AGREE: A Feature Attribution Aggregation Framework to Address Explainer Disagreements with Alignment Metrics

Craig Pirie[1,*], Nirmalie Wiratunga[1], Anjana Wijekoon[1] and Carlos Francisco Moreno-Garcia[1]

[1]*Robert Gordon University, Garthdee Road, Aberdeen, AB10 7GJ*

Abstract

As deep learning models become increasingly complex, practitioners are relying more on post hoc explanation methods to understand the decisions of black-box learners. However, there is growing concern about the reliability of feature attribution explanations, which are key to explaining machine learning models. Studies have shown that some explainable artificial intelligence (XAI) methods are highly sensitive to noise and that explanations can vary significantly between techniques. As a result, practitioners often employ multiple methods to reach a consensus on the reliability of their models, which can lead to disagreements among explainers. Although some literature has formalised and reviewed this problem, few solutions have been proposed. In this paper, we propose a novel case-based approach to evaluating disagreement among explainers and advance AGREE — an explainer aggregation approach to resolving the disagreement problem based on explanation weights. Our approach addresses the problem of both local and global explainer disagreement by utilising information from the neighbourhood spaces of feature attribution vectors. We evaluate our approach against simpler feature overlap metrics by weighting the latent space of a k-NN predictor using consensus feature importance and observing the performance degradation. For local explanations in particular, our method captures a more precise estimate of disagreement than the baseline methods and is robust against high dimensionality. This can lead to increased trust in ML models, which is essential for their successful adoption in real-world applications.

Keywords

XAI, Case Alignment, AGREE, Disagreement Problem, Feature Attribution

1. Introduction

In the preceding decade, machine learning systems have undergone significant advancements in their efficacy, albeit their adoption has been impeded by the challenging aspect of their interpretability. As such, Explainable AI (XAI) is fast becoming a prerequisite for the deployment of intelligent systems — with some countries in Europe now enforcing this by law [1]. This has

ICCBR XCBR'23: Workshop on Case-Based Reasoning for the Explanation of Intelligent Systems at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

*Corresponding author.

c.pirie11@rgu.ac.uk (C. Pirie); n.wiratunga@rgu.ac.uk (N. Wiratunga); a.wijekoon1@rgu.ac.uk (A. Wijekoon); c.moreno-garcia@rgu.ac.uk (C. F. Moreno-Garcia)

0000-0002-6799-0497 (C. Pirie); 0000-0003-4040-2496 (N. Wiratunga); 0000-0003-3848-3100 (A. Wijekoon); 0000-0001-7218-9023 (C. F. Moreno-Garcia)

spurred a surge in research dedicated to improving the transparency and accountability of AI models. Inevitably, this has spawned a number of approaches for understanding the rationale of machine learning systems such as counter-factuals, feature attribution, and natural language explanations (a review of which can be found in [2, 3, 4] respectively).

Attribution explainers are one of the popular forms of factual explanation methods used in XAI [5, 6]. These explainers provide an understanding of how a model arrived at its predictions, by identifying the most influential features (attributions) or variables that led to a particular model outcome. One example of the use of attribution explainers is in the context of a loan application [7]. Here, LIME attributions are used to explain why a particular applicant's loan was approved or rejected by highlighting the relevant factors that influenced the decision. Alternatively for image data salience maps are often used to convey the areas that conveyed most to the outcome [8, 9] (see Figure 1 below for an example).

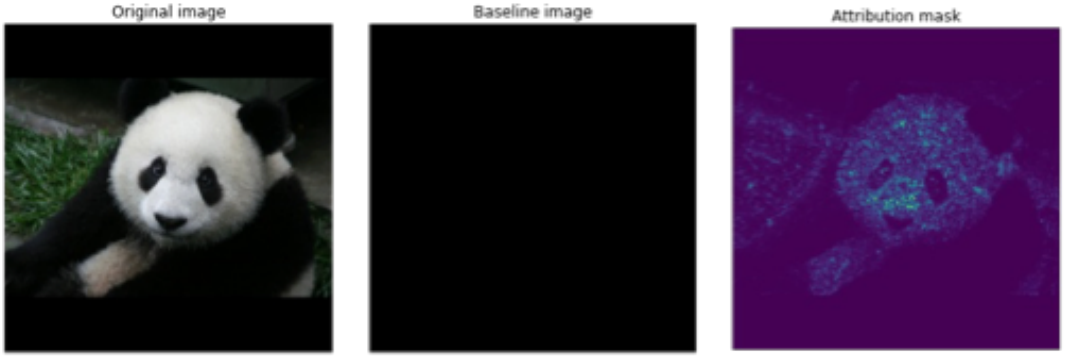

Figure 1: An example of Integrated Gradients used to explain the image classification of panda. The highlighted pixels on the right indicate the areas of the image that the model found most important according to Integrated Gradients.

Factual explainers play a crucial role in gaining the trust of humans by providing transparent and interpretable explanations for machine learning predictions [3, 6, 9, 10, 11, 12]. One of the main challenges with factual explainers is that different methods often generate different types of explanations, which can lead to discrepancies in the results [13]. For example, popular explainers such as LIME [14], SHAP [15], and Integrated Gradients (IG) [16] can all produce different feature attributions for the same model prediction. The discrepancies between different factual explainers (see example in Figure 2) can result in mistrust not only in the machine learning prediction itself but also in the explanations provided. When the explanations provided by different methods do not align, it can create confusion and skepticism among those trying to understand the model's decision-making process. This can be especially problematic when it comes to high-stakes decisions, such as in healthcare or finance, where the consequences of an incorrect prediction can be significant.

Addressing the challenge of disagreement among attribution explainers necessitates the development of an effective aggregation strategy that combines factual explanations from multiple explainer methods. While consensus voting or ranking offers some utility, it is insufficient in capturing the complex relationships between the alternative feature attributions. As such, research in Case-Based Reasoning (CBR) and Case Alignment emerges as a promising avenue

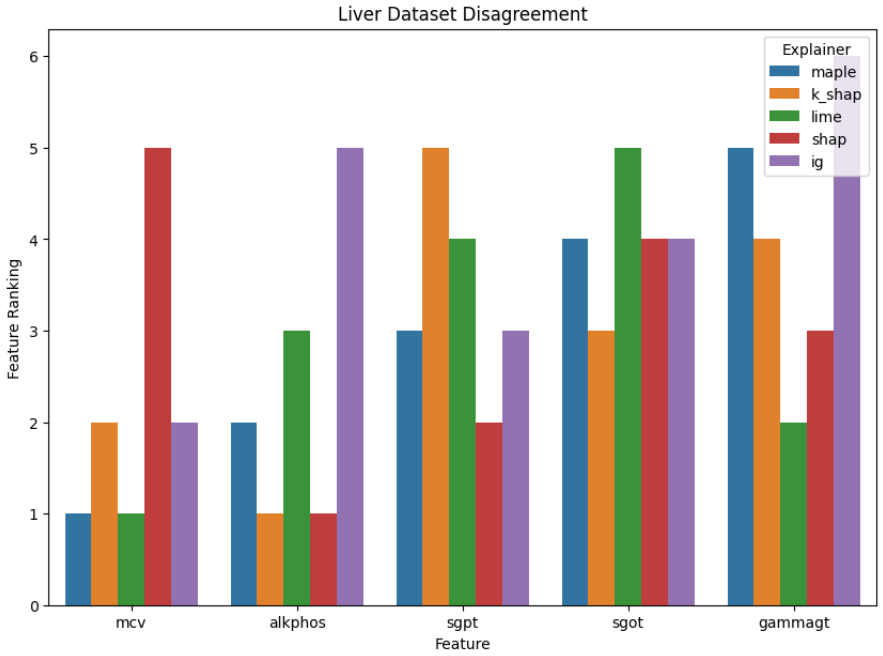

Figure 2: An example of the disagreement between feature attribution methods for explaining a neural networks prediction of liver disorders. The model was trained on the BUPA Liver Disorders dataset [17] which is available in the UCI repository.

.

for uncovering these relationships, providing a neighborhood concept that can be defined in the context of factual explainers. This paper's primary research question centers on exploring the ranking behaviour of factual explainers with regard to both local and global explanations then harnessing that to measure their relationships to identify areas of consensus. We explore how to use the alignment of neighbourhood knowledge as a means to accurately capture these relationships. Furthermore, we assign increased confidence to explanations that exhibit a higher degree of alignment with alternative feature attributions.

Accordingly, the key contributions of the paper are:

1. **Case Alignment Confidence**: A novel metric for measuring the overall agreement between an explainer against alternative explanation methods by leveraging information from local neighbourhood spaces.
2. **AGREE** — AGgregation for Robust Explanations: A framework for combining the explanations of different feature attribution explainers by exploiting alignment knowledge.

An outline of the paper is as follows: section 2 discusses a review of the related literature; sections 3 and 4 introduce the aggregation strategy using rank average and case alignment confidence strategies respectively; 6 discusses the methodology for evaluating our alignment and aggregation approach; the results are presented and discussed in section 7; and finally our research is concluded and a discussion of future work is given in 8.

2. Related Work

2.1. Inherently Interpretable Models and Explainable AI (XAI)

Inherently interpretable AI models provide clear and understandable explanations for their decisions without relying on complex feature attribution methods. These models are transparent and easy to interpret because they employ simple algorithms such as decision trees [18], linear models [19, 20], or rule-based systems [21] that allow humans to understand how they arrive at their outputs. However, there is often a trade-off between interpretability and prediction performance. Neural networks for example often learn better-performing models but are regarded as black-box learners as it is difficult to gain insight to understand their behaviour. In contrast, post-hoc explanation methods operate on opaque and complex models that are difficult to understand. Methods for post-hoc explanation differ between their access to the model (i.e. black box or access to internals), approximation of scope (i.e. global or local), search technique (i.e. perturbation or gradient) and presentation of explanation (i.e. feature-based or counterfactual) [13]. For example, perturbation methods such as LIME [14], SHAP [15], Anchors [22] and RISE [23] evaluate learners by modifying the input of a model, whether this is pixels in an image, words in a phrase, or similar elements in other data types, and observing the changes in the prediction [12]. A larger difference in the output would indicate that the perturbed feature is more important. Alternatively, gradient-based local explanations like GradCAM [24], Smoothgrad [25], Integrated Gradients [16], and Layerwise Relevance Propagation (LRP) [26] rely on the gradient between the output probabilities and the features from the input or embedding layer [27]. The prediction is used to backpropagate through the network to the input or embedding layer to estimate the feature attributions [28]. Local and global methods are distinguished by the granularity of their explanations. While global explanations summarise the behaviour of the entire model, local explanations attempt to explain on an instance-wise basis. Our work is interested in exploiting the information from the various types of explanations to obtain a more robust evaluation of black-box models.

2.2. Evaluation of Explanations and Explainer Disagreement

Typically, XAI evaluation methods involve some form of user study, which is sensible as explanations are generally user-centric. However, the approach is subjective and can be costly to undertake. Objectively evaluating explanation methods remains an active research area but various attempts have been made to quantify the effectiveness of explanations in terms of different qualities such as fidelity, interpretability, sparsity, proximity, and robustness [6, 29, 30, 31]. Empirical studies have shown that post hoc explanations can be inconsistent, unfaithful or unstable and prone to "fairwashing" [13]. Some machine learning practitioners utilise multiple different post hoc explanation methods to understand their models. Albeit, the instability of attributions poses a significant challenge — how to reach a consensus when explainers disagree. In [13], Krishna *et al.* conduct a user study to understand how machine learning practitioners resolve the problem. Astonishingly, they found that 86% of subjects either side-stepped disagreement by choosing arbitrary heuristics such as choosing their favourite method, or simply did not know how to resolve the dispute. Previous studies have proposed methods to measure disagreement in feature attribution methods [13, 32] and counterfactuals [33] by

evaluating the intersection of top-K feature vectors across two explainers. Variations of these approaches make use of auxiliary information such as sign (whether the feature had a positive or negative impact on the output) and rank (ordinal position of the feature in the vector). However, little work has been done to settle the disagreements in an intuitive manner. The closest work to ours is the study conducted by Roy et al. in [32]. Their method studies the aggregated set:

$$A = i \in S : sign(E_a, i) = sign(E_b, i) \tag{1}$$

where i is a feature in the set S of top-K most important features, E_a is the first explainer (LIME in their case) and E_b is another explainer (i.e. SHAP in their example). If both explainers can agree on the sign of the feature it is coloured green, else it is coloured red. This is a step towards explanation aggregation and does reduce the cognitive burden on the end-user when interpreting disagreement. Still, it falls short of providing a method for settling disputes. We propose an alignment-based approach to solving explainer disputes inspired by the case alignment [34] metric. Case alignment tests the assumption that *similar problems have similar solutions* in case-based reasoning applications. We posit that by forming case bases around each explainer, local neighbourhood information can be leveraged to better inform alignment measures across multiple explainers. Feature attribution vectors can then be weighted by the alignment scores to generate an aggregate explanation to present to the end user.

3. Explainer Attribution Aggregation by Rank Average

When given an instance and a prediction from a black-box model, a set of explanation attribution scores $S = [s_{ij}] \in \mathbb{R}^{n \times m}$ are obtained from n explainers (denoted as E_x) for m features. To remove the effects of differences in scale or magnitude that may exist between the attribution scores generated by the explainers, the scores from each explainer are converted to ranks, denoted by $R_i = [r_{ij}]$, where:

$$r_{ij} = \text{rank}(s_{ij}) = |\{k : s_{ik} > s_{ij}\}| + 1, j \in [1..m] \tag{2}$$

The ranking function, rank(), is applied to each element s_{ij}. It sorts and assigns ranks based on the sorted order of the attribution scores of the explainer, and handles tied scores randomly. The r_{ij} notation denotes the attribution rank by the i-th explainer for the j-th feature. In this manner, any query or case in the case base can be represented using explainer attribution ranks.

The simplest way to combine explainer attributions is to average the feature ranks. The resulting feature weights are the average row vector, obtained as follows:

$$\bar{w} = \frac{1}{n} \sum_{i=1}^{n} R_x \tag{3}$$

Here an average rank of each feature refers to the aggregated consensus explainer attribution for that feature, after taking into account the explanations provided by multiple explainers. The resulting average rank vector, $\bar{w}$, can be used as a set of feature weights for further analysis or as a baseline aggregation method.

4. Explainer Attribution Aggregation by Confidence Weighted Rank Average

Despite the advantages of rank-based aggregations, such as their simplicity and robustness against outliers, these methods inherently lack the ability to discern intricate patterns and may overlook crucial alignment relationships. This shortcoming stems from the fact that all explainers are treated as equally important in the process of generating feature weights, which may result in the loss of valuable information. However, if we were to weight feature attribution ranks based on the level of confidence of each explainer, the resulting combined explanation not only can take advantage of the strengths of multiple methods but also mitigates varying levels of performance and reliability depending on the type of black box model and the specific dataset.

4.1. Rank Overlap Alignment

For a given instance, attributions from multiple explainers can be evaluated by examining the extent of agreement with respect to the overlap in their top k features. The greater the paired overlap an explainer exhibits in relation to the others, the more confidence can be assigned to it. For a pair of explainers, E_a and E_b given all their rank assignments, R_a and R_b, a symmetrical alignment score can be derived as follows

$$\text{AlignOverlap}(E_a, E_b, k) = |\text{topk\_features}(R_a, R_b, k)| \tag{4}$$

$$\text{topk\_features}(R_a, R_b, k) = \{f \in [1, m] \mid r_{if} \leq k \text{ and } r_{jf} \leq k\}$$

Here f is a feature index and the top k feature overlap between any pair of explainers i and j is identified by considering the set of features f such that feature f is among the top k features for both explainers i and j. We iterate over all pairs of explainers i and j with $i < j$ to find the top k overlap for each pair. Note r_{ij} refers to the rank assigned by the i-th explainer to the j-th feature.

4.2. Neighbourhood Alignment

Characterising neighbourhoods in explainer attribution spaces to capture alignment provides more fine-grained information in contrast to only comparing relative ranked positions. In case alignment [34], the alignment of problem and solution spaces is compared based on the distance between the Query case, Q, and each neighbour case, C_i, in each space. The idea is that the spaces are aligned when the distances between cases are similar in both. To measure the agreement between a pair of explainers (say E_a and E_b), using this concept, we must develop a mapping method that permits each explainer to determine the representation of the query (and cases in the case bases) in two distinct representation spaces, much like problem and solution spaces used in case alignment. Thereafter as in Figure 3 neighbourhoods from two distinct spaces can be used to assess paired explainer alignment.

In order to create a paired representation for Q, the ranked representation from two explainers (such as E_a and E_b) are concatenated to form a row vector as follows: $P_{ab} = \left[R_a, R_b\right]$,

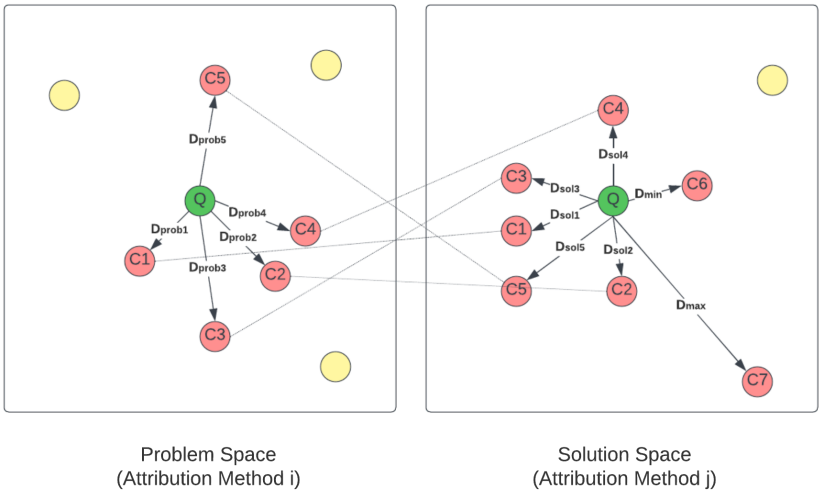

Figure 3: Visual representation of how case alignment compares the two neighbourhood spaces in our domain. The red circles represent neighbour cases, the green circles represent the query case and the yellow circles represent other distant cases in the case base that are not considered in the alignment calculation.

where an explainer pair such as E_a, E_b can be drawn from the set of attribution explainers, $\{x_1, x_2, \ldots, x_n\}$. Using the paired representation of Q, and cases, C_i, in the case base, we can use the local neighborhood alignment as a metric to assess the level of agreement between the two attribution explainer methods, E_a and E_b (as in Figure 3). Each part of the representation can be designated as the problem space (Explainer A space) or the solution space (Explainer B space). In Figure 3 below, $k = 5$ nearest neighbours, are analysed in each explainer space. An asymmetrical alignment score can be formulated for an explainer E_a, given another explainer E_b, by taking into account neighbourhood alignments as follows:

$$\text{CaseAlign}(E_a, E_b, Q) = \frac{\sum_{i=1}^{k}(1 - D^{E_a}(Q, C_i)) \cdot align(Q, C_i)}{\sum_{i=1}^{k}(1 - D^{E_a}(Q, C_i))} \tag{5}$$

The D^{E_a} notation denotes a distance computation w.r.t. to the neighbours represented according to the explainer E_a space. The nearest and farthest cases are also identified for normalisation purposes. The neighbourhood distances on the explainer E_a space are weighted by a normalised align distance, which is computed w.r.t. to explainer E_b space, as follows:

$$\text{align}(Q, C_i) = 1 - \frac{D^{E_b}(Q, C_i) - D^{E_b}_{\text{min}}}{D^{E_b}_{\text{max}} - D^{E_b}_{\text{min}}} \tag{6}$$

The degree of agreement between explainers E_a and E_b is directly proportional to the alignment score whereby the stronger the agreement the closer the alignment score is to a value of 1. The Case Alignment process is described in Figure 4.

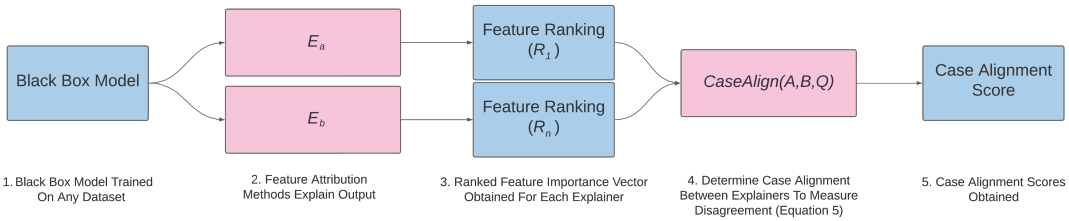

Figure 4: A visual representation of how case alignment is determined.

5. AGREE: AGgregation for Robust Explanations

To assess the alignments between a set of n explainers, we calculate all pairwise alignment scores and then aggregate them to determine a confidence level for each explainer. This confidence level can assist in arriving at a consensus on feature attribution weights (see Figure 5 for a visual aid).

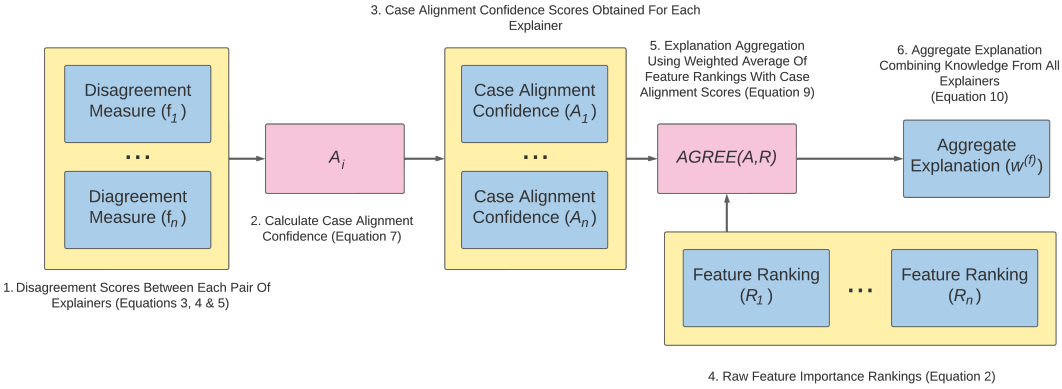

Figure 5: A visualisation of the explanation aggregation strategy.

Let $M \in \mathbb{R}^{n \times n}$ be a matrix representing the pairwise alignment relationships between a set of explainers, for a query case. An explainer confidence vector, A, is derived from M, denoting the confidence of each explainer, x_i, where for each $i \in \{1, 2, \ldots, n\}$, the value of A_i is calculated as follows:

$$A_i = \begin{cases} \frac{1}{n} \sum_{j=1}^{n} m(i,j) & \text{if } M \text{ is symmetric,} \\ \frac{1}{2n} \sum_{j=1}^{n} [m(i,j) + m(j,i)] & \text{otherwise.} \end{cases} \tag{7}$$

Here $m(i,j)$ an element of M is the pairwise alignment score between two explainer attributions for the query, i.e. $m(i,j) = f(x_i, x_j; \theta) \quad \forall i, j \in \{1, 2, \ldots, n\}$, where f is any of the explainer agreement functions. As Case Align is asymmetric, both $m(i,j)$ and $m(j,i)$ are utilised to obtain a symmetric value for A_i. Whereas since the feature overlap methods are symmetric, only $m(i,j)$ needs considered in Equation 7.

Next, we use the explainer confidence vector, A, to influence the level of importance to be assigned to each explainer's recommended feature attribution ranks, to arrive at a consensus feature attribution weight vector for the data instance as follows:

$$\bar{w}_i^{(f)} = \frac{\sum_{k=1}^{n} (A_k \cdot R_i)}{\sum_{k=1}^{n} A_k} \tag{8}$$

$$\text{AGREE}(A, R) = \bar{w}^{(f)} = \frac{1}{n} \sum_{i=1}^{n} \bar{w}_i^{(f)} \tag{9}$$

The notation $\bar{w}_i^{(f)}$ refers to the weighted attributions obtained using the alignment function f as the basis for the confidence scores in the vector A for explainer i. Note R_i is a row vector for the i-th explainer, and A_k is the confidence score of the k-th explainer.

Given a set of data instances, N, a global feature weight vector, W can be computed by averaging over all of the local weight vectors and can be used to explain a model on the global level:

$$\bar{W}^{(f)} = \frac{1}{N} \sum_{i=1}^{N} \bar{w}^{(f)} \tag{10}$$

6. Evaluation

Our evaluation strategy assumes that a feature attribution method's ability to accurately capture the significance of features within a domain can aid in model learning by providing useful feature-importance information. Therefore, by weighting the feature space of a k-NN by the agreed importance of each feature we can observe the effect on prediction performance. A stable or increased score indicates good agreement, while a dip in k-NN indicates poor agreement.

6.1. Experimental Setup

6.1.1. Datasets and AI Model

First, a set of black-box models is trained on 8 different datasets from the UCI repository[1] [35] which cover a wide range of tasks (e.g. regression and classification), domains, and data types (such as tabular and text). We chose to use different variations of a neural network for each model, as gradient-based post-hoc explanation methods are only applicable to differentiable models. A summary of the trained models can be found in Table 1 below.

6.1.2. Explainers

Five established feature attribution methods were used in the experiments to obtain a base set of explanations for the black-box models:

[1]The project code is available online at https://github.com/craigybaeb/disagreement_problem.git.

Dataset	No. Features	No. Instances	AI Task	AI Model	Accuracy/MSE
Abalone	8	4177	Regression	AutoKeras	4.6
Auto MPG	8	398	Regression	AutoKeras	12.1
IMDB	500	2500	Text Classification	MLP	80.0
Spam	500	3902	Text Classification	MLP	87.5
Cleveland	14	303	Binary Classification	AutoKeras	86.7
Liver	7	345	Binary Classification	AutoKeras	1.0
Glass	10	214	Multi-Class Classification	AutoKeras	1.0
Wine	13	178	Multi-Class Classifaction	AutoKeras	94.4

Table 1
Summary of the datasets and algorithms used for training the black-box models.

LIME [14] is a model-agnostic feature importance explanation method that implements a surrogate model around a data instance to estimate how each feature contributed to the black-box model output. LIME creates a set of perturbations within the instance's neighbourhood and annotates them using the black-box model. This newly labeled dataset is used to create a linear interpretable model (e.g. a weighted linear regression model). The resulting surrogate model is interpretable and only locally faithful to the black-box model (i.e. correctly classifies the input instance, but not all data instances outside its immediate neighbourhood). The new interpretable model is used to classify the data instance and an explanation of the predicted class is formed by obtaining the weights that indicate how each feature contributed to the outcome.

SHAP [15] is a model-agnostic feature relevance explainer with theoretical guarantees about consistency and local accuracy from game theory and is based on the Shapley regression values [36]. Shapley values are calculated by creating linear models using subsets of features present in a case base, X. More specifically, a model is trained with a subset of features of size, m', and another model is trained with a subset of features of size, $m' + \hat{m}$. Here, $m' + \hat{m} <= m$, and the second model additionally includes a set of features, $\hat{m}$, selected from the set of features that were left out in the first model. A set of such model pairs is created for all possible feature combinations. For a given data instance that needs to be explained, the prediction differences of these model pairs are averaged to find the explainable feature relevance weights. While **Kernel SHAP** is the vanilla implementation of SHAP, there are multiple alternative methods to approximate Shapley values proposed in the literature, namely **Deep SHAP**, BayesSHAP and TreeSHAP [15, 37, 38]. For example, Deep SHAP combines the intuition of SHAP with Deep LIFT [39] to exploit additional information about deep neural networks. Deep LIFT can approximate SHAP values by assuming that the input features are independent and that the deep neural network is linear. As a result, an approximation of SHAP values can be obtained faster than that of other methods for deep models.

Integrated Gradients is a gradient-based approach to finding feature attribution weights [40]. An attribution is calculated as the sum of gradients on data points occurring at sufficiently small intervals along the straight-line path from a baseline, to the query. In practice, a large number of perturbations is preferred because the summation of gradients is a

discrete approximation of continuous integration as discussed in [40]. We chose 50 perturbations and an all-zero instance as the baseline for all datasets when calculating Integrated Gradients. It is duly noted that this is not favourable in all contexts as this could lead to null attribution. However, the literature indicates that the selection of a baseline is an open research question with no ideal solution at present [41].

MAPLE [42] is an acronym for Model Agnostic SuPervised Local Explanations and combines the ideas of SILO [43] for local linear modeling and DStump [44] for feature selection. It centers around the use of a random forest which allows it to be used as an inherently interpretable predictor or as a standalone black-box explainer. SILO defines a local neighborhood by assigning a weight to each training point depending on how frequently that point exists in the same leaf node as the given point across all trees in the random forest. To obtain feature importances, it uses the same approach as DStump which works by summing the impurity reductions of each root node in a tree where a split was made, adjusting for the number of points in the node then averaging over the forest.

6.1.3. Alignment Measures

We compare our Case Alignment Score (Section 4) with a simple mean of feature rankings (**AVG**) and the 6 feature agreement methods proposed in [13]: **FA** = Feature Agreement; **SA** = Sign Agreement; **RA** = Rank Agreement; **SRA** = Signed Rank Agreement; **RC** = Rank Correlation; and **PRA** = Pairwise Rank Alignment.

The agreement of both local and global explanations is compared. We test AGREE using global explanations to measure its ability to capture the global feature importance of the model (Equation 8). Whereas we evaluate local explanations to scrutinise the aggregation strategy on a more granular level (Equation 9).

6.1.4. k-NN Variants

The explanations gathered are then used to weigh the k-NN feature space using a weighted Euclidean distance function, such that a k-NN may be represented by k-NN($\bar{w}^{(f)}$). f is either an aggregate explanation using case alignment, mean importance, or any of the feature overlap methods, an individual base explainer or simply a function returning a vector of length M where each element = 1 for an unweighted k-NN. The non-weighted k-NN is used as a baseline and k neighbours is set to 5 for all 14 experiments.

Given the weights from an aggregate explanation (either global or local) we calculate weighted Euclidean distances between two cases x and y as in Equation 11:

$$dist(x, y) = \sqrt{\sum_{i=1}^{m} \bar{w}^{(f)}(x_i - y_i)^2} \tag{11}$$

Mean Squared Error (MSE) is used to quantify the performance of the aggregate explanations for regression datasets, whereas accuracy is used to evaluate classification explanations.

Dataset	UW	CA	AVG	FA	SA	RA	SRA	RC	PRA	L	DS	KS	IG	MPL
Abalone	4.89	5.04	5.06	**4.96**	**4.96**	**4.96**	**4.96**	5.02	5.00	4.93	5.01	5.01	5.05	4.97
Auto MPG	18.69	18.94	19.04	**18.93**	**18.93**	19.00	19.00	18.91	18.93	19.94	19.06	19.24	19.14	16.89
IMDB	0.67	0.67	0.67	0.66	0.66	**0.71**	**0.71**	0.68	0.67	0.67	0.66	0.65	0.66	0.67
Spam	0.87	**0.84**	**0.84**	**0.84**	**0.84**	**0.84**	**0.84**	**0.84**	**0.84**	0.87	0.87	0.86	0.85	0.85
Cleveland	0.60	**0.57**	**0.57**	0.53	0.53	0.50	0.50	**0.57**	0.53	0.57	0.57	0.57	0.57	0.57
Liver	0.64	0.77	**0.84**	**0.84**	**0.84**	**0.84**	**0.84**	0.73	0.75	0.74	0.79	0.78	0.64	0.84
Glass	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	1.00	1.00	1.00	0.83
Wine	**0.83**	**0.83**	**0.83**	0.78	0.78	0.78	0.78	**0.83**	0.78	0.72	0.94	0.83	0.89	0.72

Table 2
Results of Global Experiments.

Dataset	UW	CA	AVG	FA	SA	RA	SRA	RC	PRA	L	DS	KS	IG	MPL
Abalone	4.89	4.98	5.00	5.05	5.05	5.10	5.10	**4.88**	5.00	4.93	5.10	5.02	5.11	5.00
Auto MPG	18.69	18.9	19.32	**16.93**	**16.93**	18.29	18.29	17.28	17.71	18.25	20.10	21.81	20.68	18.73
IMDB	0.67	**0.69**	0.66	0.66	0.67	0.68	0.68	0.63	0.66	0.60	0.70	0.62	0.67	0.69
Spam	0.87	0.86	**0.87**	0.85	0.85	0.85	0.85	0.84	0.85	0.86	0.86	0.86	0.85	0.85
Cleveland	0.60	**0.60**	0.53	**0.60**	**0.60**	0.43	0.43	0.53	**0.60**	0.57	0.57	0.63	0.53	0.53
Liver	0.64	0.84	0.77	**0.86**	**0.86**	0.84	0.84	0.68	0.83	0.73	0.82	0.83	0.73	0.84
Glass	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.98	**1.0**	1.00	1.00	1.00	1.00	0.83
Wine	**0.83**	**0.83**	0.78	0.78	0.78	**0.83**	**0.83**	0.78	0.78	0.78	0.94	0.78	0.94	0.83

Table 3
Results of Local Experiments.

7. Results

7.1. Explanation Fidelity

The results of the global explanation experiments (shown in Table 2) suggest that the AGREE method using any of the feature overlap methods performs equally as well in terms of fidelity (faithfulness of the explanation to the model) as taking a simple average of feature rankings. AGREE based on case alignment falls short by one dataset. The closeness in results could suggest that the loss of information brought on by the global aggregation increases the instability of explanations. Therefore, we look to local explanations as a better solution.

AGREE appears more promising for local explanations (Table 3), beating or matching the performance of the k-NN using average weighting in all cases bar one dataset (although this is the Glass dataset which seems too simple for evaluation). Case alignment aggregations appear to perform on par with feature overlap metrics proposed by [13], in terms of explanation fidelity. These results are promising as they suggest that overall, the intuition provided by our aggregation strategy is more intuitive than taking a simple arithmetic mean.

7.2. Robustness of Disagreement Measures

Figure 6 shows an example of agreement matrices for a local explanation in the IMDB dataset ($N = 500$ features) using feature agreement to represent the overlap metrics defined in [13] against our case align approach. We observe that as the number of features rises, alignment

measured via feature agreement tends to drop significantly, to almost zero in many cases. This is not the case for case alignment, which stays relatively stable as N increases. This behavior is intuitive, owing to the fundamental nature of both metrics. Feature agreement is sensitive to k features selected for comparison — k increases, the likelihood that these features do not agree increases. Whereas case align takes into consideration all features in a query to compare explanations and therefore remains relatively unaffected by the number of features in an instance. In general, across all the datasets, case alignment appears to be more robust against a rise in dimensionality. This indicates to us that our method may be favourable in realistic domains where the number of features is often high.

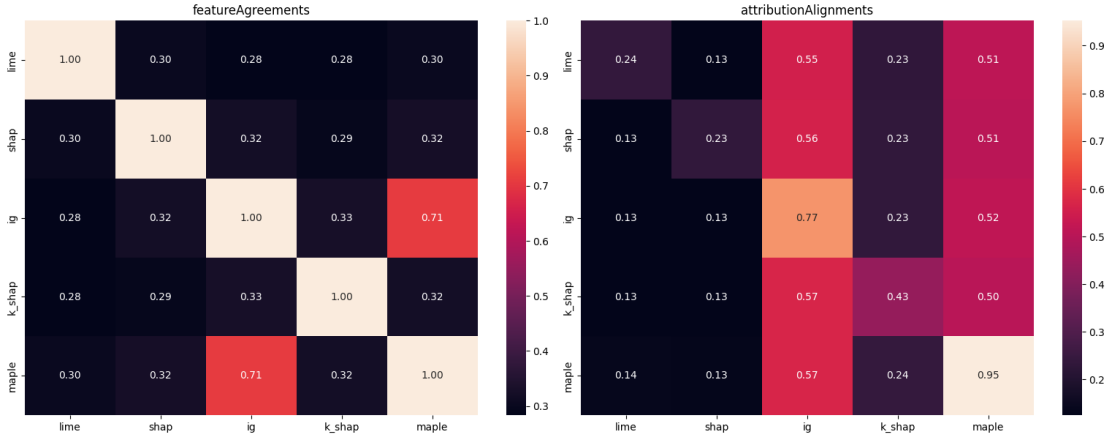

Figure 6: Disagreement matrices for an instance in the IMDB dataset (N = 500 features). The Feature Agreement matrix (left) shows the symmetric nature of the metric (note the diagonal is 1 for each pair) but the low levels of disagreement across the board. The Case Alignment matrix (right) shows the asymmetric nature of case align but indicates it can maintain a higher level of agreement in higher dimensions.

8. Conclusion & Future Work

Two key contributions were presented in this work: a novel method for evaluating disagreement between multiple explainers based on local neighbourhood alignment; and AGREE — a novel explanation aggregation method to formulate robust explanations from the knowledge of various explainer methods. We evaluate Case Alignment as a measure of disagreement, and the AGREE method by weighting a k-NN by explanations and observing the degradation or enhancement of performance. Our approach to measuring disagreement was found to be more robust than previous feature overlap methods and when settling local disagreements the measure is able to capture local information to better inform the aggregate explanation using AGREE. For global disagreements, our experiments found that AGREE does not significantly improve upon simply taking the average of feature rankings to establish an aggregate importance vector. AGREE was found to outperform (or match) simply taking the mean feature ranking across all local explanations in 87.5% of datasets. Another advantage of AGREE is that it is agnostic to both the explanation methods and disagreement measures used.

We plan to continue this study by further evaluating AGREE in additional domains and modalities (such as time-series and image data). In future work, we also propose to extend AGREE to solve disputes between counterfactuals, as they have been neglected in this paper but are a popular method for explanation. These studies will be supplemented by working with our partners in the oil and gas industry to explain anomalies in early warning time-series systems on offshore platforms. In doing so, we will conduct a more complete evaluation through further quantitative and qualitative studies while evaluating AGREE explanations on the human-interpretability level.

References

[1] M. K. Belaid, E. Hüllermeier, M. Rabus, R. Krestel, Do we need another explainable ai method? toward unifying post-hoc XAI evaluation methods into an interactive and multi-dimensional benchmark, 2022. arXiv:2207.14160.

[2] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.

[3] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[4] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on XAI and natural language explanations, Information Processing & Management 60 (2023) 103111.

[5] S. R. Islam, W. Eberle, S. K. Ghafoor, M. Ahmed, Explainable artificial intelligence approaches: A survey, arXiv preprint arXiv:2101.09429 (2021).

[6] S. Jesus, C. Belém, V. Balayan, J. a. Bento, P. Saleiro, P. Bizarro, J. a. Gama, How can i choose an explainer? an application-grounded evaluation of post-hoc explanations, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 805–815. URL: https://doi.org/10.1145/3442188.3445941. doi:10.1145/3442188.3445941.

[7] S. Upadhyay, V. Isahagian, V. Muthusamy, Y. Rizk, Extending lime for business process automation, arXiv preprint arXiv:2108.04371 (2021).

[8] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[9] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, arXiv preprint arXiv:2006.11371 (2020).

[10] E. M. Kenny, E. D. Delaney, D. Greene, M. T. Keane, Post-hoc explanation options for XAI in deep learning: The insight centre for data analytics perspective, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III, Springer, 2021, pp. 20–34.

[11] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE transactions on neural networks and learning systems 32 (2020) 4793–4813.

[12] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognition Letters 150 (2021) 228–234. URL: https://www.sciencedirect.com/science/article/pii/S0167865521002440. doi:https://doi.org/10.1016/j.patrec.2021.06.030.

[13] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, H. Lakkaraju, The disagreement problem in explainable machine learning: A practitioner's perspective, arXiv preprint arXiv:2202.01602 (2022).

[14] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, CoRR abs/1602.04938 (2016). URL: http://arxiv.org/abs/1602.04938. arXiv:1602.04938.

[15] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. URL: https://arxiv.org/abs/1705.07874. doi:10.48550/ARXIV.1705.07874.

[16] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, CoRR abs/1703.01365 (2017). URL: http://arxiv.org/abs/1703.01365. arXiv:1703.01365.

[17] Liver Disorders, UCI Machine Learning Repository, 1990. DOI: https://doi.org/10.24432/C54G67.

[18] S. R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE transactions on systems, man, and cybernetics 21 (1991) 660–674.

[19] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1996) 267–288.

[20] P. McCullagh, J. A. Nelder, Generalized linear models, volume 37 of, Monographs on statistics and applied probability (1989).

[21] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404.

[22] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

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

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[25] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint arXiv:1706.03825 (2017).

[26] 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.

[27] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, N. C. Bouaynaya, Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks, IEEE Signal Processing Magazine 39 (2022) 73–84.

[28] K. Abhishek, D. Kamath, Attribution-based XAI methods in computer vision: A review, arXiv preprint arXiv:2211.14736 (2022).

[29] 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.

[30] E. Amparore, A. Perotti, P. Bajardi, To trust or not to trust an explanation: using leaf to evaluate local linear XAI methods, PeerJ Computer Science 7 (2021) e479.

[31] P. Lopes, E. Silva, C. Braga, T. Oliveira, L. Rosado, XAI systems evaluation: A review of

human and computer-centred methods, Applied Sciences 12 (2022) 9423.
[32] 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), 2022, pp. 444–448. doi:10.1109/ICSME55016.2022.00056.
[33] D. Brughmans, L. Melis, D. Martens, Disagreement amongst counterfactual explanations: How transparency can be deceptive, arXiv preprint arXiv:2304.12667 (2023).
[34] M. Raghunandan, N. Wiratunga, S. Chakraborti, S. Massie, D. Khemani, Evaluation measures for TCBR systems, in: Advances in Case-Based Reasoning: 9th European Conference, ECCBR 2008, Trier, Germany, September 1-4, 2008. Proceedings 9, Springer, 2008, pp. 444–458.
[35] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.
[36] L. S. Shapley, A value for n-person games, in: Contributions to the Theory of Games, 1953, pp. 307–317.
[37] D. Slack, A. Hilgard, S. Singh, H. Lakkaraju, Reliable post hoc explanations: Modeling uncertainty in explainability, Advances in neural information processing systems 34 (2021) 9391–9404.
[38] S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv preprint arXiv:1802.03888 (2018).
[39] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3145–3153. URL: https://proceedings.mlr.press/v70/shrikumar17a.html.
[40] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
[41] P. Sturmfels, S. Lundberg, S.-I. Lee, Visualizing the impact of feature attribution baselines, Distill 5 (2020) e22.
[42] G. Plumb, D. Molitor, A. S. Talwalkar, Model agnostic supervised local explanations, Advances in neural information processing systems 31 (2018).
[43] A. Bloniarz, A. Talwalkar, B. Yu, C. Wu, Supervised neighborhoods for distributed nonparametric regression, in: Artificial Intelligence and Statistics, PMLR, 2016, pp. 1450–1459.
[44] J. Kazemitabar, A. Amini, A. Bloniarz, A. S. Talwalkar, Variable importance using decision trees, Advances in neural information processing systems 30 (2017).