`: The logarithm of the curvature of the time series representing the vessel's distance from its starting point. Curvature is calculated as the ratio of the sum of absolute consecutive changes in the time series to the amplitude of the values. Values close to 1 indicate nearly straight motion, whereas higher values reflect more frequent turning.

`DistStartTrendAngle`: The angle of the linear trend fitted to the time series of the vessel's distance from its starting location. Larger angles indicate stronger movement away from the starting point, while smaller angles suggest slower movement or a return toward the origin.

`Log10DistStartTrendDevAmplitude`: The logarithm of the amplitude of deviations from the trend line of the distance from the starting point. This measure captures trajectory tortuosity, indicating the extent to which the vessel's path deviates from a straight-line trend. Higher values correspond to more irregular or zigzag motion.

Port proximity: `MaxDistPort`, `Log10MinDistPort`. These features represent the maximum and log-transformed minimum distances to the nearest port and are useful for identifying anchoring behaviour and port-related manoeuvrers.

The features `Log10Curvature` and `Log10DistStartTrendDevAmplitude` together characterise the complexity of a trajectory segment. While `Log10Curvature` reflects global movement patterns, trajectory tortuosity captures local variations such as zigzagging. Combined, these measures help differentiate between steady outward motion, looping behaviour, and localised manoeuvring.

The overall accuracy of the trained random forest model was 0.97. The random forest model used consists of 100 decision trees where each was trained on a bootstrapped sample of the training data. Once the rules had been extracted and cleaned, there was a total of 9515 rules that were used for further analysis.

4.1. Interactive Exploration

Three attribution methods; TreeSHAP, KernelSHAP, and LIME were used to evaluate the explanations along with the explanations from the extracted rules (Fig. 2). Overall, the methods show good alignment across most features and classes, particularly where feature attributions are stronger in magnitude. In these regions, all techniques assign predominantly positive attributions to similar features, and the rule counts (square sizes) show corresponding increases, indicating that both the attribution-based and rule-based methods identify the same decision boundaries. This convergence is most evident for the speed features in all classes. For example, to predict Class 1 (Forward Movement), higher values of `SpeedMinimum`, `SpeedQ1`, `SpeedMedian`, and `SpeedQ3` are consistently flagged as positive predictors in all methods, with larger squares and red shading appearing together in the higher feature bins. This gives a high degree of confidence that vessel speed is a robust and reliable indicator of forward movement, and one that aligns naturally with domain expectations.

However, there are also contradictions, particularly where the feature attribution magnitudes are smaller. For example, TreeSHAP shows strong negative attributions (blue shading) for certain features while LIME assigns positive or neutral values to those same features. This discrepancy is particularly evident in the lower feature rows where the attribution strengths are smaller. An example of this contradiction occurs in the `Log10Curvature` feature for Class 1. TreeSHAP suggests that low values of this feature negatively influence the Forward Movement prediction, while KernelSHAP, LIME and the rule-based extractor indicate the opposite showing that low curvature is a positive indicator of forward movement. The interactive visualisation makes this contradiction immediately apparent, allowing the user to cross-reference it with domain knowledge. In practice, a vessel moving forward in a straight path would be expected to have low curvature, so the majority of explanations are more consistent with physical reality. This suggests that in this instance, TreeSHAP's globally computed SHAP values may be capturing a complex interaction effect that does not reflect the intuitive interpretation of the feature, highlighting the risk of relying on any single explanation method without cross-validation.

These contradictions likely stem from fundamental differences in methodology: TreeSHAP computes exact SHAP values based on the tree structure itself, while LIME approximates the model locally using linear surrogates. When the decision boundary is highly non-linear or involves complex feature interactions (as suggested by varying rule counts), LIME's local linear approximation may diverge significantly from TreeSHAP's global game-theoretic approach. The varying square sizes also suggest that in regions with many rules (larger squares), the methods have lower agreement, indicating that model complexity in certain feature spaces leads to explanation instability.

Based on the observations made by comparing the explanation methods with domain knowledge, some generalisations on where to trust the explainers and where not to can be made. For example, where the feature attributions are stronger in magnitude, the methods seem to agree more. This suggests that when a feature has a

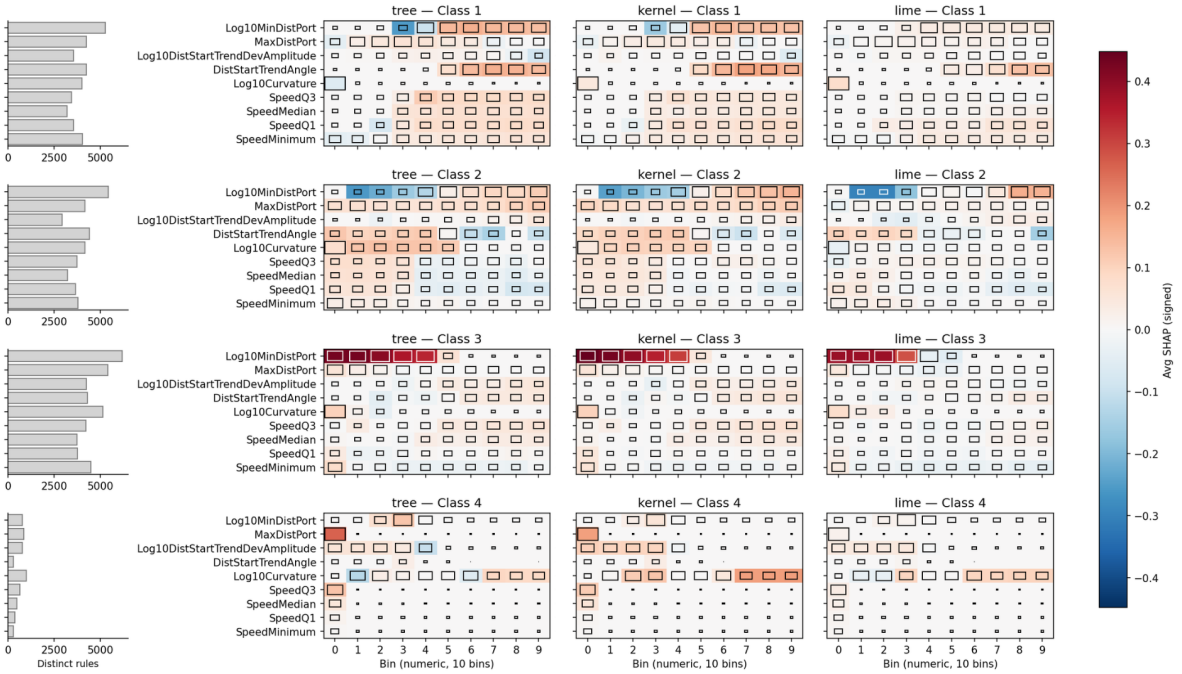

Figure 2: Hybrid rule–attribution explanation view. Each subplot shows one attribution method (TreeSHAP, KernelSHAP, LIME) as a heatmap over discretised feature bins (y-axis: features, x-axis: feature bins; colour: attribution value). Square size encodes the number of rules in each bin, revealing where rule-based and attribution-based explanations agree or diverge; the number of bins can be interactively adjusted.

dominant and clear cut impact on the prediction, the methods are more trustworthy.

4.2. Feedback-driven refinement

The exploration in Section 4.1 provides a basis for refining both the model and its explanations. The visualisation of rules and attributions flagged explanations inconsistent with domain understanding of vessel behaviour. These flagged regions become feedback items that can be acted upon in several ways:

Data- and label-level refinement: Experts can inspect trajectory segments associated with suspicious bins and correct or improve labels.

Rule-level refinement: Where contradictions persist but domain logic is clear, experts can propose constraints to guide post-hoc rule pruning/merging or provide soft constraints for model training.

Explainer-level calibration: If a specific explainer systematically disagrees with both domain knowledge and other methods, its outputs can be down-weighted or excluded from the hybrid view for those conditions; conversely, experts may rely more strongly on particular methods in regions where alignment is consistently high.

The visual analytics component supports this feedback loop by allowing users to compare before and after states of the model and explanations and, using the Rule–attribution view, to check whether contradictions have been mitigated or new inconsistencies have emerged.

5. Conclusions

In this paper, we argue that relying on a single XAI approach is insufficient for trustworthy, human-aligned AI systems. To address this, we introduce a hybrid approach that combines rule-based explanations with feature attribution methods, presented through an interactive visual analytics interface for unified comparison and evaluation. A case study on vessel movement pattern recognition illustrated that where feature influences were strong, explanation methods converged reliably. Conversely, in cases of subtle or non-linear interactions, methods diverged, increasing the risk of misinterpreting model behaviour. The interactive visualisation approach provides several practical benefits:

- Agreements and discrepancies between explanation methods are shown simultaneously, reducing the need for manual analysis.
- Users can drill down into specific features and bins and cross-reference against domain knowledge.
- Contradictions that would likely go unnoticed in isolation are made visible and become meaningful points for investigation.

This underscores the importance of integrating interactive visual analytics into explanation validation processes. Future work will expand the approach to additional domains and develop methods to automatically flag contradictions for expert review, enhancing the feedback loop between human insight and visualisation.

References

[AAA*24] ANDRIENKO N., ANDRIENKO G., ARTIKIS A., MANTENOGLOU P., RINZIVILLO S.: Human-in-the-loop: visual analytics for building models recognising behavioural patterns in time series. *IEEE Computer Graphics and Applications* (2024). doi:10.1109/MCG.2024.3379851. 4

[AAAW22] ANDRIENKO N., ANDRIENKO G., ADILOVA L., WROBEL S.: Visual analytics for human-centered machine learning. *IEEE Computer Graphics and Applications 42*, 1 (2022), 123–133. doi:10.1109/MCG.2021.3130314. 2

[BGG*23] BODRIA F., GIANNOTTI F., GUIDOTTI R., NARETTO F., PEDRESCHI D., RINZIVILLO S.: Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery 37*, 5 (2023), 1719–1778. doi:10.1007/s10618-023-00933-9. 2

[CKK24] CHATZIMPARMPAS A., KUCHER K., KERREN A.: Visualization for trust in machine learning revisited: The state of the field in 2023. *IEEE Computer Graphics and Applications 44*, 3 (2024), 99–113. doi:10.1109/MCG.2024.3360881. 1

[CKR*25] CAPPUCCIO E., KATHIRGAMANATHAN B., RINZIVILLO S., ANDRIENKO G., ANDRIENKO N.: Integrating human knowledge for explainable ai. *Machine Learning 114*, 11 (2025), 250. doi:10.1007/s10994-025-06879-x. 2

[CMK23] CHATZIMPARMPAS A., MARTINS R. M., KERREN A.: Visruler: Visual analytics for extracting decision rules from bagged and boosted decision trees. *Information Visualization 22*, 2 (2023), 115–139. doi:https://doi.org/10.1177/14738716221142005. 2

[CvW20] COLLARIS D., VAN WIJK J. J.: Explainexplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)* (2020), pp. 26–35. doi:10.1109/PacificVis48177.2020.7090. 2

[GR19] GUIDOTTI R., RUGGIERI S.: On the stability of interpretable models. In *2019 international joint conference on neural networks (IJCNN)* (2019), IEEE, pp. 1–8. doi:10.48550/arXiv.1810.09352. 2

[HMKL23] HOFFMAN R. R., MUELLER S. T., KLEIN G., LITMAN J.: 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. 2

[LEC*20] LUNDBERG S. M., ERION G., CHEN H., DEGRAVE A., PRUTKIN J. M., NAIR B., KATZ R., HIMMELFARB J., BANSAL N., LEE S.-I.: From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence 2*, 1 (2020), 56–67. doi:10.1038/s42256-019-0138-9. 3

[LGM20] LIAO Q. V., GRUEN D. M., MILLER S.: Questioning the AI: informing design practices for explainable AI user experiences. *CoRR abs/2001.02478* (2020). URL: http://arxiv.org/abs/2001.02478, arXiv:2001.02478. 1

[LL17] LUNDBERG S. M., LEE S.-I.: A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., p. 4768–4777. doi:10.5555/3295222.3295230. 1, 2, 3

[LWH*22] LEE S., WANG X., HAN S., YI X., XIE X., CHA M.: Self-explaining deep models with logic rule reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2022), NIPS '22, Curran Associates Inc. doi:10.5555/3600270.3600502. 2

[LZED25] LAGO M. A., ZAMZMI G., EICH B., DELFINO J. G.: Evaluating explainability: A framework for systematic assessment and reporting of explainable ai features. *arXiv preprint arXiv:2506.13917* (2025). doi:10.3390/bioengineering13010111. 2

[MCLM24] MAÇÃS C., CAMPOS J. R., LOURENÇO N., MACHADO P.: Visualisation of random forest classification. *Information Visualization 23*, 4 (2024), 312–327. doi:https://doi.org/10.1177/14738716241260745. 2

[MFP*23] MORANDINI S., FRABONI F., PUZZO G., GIUSINO D., VOLPI L., BRENDEL H., BALATTI E., DE ANGELIS M., DE CESAREI A., PIETRANTONI L., ET AL.: Examining the nexus between explainability of AI systems and user's trust: A preliminary scoping review. In *CEUR Workshop Proceedings* (2023), vol. 3554, CEUR-WS, pp. 30–35. 2

[MJEP*21] MARCÍLIO-JR W. E., ELER D. M., PAULOVICH F. V., RODRIGUES-JR J. F., ARTERO A. O.: Explorertree: a focus+ context exploration approach for 2d embeddings. *Big Data Research 25* (2021), 100239. doi:10.1016/j.bdr.2021.100239. 2

[MQB18] MING Y., QU H., BERTINI E.: Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 342–352. doi:10.1109/TVCG.2018.2864812. 2

[NP21] NETO M. P., PAULOVICH F. V.: Explainable Matrix - Visualization for Global and Local Interpretability of Random Forest Classification Ensembles . *IEEE Transactions on Visualization & Computer Graphics 27*, 02 (Feb. 2021), 1427–1437. doi:10.1109/TVCG.2020.3030354. 2

[PSGH*21] POURSABZI-SANGDEH F., GOLDSTEIN D. G., HOFMAN J. M., WORTMAN VAUGHAN J. W., WALLACH H.: Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2021), CHI '21, Association for Computing Machinery. URL: https://doi.org/10.1145/3411764.3445315, doi:10.1145/3411764.3445315. 1

[RSG16] RIBEIRO M. T., SINGH S., GUESTRIN C.: "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* (New York, NY, USA, 2016), KDD '16, Association for Computing Machinery, p. 1135–1144. doi:10.1145/2939672.2939778. 1, 2, 3

[SGB*25] SILVA P., GUARDIEIRO V., BARR B., SILVA C., NONATO L. G.: Visagreement: Visualizing and exploring explanations (dis)agreement. *IEEE Transactions on Visualization and Computer Graphics 31*, 10 (2025), 7862–7875. doi:10.1109/TVCG.2025.3558074. 2

[XCD*22] XENOPOULOS P., CHAN G., DORAISWAMY H., NONATO L. G., BARR B., SILVA C.: Gale: Globally assessing local explanations. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022* (25 Feb–22 Jul 2022), Cloninger A., Doster T., Emerson T., Kaul M., Ktena I., Kvinge H., Miolane N., Rieck B., Tymochko S., Wolf G., (Eds.), vol. 196 of *Proceedings of Machine Learning Research*, PMLR, pp. 322–331. URL: https://proceedings.mlr.press/v196/xenopoulos22a.html. 2

[ZWLC18] ZHAO X., WU Y., LEE D. L., CUI W.: iforest: Interpreting random forests via visual analytics. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 407–416. doi:10.1109/TVCG.2018.2864475. 2