Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Doctoral Consortium Track

Visual Analytics for Guiding Feature Attribution Method Selection

Priscylla Silva
University of São Paulo, Brazil
priscylla.silva@usp.br

Abstract

Feature attribution methods explain model predictions by assigning importance to input features. Choosing the most suitable method is difficult due to inconsistencies and a lack of universal solutions. Our research focuses on proposing tools and methodologies for evaluating, comparing, and selecting these methods.

1 Introduction

The use of machine learning models is growing in our society, but high-stakes domains like healthcare demand explainable AI (XAI) to justify model decisions. Feature attribution methods provide a means to explain predictions by scoring each input feature based on its influence on a specific outcome. With many feature attribution methods available, a key question arises: How do we select the most suitable one for a specific context? This selection process presents several challenges, including:

C1. Different explanation methods may generate conflicting explanations for the same model instance [Krishna *et al.*, 2024].

C2. No universal method performs optimally across all models [Han *et al.*, 2022].

C3. Evaluation metrics for assessing explanations vary in focus and can produce conflicting results [Chen *et al.*, 2024].

C4. Machine learning practitioners often select methods based on personal preferences instead of systematic approaches [Krishna *et al.*, 2024].

2 Current Research Contributions

Our work focuses on two projects related to the research challenges mentioned in the previous section. The first project studies disagreements between feature attribution methods (C1), and the second addresses C2, C3, and C4. Both projects focus on analyzing feature attribution methods in the context of tabular data.

2.1 Conflicting Explanations

Collaborating with industry practitioners and XAI researchers, we conducted meetings and interviews to discuss and explore several scenarios where disagreement between feature attribution methods occurred. Through these interactions, we identified three key research questions:

R1. Is there a relationship between disagreement and the quality of the explanations?

R2. Is there a relationship between disagreement and the accuracy of the model?

R3. Are there specific features responsible for causing disagreement?

To investigate these research questions, we developed Visagreement [Silva *et al.*, 2025], a visualization-assisted methodology designed to help researchers explore the problem of disagreement between feature attribution methods. Visagreement helps researchers generate hypotheses and insights into disagreement patterns.

Visagreement consists of a visualization tool and an integrated Python library. The library was designed to streamline the data processing step and the execution of explanation methods. Once the data is processed, users can utilize the tool to perform their analyses. Visagreement allows users to apply their own models and datasets to conduct customized investigations.

The primary contributions of this work are as follows:

- We proposed the concept of (dis)agreement space and introduced an interactive visualization that enables users to easily identify and analyze instances where the methods most agree and disagree.
- We developed an open-source tool to support researchers in investigating hypotheses related to the disagreement problem.

To evaluate Visagreement, we conducted three case studies with seven datasets (one case study for each research question) and one user study.

Case Study 1. We investigated whether disagreement correlates with explanation quality. We used two metrics to assess quality but found no correlation between the metrics and the level of agreement or disagreement among the methods.

Case Study 2. This case study explored the relationship between disagreement and model performance. Through the

visualizations available in the tool, we observed two phenomena across various models and datasets: *i)* Instances where the methods disagreed most often corresponded to instances where the model performed significantly worse than those where the methods agreed the most; *ii)* Furthermore, we observed that disagreement among attribution methods was disproportionately higher on negative-class instances (label 0).

Case Study 3. In this case study, we analyzed whether specific features or subsets of features were more influential in causing disagreements. Our analyses, however, did not identify any consistent pattern that would suggest such a relationship exists.

In keeping with the goal of Visagreement — to support hypothesis generation — we used the findings from case study 2 to investigate the hypothesis that there is a relationship between model performance and the level of disagreement between feature attribution methods. An initial study was conducted with two datasets from the education domain, and the results revealed a strong correlation between model performance and disagreement levels [Silva *et al.*, 2024]. This investigation demonstrates how Visagreement effectively facilitates hypothesis-driven research.

Ongoing projects and future plans. We are studying more models and datasets for robust results. Our tools focus on tabular data but can adapt to images and text in future work.

2.2 Choosing Feature Attribution Methods

Many works in the literature propose metrics to evaluate and compare feature attribution methods. However, there is a gap in selecting the most appropriate method (C4). Choosing the right method is crucial because no single method consistently performs optimally across all models (C2), and the diversity of metrics and measured properties can lead to confusion when making a choice (C3).

To address this gap, we conducted a literature review to analyze existing approaches that support the selection, comparison, and evaluation of feature attribution methods. Our primary focus was on tools and frameworks, both technological and conceptual. The analysis revealed a variety of works for individual evaluation or pairwise comparison between methods. However, we identified a notable gap: none of the reviewed works provided comprehensive support for all three functionalities—evaluation, comparison, and selection —simultaneously.

Based on these findings, we developed a visual analytics Python library that assists data analysts in comparing, selecting, and visualizing different model explainability methods. This library enhances the decision-making process by integrating intuitive visual representations with interactive charts, following a human-AI collaboration approach. The library was evaluated through a user study with 10 participants, who reported that the visualizations provided sufficient grounds for selecting a method or a smaller subset of methods to present to a domain expert for the final decision. We wrote a paper with these results that is currently under review.

Inspired by the results from the literature review and the feedback from the library evaluation, we identified the need for a structured guide to assist in choosing the most suitable feature attribution method. Therefore, we proposed a conceptual framework designed to guide the decision-making process. The framework consists of three layers:

- Model-centric layer: Ensures explanations faithfully reflect the model's behavior (e.g., fidelity metrics).
- Explanation-centric layer: Assesses explanation quality (e.g., robustness, interpretability).
- Human-centric layer: Evaluates usefulness for end-users (e.g., trust, clarity via user studies).

We proposed a bottom-up approach to the framework to ensure that the chosen method generates explanations that are faithful (model-centric), robust (explanation-centric), and useful (human-centric) in that order. We organized the literature review and the framework into a survey paper, which is also currently under review.

Ongoing projects and future plans. Based on the proposed framework, we are developing a comprehensive toolkit that integrates a complete pipeline for selecting the most suitable explanation method. This toolkit will support the entire process, from applying quantitative metrics to conducting user studies with domain experts. Two key challenges in this development are reducing the search space, as the numerous existing methods and metrics make exhaustive evaluation impractical, and resolving conflicts between metrics within the same framework layer to ensure consistent method selection.

Acknowledgments

This work is sponsored by Fapesp #22/03941-0 and #23/05783-5.

References

[Chen *et al.*, 2024] Zixi Chen, Varshini Subhash, Marton Havasi, Weiwei Pan, and Finale Doshi-Velez. What makes a good explanation?: A harmonized view of properties of explanations, 2024.

[Han *et al.*, 2022] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. In *Advances in Neural Information Processing Systems*, volume 35, pages 5256–5268, New Orleans, LA, USA, 2022. Curran Associates, Inc.

[Krishna *et al.*, 2024] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024.

[Silva *et al.*, 2024] Priscylla Silva, Claudio Silva, and Luis Gustavo Nonato. Exploring the relationship between feature attribution methods and model performance. In *Proceedings of the 2024 AAAI Conference on Artificial Intelligence*, volume 257 of *Proceedings of Machine Learning Research*, pages 29–38. PMLR, 26–27 Feb 2024.

[Silva *et al.*, 2025] Priscylla Silva, Vitoria Guardieiro, Brian Barr, Claudio Silva, and Luis Gustavo Nonato. Visagreement: Visualizing and exploring explanations (dis)agreement. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–14, 2025.