The Dynamics of Explainability: Diverse Insights from SHAP Explanations using Neighbourhoods

Urja Pawar, Ruairi O'Reilly, Christian Beder and Donna O'Shea

Munster Technological University, Cork, Ireland

Abstract

This paper presents a discussion on utilising a dashboard tool to enhance the interpretability of SHapley Additive exPlanations (SHAP) in healthcare Artificial Intelligence(AI)-based applications. Despite SHAP's potential to demystify an AI model's decisions, interpreting SHAP values remains challenging, especially when considering different data neighbourhoods [1]. This issue is particularly critical in healthcare, where decision-making requires high precision and clarity. We demonstrate three use cases that can effectively demonstrate the utility of interactive neighborhood exploration. The first compares SHAP explanations in two similar patient neighbourhoods with different classifications, offering unique insights into features that influence classification changes. The second use case focuses on "feature freezing" which isolates certain features to better understand their impact. This can enable highlighting diagnostic tests considered important by a Machine Learning (ML) model for a specific population of patients (e.g., patients of the same age). The final use case demonstrates the relationship between sufficient features for a given classification and the importance ranking by SHAP.

Keywords

Explainable AI, Neighbourhoods, SHAP explanations

1. Introduction

The field of Explainable AI (XAI) is dedicated to enhancing the transparency and trustworthiness of AI systems [2]. One state-of-the-art XAI framework is SHAP, which highlights the influence of dataset features on AI model predictions by assigning importance scores(SHAP values) and is based on a solid theoretical foundation [3]. However, interpreting SHAP values is challenging due to their dependence on the sample sets or "neighbourhoods" used during analysis, complicating their application in the precision-critical field of healthcare [4, 5, 1].

In SHAP's analysis, different samples are constructed by mixing the feature values from the input sample(to be explained) and the neighbourhood samples. Since different neighbourhoods produce varying explanations, a layer of complexity and ambiguity is introduced when interpreting SHAP values. When only one of these varying explanations is presented in a standalone manner, it can lead to an inconsistent interpretation of the model's behaviour. Furthermore, the assigned importance scores by SHAP are relative to feature values in the training data, complicating their interpretation against a fixed baseline. Presenting multiple explanations collectively to users enables understanding how distinct feature values influence the model's decisions, enhancing insights into the model's behaviour.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

urja.pawar@mycit.ie (U. Pawar); Ruairi.OReilly@mtu.ie (R. O'Reilly); Christian.Beder@mtu.ie (C. Beder); Donna.OShea@mtu.ie (D. O'Shea)

0009-0009-2902-9425 (U. Pawar); 0000-0001-7990-3461 (R. O'Reilly); 0000-0001-8485-019X (C. Beder); 0000-0002-2437-3106 (D. O'Shea)

Previous studies have explored the impacts of neighbourhoods on SHAP's explanations [4, 6]. However, fewer studies have focused on providing varying explanations by exploring different neighbourhoods and interpreting these explanations in the context of a use case. There is an emerging need for an interactive exploration of these neighbourhoods, enabling users to adjust various parameters such as distance metrics, handle data imbalance, and allow feature value ranges in a neighbourhood. Additionally, fundamental explanation blocks, such as sufficient feature sets, provide specific insight regarding an ML model. Sufficient sets of features can preserve a classification regardless of the values of other features.

To address the challenges of clarity and consistency in interpreting SHAP values across neighbourhoods, this work presents a discussion for an interactive analysis of AI decision-making processes. Our tool enables users to explore how different types of samples impact the explanations and their interpretation. This holistic interpretation will facilitate a deeper understanding of model decisions, enhancing SHAP-driven insights. To demonstrate the utility of the proposed approach, we focused on the following three insights:

- **Similar vs different patient records:** SHAP values using similar and differently classified patient records are compared to highlight key influences on classification confidence.
- **Feature Freezing:** SHAP values are analysed when certain features are held constant in a neighbourhood for understanding feature significance within specific patient groups.
- **Sufficient Sets:** The relationship between sufficient feature sets and their importance rankings across various neighbourhoods is analysed.

2. Related Work

SHAP [3] calculates the impact of features on model predictions by averaging the changes when a feature is added or removed, using training set values for replacing "removed" features. This approach forms a 'neighbourhood' around an input, x, where SHAP analyses the perturbed samples formed by combining attributes from x and training data.

Authors in [4] demonstrated how XAI frameworks - SHAP and LIME can be *fooled* to generate FI scores that do not reflect the actual learned representation of AI/ML models. The ML models might be biased, but if specific neighbourhoods based on adversarial examples are used, the explanations generated can be used to hide the biases in models.

Various frameworks have been designed to improve the transparency of ML models via interactive exploration. For instance, "modelStudio"[7], "explAIner" [8], and "InterpretML"[9] provided an interactive environment that allows users to explore and understand predictions through various explanations. Complementing these tools, "The what-if tool" [10] focused on interactive probing of models, providing users with the means to test hypothetical scenarios and observe how model predictions change with varying input data for understanding model sensitivity and decision boundaries. However, none of the dashboards explores the neighbourhoods used by XAI frameworks to understand the context of model predictions, which is crucial for evaluating the robustness and applicability of explanations.

To explore the impact of various neighbourhoods on the explanations produced by surrogate XAI frameworks such as LIME, [6] developed an interactive dashboard - Explain-Explore. However, no exploratory tool has been proposed in the literature to interpret SHAP's explanations with respect to distinct neighbourhoods. In [1], a comprehensive analysis of the impact of

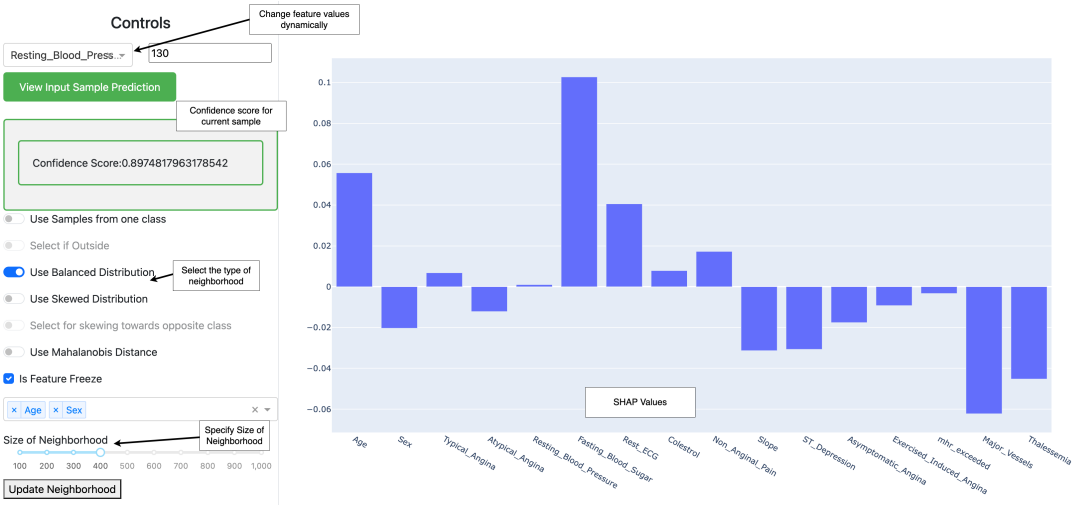


Figure 1: Proposed Interactive Dashboard

neighbourhoods on SHAP and LIME was conducted. This work presents an extension on [1] by enabling an interactive analysis of neighbourhoods with SHAP.

3. Methodology

This proposed work is implemented with the Dash[1] and Plotly[2] libraries in Python. The subsequent sections outline the methodology for building this tool, detailing the types of neighbourhoods and the use cases for utility analysis.

The workflow of our dashboard creation starts with the selection and pre-processing of the required dataset, followed by model training. SHAP values for the model's predictions are generated by analysing an input sample x and the trained classifier. Initially, the neighbourhood for calculating the values is derived from the training dataset. The dashboard also presents the classification output and the associated confidence score. A critical step is identifying minimal subsets of features that are sufficient for determining the model's predictions. A detailed description of these sets is provided in Section 3.3.

As shown in Figure 1, the proposed tool plots SHAP values and displays confidence scores and sufficient feature sets. Users can modify input sample values, observe feature importance and confidence score changes, and choose neighbourhood settings. Updates in neighbourhood settings re-calibrate SHAP calculations with updated scores. Neighbourhood information (e.g., mean feature values) is accessible via hover-over tooltips on the bar charts. The time complexity for generating SHAP explanations is directly proportional to the number of features in the dataset. In future work, we will explore optimisation techniques that can be used with SHAP for generating explanations faster[11].

3.1. Dataset and Classifier

In this study, we used the Heart Disease dataset from the UCI repository [12] for heart disease prediction using individual health records. This dataset was selected to demonstrate different

[1] https://dash.plotly.com/
[2] https://plotly.com/python/

use cases about the medical features in tabular datasets and includes 14 attributes related to heart disease, spanning initial and detailed medical tests, with 297 records where 160 are classified as negative and 137 as positive for heart disease. Features such as age, sex, types of angina, blood sugar, blood pressure, and cholesterol levels, which are standard initial tests before more advanced examinations, are included [13]. In one of our use cases, we freeze these standard features in a neighbourhood to showcase how SHAP's revised rankings enhance model understanding. The presented approach is extensible to any other tabular medical dataset.

For the model, a Support Vector Machine (SVM) classifier was employed due to its robustness in handling both linear and non-linear classifications suited for the complex patterns in heart disease data. The classifier achieved an accuracy of 86.67% using 5-fold cross-validation. However, the primary focus of this work is not on model accuracy but on the application of SHAP, a model-agnostic framework. This emphasis allows us to concentrate on the interpretability and utility of SHAP explanations, regardless of the model's accuracy metrics.

3.2. Neighbourhoods

As outlined earlier, SHAP constructs contexts by generating perturbed samples by randomly replacing original input feature values with those from neighbourhood samples. The standard version of SHAP uses training data and is referred to as "standard". Further explored neighbourhoods are described below, based on the sample's classifications:

1. **Balanced/Skewed**: Balanced/skewed distributions of classified perturbed samples
2. **Similar versus Different Classifications**: Composing samples that either share the same classification as the input (referred to as $inside$) or differ from it ($outside$)
3. **Using Mahalanobis Distance for Locality**: Neighborhood samples based on Mahalanobis distance metric
4. **Freezing the Features**: Specific features can be fixed across samples to have the same value as the input sample to isolate and understand the effects of other varying features.

3.3. Utility Analysis

This section describes use cases to demonstrate how multiple insights can enhance clinicians' understanding of an AI/ML model.

Similar vs Different Patients: SHAP explanations for patient records under two neighbourhoods - outside and inside, are compared to the standard and balanced settings. We discuss a use-case using a specific patient record and then provide Kendall correlation between feature rankings in standard SHAP versus different neighbourhoods to demonstrate the dissimilarity. Presenting different rankings, we discuss how different types of samples affect SHAP values and how to interpret them with respect to the classification results.

Feature Freezing: Specific neighbourhoods are constructed, keeping some features constant across all samples. This enables examining how particular features impact groups of patients who share characteristics like age or gender. By keeping the basic set of features constant, clinicians can better understand how specific advanced tests influence the outcome. Kendall correlation between the two rankings of variable features - with frozen neighbourhood and with standard SHAP - are provided to demonstrate the dissimilarity in their rankings.

Sufficient Sets: The relationship between sufficient feature sets and high-ranked features by SHAP is studied. The sufficient sets of features are subsets that keep the classification as it is in a neighbourhood even when other features were varied [14]. The percentage of times, when the top-3 ranked features by SHAP under different neighbourhoods, are also the sufficient subset of features is presented. As SHAP values can be positive/negative, the results are categorised based on whether these top-3 features are negatively scored, positively scored, or overall top-3 features. The code to reproduce results is available on Github[3].

4. Results

To demonstrate the utility of our dashboard, specific patient records are considered and the SHAP plots are shown for discussions.

4.1. Similar vs Different Patient Records

Table 1
Patient record selected for use-case: Similar vs different records

Age	Sex	Typical Angina	Atypical Angina	Resting BP	Blood Sugar	Rest ECG	Choles - terol	Non-Anginal Pain	Slope	ST De-pression	Asymptoma -tic Angina	Exercised In-duced Angina	MHR Ex-ceeded	Major Vessels	Thala - ssemia
41.0	0.0	0.0	1.0	130.0	0.0	2.0	204.0	0.0	1.0	1.4	0.0	0.0	0.0	1.0	3.0

We analysed SHAP values for a patient record (detailed in Table 1) under four settings: standard, balanced, inside, and outside. Figure 2a shows the standard SHAP values for classifying the patient as healthy ($f(x) = 0$). Figure 2b illustrates the values in a balanced neighbourhood, and Figures 2c and 2d display the values using the outside and inside settings, respectively.

The figures highlight variations in feature rankings across different settings, each providing a distinct interpretation. This analysis focuses on the impact of two specific features, major vessels and fasting blood sugar, on classification probabilities. Changing the number of major vessels from 1 to 0 increases classification probability from 0.72 to 0.90 for class 0, whereas a change from 1 to 2 alters the classification. This shows how the number of major vessels influences classification. The mean value of major vessels in training data is 1. Still, the standard SHAP values in Figure 2a do not explain why there is a positive contribution associated with major vessels equal to 1. Conversely, the mean value of major vessels in “outside" is 2, making it easy to interpret that an increase in the value alters (negative impact shown in Figure 2c) the classification. This is also evident in Figure 2b, however, with less strength due to the combined effect from samples of both classes. For fasting blood sugar, altering the value from 0 to 1 improves the prediction probability to 0.87 for class 0. This significant effect is captured in “inside" (Figure 2d) as the mean value of fasting blood sugar in this neighborhood is 1.

Kendall correlation coefficients reveal discrepancies in SHAP rankings across different neighbourhood settings. Specifically, the correlation between standard and balanced neighbourhood settings is 0.669, between standard and outside settings is 0.694, and between standard and inside settings is 0.671. These variations highlight how different neighbourhood contexts yield different insights that can be used for clearer interpretations. In the context of healthcare, these insights can significantly enhance patient care. By understanding how specific changes in

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

(a) Standard SHAP scores

(b) SHAP scores with “balanced"

(c) SHAP scores with “outside"

(d) SHAP scores with “inside"

Figure 2: SHAP scores under different neighborhood settings

Neighbourhoods	Standard	Balanced	Outside	Inside
Mean values (Major Vessels)	1	1	2	0
Mean Values (Blood Sugar)	0	1	0	1

Table 2
Mean Values of Major vessels and Blood sugar across different neighbourhoods

patient data affect health outcomes using the ML model, clinicians can interpret the knowledge of the ML model more effectively.

4.2. Feature Freezing

Age	Sex	Typical Angina	Atypical Angina	Resting BP	Blood Sugar	Rest ECG	Choles - terol	Non-Anginal Pain	Slope	ST De-pression	Asymptoma -tic Angina	Exercised In-duced Angina	MHR Ex-ceeded	Major Vessels	Thala - ssemia
50.0	0.0	0.0	0.0	120.0	0.0	0.0	219.0	1.0	2.0	1.6	0.0	0.0	0.0	1.0	3.0

Table 3
Patient record selected for use-case: feature freezing

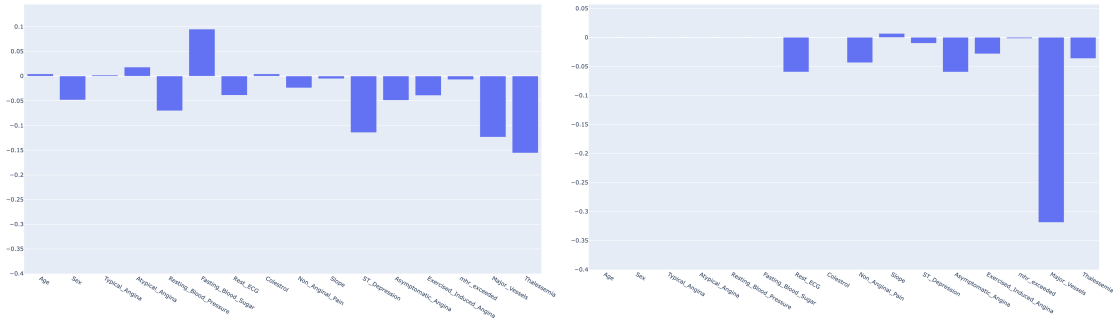
A patient record, which includes basic medical history and basic diagnostic tests like age, sex, presence of typical or atypical angina, fasting blood sugar, blood pressure, and cholesterol levels, is selected for analysis as described in Table 3. We keep these values constant across a local neighbourhood to assess the impact on the rankings of the remaining features.

Figure 3a shows the standard SHAP values, while Figure 3b displays the values when basic features are held constant (“frozen"). There’s a noticeable shift: Thalassemia, an advanced medical test, gains more importance when basic features are frozen compared to its lesser significance in standard settings. The complex interactions between features may obscure Thalassemia’s impact in standard settings compared to when the basic medical features are frozen, highlighting the significance of the test.

Sufficient sets identification (%)	top-3 positive	top-3 negative	top-3 overall
Standard	20%	27%	46%
Balanced	20%	63%	60%
Inside	20%	60%	6%
Outside	20%	66%	86%

Table 4
Percentage of patient records where sufficient sets were identified by top-3 features

The Kendall correlation between the rankings of “unfrozen" features in standard and frozen settings was found to be 70%, indicating notable dissimilarity. This use case can enable clinicians to clearly understand and amplify the relevance of diagnostic tests like Thalassemia with respect to specific patients.



(a) SHAP Scores with standard settings (b) SHAP Scores with Frozen Features

Figure 3: Different SHAP rankings on freezing features

4.3. Sufficient Feature Sets

An experiment was conducted to observe and associate the sufficient feature sets with SHAP feature rankings under different neighbourhood settings to note how many of the top-3 features form a sufficient set to maintain a given classification. As shown in Table 4, the standard SHAP scores could identify sufficient sets in only 46% of the patient records. However, with $outside$ settings, SHAP’s analysis focuses more on patients with a different classification and identifies sufficient sets in 86% of the patients when the top-3 overall features are considered. This shows that to identify features sufficient to maintain a classification, SHAP should be provided with more examples from different classifications to highlight impactful features with respect to the current classification.

5. Conclusions

This paper presented an interactive approach to improve the interpretability of SHAP in healthcare by exploring the impact of different neighbourhoods. The utility of this dashboard is demonstrated through three specific use cases, showing its effectiveness in providing detailed insights into SHAP explanations. Future work will focus on expanding the dashboard’s functionalities and improving its user interface. It will include adding more diverse XAI techniques and exploring further neighbourhood definitions to encompass a more comprehensive array of clinical scenarios. We also plan to make the tool available as a pip package and conduct user studies with healthcare professionals to refine and enhance the dashboard’s features based on real-world feedback.

References

[1] U. Pawar, C. Beder, O. Ruairi, O. Donna, et al., On the impact of neighbourhood sampling to satisfy sufficiency and necessity criteria in explainable ai, in: Causal Learning and Reasoning, PMLR, 2024, pp. 570–586.

[2] H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating shapley values of local components, arXiv preprint arXiv:1911.11888 (2019).

[3] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[4] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186. URL: https://doi.org/10.1145/3375627.3375830. doi:10.1145/3375627.3375830.

[5] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. Friedler, Problems with shapley-value-based explanations as feature importance measures, in: International Conference on Machine Learning, PMLR, 2020, pp. 5491–5500.

[6] D. Collaris, J. J. van Wijk, Explainexplore: Visual exploration of machine learning explanations, in: 2020 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2020, pp. 26–35.

[7] H. Baniecki, P. Biecek, modelstudio: Interactive studio with explanations for ml predictive models, Journal of Open Source Software 4 (2019) 1798.

[8] T. Spinner, U. Schlegel, H. Schäfer, M. El-Assady, explainer: A visual analytics framework for interactive and explainable machine learning, IEEE transactions on visualization and computer graphics 26 (2019) 1064–1074.

[9] H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv e-prints (2019) arXiv–1909.

[10] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (2019) 56–65.

[11] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, R. Ranganath, Fastshap: Real-time shapley value estimation, in: International Conference on Learning Representations, 2021. URL: https://doi.org/10.48550/arXiv.2107.07436.

[12] P. Robert Detrano, M.D., Uci. 2010. v. a. medical center, long beach and cleveland clinic foundation (1988). URL: https://archive.ics.uci.edu/ml/datasets/heart+disease.

[13] U. Pawar, C. T. Culbert, R. O'Reilly, Evaluating hierarchical medical workflows using feature importance, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 265–270.

[14] S. Galhotra, R. Pradhan, B. Salimi, Explaining black-box algorithms using probabilistic contrastive counterfactuals, in: Proceedings of the 2021 International Conference on Management of Data, 2021, pp. 577–590.