

A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME

Ahmed M. Salih,* Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz

eXplainable artificial intelligence (XAI) methods have emerged to convert the black box of machine learning (ML) models into a more digestible form. These methods help to communicate how the model works with the aim of making ML models more transparent and increasing the trust of end-users in their output. SHapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic Explanation (LIME) are two widely used XAI methods, particularly with tabular data. In this perspective piece, the way the explainability metrics of these two methods are generated is discussed and a framework for the interpretation of their outputs, highlighting their weaknesses and strengths is proposed. Specifically, their outcomes in terms of model-dependency and in the presence of collinearity among the features, relying on a case study from the biomedical domain (classification of individuals with or without myocardial infarction) are discussed. The results indicate that SHAP and LIME are highly affected by the adopted ML model and feature collinearity, raising a note of caution on their usage and interpretation.

accompanied by complexity in understanding how these models work, why the models make a specific decision, what features/regions are most influencing the model output, and the degree of certainty the model has in the generated outcome. All of these questions and more are raised by the end-users, especially when advanced models including deep neural networks are implemented. Accordingly, a new field of research has emerged named eXplainable artificial intelligence (XAI) aiming at demystifying “black box” models into a more comprehensible form.^[5] XAI is indispensable to increase the model transparency and the trust of end-users in the model outcome.^[6,7] Such additional reassurances are essential for the wide implementation of such models, particularly in high-risk fields such as healthcare.

However, specific aspects such as model-dependency and collinearity across the features might affect the quality of the XAI outcome. In this perspective article, we aim to reveal how model-dependency and the presence of collinearity affect the XAI outcome. Moreover, we use a case study from the biomedical domain to examine the effects of the


1. Introduction

Machine learning (ML) and deep learning methods have shown great success in a variety of domains, e.g., biology,^[1] medicine,^[2] economy,^[3] and education.^[4] However, such success is

A. M. Salih, Z. Raisi-Estabragh, S. E. Petersen
 William Harvey Research Institute
 NIHR Barts Biomedical Research Centre
 Queen Mary University of London
 London E1 4NS, UK
 E-mail: a.salih@leicester.ac.uk

A. M. Salih, Z. Raisi-Estabragh, S. E. Petersen
 Barts Heart Centre
 St. Bartholomew's Hospital
 Barts Health NHS Trust
 E1 1BB West Smithfield, London, UK

A. M. Salih
 Department of Population Health Sciences
 University of Leicester
 Leicester LE1 7RH, UK

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202400304>.

© 2024 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202400304

A. M. Salih
 Department of Computer Science
 Faculty of Science
 University of Zakho
 Kurdistan Region, Zakho 42002, Iraq

I. B. Galazzo, G. Menegaz
 Department of Engineering for Innovation Medicine
 University of Verona
 37129 Verona, Italy

P. Radeva, K. Lekadir
 Department of de Matemàtiques i Informàtica
 University of Barcelona
 08007 Barcelona, Spain

S. E. Petersen
 Health Data Research
 UK

S. E. Petersen
 British Library
 Alan Turing Institute
 96 Euston Rd., London NW1 2DB, UK

forementioned issues on two of the most common XAI methods. In addition, another case study was used to reveal how the XAI methods can be implemented and what are the possible solutions to overcome their limitations in terms of model-dependency and collinear features.

2. XAI

Several approaches have been proposed as XAI methods dealing with a variety of data and model types, aiming at explaining the models outputs locally and globally. Among these, SHapley Additive exPlanations (SHAP)^[8] and Local Interpretable Model Agnostic Explanation (LIME)^[9] represent the two most popular XAI methods based on the current literature in different domains.^[10] To further substantiate this and inspired by ref. [10], we considered the GitHub Star, an index used for quantifying the popularity of tools on GitHub and representing appreciation and usage of tools/projects. Moreover, most of the developers consider the stars before using a specific tool.^[11] Based on these considerations, we collected the GitHub Star for 10 popular XAI methods. As reported in **Figure 1**, SHAP and LIME represent the most exploited methods, both featuring an increasing number of stars. Accordingly, they were considered in this perspective for the discussion of the outcomes relying on different models in two case studies.

SHAP^[8] is an XAI method based on game theory. It aims at explaining any model by considering each feature (or predictor) as a player and the model outcome as the payoff. SHAP provides local and global explanations, meaning that it has the ability to explain the role of the features for all instances and for a specific instance. LIME^[9] is another XAI method that aims at explaining how the model works locally for a specific instance in the model. To this end, it approximates any complex model and transfers it to a local interpretable model for a specific instance. **Table 1** shows a direct comparison between both methods using different metrics. The tables show that SHAP has some advantages over LIME. SHAP considers different combinations to calculate the feature attribution while LIME fits a local surrogate model. Moreover, SHAP provides both global and local explanations while LIME is limited to local explanations only. In addition, SHAP might have the ability to

detect nonlinear associations (depending on the used model) while LIME fails to capture such associations because it fits a local linear model. In terms of visualization, SHAP generates several plots reporting the outcomes both locally and globally while LIME generates one plot per instance. Finally, LIME is much faster than SHAP, especially with tree-based models.

Besides the self-explaining properties mentioned in Table 1, it is worth pointing out that features collinearity and nonlinear dependency across features still impact the outcomes of both methods, limiting their reliability and, in consequence, trust. As for collinearity, even though in SHAP this issue is attenuated by the interplay of features in and across coalitions, it still remains unsolved. In particular, the Shapley method suffers from the inclusion of unrealistic data instances when features are correlated. To simulate that a feature value is missing from a coalition, it is marginalized, and missing values are obtained by sampling from the feature's marginal distribution. However, this makes sense only if features are uncorrelated.^[12] In LIME, the features are treated as if they were independent, calling for new solutions accounting for their interplay. Along the same line, nonlinear dependencies among features cannot be accounted for by LIME locally, being the local and linear surrogate model. Despite the limitations of SHAP and LIME in terms of uncertainty estimates, generalization, nonlinear dependencies (with LIME), feature dependencies, and inability to infer causality,^[13] they hold substantial value for explaining and interpreting complex ML models.

However, does the end-user understand how these XAI methods work? And why do they identify specific features as more informative than others? Is it enough for the end-user to know that these features are more informative because they improve the model output without knowing how the XAI method came up with such results? For example, when SHAP assigns a high/low score for a feature, does the end user know how this score was calculated? SHAP and LIME perform many analyses in the background and solve complex equations to come up with their explanations. In many settings, complex models will be interpreted by nonexpert end-users, who may find understanding the working of XAI methods challenging. It is not expected that the end-users from different domains understand every minutiae of XAI methods, but it is vital that they are aware of the general framework of the XAI method used. While XAI methods aim at unveiling the complexity of complex black box models, they suffer from the same issue, in that their usefulness may be limited by the difficulty in understanding their outputs. In this perspective piece, we discuss SHAP and LIME XAI methods, highlighting their underlying assumptions with the aim of helping the end-users to grasp their key concepts appropriately. We will also present some notions to increase the understanding of XAI methods and promote their appropriate usage by the researcher community.

2.1. SHAP

SHAP is a post-hoc model-agnostic method that can be applied to any ML model.^[8] It is based on game theory which calculates the contribution of each player to the payout. In ML models, the players and the payout are substituted by features and the model outcome, respectively. SHAP calculates a score for each feature in the model, which represents its weight to the model output.

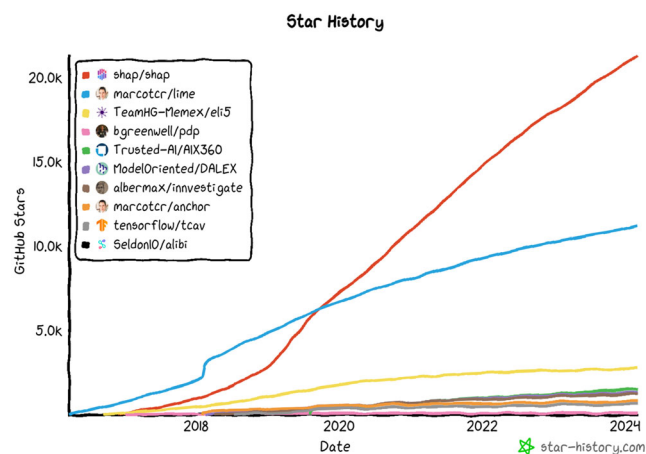


Figure 1. GitHub Star for 10 common XAI methods.

Table 1. Comparison between SHAP and LIME.

| Metrics | SHAP | LIME |
|----------------------------|---|---|
| Concept | Applies to the model as-is | Fits a local surrogate model to explain the complex model |
| Theory | Additive feature attribution based on game theory | Feature perturbation method |
| Type | Post-hoc model-agnostic | |
| Data type | Images, tabular data, and signals | |
| Explanation | Global, local | Local |
| Collinearity consideration | Not in the original method | No |
| Nonlinear decision | Depends on the used model | Incapable |
| Computing time | Higher | Lower |
| Visualization | Waterfall, beeswarm, and summary plots | One single plot |

To calculate the scores, it considers all combinations between the features (i.e., coalitions) to cover all cases where all features and a subset of features are used in the model. Due to the increases of computational complexity of SHAP when the number of features increases, an approximation has been proposed, named Kernel SHAP.^[8]

SHAP has been applied widely in a variety of domains to explain models' outcomes, either locally or globally.^[14–17] However, there are some important points the end-users should be aware of when applying SHAP. First, SHAP is a model-dependent method. This means that the SHAP outcome depends on the ML model used for the classification/regression task,

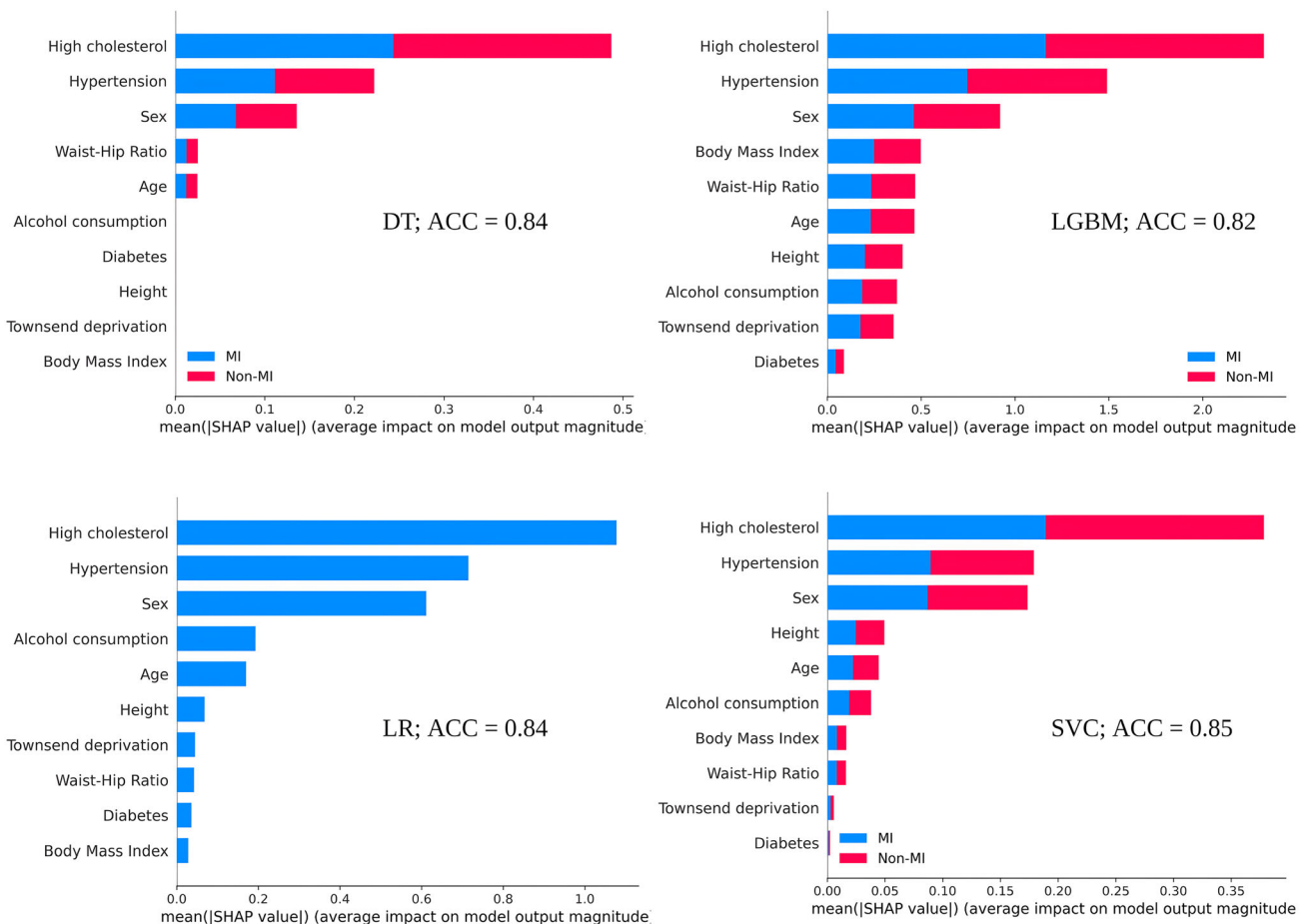


Figure 2. SHAP output to explain the four models globally. DT: decision tree; LGBM: light gradient-boosting machine; LR: logistic regression; SVC: support vector machines classifier; ACC: accuracy; MI: myocardial infarction.

which will possibly lead to different explainability scores. Accordingly, when different models are applied, to the same task using the same data, the top features identified by SHAP may differ between ML models.

To illustrate the model-dependency point, we used four ML models to classify 1500 subjects (20% test) from the UK Biobank into individuals with myocardial infarction (MI) and controls (non-MI). The included models are decision tree (DT), logistic regression (LR), light gradient-boosting machine (LGBM), and support vector machines classifier (SVC). Ten different variables were considered as features in the models. These models were implemented using Python (version 3.11.4), Scikit-learn library (version 1.3.0), and the codes of SHAP on GitHub. The code of the current perspective is available at (<https://github.com/amaa11/NMR>).

The order of the important features of the four classification models is reported in **Figure 2**. The figure ranks the features in order of importance based on their effect toward the model outcome. It can be appreciated that there is agreement for the top three most informative features among the tested models. However, there is a notable variation in the order of the remaining seven features. For instance, body mass index is the least important one in DT and LR, while it is the third in the LGBM model and the seventh in the SVC model. The position of alcohol consumption and Waist-Hip ratio similarly varies across the models. In addition, the last five features have a SHAP score close to zero in DT model, indicating they do not affect the model output. It is worth noting that despite the observed variance in feature order, the accuracy is comparable across the four classification models.

Second, another potential pitfall is related to the misinterpretation of the scores or SHAP values. The assigned scores do not represent the weight of the features with respect to the outcome, as their importance is encoded in the ranking. The end-users should focus on the order of the features which represent their significance. Third, SHAP is not protected against biased classifier and might generate unrealistic explanations that do not capture the underlying biases.^[18] Finally, SHAP assumes the features are independent, thus that there is no correlation between the variables included in the ML models. In the considered case study, most of the features are collinear including high cholesterol and body mass index. Such an assumption will affect the assigned score (weight) for each feature. Indeed, some features might be assigned a low score despite being significantly associated with the outcome. This is because they do not improve the model performance due to their collinearity with other features whose impact has already been accounted for. Although there are some works that tried to deal with the issue of collinearity,^[19,20] yet the proposed methods are either limited to a local explanation^[19] or the explanation is user-dependent.^[20] Another approach was proposed to assess the stability of the list of informative features generated by XAI methods, particularly when the features are collinear.^[21] The method calculates a value named normalized movement rate (NMR) which assesses how the order of the features will be affected when the top features are removed from the model iteratively. The smaller the NMR, the more the list of informative features is stable. The authors of NMR extended their work by presenting a new method to address the collinearity issues with XAI methods. The method is named

modified index position (MIP). It takes the outcomes (e.g., list of informative predictors) of any XAI (e.g., SHAP, LIME) and reorders them considering the multicollinearity.^[22] Unlike,^[19,20] the method does not require any intervention from the user and can be applied to any model. It works similarly to NMR by iteratively removing the top feature and retraining and testing the model. Thereafter, it examines how the features are reordered in the model which implies the effect of collinearity. More details on the method and how can be applied reflects at (<https://github.com/amaa11/MIP>).

2.2. LIME

LIME is a model-agnostic local explanation method.^[9] It explains the influence of each feature on the outcome for a single subject.

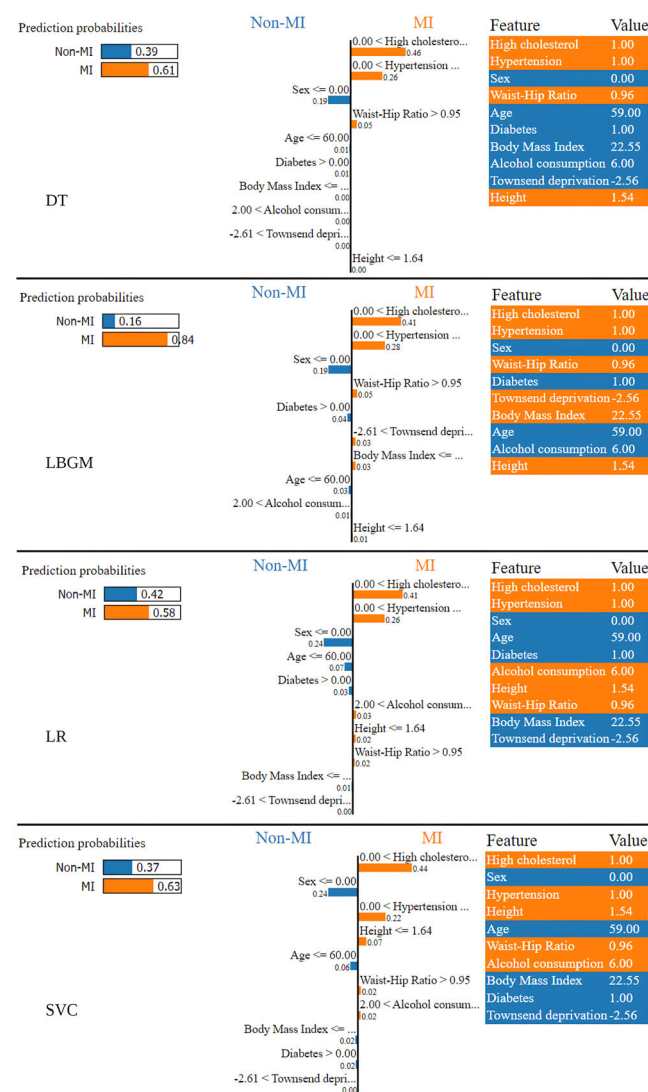


Figure 3. LIME output to explain the model locally for the same instance using four classifiers. DT: decision tree; LGBM: light gradient-boosting machine; LR: logistic regression; SVC: support vector classifier; MI: myocardial infarction.

In the classification models, it shows the probability that the subject might belong to any class. In addition, it shows the contribution of each feature in each class with a visualized plot.

However, LIME converts any model into a linear local model, and then reports the coefficient values which represent the weights of the features in the model. In other word, if the user applies some models that take into account the nonlinearity between features and the outcome, this might be missing in the explanation generated by LIME. This is because the nonlinearity is lost in the surrogate model. In addition, LIME is a model-dependent method, meaning the used model will affect the outcome of LIME for the same task and dataset. As for SHAP, we used the same case study to evaluate the list of informative features associated with the four classifiers. **Figure 3** shows the output of LIME for a representative subject. The first part of the plot (left) shows the probability that the subject is classified as control (Non-MI) or with MI in each of the used models. The second part (middle) shows the weight, i.e., coefficient value, of each feature in the local linear model, while the last part on the right shows the actual value of each feature. Moreover, the plot shows the features contribution toward each class based on the assigned color. In this case, the probability belonging to one or the other class is different for each of the used models. It shows that LGBM is the most certain, while DT is the least. In addition, the plot shows that the same feature is contributing to different classes across the tested models. For example, alcohol consumption

contributes to the MI class in LR and SVC, while it contributes to the non-MI class in the DT and LGBM.

Body mass index and Townsend deprivation contribute to the MI class in the LGBM model, while they contribute to the on-MI class in the other three models. In addition, the used features have similar effect sizes although four different models were used. This is due to the fact that LIME generates and approximates a local linear model and then reports the weights of the features.

Concerning collinearity, the interpretation of the weights generated by LIME indicates that an increase/decrease per one unit change in the feature will lead to an increase/decrease in the outcome while other features are kept unchanged. Such an assumption is not realistic with collinear data where groups of features might change simultaneously. It is indeed the correct interpretation for the coefficient values in linear models. But because they are generated by LIME, the user might think that they have more power and meaning than the classical coefficient values in the ML models. Finally, similarly to SHAP, LIME can be fooled by biased classifiers, leading to explanations that do not reflect or represent the biases.^[18]

3. A Case Study

The following case study illustrates the limitations of SHAP in terms of model dependency and collinearity, and the possible

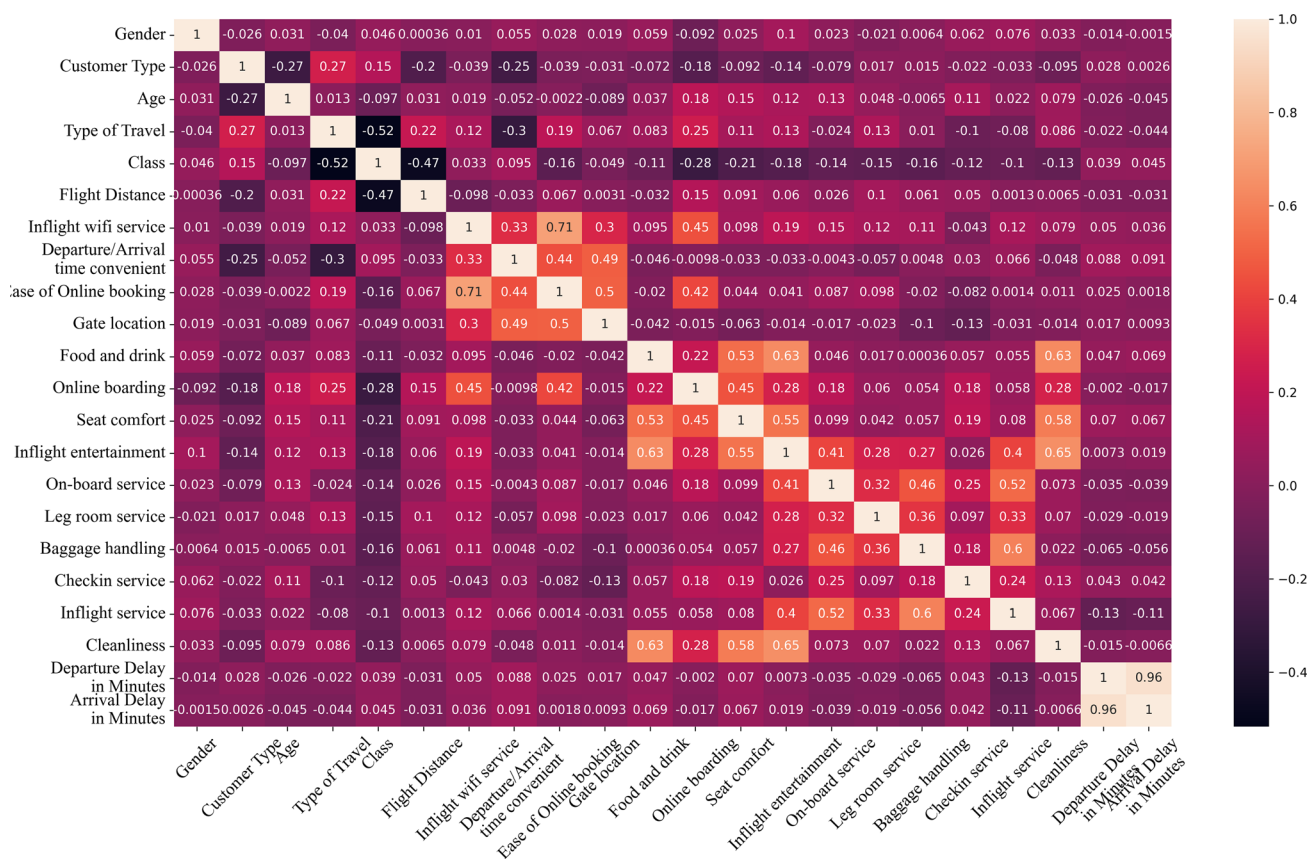


Figure 4. Correlation heatmap.

available solutions to overcome them. The case study can be extended to LIME as well as to any other XAI method. Airline Passenger Satisfaction data from Kaggle for 500 subjects (satisfied, $n = 250$) was used in the case study. Out of these, 22 features were used to predict whether the passenger was satisfied or not. Four classifications models were used that are LGBM, LR, DT, and SVC. The data were divided into training and testing (20%). The default parameters of each model were used. Thereafter, SHAP was applied to identify the most informative predictors for each model. **Figure 4** shows the correlation heatmap of the used features in the model. The figures show that there is collinearity between some features which will affect the outcome of XAI.

Table 2 shows the most informative features in each model generated by SHAP. It is noted that each model generated a different list of informative features although their accuracy was relatively similar apart from LGBM for which it is higher. It is worth mentioning that we cannot be certain that the LGBM is better than the other models because we used the default parameters of each model, and applying hyperparameter tuning might produce different accuracy for each model. The variation in the list was even observed in the top one where two models identified class as the most important one while the other two identified other features. The question is which one of these lists to consider given that the data are collinear and each model presented a

different list generated by SHAP. Some might argue that we should consider the outcome of SHAP with LGBM as LGBM reached the highest accuracy among other models. However, in the previous case, we showed that the accuracy of the models might be comparable with some data. NMR helps to examine which one of these models produced a more stable list against the collinearity. We have applied NMR to each model which produced the following results: LGBM: 0.231, LR: 0.275, DT: 0.445, and SVC: 0.273. Accordingly, LGBM has the lowest NMR value which indicates that the corresponding outcome is the most robust. NMR shows which model is more stable but it does not enhance the outcomes of SHAP to consider the collinearity. MIP then can be used to modify the outcome of SHAP and to obtain a list of informative features that consider the dependency among the features. The outcome of MIP for LGBM (with the smallest NMR value) alongside the SHAP outcome is explained in **Table 3**. The table shows that there is variation in each list. For example, Ease of Online booking is the fifteenth in the SHAP list while it is the fifth when MIP was applied.

We have applied MIP to produce a global list of informative features. In a similar way, if the aim is to provide a local list of informative features, then XAI should be applied locally for a specific subject followed by MIP. In addition, a local explanation for a specific individual can also be produced by applying the proposed method^[19] that modifies SHAP to consider collinearity.

Table 2. List of informative features produced by SHAP. LGBM: light gradient-boosting machine; LR: logistic regression; DT: decision tree; SVC: support vector machines classifier; ACC: accuracy.

| LGBM (ACC: 0.91) | LR (AC: 0.85) | DT (ACC: 0.84) | SVC (ACC: 0.86) |
|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Inflight wifi service | Class | Online boarding | Class |
| Type of travel | Online boarding | Inflight wifi service | Online boarding |
| Online boarding | Cleanliness | Type of travel | Type of travel |
| Class | Seat comfort | Class | Seat comfort |
| Cleanliness | Arrival delay in minutes | Cleanliness | Cleanliness |
| On-board service | Inflight wifi service | Age | Inflight wifi service |
| Departure/arrival time convenient | Customer type | Legroom service | Legroom service |
| Baggage handling | Departure delay in minutes | Customer type | Food and drink |
| Legroom service | Ease of online booking | Inflight entertainment | On-board service |
| Food and drink | Type of travel | Baggage handling | Arrival delay in minutes |
| Age | Gender | Gender | Ease of online booking |
| Customer type | On-board service | On-board service | Departure delay in minutes |
| Flight distance | Flight distance | Arrival delay in minutes | Gate location |
| Arrival delay in minutes | Legroom service | Gate location | Gender |
| Ease of online booking | Age | Flight distance | Inflight service |
| Seat comfort | Departure/arrival time convenient | Seat comfort | Baggage handling |
| Gate location | Inflight entertainment | Departure delay in minutes | Check-in service |
| Inflight service | Food and drink | Ease of online booking | Customer type |
| Check-in service | Inflight service | Inflight service | Flight distance |
| Departure delay in minutes | Gate location | Food and drink | Departure/arrival time convenient |
| Gender | Check-in service | Check-in service | Inflight entertainment |
| Inflight entertainment | Baggage handling | Departure/arrival time convenient | Age |

Table 3. List of informative features produced by SHAP and modified by MIP.

| SHAP | MIP |
|-----------------------------------|-----------------------------------|
| Inflight wifi service | Inflight wifi service |
| Type of travel | Online boarding |
| Online boarding | Type of travel |
| Class | Class |
| Cleanliness | Ease of online booking |
| On-board service | On-board service |
| Departure/arrival time convenient | Cleanliness |
| Baggage handling | Arrival delay in minutes |
| Legroom service | Inflight entertainment |
| Food and drink | Legroom service |
| Age | Food and drink |
| Customer type | Flight distance |
| Flight distance | Departure/arrival time convenient |
| Arrival delay in minutes | Seat comfort |
| Ease of online booking | Baggage handling |
| Seat comfort | Age |
| Gate location | Departure delay in minutes |
| Inflight service | Inflight service |
| Check-in service | Gate location |
| Departure delay in minutes | Check-in service |
| Gender | Customer type |
| Inflight entertainment | Gender |

4. Recommendations

SHAP and LIME are two popular XAI methods that aid in understanding ML models in different research fields. They have been implemented in some sensitive domains^[23–25] where misinterpreting the outcomes might be very expensive or critical. Data scientists who are working daily on ML and XAI tend to overtrust the explanations generated by XAI methods and do not accurately understand the visualized output of the XAI methods,^[26] that could result in a misuse of the interpretability tools.

It is crucial that SHAP results are presented alongside the corresponding output plots, presenting them with a simple language to explain the outcomes and the assumptions behind SHAP (e.g., features are independent and the outcomes are model-dependent). Moreover, if possible, the end-users should implement different ML models when dealing with collinear features to compare the SHAP outcomes across models and evaluate their robustness. Using post-hoc proxies such as the NMR^[21] value would be useful to select the model that presents the more stable list of informative features generated by any XAI method. MIP^[22] then can be used to enhance the outcome of XAI in the presence of collinear features if the aim is to explain the model globally. In contrast, if the aim is to explain the model locally for a single instance or subgroup of individuals, then MIP^[22] and approximated SHAP value (shapr)^[19] can be implemented.

This is because MIP can be applied to any XAI method and shapr is a modified version of SHAP, and both take into account the collinearity among the features. In addition, converting the scores of SHAP of each feature of the model (especially in classification models) into a more digestible form would increase the understanding of the score and ultimately the method itself. It is worthy to note that LIME provides explanation regarding the local model linearity with the model outcome as the users might not be familiar with the concept behind LIME. The users will be more aware and understand the outcome when a simple language accompanies the outcome. Moreover, the explanation of LIME might be different using the same model, but for other instances. In other words, the interpretation of LIME only applies to one subject and cannot be used or considered as a general interpretation for the whole model. Finally, GraphLIME^[27] was proposed as an updated version of LIME to explain graph-based models where nonlinear association is more appropriately considered.

5. Conclusion

In the current perspective, we discussed two widely used XAI methods especially with tabular data. The highlighted and discussed points are very significant and critical to be considered when XAI methods are implemented in any domain. Considering the end-users are not from a technical background, it is needful that they are aware of these issues to use the methods most appropriately.

Acknowledgements

K.L. and G.M. contributed equally to this work. A.M.S. is supported by a British Heart Foundation project grant (PG/21/10619). I.B.G. and G.M. acknowledge support from Fondazione CariVerona (Bando Ricerca Scientifica di Eccellenza 2018, EDIPO project - reference number 2018.0855.2019). Z.R.-E. recognizes the National Institute for Health and Care Research (NIHR) Integrated Academic Training Programme which supports her Academic Clinical Lectureship post and was also supported by the British Heart Foundation Clinical Research Training Fellowship no. FS/17/81/33318. S.E.P. acknowledges support from the National Institute for Health and Care Research (NIHR), Biomedical Research Centre at Barts and has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 825903 (euCanSHare project). This article is supported by the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH), which is funded by the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI). Views expressed are those of the authors and not necessarily those of the AI4VBH Consortium members, the NHS, Innovate UK, or UKRI.

Conflict of Interest

SEP provides consultancy to Cardiovascular Imaging Inc, Calgary, Alberta, Canada. The remaining authors have no disclosures.

Keywords

collinearity, interpretability, Local Interpretable Model Agnostic Explanation, SHapley Additive exPlanations, eXplainable artificial intelligence

Received: April 16, 2024

Revised: May 18, 2024

Published online: June 27, 2024

- [1] B. Richards, D. Tsao, A. Zador, *Cell* **2022**, 185, 2640.
- [2] P. Hamet, J. Tremblay, *Metabolism* **2017**, 69, S36.
- [3] E. A. Gyasi, H. Handroos, P. Kah, *Procedia Manuf.* **2019**, 38, 702.
- [4] K. Zhang, A. B. Aslan, *Comput. Educ.: Artif. Intell.* **2021**, 2, 100025.
- [5] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, *Sci. Rob.* **2019**, 4, eaay7120.
- [6] L. Szabo, Z. Raisi-Estabragh, A. Salih, C. McCracken, E. R. Pujadas, P. Gkontra, M. Kiss, P. Maurovich-Horvath, H. Vago, B. Merkely, A. M. Lee, *Front. Cardiovasc. Med.* **2022**, 9, 1016032.
- [7] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Daz-Rodriguez, F. Herrera, *Inf. Fusion* **2023**, 99, 101805.
- [8] S. M. Lundberg, S.-I. Lee, *Advances in Neural Information Processing Systems*, The MIT Press Number 12, Vol. 30, **2017**.
- [9] M. T. Ribeiro, S. Singh, C. Guestrin, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, USA* **2016**, pp. 1135–1144.
- [10] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, in *Int. Workshop on Extending Explainable AI Beyond Deep Models and Classifiers* **2022**, pp. 13–38.
- [11] H. Borges, M. Valente, *J. Syst. Software* **2018**, 146, 112.
- [12] I. Learning, *A Guide for Making Black Box Models Explainable* **2019**.
- [13] C. Molnar, G. König, J. Herbringer, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, in *xxAI-Beyond Explainable AI: Int. Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Springer, New York, NY **2022**, pp. 39–68.
- [14] M. V. Garca, J. L. Aznarte, *Ecol. Inf.* **2020**, 56, 101039.
- [15] Y. Kim, Y. Kim, *Sustainable Cities Soc.* **2022**, 79, 103677.
- [16] I. Ullah, K. Liu, T. Yamamoto, M. Zahid, A. Jamal, *Int. J. Energy Res.* **2022**, 46, 15211.
- [17] K. K. Pabodha, M. Kannangara, W. Zhou, Z. Ding, Z. Hong, *J. Rock Mech. Geotech. Eng.* **2022**, 14, 1052.
- [18] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New York, NY **2020**, pp. 180–186.
- [19] K. Aas, M. Jullum, A. Løland, *Artif. Intell.* **2021**, 298, 103502.
- [20] M. Mase, A. B. Owen, B. Seiler, *arXiv preprint arXiv:1911.00467*, **2019**.
- [21] A. Salih, I. B. Galazzo, F. Cruciani, L. Brusini, P. Radeva, in *2022 IEEE Int. Conf. Image Processing (ICIP)*, IEEE, Piscataway, NJ **2022**, pp. 4003–4007.
- [22] A. Salih, I. Galazzo, Z. Raisi-Estabragh, S. Petersen, G. Menegaz, P. Radeva, *IEEE J. Biomed. Health Inf.* **2024**.
- [23] N. George, E. Moseley, R. Eber, J. Siu, M. Samuel, J. Yam, K. Huang, L. A. Celi, C. Lindvall, *PLoS One* **2021**, 16, 0253443.
- [24] A. D. Haimovich, N. G. Ravindra, S. Stoytchev, H. P. Young, F. P. Wilson, D. Van Dijk, W. L. Schulz, R. A. Taylor, *Ann. Emerg. Med.* **2020**, 76, 442.
- [25] Y. Zhang, Y. Weng, J. Lund, *Diagnostics* **2022**, 12, 237.
- [26] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, J. W. Vaughan, in *Proc. 2020 CHI Conf. Human Factors in Computing Systems*, New York, NY **2020**, pp. 1–14.
- [27] Q. Huang, M. Yamada, Y. Tian, D. Singh, Y. Chang, *IEEE Trans. Knowl. Data Eng.* **2022**, 35, 6968.