On the Impact of Neighbourhood Sampling to Satisfy Sufficiency and Necessity Criteria in Explainable AI

Urja Pawar URJA.PAWAR@MYCIT.IE
Christian Beder CHRISTIAN.BEDER@MTU.IE
Ruairi O'Reilly RUAIRI.OREILLY@MTU.IE
Donna O'Shea DONNA.OSHEA@MTU.IE
Munster Technological University, Ireland

Editors: Francesco Locatello and Vanessa Didelez

Abstract

In the context of Machine Learning(ML) and Artificial Intelligence (AI), the concepts of sufficiency and necessity of features offer nuanced perspectives on the cause-and-effect relationships underlying a model's outputs. These concepts are, therefore, essential in Explainable AI (XAI) as they can provide a more holistic understanding of a "black-box" AI model. Addressing this need, our study explored the relationships between the XAI's explanations and the sufficiency and necessity of features in data. This is achieved by emphasising the impact of neighbourhoods, which are central in generating explanations. By analysing a diverse set of neighbourhoods, we highlighted how they influence the alignment between the feature rankings by XAI and the measures of sufficiency and necessity. This work offers two contributions. First, it provides a comprehensive discussion on how XAI frameworks relate to sufficiency and necessity with respect to their operating neighbourhoods; and second, it empirically demonstrates the effectiveness of these neighbourhoods in conveying the sufficiency and necessity of features by the XAI frameworks.

Keywords: Explainable AI, Necessity, Sufficiency, Neighborhoods, Evaluation

1. Introduction

The measures of sufficiency and necessity (S/N) are fundamental in understanding the cause-and-effect relationships within AI decision-making processes Watson et al. (2021). The causality is examined by assessing the impact of features in the data (the causes) on the outcome by an ML model (the effect) Kommiya Mothilal et al. (2021). S/N are specific types of causal conditions — a necessary condition is one that must be present for a certain outcome to occur, while a sufficient condition ensures the outcome if present Watson et al. (2021). Understanding the causal effects of features on a model's outcomes is critical, particularly when AI models are used in sensitive sectors like healthcare and finance Gade et al. (2020). Explainable Artificial Intelligence (XAI) serves as a bridge between the complexity of AI models and human understanding, making the decision-making processes of these models transparent Ribeiro et al. (2016). While XAI frameworks offer various types of explanations, they lack in directly explaining whether a feature is necessary or sufficient for a particular outcome Kommiya Mothilal et al. (2021).

Popular model-agnostic frameworks in XAI include Feature Importance (FI) frameworks such as LIME (Local Interpretable Model-agnostic Explanations) Ribeiro et al. (2016), SHAP (Shapley Additive Explanations) Palacio et al. (2021), and CounterFactual(CF) frameworks such as DiCE (Diverse Counterfactual Explanations) Chou et al. (2022). The two main factors while generating explanations are the methodology of the framework and the local neighbourhoods used by the framework Han et al. (2022). To explicitly convey S/N measures, their relationship with popular XAI frameworks has been discussed in many works Watson et al. (2021); Balkir et al. (2022). However, there was a lack of discussion on how these neighbourhoods in XAI associate with the S/N measures of the features. Furthermore, the unified representations used for S/N measures and the XAI explanations did not explicitly distinguish their core methodology and the neighbourhoods.

To evaluate existing XAI frameworks based on how well they convey S/N, Kommiya Mothilal et al. (2021) empirically demonstrated that the top-ranked features by XAI frameworks are neither necessary nor sufficient. However, the study was limited to the standard versions of XAI frameworks and didn't investigate other neighbourhoods with the frameworks. Furthermore, the proposed methodology assessed the S/N of individual top-3 features versus the rest of the features. However, for a more robust evaluation, the assessment should consider each feature present in the data and also account for the relative ranking among features.

Overall, there is a lack of discussions and experimental evaluations that explore the influence of neighbourhoods on feature rankings, as well as their links to S/N measures. This gap presents an opportunity for further research in enhancing our understanding of widely used model-agnostic XAI techniques, and this study aims to address this gap. Our research offers two main contributions. First, we provide a clear, unified discussion that underscores the importance of neighbourhoods in XAI frameworks and their connection to S/N measures. Second, we conduct empirical tests using the Kendall correlation to reveal how neighbourhoods affect the way these frameworks interpret the importance of features in terms of their sufficiency and necessity.

2. Related Work

Kommiya Mothilal et al. (2021) theoretically discussed SHAP, LIME, and DiCE frameworks with respect to the S/N measures and experimentally evaluated them for their effectiveness in identifying sufficient and necessary features. The study concluded that none of the frameworks explicitly highlight the most sufficient or necessary features, and the frameworks should be utilised as complementary tools to understand an ML model's operation from different perspectives. However, the study didn't discuss and explore the relationship between the explanations and the neighbourhood samples that are used for the analysis and how this could impact the evaluation.

The relationship between XAI frameworks and concepts of S/N has been discussed comprehensively using a unifying framework by Watson et al. (2021). To convey S/N, a novel XAI framework - Local Explanations via Necessity and Sufficiency (LENS) was proposed by Watson et al. (2021). This study also included experimental evaluations to compare LENS with other XAI frameworks, focusing on the identification of sufficient and necessary features. However, it did not explore the significant aspect of how neighbourhood samples affect the evaluations. This step is important, as multiple studies, such as Slack et al. (2020); Rasouli and Yu (2020); Ribeiro et al. (2016), have emphasised the considerable influence of neighbourhood sampling on the generated explanations.

Similarly, the study by Balkir et al. (2022) introduced a novel framework to provide two distinct scores related to S/N, as opposed to using a single metric for feature importance. These complementary scores were shown to be effective in detecting model biases and offering clearer interpretations. In this paper, we adopt the methodology from Balkir et al. (2022); Galhotra et al. (2021) to compute S/N scores as the approach can be well integrated with tabular datasets.

The importance of neighbourhoods XAI frameworks is highlighted and formalised by Han et al. (2022), which introduces a Local Function Approximation (LFA) framework. This LFA framework shows that various XAI frameworks can be viewed as specific instances that utilise different local neighbourhoods and loss functions. To effectively evaluate the performance of popular XAI frameworks in conveying S/N measures, it's essential to rigorously assess key elements such as neighbourhoods. This ensures that explanations from popular XAI frameworks are more purposeful and contribute to the wider adoption of AI systems Freiesleben and König (2023).

3. Discussion - S/N measures and Explainable AI

In this section, we define and represent S/N measures and the FI scores by the selected XAI frameworks in a unified format using Lebesgue integrals. The neighbourhoods used by different XAI frameworks are highlighted, and their association with the S/N measures is discussed.

Notations and Background Information: An input sample u is classified as $f(u) = y$ by an ML classifier f, with the actual prediction value denoted by $\hat{f}(u)$. x_j represents the values of j set of features in a sample x and intervening x to alter some feature values results in x'. XAI frameworks generate local explanations using neighbourhoods around the input sample. In this study, the local neighbourhood is defined by the Mahalonobis distance of samples. In a neighbourhood, samples are selected based on specific criteria and analysed to explain an individual classification. The distribution of the selected samples within a neighbourhood of x using specific criteria is denoted by $R(x)$. This $R(x)$ can be defined by various means, such as marginal or conditional distributions, as it represents the distribution following specific interventions/perturbations after the samples are drawn from a predetermined distribution Watson et al. (2021). In this study, we assume features are independent and can be changed individually. Since the selected XAI frameworks don't account for causal relationships among features, this is noted for future work.

3.1. Sufficiency and Necessity (S/N)

The methodology to compute S/N scores of the features is adopted from Balkir et al. (2022); Galhotra et al. (2021) as the approach can be well integrated with tabular datasets. We have used binary classification examples to clearly demonstrate the concept of neighbourhoods in XAI. For representing the sufficiency of a feature (or a set of features), *j*, we define equation 1, where A is the local neighbourhood and $R(x)$ is its distribution. If $x'_j = u_j$,

$$S_j(f, u) = \int_A I(f(x') = y) R(x) dx \tag{1}$$

For an input-output pair (u, y), S_j is the integral of an indicator function $I(f(x') = y)$ that is 1 if $f(x') = y$ and 0, otherwise. This is to count how many samples,x, are at the same classification as the input's (y) when the values of j feature(s) are changed to that of the original input's $(x_j \leftarrow u_j)$. To ensure that the classification y is due to the specific intervention with original feature values (u_j), we consider samples from a different classification to count how many of them arrived at the same classification as y after the intervention. The sufficiency is high if most samples in the neighbourhood change their classification to be the same as y. Therefore, the S_j measure informs the extent to which the j feature value is causing the current classification. The $R(x)$ represents the conditional probability of selecting a sample for estimation, given the preexisting distribution and the conditions $x_j \neq u_j$ and $f(x) \neq y$. In cases where $x_j = u_j$ or $f(x) = y$, $R(x)$ is set to zero, as such samples do not contribute any valuable information about feature j in maintaining the classification.

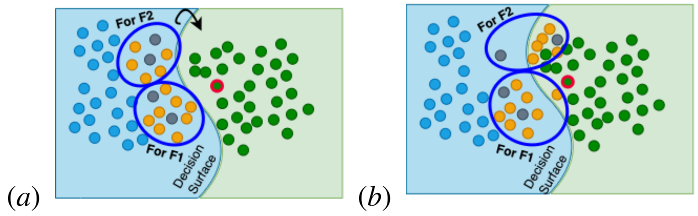


Figure 1: a) Neighbourhoods for calculating S for - F1 and F2. b) After intervention, F2's samples shifted heavily towards the same class (green) as compared to F1's.

As shown in figure 1(a), input u is highlighted in red, and the neighbourhood to estimate the sufficiency of two features $F1$ and $F2$ is highlighted in blue ovals. The yellow dots represent training samples and grey ones represent perturbed samples. Notice that the input sample (green class) and neighbourhood (blue class) lie on opposite sides of the decision boundary. If the sufficiency of a feature is high, changing values of that feature in the neighbourhood should bring many samples back to the same class as input, as shown for feature $F2$ in figure 1(b).

Similarly, we represent the necessity measure of a feature (or a set of features) j, in equation 2. If $x'_j \neq u_j$,

$$N_j(f, u) = \int_A I(f(x') \neq y) R(x) dx \tag{2}$$

Necessity measure N_j is the integral over an indicator function $I(f(x') \neq y)$ that is 1 if $f(x') \neq y$ and 0 otherwise. This counts how many samples change their classification when the j feature values are changed. The necessity is high if the majority of samples in the neighbourhood change their classification to be different from that of u. Therefore, the N_j measure informs the extent to which the change in j feature value causes a change of classification. The values in j are changed to any other values from the training data such that the new samples are within a local neighborhood defined using the Mahalonobis distance. The $R(x)$ represents the conditional probability of selecting a sample, given the preexisting distribution and the conditions $x_j = u_j$ and $f(x) = y$. For samples where $x_j \neq u_j$ or $f(x) \neq y$, $R(x)$ is set to zero, as they do not contribute meaningful information for assessing necessity.

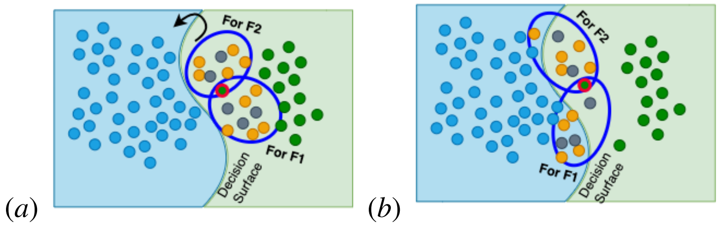


Figure 2: a) Neighbourhoods for calculating N of features - F1 and F2. b)After intervention, F1's samples shifted heavily towards the opposite class (blue) as compared to F2's .

In figure 2, the input and the neighbourhoods to calculate the necessity for features $F1$ and $F2$ lie on the same side of the decision boundary. If the necessity of a feature is high, changing its value should shift a majority of the samples to the opposite class (blue), as shown for the feature $F1$ in figure 2(b).

S/N are very interlinked concepts as the feature that has a value that is sufficient in keeping a classification can also have a different value that is necessary for the opposite classification to occur. The neighbourhood selection for both S/N is based on classification change and this guided us to specifically explore neighbourhoods based on the number of same/different classes around an input sample. This is elaborated later in section 4.2.

3.2. Explainable AI

The two most commonly used XAI techniques are FI and CFs Antoniadi et al. (2021). The following section discusses the widely used XAI frameworks and their relationship with the S/N measures.

3.2.1. Feature Importance (FI)

FI estimates the contribution of each feature in a classification and determines the importance of each feature by assigning scores Ribeiro et al. (2016). For empirical evaluations, popular open-source FI frameworks - SHapley Additive Explanations (SHAP)[1] and Locally Interpretable Model-agnostic Explanations (LIME)[2], are considered in this evaluation.

SHAP To explain an input-output pair (u, y), SHAP Lundberg and Lee (2017) estimates the average marginal contribution of an individual j^{th} feature considering all possible feature subsets. This is done by analysing prediction difference by $\hat{f}$ on including (x_{+j}) and excluding(x_{-j}) the j^{th} feature value in different feature subsets. Here, exclusion means replacing the feature value by sampling from training data Palacio et al. (2021). The feature subsets are also constructed by including the original value of the features belonging to the subset and replacing the values of other features from the training data samples. SHAP can be approximated using Monte-Carlo sampling as proposed in Štrumbelj and Kononenko (2014). The kernel used to weigh the samples is based on the size of the feature subset used, i.e., how many feature values are common between the original input and the sample. Higher weights are assigned to samples that have either few or too many common features.

The neighbourhood in SHAP is generated after perturbing input u using training data and weighting the samples using the kernel. Here, $R(x)$ describes the kernel-weighted conditional probability based on a given distribution that is adjusted for interventional effects due to feature perturbations. Based on the kernel, higher values of $R(x)$ are assigned to samples with very small or very large replacement of features. Very small subsets of features enable an understanding of the necessity of jth feature because the rest of the features are kept constant (small subset = few features replaced). Very large subsets enable an understanding of its sufficiency as almost all features are replaced, and $\hat{f}(x_{+j}) - \hat{f}(x_{-j})$ captures the effect of jth feature towards a prediction in all samples. We define SHAP scores using equation 3. Here, $SHAP_j(f, u)$ represents how much a specific feature j is contributing overall towards a prediction of u by f.

[1] https://github.com/slundberg/shap

[2] https://github.com/marcotcr/lime

$$SHAP_j(f,u) = \int_A (\hat{f}(x_{+j} - \hat{f}(x_{-j}))R(x)dx \tag{3}$$

As shown in figure 3(a), neighbourhood for SHAP for explaining an input (highlighted in red) are a set of samples(highlighted in grey) that contain a combination of feature values from original input and training samples. These perturbed samples with few or many common feature values are given higher weights represented by their size.

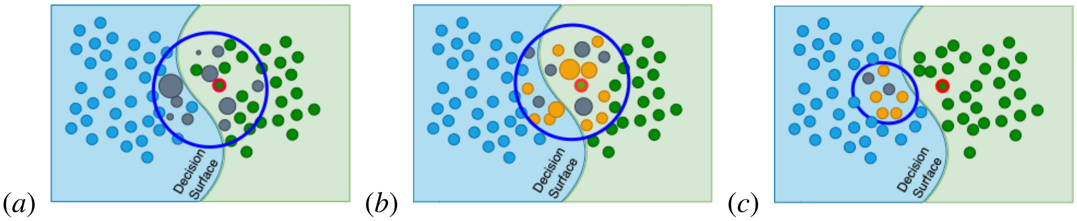


Figure 3: Neighbourhood representations a) SHAP, b) LIME, and c) DiCE.

LIME To explain an input-output pair $(u, f(u))$ locally, LIME Ribeiro et al. (2016) estimates the FI scores by generating a local neighbourhood around u and training different interpretable linear models G on the neighbourhood samples weighted using a distance metric, and their classifications by f as ground truth. The coefficients of the linear model $g \in G$ that best approximates f provide the FI scores. These FI scores can be approximated using equation 4. $LIME_j(f,u)$ represents how much a specific individual feature value contributes to a prediction of u by f with respect to the local samples. If $x_j \neq u_j$,

$$LIME_j(f,u) = \int_A \frac{(\hat{f}(x) - \hat{f}(u))}{|x_j - u_j|} R(x)dx \tag{4}$$

Small changes in feature values and large changes in the model's prediction cause the larger magnitude of FI scores. Here, $R(x)$ serves as a weight that incorporates the conditional probability of a sample given its perturbation from u and the condition $x_j \neq u_j$, with an additional weighting based on the sample's proximity to u. Note that R(x) is 0 for samples where $x_j = u_j$.

The intuition behind LIME is based on linear regression. In a binary setting, it estimates the probability of a data point being classified as positive (1) and assigns coefficients accordingly. Features with positive coefficients indirectly inform how much *sufficient* their value is in keeping a positive classification and how necessary a feature value is for a negative classification, as a minor deflection in feature values can lead to a positive classification. This is vice versa for features with negative coefficients. As shown in figure 3(b), LIME can use a local neighbourhood using training samples (highlighted in yellow) as well as perturbed samples (highlighted in grey).

3.2.2. COUNTERFACTUALS (CFS)

CF explanations describe the smallest change to the feature values that changes the classification to another class Poyiadzi et al. (2020). In Wachter et al. (2017), a loss function to minimise the distance between the input sample and CF explanation while keeping a different classification was proposed, and CFs generated using this classical technique are termed Watchter CFs. In Mothilal et al. (2020), an additional term is added to the loss function to ensure diversity in the CF explanations generated (with values of multiple features to change). The CF explanations generated using the diversity-based loss function are called Diverse Counterfactual Explanations (DiCE). In this empirical study, the DiCE framework is used for the evaluation as it enables the generation of a diverse range of CFs that can be used for the analysis of CF's association with S/N.

To compare CF explanations with S/N measures, FI scores are generated by observing the number of times a feature is changed in the K number of CFs Kommiya Mothilal et al. (2021). If $I(x_j \neq u_j)$ is the indicator function that is 1 if the feature j changes it's value in a CF and 0 otherwise, FI scores using CFs are represented by equation 5. $CF\_FI_j(f, u)$ represents the extent to which a change in a feature j will occur while altering the classification with minimum changes.

$$CF\_FI_j(f, u) = \int_A I(x_j \neq u_j) R(x) dx \tag{5}$$

where A is the neighbourhood of u and $R(x)$ represents the conditional probability of selecting a sample from a given distribution that has a different classification than that of the input u. $R(x) = 0$ for samples that belong to the same class as $f(u)$. This approach is useful to effectively interpret information from the CFs by capturing the likelihood that a specific feature change resulted in a change of classification in a given local neighbourhood Von Kügelgen et al. (2023).

As shown in figure 3(c), multiple CFs (CF1 and CF2) are possible for explaining an instance u. Both CF1 and CF2 are close to input u and lie on the other side of the decision boundary. The local neighbourhood (blue oval) in CFs has samples that are classified differently than the original input. CFs directly inform about the necessity of features. If a feature is repeatedly prompted to change for changing a classification, it implies a higher necessity of that feature.

4. Empirical Evaluation

In this study, we aim to evaluate XAI frameworks using various neighbourhood types to convey S/N measures of features. To achieve this, we first detail the components used in the experimental methodology.

4.1. Datasets and Classifier

We utilised three medical tabular datasets; all focused on binary classification. First is the cervical cancer dataset containing 1,256 records and 24 features Kelwin Fernandes and Fernandes. Second, is the heart disease dataset, with 297 records and 13 features Robert Detrano, and the third is the diabetes dataset, with 768 records, consisting of 7 features Smith. The three binary classification datasets are chosen for a consistent methodology and simplified calculations of the S/N measures and the explanations. We employed a Support Vector Machine (SVM) classifier with a linear kernel for our experiments. The accuracy scores achieved by the SVM classifier for cervical cancer, heart disease, and diabetes datasets are 97.1%, 85.18%, and 75.6%, respectively. Since our paper primarily aims to assess model-agnostic XAI frameworks across various neighbourhoods, the selection of the classifier isn't the main point of our discussion.

4.2. Neighbourhoods

As discussed in section 3.2, XAI frameworks construct their neighbourhoods - labelled as the $standard$. Since S/N measures are calculated from changes in classification, we examined different neighbourhoods characterised by distinct classification results. All the neighbourhoods used random perturbation on the dataset using values from the training data. We used the Mahalonobis distance to define locality and select those perturbed samples that are closer to the input, based on the covariance established by the training data. With a specific type of neighbourhood, the appropriate conditional probability holds to reflect in $R(x)$.

1. **Perturbed**$(perturb)$: In this, to create a local neighbourhood, the input is not perturbed randomly but by altering features progressively: starting with one, then pairs, and progressing to larger sets. This ensures that more samples have similar feature values.
2. **Balanced**: This neighborhood contains a balanced ratio of samples from each class.
3. **Skewed opposite** ($skewed(opp)$): This involves a skewed sample set: 75% from different classes and 25% from the input's class.
4. **Skewed similar** ($skewed(same)$): In this, the skew is towards the input sample's class: 75% of samples are from the input's class and 25% from different classes.
5. **Restricted Outside (**$outside$**)**: This contains samples classified differently than the input sample.
6. **Restricted Inside (**$inside$**)**: This contains samples classified the same as the input sample.

Based on the above taxonomy, multiple neighbourhoods were generated. For example, $perturb\_balanced$ uses the specific perturbation described above and selects an equal number of samples from both (0/1) classes.

Application of neighbourhoods: The method for scoring features varies among XAI frameworks, so not all neighbourhoods fit every framework. For LIME, all neighbourhoods except $inside$ and $outside$ work, as it needs samples from both classes to train the interpretable classifier for approximating a feature's importance. SHAP can use all types of neighbourhoods as it uses them reconstruct its own one as discussed in the previous section. Conversely, DiCE needs an $outside$ restricted setting, so only corresponding neighbourhood combinations are considered.

4.3. Experimental Methodology

Our detailed methodology is illustrated in figure 4. Initially, the dataset is used to train the black-box classifier. Post this training phase, each test data sample, along with its associated neighbourhood defined by the "neighbourhood type", is processed by the three XAI frameworks - SHAP, LIME and DiCE, and the S/N measures computation block. The S/N scores are computed for individual features, and rankings are generated. While SHAP and LIME generate explanations in the form of FI scores, scores from the DiCE CFs are estimated by observing changes in feature values across the nearest K-CFs, as discussed in section 3.2.2. Based on iterative experimentation, we chose a K value of 200 to achieve an optimal correlation between CFs and S/N scores. We considered the

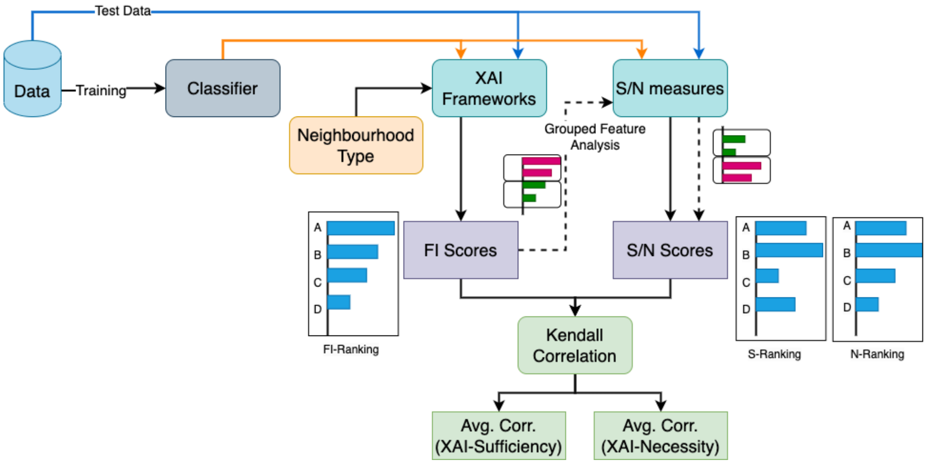


Figure 4: Methodology for evaluating XAI for S/N using various neighbourhoods.

magnitude of FI scores to rank features as a higher magnitude implies a feature's stronger influence on the prediction, which can be interpreted in the context of its S/N.

To measure the alignment between the feature rankings (derived from FI scores' magnitude) and S/N rankings, we used the Kendall tau correlation coefficient. A high correlation coefficient indicates a strong alignment of XAI's feature rankings with the S/N measures of individual features. These results are aggregated for every combination of neighbourhoods and XAI frameworks across datasets. For simplification, we present the average correlation using different types of neighbourhoods to compare and analyse their impact.

Grouped Feature Analysis: Based on the neighbourhood that shows the best correlation with the S/N rankings, the correlation is again analysed with respect to subsets of features. This analysis provides insights into how the top-n ranked features associate with their collective S/N. By evaluating both individual and grouped feature rankings; we aim to present a more layered understanding of the feature rankings provided by the XAI frameworks. As denoted by the dotted lines in figure 4, the analysis evaluates n-sized feature subsets for their S/N. For instance, if n equals 2, the top-2 ranked features (ranked 1st and 2nd), followed by the subsequent two features (ranked 2nd and 3rd, and so on), are assembled and ranked. This grouped ranking is compared to the collective S/N of these feature subsets of size - 2 via the Kendall correlation. This is repeated for n =1,2,3 and 4. The code to reproduce results using our methodology is available on Github[3].

5. Results and Discussion

In this section, we present the average correlation between S/N measures and the feature rankings (individual and grouped) from the XAI frameworks - SHAP, LIME, and DiCE. Our goal is to evaluate this correlation across various neighbourhoods and pinpoint those best at conveying the S/N measures. Subsequent subsections delve into the correlation analysis for each XAI framework.

[3] Link to the code - https://github.com/UrjaPawar/NeceSuffXAI

5.1. Correlations of Individual Feature Rankings with S/N Measures

In this section, the mean correlation between individual feature rankings by each XAI framework and S/N scores are presented using plots with neighbourhoods listed on the x-axis and correlation values on the y-axis.

Correlations with SHAP rankings Figures 5(a) and (b) show the average Kendall τ between SHAP rankings and S and N rankings, respectively, across the three datasets. Both the results are similar to each other because S/N measures are closely linked.

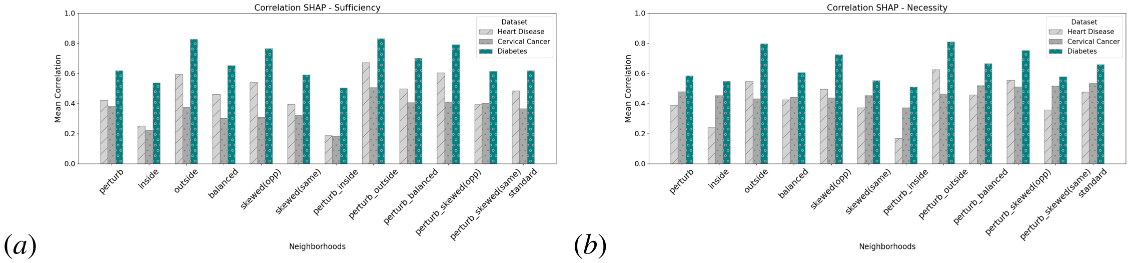
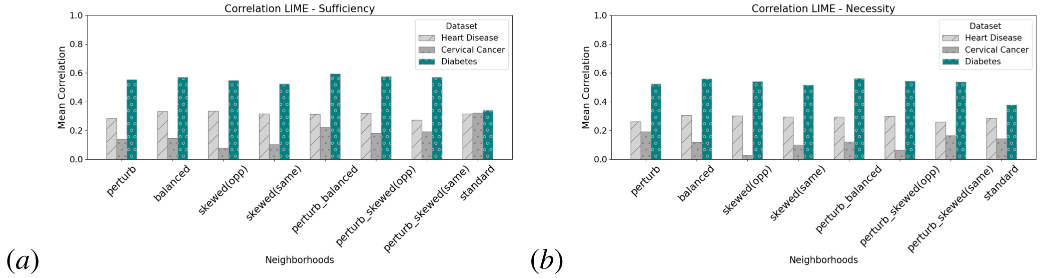


Figure 5: Mean Kendall correlation coefficient (τ) between SHAP and a)Sufficiency, and b)Necessity rankings in different neighbourhoods across datasets.

Across the datasets, the mean correlations with S/N using various neighbourhoods showed similar patterns, and we discuss the results derived from averaging these values from all datasets. For sufficiency in figure 5 (a), neighbourhoods with $perturb$ show around 4.47% higher average correlation than those without $perturb$. The $balanced$ neighbourhoods show 10.90% lower average correlation than those $skewed$ towards the opposite class and 17.31% higher than those $skewed$ towards the same class. On average, $skewed(opposite)$ showed 31.66% higher correlation than $skewed(same)$. An important observation is that the neighbourhoods based on $inside$ showed 51.35% lower average correlation than $outside$. The highest correlation was observed in $perturb\_outside$ that showed approximately 30% higher correlation as compared to the $standard$ SHAP. This is intuitive as the concept of sufficiency is based on the notion of keeping the classification as it is, and derivation of the scores in $outside$ neighbourhoods is done by estimating the FI in achieving the same classification when compared to feature values that caused an opposite classification.

Similarly, correlations with the necessity scores (figure 5 b) were also the highest with $perturb\_outside$ with $outside$ neighbourhoods performing 39.14% higher than $inside$ ones, on average. As necessity measures the impact of features on changing the classification, the $outside$ neighbourhoods enable an analysis of high-impact feature values that resulted in classification change with respect to any change in their own values.

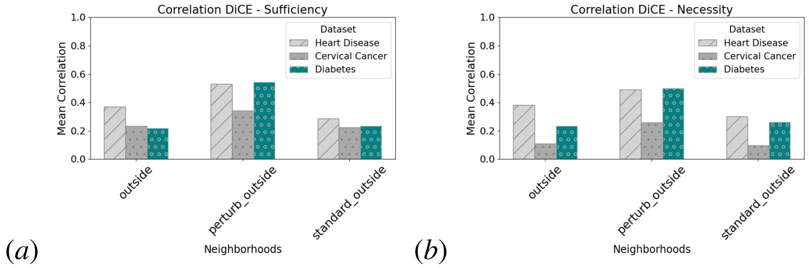
Correlations with LIME Rankings Figure 6(a) and (b) present the average correlation between LIME rankings and S and N rankings, respectively. For sufficiency, there was no significant difference in the correlations between $perturb$ neighbourhoods and those without $perturb$. However, on average, $balanced$ neighbourhoods showed 11.77% higher average correlation than $skewed$ to the opposite class and 6.59% higher than those $skewed$ to the same class. While the difference is not very significant, it shows that a balanced neighbourhood provides insights into a local interpretable



(*a*) (*b*)

Figure 6: Mean Kendall correlation coefficient (τ) between LIME and a)Sufficiency, and b)Necessity rankings in different neighbourhoods across datasets.

model for identifying sufficient features that don't lead to a change of classification when other feature values are changed. Overall, the $perturb\_balanced$ neighbourhood showed the best average correlation across the datasets with 33% higher average correlation as compared to the $standard$ LIME. Similarly, for necessity, the $perturb\_balanced$ neighbourhood provided the best correlations such that the local interpretable model can effectively capture the necessary feature values that drive the classification change.



(*a*) (*b*)

Figure 7: Mean Kendall Tau correlation Coefficient (τ) between DiCE and a)Sufficiency, and b)Necessity rankings in different neighbourhoods across datasets

Correlations with DiCE Rankings: Figure 7(a) and (b) present the average correlation between DiCE importance scores and S and N scores, respectively. For S/N, $perturb\_outside$ neighbourhood showed the best average correlation (150% higher than $standard$ DiCE) across the datasets as the perturbation ensures more samples with common feature values and, therefore, a more local neighbourhood for estimating individual features' S/N with respect to the classification.

5.2. Grouped Feature Analysis

To understand the correlation of top-n ranked features and their collective S/N scores, the correlations were also calculated for feature sets using the best-performing neighbourhoods as mentioned in the section 4.3. Figure 8(a) and (b) presents the average correlation between the XAI frameworks with their corresponding best-performing neighbourhood and S/N rankings of the feature sets. For SHAP and DiCE - $perturb\_outside$ neighbourhood was selected, and for LIME, $perturb\_balanced$

was selected. Note that, among the three XAI frameworks, SHAP provides the highest average correlation for individual feature S/N (with n=1). This can be attributed to SHAP's kernel assigning a higher weight to samples with very few or too many common features that enable the ranking to correlate with the individual S/N of the the feature.

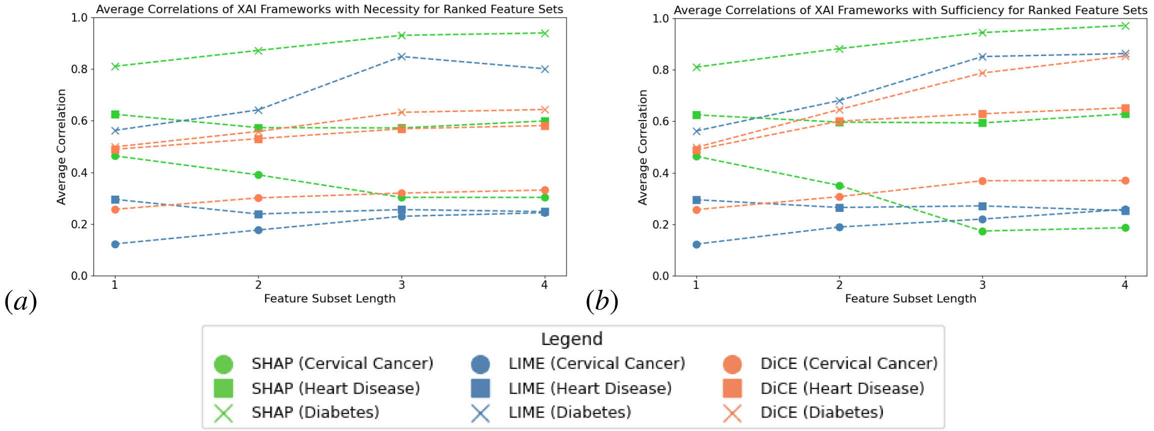


Figure 8: Mean correlation coefficient between rankings of feature subsets using optimal neighbourhoods in XAI frameworks and a) necessity scores b) sufficiency scores

As shown in figure 8(a), the average correlation of LIME and DiCE rankings with necessity rankings was almost the same as the size of feature subsets was increased. This is intuitive given the strong correlation of XAI frameworks with the necessity of using their respective neighbourhoods. If a high-ranked feature was indeed *necessary* for the classification, then changing even one feature in the subsets of high-ranked features should change the classification. SHAP's correlation decreased slightly at the feature subset length =2 due to SHAP's kernel facilitating the ranking based on individual feature's importance. Generally, if high-ranked features by XAI frameworks are necessary, then their subsets are also necessary.

For sufficiency, in figure 8(b), the average correlations of LIME and DiCE slightly increased with the increasing size of feature subsets as the subsets of sufficient features are sufficient as well. Similar to necessity, SHAP's correlation with sufficiency didn't increase with larger feature subsets as it's kernel provides importance score with respect to the whole set of features and not subsets of features. Generally, if high-ranked features by XAI frameworks are sufficient, then their subsets are also sufficient.

6. Limitations and Future work

Our study offers valuable insights into the role of neighbourhoods in XAI frameworks in relation to the S/N rankings of the features. However, there are certain limitations to our work. First, our methodology assumes causal independence among input features and did not capture the complexity of causal inter-dependencies that exist in real-world systems. Second, our research relied on a limited range of datasets. Third, the study focused on a specific set of neighbourhoods for generating explanations. Although these were carefully chosen, they do not cover the entire spectrum of possible neighbourhoods that could be used with XAI frameworks. In future work, we aim to address these limitations by investigating methods to incorporate causal relationships, expanding the datasets and exploring additional types of neighbourhoods to present more robust conclusions and granular insights.

7. Conclusions

In this study, we assessed the effectiveness of three prominent XAI frameworks — SHAP, LIME, and DiCE — correlating with S/N measures. Our study offers valuable insights into how different types of neighbourhoods influence the rankings by XAI frameworks. Our findings suggest that conveying the S/N of features by XAI frameworks can be improved using specific neighbourhoods, underscoring their utility in XAI frameworks. For conveying individual feature sufficiency and necessity, SHAP performed best with neighbourhoods based on samples belonging to the opposite class. LIME performed best with well-balanced neighbourhoods, and DiCE with more locally defined neighbourhoods. Among the three widely used XAI frameworks, SHAP is the most effective in conveying the S/N of individual features. Additionally, our investigation into grouped sets of features has shown that if top-ranked features are individually necessary/sufficient, their subsets are also necessary/sufficient for a given classification.

The proposed methodology can be extended to multi-class classification of tabular datasets as the concepts of sufficiency and necessity focus only on two types of classes - one associated with the input sample and ones that are not. The methodology can also be used to observe correlations among evaluation metrics such as metrics related to fidelity that are often not aligned with each other. This can enable us to identify the types of input samples (the neighborhoods) for which these metrics demonstrate stronger correlation with each other, thereby indicating a higher level of trustworthiness in these metrics.

Our research makes two important contributions to the XAI domain. First, it presents an easy-to-follow discussion on the interplay between XAI frameworks and the S/N measures, particularly focusing on the role of neighbourhoods. Second, we provide empirical evidence confirming the effectiveness of certain types of neighbourhoods in conveying S/N measures. These findings are important to use XAI in practical settings, as they offer a more nuanced view of how black-box models make particular decisions.

Acknowledgments

This publication has emanated from research co-sponsored by McKesson and Science Foundation Ireland under Grant number SFI CRT 18/CRT/6222.

References

Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. 11(11):5088, 2021.

Esma Balkir, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. *arXiv preprint arXiv:2205.03302*, 2022.

Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

Timo Freiesleben and Gunnar König. Dear xai community, we need to talk! fundamental misconceptions in current xai research. *arXiv preprint arXiv:2306.04292*, 2023.

K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly. Explainable ai in industry: practical challenges and lessons learned. *Companion Proceedings of the Web Conference 2020*, 2020. doi: 10.1145/3366424.3383110.

Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems*, 35:5256–5268, 2022.

Jaime S. Cardoso Kelwin Fernandes and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. *Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017*. URL https://archive.ics.uci.edu/ml/datasets/Cervical+cancer.

Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Sheraz Ahmed, Jörn Hees, and Andreas Dengel. Xai handbook: towards a unified framework for explainable ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3766–3775, 2021.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Peyman Rasouli and Ingrid Chieh Yu. Explan: Explaining black-box classifiers using adaptive neighborhood generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Ph.D. Robert Detrano, M.D. Uci. 2010. v. a. medical center, long beach and cleveland clinic foundation. URL https://archive.ics.uci.edu/ml/datasets/heart+disease.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

Everhart J.E. Dickson W.C. Knowler W.C. Johannes R.S. Smith, J.W. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press 1988*. URL https://archive.ics.uci.edu/ml/support/diabetes.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning*, pages 177–196. PMLR, 2023.

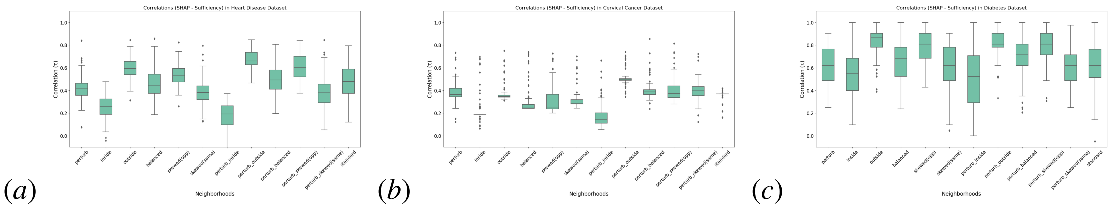
Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

David S Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. In *Uncertainty in Artificial Intelligence*, pages 1382–1392. PMLR, 2021.

Appendix A. Additional Results

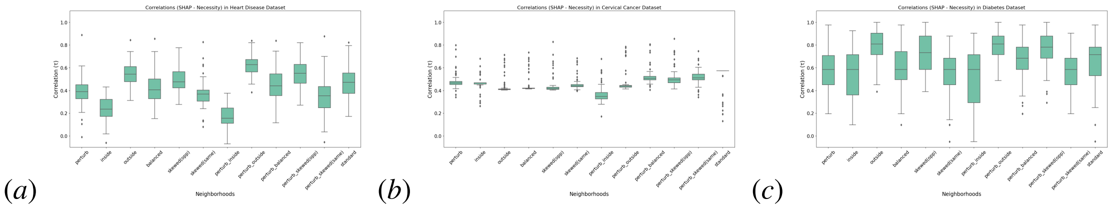
The distribution of the correlation τ in each dataset was analysed and a higher variance was observed in all the results of the diabetes dataset as compared to much lower variance in the results of the cervical cancer dataset. This can be attributed to the number of features. With lesser number of features each feature has a greater influence on the Kendall correlation that can lead to more fluctuations in the correlation values. The diabetes dataset has only 7 features while cervical cancer used 24 features.

Table 1 and 2 present the average correlation with sufficiency and necessity, respectively, across various selections of neighbourhoods.



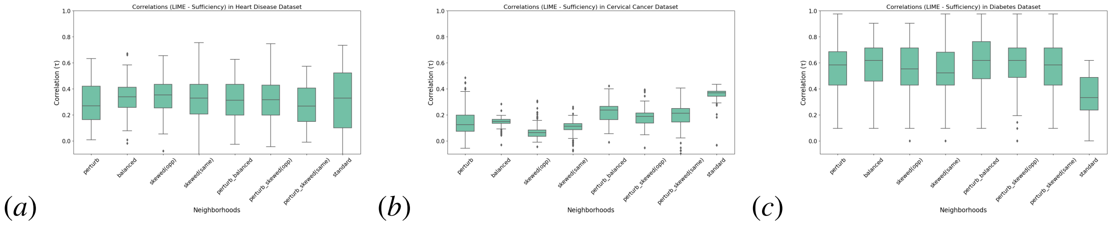
(*a*) (*b*) (*c*)

Figure 9: Mean Kendall Tau correlation Coefficient (τ) between SHAP and sufficiency Scores in different neighbourhoods across datasets



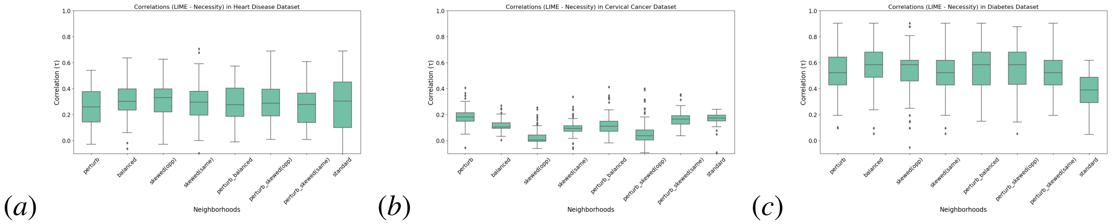
(*a*) (*b*) (*c*)

Figure 10: Mean Kendall Tau correlation Coefficient (τ) between SHAP and necessity Scores in different neighbourhoods across datasets



(*a*) (*b*) (*c*)

Figure 11: Mean Kendall Tau correlation Coefficient (τ) between LIME and Sufficiency Scores in different neighbourhoods across datasets



(*a*) (*b*) (*c*)

Figure 12: Mean Kendall Tau correlation Coefficient (τ) between LIME and Necessity Scores in different neighbourhoods across datasets

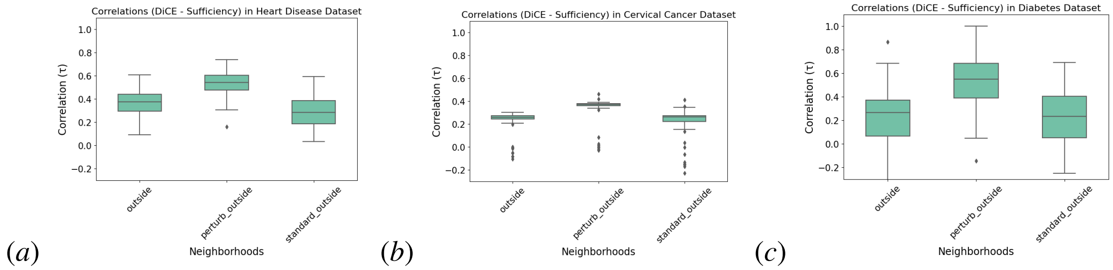


Figure 13: Mean Kendall Tau correlation Coefficient (τ) between LIME and Sufficiency Scores in different neighbourhoods across datasets

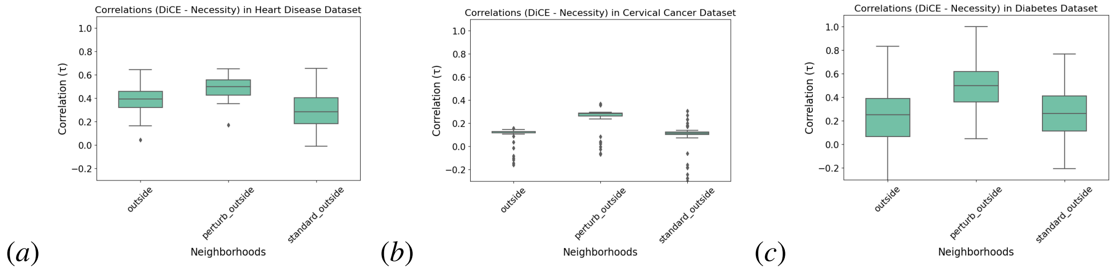


Figure 14: Mean Kendall Tau correlation Coefficient (τ) between DiCE and Necessity Scores in different neighbourhoods across datasets

XAI	Datasets	Avg. Corr	Restricted inside/outside		Random/Perturbed		Balanced/Skewed		
			Inside	Outside	Random	Perturbed	Balanced	Skewed(Opposite)	Skewed(Similar)
SHAP	Heart Disease	0.46	0.21	0.62	0.44	0.47	0.48	0.56	0.40
	Cervical Cancer	0.33	0.17	0.44	0.29	0.37	0.38	0.42	0.32
	Diabetes	0.67	0.17	0.44	0.66	0.37	0.38	0.42	0.32
LIME	Heart Disease	0.30	NA	NA	0.32	0.28	0.31	0.31	0.30
	Cervical Cancer	0.17	NA	NA	0.15	0.19	0.19	0.11	0.15
	Diabetes	0.57	NA	NA	0.55	0.59	0.61	0.59	0.60
DICE	Heart Disease	0.42	NA	NA	0.35	0.51	NA	NA	NA
	Cervical Cancer	0.27	NA	NA	0.23	0.34	NA	NA	NA
	Diabetes	0.35	NA	NA	0.22	0.54	NA	NA	NA

Table 1: Average correlation with sufficiency scores across various types of neighbourhoods

XAI	Datasets	Avg. Corr	Restricted inside/outside		Random/Perturbed		Balanced/Skewed		
			Inside	Outside	Random	Perturbed	Balanced	Skewed(Opposite)	Skewed(Similar)
SHAP	Heart Disease	0.45	0.20	0.61	0.45	0.46	0.47	0.55	0.39
	Cervical Cancer	0.44	0.37	0.45	0.44	0.44	0.47	0.46	0.42
	Diabetes	0.65	0.50	0.80	0.65	0.66	0.66	0.74	0.56
LIME	Heart Disease	0.28	NA	NA	0.28	0.26	0.29	0.29	0.28
	Cervical Cancer	0.16	NA	NA	0.16	0.18	0.18	0.12	0.16
	Diabetes	0.56	NA	NA	0.56	0.56	0.58	0.57	0.58
DICE	Heart Disease	0.42	NA	NA	0.45	0.50	NA	NA	NA
	Cervical Cancer	0.16	NA	NA	0.18	0.25	NA	NA	NA
	Diabetes	0.35	NA	NA	0.37	0.51	NA	NA	NA

Table 2: Average correlation with necessity scores across various types of neighbourhoods