Towards Ensemble Explanation Strategies and Solving the Disagreement Problem in XAI

Craig Pirie[1]

[1]*Robert Gordon University (RGU), Aberdeen, U.K.*

Abstract

Explainable Artificial Intelligence (XAI) aims to enhance trust, transparency, and accountability in AI systems, but disagreement between explanation methods remains a critical challenge. This thesis tackles the disagreement problem faced by ensemble explanation strategies to provide coherent, robust, and faithful explanations. For feature attribution explainers (RQ1), we propose AGREE, a robustness-based framework that aggregates explanations by weighting them according to robustness. To extend this approach to image data, we propose a region-based method for enhancing robustness and interpretability using human-interpretable regions and showcase its ability in a neonatal pain classification task. Such approaches were found to improve the interpretability and consistency of explanations. For counterfactual explanations (RQ2), we develop GeneticMCF and EnsembleMCF to minimise disagreement while balancing proximity and sparsity. Ongoing work involves using NSGA-II to strike a better balance between agreement and diversity. Finally, for RQ3, we propose using Case-Based Reasoning (CBR) to integrate feature attribution and counterfactual explanations. This work advances XAI by offering systematic methods to reconcile disparate explanations, improving robustness, interpretability, and coherence.

Keywords

Explainer Disagreement, Ensemble XAI, Evaluation of XAI, Fairness & Interpretability

1. Introduction

1.1. Explainable Artificial Intelligence (XAI) Paradigms

Transparent Artificial Intelligence (AI) has gained increasing importance due to concerns over accountability, trust, and compliance with regulations such as the General Data Protection Regulation (GDPR), which grants individuals the ‘right to an explanation’ [1]. Explainable AI (XAI) aims to address these concerns by enhancing the interpretability of complex AI decision-making processes, thereby fostering trust and enabling effective oversight [2].

Common XAI paradigms include feature attribution, semi-factual, and counterfactual explanations. Feature attribution methods identify influential features contributing to model predictions and include SHAP [3], Layerwise Relevance Propagation (LRP) [4], Integrated Gradients [5], Guided Backpropagation (GBP) [6], and Vanilla Gradients [7]. These methods fall into perturbation-based, gradient-based, and decomposition-based approaches. Perturbation-based methods like Kernel SHAP approximate Shapley values through local perturbations, while gradient-based methods (e.g., Integrated Gradients) leverage gradients to determine importance. Decomposition-based methods like LRP redistribute the model’s output backward, preserving relevance through the network’s structure.

Semi-factual explanations explore scenarios where substantial changes to input features yield the same prediction, helping to examine a model’s decision boundaries [8]. In contrast, counterfactual explanations describe minimal modifications that alter the outcome, closely aligning with human ‘what-if’ reasoning [9]. These explanations are particularly useful for generating actionable recommendations in domains like healthcare and finance. WachterCF [10], one of the earliest methods, uses an optimisation

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

*Corresponding author.

c.pirie11@rgu.ac.uk (C. Pirie)
https://rgu-repository.worktribe.com/person/1065030/craig-pirie (C. Pirie)
0000-0002-6799-0497 (C. Pirie)

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

algorithm to minimise distance between the original instance and the counterfactual, focusing on proximity and validity without considering diversity.

Subsequent methods have expanded this foundation by introducing additional objectives to enhance counterfactual quality. DiCE [11] employs an optimisation-based approach to generate diverse counterfactuals that maximise proximity and validity, enabling users to explore various scenarios. NICE [12] focuses on identifying nearest unlike neighbours to maintain high validity and coverage near decision boundaries. DisCERN [13] integrates neighbourhood-based counterfactual generation with feature attribution, producing ordered, actionable recommendations by prioritising features based on their importance.

While an abundance of XAI methods have been proposed in the literature, a standardised approach for evaluating their quality remains lacking. The Co-12 framework attempts to unify these evaluation efforts by defining desirable properties that explanations should satisfy. The following section discusses existing evaluation metrics and how they relate to the Co-12 framework.

1.2. Desirable XAI Properties and Their Related Evaluation Metrics

The Co-12 framework [14] consolidates various desirable explanation properties such as Correctness (whether the explanation accurately reflects the inner workings of the model), Consistency (how deterministic the explanation is), and Completeness (how well the explanation describes the model). Sensitivity and Infidelity, proposed by [15], are two metrics commonly used to evaluate these properties. Sensitivity measures the robustness of explanations under slight variations, while Infidelity assesses how accurately the explanation represents the model's decision-making process.

Proximity, Sparsity, and Diversity are all important properties when evaluating counterfactual explanations. Proximity measures the similarity between the query and the generated counterfactual, making the proposed changes more actionable and realistic, which aligns with the Co-12 property of Connectedness. Sparsity measures the number of features that need to be modified, contributing to the Co-12 property of Compactness by favouring simpler and more interpretable explanations. Diversity, on the other hand, refers to generating multiple valid and distinct counterfactuals to provide users with a range of options for recourse, allowing them to select the most suitable pathway according to their needs and constraints. This aligns with the Co-12 property of Context, as it broadens the scope of interpretability by presenting a variety of feasible actions rather than a single solution.

Among these properties, Coherence is particularly relevant to this work, as it concerns the alignment of explanations from different sources. Agreement, a key aspect of coherence, measures how well different explainers align in their feature attributions or recommended changes. However, when coherence is compromised due to conflicting insights from various explainers, the 'Disagreement Problem' emerges.

1.3. The Disagreement Problem in XAI

The 'Disagreement Problem' [16] occurs when different explainers produce conflicting insights about feature importance, their relative rankings, or the recommended changes needed to alter outcomes. For instance, factual explainers like SHAP and LRP may provide contradictory explanations about which features are most influential, while counterfactual explainers may suggest different, sometimes opposing, actions to achieve the same desired outcome. Figure 1 illustrates the extent of disagreement among multiple feature attribution explainers the breast cancer dataset. Furthermore, while diversity in counterfactual explanations is often desirable for providing multiple pathways to achieve a desired outcome, it can be exploited for 'fairwashing' by selectively presenting favourable explanations that conceal biased decision-making [17].

This inconsistency undermines trust in AI systems, making it challenging to reach a consensus on which insights are reliable. Therefore, it is essential to develop methods that can unify explainer outputs and mitigate disagreements. In this paper, we address the Disagreement Problem by investigating how to combine different feature attribution and counterfactual explainers to produce consistent, accurate,

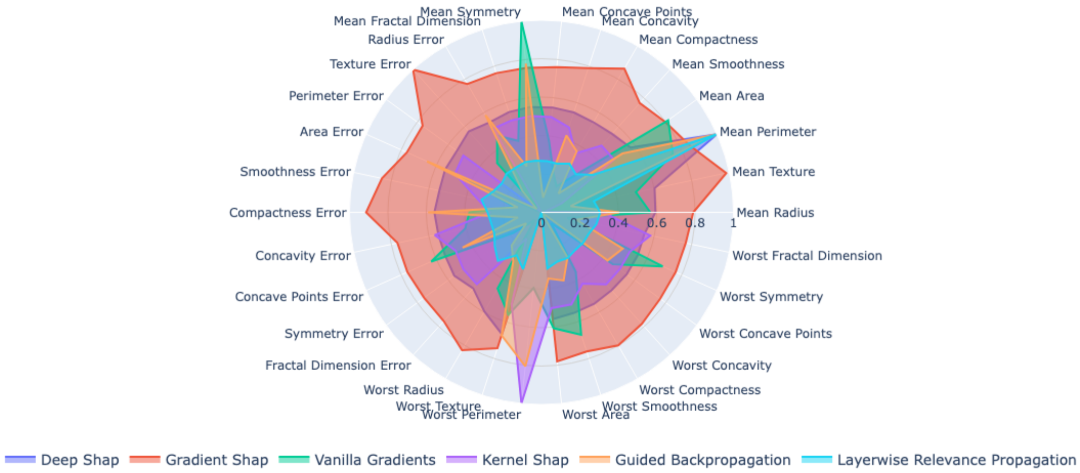

Figure 1: Example of the disagreement between an ensemble of 6 feature attribution explainers regarding the most important features for a breast cancer diagnosis.

and trustworthy explanations. Our approach focuses on evaluating explainer contributions using metrics such as robustness for factual explainers, and proximity, sparsity, and validity for counterfactual explainers. By aggregating explanations in a principled manner, we aim to provide a coherent and reliable interpretation of model decisions.

2. Related Work

The disagreement problem in XAI has recently gained attention as various explainers often produce conflicting insights. Krishna et al. [16] were the first to empirically study this problem, revealing that machine-learning practitioners struggled to resolve discrepancies and often resorted to subjective preferences such as choosing the most recent or preferred explainer. They also introduced methods for quantifying disagreement by comparing top-k feature sets across explainers.

Building on these metrics, Roy et al. [18] proposed filtering explanations by reporting only the top-k features with shared attribution signs to reduce disagreement. In contrast, Schwarzschild et al. [19] addressed the problem preemptively by introducing a “consensus loss” during model training, aimed at promoting coherence among downstream explainers. Laberge et al. [20] suggested that disagreement arises from differing assumptions about feature interactions, proposing regional explanations based on an ‘FD-tree’ to partition the feature space and enhance consensus. Ensemble techniques have also been applied to XAI to enhance explanation quality. For example, [21] combine SHapley Additive exPlanations (SHAP) [3] and Grad-CAM++ [22] through a stacking approach, resulting in explanations that are more consistent and better aligned with user expectations.

While these approaches address the disagreement problem primarily for feature attribution explainers, Brughmans et al. [17] focus on counterfactual explanations, conducting a large-scale empirical study involving 40 datasets, 12 explanation-generating methods, and two black-box models, resulting in over 192,000 explanations. Their findings highlight significant disagreement between methods, which can be exploited for fairwashing by selecting explanations that obscure biased decisions. The study suggests that disagreement is predominantly influenced by dataset characteristics and the choice of counterfactual algorithm.

Although existing approaches have made progress in mitigating the disagreement problem, they primarily focus on aligning explanations without evaluating their quality within the ensemble. Our work addresses this gap by assessing explanation quality using metrics such as robustness, proximity, sparsity, and validity to reconcile disagreements. However, this will require improving existing metrics like sensitivity, which often rely on noise-based samples that can be out-of-distribution and skew robustness measurements. To address this, we employ a nearest-neighbour-based perturbation method to generate more realistic samples, ensuring a fairer and more accurate evaluation of explanation robustness. Additionally, we propose that case-based reasoning [23] and optimisation algorithms, such as genetic algorithms [24] and NSGA-II [25], are well-suited to this problem. Case-based reasoning provides a structured way of comparing new explanations to previously encountered cases, enabling the identification of explanations that are most consistent with past successful cases. Meanwhile, optimisation algorithms such as NSGA-II allow us to balance competing objectives like diversity, agreement, and proximity, providing a systematic way to aggregate and reconcile conflicting explanations.

3. Research Questions

We are interested in alleviating the disagreement problem for both feature-attribution and counterfactual explainers. This is important because resolving disagreement between these methods can provide end-users with clearer, more consistent explanations, ultimately enhancing their trust and understanding of AI systems. Accordingly, we pose the following research questions (RQs):

RQ1 (*Answered*): Can the link between robustness and disagreement be exploited by means of aggregation to increase the alignment between, and faithfulness of, feature attribution explanations in XAI?

RQ2 (*Partially Answered*): What measures can be implemented to find an optimal agreement-diversity trade-off and minimise the disagreement between counterfactual XAI explainers, so that fairwashing is mitigated and end-users can build stronger mental models of automated decision-making systems?

RQ3 (*Pending Investigation*): How can the disagreement between feature attribution and counterfactual explanations in XAI be quantified and reconciled to leverage their complementary strengths for improved interpretability?

RQ1 has been thoroughly addressed, and the findings are expected to be disseminated soon. Regarding RQ2, significant progress has been made by framing the problem as an optimisation task. While initial findings using a single-objective genetic algorithm were promising, a multi-objective approach has now been implemented in an attempt to balance competing metrics more effectively. Evaluation of this approach is planned for the near future. RQ3 remains unexplored at this stage.

4. Methods

4.1. AGREE — Robust Ensemble Feature Attribution Explanations (RQ1)

To address feature attribution disagreement (RQ1), we introduce AGREE (Aggregation for Robust Explanation Experiences). Our previous work [26] proposed AGREE by aggregating explanations based on agreement using case alignment, a new metric where agreement is calculated by comparing explanations within neighbourhoods of similar instances. This approach aimed to capture agreement by aligning explanations that produced similar feature attributions for neighbouring cases. While this approach effectively captured agreement between local explanations, it only indirectly captured robustness, limiting its ability to weight explanations according to quality, making it less reliable.

To overcome this, we now aggregate explanations using a robustness metric that relies on generating perturbations along a specific path between the instance being explained and its Most Distant Neighbour (MDN, see [8]) — the instance of the same class closest to the Nearest Unlike Neighbour (NUN). This robustness assessment involves comparing similarity curves between the original instance and its perturbations against their corresponding explanations, using two metrics: *Robustness Area* for area difference and *Fréchet Similarity* for curve trajectory similarity. The scores are then combined and normalised to weight explainers by their robustness, ensuring that more robust explanations contribute more significantly to the aggregated explanation, thereby improving reliability and consensus.

To extend this work to image data, we build on Laberge et al.'s findings that region-based explanations reduce disagreement by simplifying feature interactions and improving alignment [20]. We applied this approach using manually segmented images to analyse agreement at the region level, comparing human-interpretable concepts rather than fine-grained pixels where disagreement is less meaningful. This step, incorporated as a precursor to AGREE, aims to enhance interpretability, faithfulness, and robustness of image-based ensemble explanations.

4.2. Minimising Disagreement with Meta-counterfactual Explanations (RQ2)

For RQ2, we propose 'meta-counterfactuals' (MCF) to address disagreement between counterfactual explainers using optimisation techniques like Genetic Algorithms (GeneticMCF). Disagreement is quantified by differences in the size and direction of actions recommended with respect to the query, where size refers to the magnitude of change applied to each feature relative to the query (e.g., a salary increase of $10,000) and direction refers to whether the change should increase, decrease, or remain unchanged. This metric is integrated into the optimiser to minimise disagreement within diverse explanations or between baseline explainers, reducing fairwashing' and enhancing user interpretability.

Additionally, we introduce EnsembleMCF, which aggregates multiple baseline counterfactuals by ranking them based on proximity, sparsity, and disagreement. This consensus-based approach improves coherence and interpretability without compromising essential counterfactual properties. While EnsembleMCF establishes a coherent MCF, GeneticMCF offers further improvement by directly minimising disagreement over a broader search space.

4.3. A Case-based Approach to Reduce Disagreement Between Feature Attribution and Counterfactual Explainers (RQ3)

We propose to leverage Case-Based Reasoning (CBR) to reduce the disagreement between counterfactual and feature attribution explainers (RQ3) by identifying feature attributions that are more likely to support a given counterfactual explanation. A critical first step involves developing a suitable method for measuring agreement between the two explanation types. The underlying idea is that similar past cases demonstrating agreement between feature attribution and counterfactual explanations can guide the formation of a likely consensus ensemble. A case in this context would consist of an instance, its associated counterfactual explanation, the corresponding model prediction, and the feature attribution explanations generated for that instance. While the feature attribution explanation will still need to be generated on the model to ensure faithfulness, the use of CBR aims to reduce the overhead associated with running and comparing every possible feature attribution method. Instead, by comparing against past cases, we can prioritise those explainers most likely to produce compatible explanations, thereby enhancing efficiency and promoting coherence.

5. Preliminary Findings

As discussed in Section 3, we have made substantial progress towards addressing RQs 1 and 2. While work on the multi-objective approach is still ongoing, our preliminary results from the single-objective approach, which remain unpublished, provide a partial answer to RQ2. A comprehensive overview of the preliminary findings from each RQ follows:

5.1. Feature Attribution (RQ1)

5.1.1. Experiment 1: Perturbation Comparison

We evaluated AGREE's robustness metric by comparing our nearest-neighbour-based perturbations (anchored by the MDN) against noise-based methods commonly used in sensitivity analysis. Using a randomly selected subset of 50 instances, we generated 100 perturbations per instance and assessed their realism via the Wasserstein distance between the original and perturbed distributions. Results showed that our approach produced more realistic samples, which should provide a fairer evaluation and more accurate robustness estimation.

5.1.2. Experiment 2: Faithfulness Evaluation

We assessed the faithfulness of original and aggregated explanations using infidelity as the primary metric. Several aggregation strategies were compared: selecting the most robust explanation (max), mean, median, and our weighted average approach incorporating robustness weights. Weighted average and mean consistently performed best, with similar performance when robustness scores were equal, suggesting the choice between the two is trivial in such cases.

5.1.3. Experiment 3: Real-World Application

We applied AGREE to a real-world oil and gas dataset, evaluating both infidelity and disagreement metrics. Results indicated that reducing disagreement can better help detect false alerts in anomaly detection. The weighted average approach performed well, demonstrating the practicality and robustness of AGREE in real-world applications. Overall, these findings highlight AGREE's effectiveness in improving explanation faithfulness and reducing disagreement, particularly in real-world use cases.

5.1.4. Experiment 4: Region-Based Explanation Comparison

We explored the potential of region-based explanations for neonatal pain assessment, where disagreement is measured at the region level rather than pixel level [27]. By comparing explainers (Grad-CAM, Integrated Gradients) against human attention heatmaps using predefined region masks, we found that region-based explanations provide higher agreement and are more interpretable compared to pixel-based methods, particularly for pain-related features (nose, mouth, between eyebrows). While this experiment was not integrated with AGREE, it suggests that region-level aggregation could enhance agreement and interpretability if incorporated in the future.

5.2. Counterfactual Methods (RQ2)

5.2.1. Experiment 1: Baseline Comparison

We evaluated meta-counterfactuals generated by EnsembleMCF and GeneticMCF against four baseline explainers (DiCE, NICE, DisCERN, and WachterCF) across five health and finance datasets. Metrics used include disagreement, proximity, and sparsity. EnsembleMCF achieved superior agreement without compromising proximity or sparsity, while GeneticMCF improved proximity and sparsity but struggled with disagreement. Results are based on 20 repeated trials with a 70-30 train-test split and pairwise t-tests.

5.2.2. Experiment 2: Qualitative Analysis

We conducted a qualitative analysis to assess the plausibility, interpretability, and coherence of example meta-counterfactuals. EnsembleMCF consistently provided coherent and interpretable explanations, whereas GeneticMCF improved certain cases but often faced challenges due to the diversity-agreement trade-off. The findings suggest that multi-objective approaches, such as NSGA-II [25], should be further investigated to balance the competing objectives of agreement, proximity, and sparsity.

6. Conclusion and Next Steps

This work addresses the disagreement problem in XAI by developing ensemble explanation strategies aimed at enhancing coherence, robustness, and faithfulness. For RQ1, we proposed AGREE, a robustness-based aggregation framework that quantifies explanation stability via perturbations to produce consensus explanations. This approach effectively improves faithfulness and alignment across multiple feature attribution explainers, particularly when applied to region-based explanations. For RQ2, we introduced meta-counterfactuals, developing GeneticMCF and EnsembleMCF to aggregate counterfactual explanations, highlighting the importance of balancing diversity and agreement. Finally, for RQ3, we proposed integrating feature attribution and counterfactual explanations using Case-Based Reasoning to produce coherent multi-faceted explanations. The immediate next steps for this research are:

- **Multi-objective Meta-Counterfactual Generation (RQ2):** Evaluate the effectiveness of the multi-objective approach using an NSGA-II optimiser, incorporating a plausibility term to enhance robustness. Ablation studies will be conducted to fine-tune parameters and assess the trade-off between diversity and agreement.

- **Addressing Disagreement Between Feature Attribution and Counterfactual Explainers (RQ3):** Implement and assess the proposed CBR-guided system for integrating feature attribution and counterfactual explanations. Evaluation will include comparing disagreement levels with and without CBR guidance, as well as user studies to assess coherence and practical utility.

This work contributes to the field of XAI by providing a systematic framework for aggregating explanations from different sources, improving robustness and coherence while mitigating the disagreement problem. The introduction of robustness-weighted aggregation, meta-counterfactual generation, and the proposed integration of feature attribution and counterfactual explanations form a comprehensive approach to enhancing interpretability and reliability in XAI.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] L. Edwards, M. Veale, Enslaving the algorithm: From a “right to an explanation” to a “right to better decisions”?, IEEE Security & Privacy 16 (2018) 46–54.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information fusion 58 (2020) 82–115.

[3] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[5] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[6] J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: ICLR (workshop track), 2015.

[7] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, in: Proceedings of the International Conference on Learning Representations (ICLR), ICLR, 2014.

[8] S. Aryal, M. T. Keane, Even if explanations: prior work, desiderata & benchmarks for semi-factual XAI, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023, pp. 6526–6535.

[9] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI), in: Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28, Springer, 2020, pp. 163–178.

[10] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[11] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 607–617.

[12] D. Brughmans, P. Leyman, D. Martens, NICE: an algorithm for nearest instance counterfactual explanations, Data Mining and Knowledge Discovery (2023) 1–39.

[13] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, D. Corsar, Discern: Discovering counterfactual explanations using relevance features from neighbourhoods, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2021, pp. 1466–1473.

[14] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Computing Surveys 55 (2023) 1–42.

[15] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, On the (in) fidelity and sensitivity of explanations, Advances in neural information processing systems 32 (2019).

[16] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, H. Lakkaraju, The disagreement problem in explainable machine learning: A practitioner's perspective, Transactions on Machine Learning Research (2024). URL: https://openreview.net/forum?id=jESY2WTZCe.

[17] D. Brughmans, L. Melis, D. Martens, Disagreement amongst counterfactual explanations: how transparency can be misleading, Top (2024) 1–34.

[18] S. Roy, G. Laberge, B. Roy, F. Khomh, A. Nikanjam, S. Mondal, Why don't XAI techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions, in: 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, 2022, pp. 444–448.

[19] A. Schwarzschild, M. Cembalest, K. Rao, K. Hines, J. Dickerson, Reckoning with the disagreement problem: Explanation consensus as a training objective, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023, pp. 662–678.

[20] G. Laberge, Y. B. Pequignot, M. Marchand, F. Khomh, Tackling the XAI disagreement problem with regional explanations, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 2017–2025.

[21] L. Zou, H. L. Goh, C. J. Y. Liew, J. L. Quah, G. T. Gu, J. J. Chew, M. P. Kumar, C. G. L. Ang, A. W. A. Ta, Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and covid-19 respiratory infections, IEEE Transactions on Artificial Intelligence 4 (2023) 242–254. doi:10.1109/TAI.2022.3153754.

[22] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, 2018, pp. 839–847.

[23] J. M. Schoenborn, R. O. Weber, D. W. Aha, J. Cassens, K.-D. Althoff, Explainable case-based reasoning: a survey, in: AAAI-21 workshop proceedings, 2021.

[24] S. Katoch, S. S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future, Multimedia tools and applications 80 (2021) 8091–8126.

[25] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE transactions on evolutionary computation 6 (2002) 182–197.

[26] C. Pirie, N. Wiratunga, A. Wijekoon, C. F. Moreno-Garcia, Agree: A feature attribution aggregation framework to address explainer disagreements with alignment metrics (2023).

[27] C. Pirie, L. A. Ferreira, G. de Almeida Sá Coutrin, L. P. Carlini, C. F. Moreno-Garcia, M. C. de Moraes Barros, R. Guinsburg, C. E. Thomaz, R. N. Orsi, N. Wiratunga, Understanding disagreement between humans and machines in XAI: Robustness, fidelity, and region-based explanations in automatic neonatal pain assessment, in: Proceedings of The 3rd World Conference on eXplainable Artificial Intelligence, Springer, 2025. Accepted, camera-ready copy in preparation.