The Scientific Temper (2024) Vol. 15 (2): 2225-2230
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.35
https://scientifictemper.com/

RESEARCH ARTICLE

A novel method for developing explainable machine learning framework using feature neutralization technique

Krishna P. Kalyanathaya[1*], Krishna Prasad K[2]

Abstract

The rapid advancement of artificial intelligence (AI) has led to its widespread adoption across various domains. One of the most important challenges faced by AI adoption is to justify the outcome of the AI model. In response, explainable AI (XAI) has emerged as a critical area of research, aiming to enhance transparency and interpretability in AI systems. However, existing XAI methods facing several challenges, such as complexity, difficulty in interpretation, limited applicability, and lack of transparency. In this paper, we discuss current challenges using SHAP and LIME metrics being popular methods for explainable AI and then present a novel approach for developing an explainable AI framework that addresses these limitations. This novel approach uses simple techniques and understandable human explanations to provide users with clear and interpretable insights into AI model behavior. Key components of this approach include model-agnostic interpretability, a newly developed explainable factor overcoming the challenges of current XAI methods and enabling users to understand the decision-making process of AI models. We demonstrate the effectiveness of the new approach through a case study and evaluate the framework's performance in terms of interpretability. Overall, the new approach offers enhanced transparency and trustworthiness in AI systems across diverse applications.

Keywords: Artificial intelligence, Machine learning, Explainable AI, XAI, Feature neutralization, LIME, SHAP.

Introduction

Recently, artificial intelligence (AI) systems have demonstrated remarkable capabilities across various domains such as e-commerce, healthcare, finance and many others. There is an increasing demand for safe, secure and trustworthy use of AI systems in the business domains. Explainable artificial intelligence (XAI) is a powerful tool to make AI applications safe, secure and trustworthy use of AI to demystify the complex inner workings of AI models and make their outputs more accessible to human understanding.

This paper explores the motivation behind the development of a new framework, the challenges encountered, and the potential impact on user engagement with AI systems. In the subsequent sections, we delve into the design principles that guide the development of novel methods, the integration of XAI techniques to ensure interpretability, and practical considerations for real-world applications. Through this work, we aim to contribute to the ongoing discourse on XAI, providing a tangible solution that bridges the gap between the complexity of AI models and the need for transparent, user-friendly interactions (Arrieta, A. *et al*., 2020).

XAI has emerged as a critical component in the development and deployment of artificial intelligence systems. As AI models grow in complexity, there is an increasing demand for transparency and interpretability to bridge the gap between the model's decisions and human understanding. XAI frameworks need to be designed with the end-user in mind. The challenge lies in presenting complex technical explanations in a comprehensible manner. The design should consider the diverse backgrounds and expertise of users, ensuring that the explanations cater to both technical and non-technical audiences. XAI frameworks typically consist of several

[1]College of Computer Science and Information Science, Srinivas University, Mangalore, India.

[2]Department of Cyber Security and Cyber Forensic, Institute of Engineering and Technology, Srinivas University, Mukka, Mangalore, India.

***Corresponding Author:** Krishna P. Kalyanathaya, College of Computer Science and Information Science, Srinivas University, Mangalore, India., E-Mail: krishna.prakash.kk@gmail.com

How to cite this article: Kalyanathaya, K. P. and Prasad, K. K. (2024). A novel method for developing explainable machine learning framework using feature neutralization technique. The Scientific Temper, **15**(2):2225-2230.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.35

Source of support: Nil

Conflict of interest: None.

Received: 08/04/2024 **Accepted:** 10/05/2024 **Published** : 15/06/2024

components such as a model interpreter, explanations generator, query engine and visualizations designed to enhance the interpretability and transparency of machine learning models (Kalyanathaya and Krishna Prasad, 2022).

In our previous work (Kalyanathaya and Krishna Prasad, 2022), we built a concept diagram of XAI framework that includes a query engine, model interpreter and model predictor to explain the model outcomes. This will explain the logical relationship between the observations (new data) and outcomes. The objective is to explain the outcome of the ML model in user user-friendly interface to business users.

An Overview of SHAP and LIME

Shapley additive explanations

Shapley additive explanations (SHAP) is a method used for explaining individual predictions from machine learning models. It is based on Shapley values from cooperative game theory, providing a theoretically grounded approach to understanding the impact of each feature on the model's output. SHAP values assign a contribution to each feature, indicating how much it influences the prediction compared to a baseline. This method offers global insights into feature importance and local explanations for individual predictions, enabling users to interpret complex models more effectively and build trust in their decisions (Lundberg, S. M., and Lee, S. I., 2017).

Local interpretable model-agnostic explanations

Local interpretable model-agnostic explanations (LIME) is another popular technique for explaining individual predictions, particularly for black-box models. It generates local approximations of the model's behavior around a specific prediction by perturbing the input data and observing changes in the output. LIME builds interpretable surrogate models, such as linear models or decision trees, to approximate the complex model locally. These surrogate models provide insights into how different features contribute to the prediction for a given instance, enhancing interpretability and transparency. LIME is widely used for its simplicity and flexibility in explaining a wide range of models (Molnar, C. 2022).

Following are the summaries of the study of the most recent literature:

- *Oblizanov, A. et al., 2023*

This research explores evaluation metrics for explainable AI global methods using synthetic data, shedding light on the challenges and advancements in assessing the performance of interpretable models and contributing to the understanding of their effectiveness and limitations. The authors propose the evaluation methods must be based on accuracy features, must have stable distribution and must be instance-guided. The study indicated that the accuracy of both SHAP and LIME methods degraded as the correlation coefficient between input features increased.

- *Huang, X. and Marques-Silva, J., 2023*

Two papers critique the adequacy and suitability of Shapley values for explainability, raising concerns and providing alternative viewpoints on their effectiveness in interpreting machine learning models.

- *Krishna, S. et al., 2022*

The paper delves into the inconsistency problem in explainable ML, reflecting on practitioners' perspectives and highlighting the implications for model interpretability and trustworthiness in real-world applications.

- *Fernández-Loría, C. et al., 2020*

The paper proposes the counterfactual approach to explaining data-driven decisions made by AI systems, presenting a novel perspective on interpretability that aims to enhance transparency and accountability in algorithmic decision-making processes.

- *Slack, D. et al., 2020*

The study investigates adversarial attacks on post hoc explanation methods like SHAP and LIME, revealing vulnerabilities that could undermine the reliability and trustworthiness of explanations generated by these techniques.

- *Janzing, D. et al., (2020)*

Addressing the causal problem, Janzing *et al*. discuss challenges in quantifying feature relevance for explainable AI, highlighting potential limitations in the interpretation of SHAP values and other methods.

- *van der Horst, J. et al., 2020*

The critical review examines SHAP values, offering insights into their strengths and weaknesses and suggesting areas for further research and improvement in understanding their role in model interpretability.

Each summary encapsulates the key points and contributions of the respective literature while maintaining uniqueness and brevity.

While SHAP has been primarily applied to tabular data, its applicability to other data types, such as text, images, or time-series data, may be limited. Adapting SHAP to handle these data types effectively remains an area of ongoing research and development. Despite these drawbacks, SHAP remains a valuable tool for providing model-agnostic explanations in XAI. Addressing these limitations and developing techniques to improve the scalability, interpretability, and robustness of SHAP will be crucial for advancing the field of explainable AI (Agarwal, N. and Das, S. 2020).

Our study summarizes the following challenges that exists in the current approaches used by methods and tools for interpretability and explainability. While SHAP (SHapley Additive exPlanations) metrics are widely used in XAI to provide insights into the contribution of individual features to model predictions, they also have certain drawbacks. Some of the drawbacks of SHAP metrics include:

Computational Complexity

SHAP computations can be computationally expensive, especially for complex models and large datasets. As SHAP values require evaluating model predictions for every possible combination of features, the computational burden increases exponentially with the number of features and instances. This can limit the scalability of SHAP, particularly in real-time or resource-constrained applications.

Interpretability Challenges

While SHAP values provide explanations for individual predictions, interpreting these values can be challenging, especially for non-technical users. Understanding the impact of multiple features on a prediction and the interactions between them may require domain expertise and may not always be intuitive.

Dependence on Model Complexity

SHAP values may not always provide meaningful explanations for highly complex models, such as deep neural networks. In such cases, interpreting SHAP values may be less straightforward, and the explanations may not fully capture the underlying decision-making process of the model.

High-Dimensional Data

SHAP may face limitations when dealing with high-dimensional data, where the number of features is large relative to the number of instances. In such scenarios, interpreting SHAP values and identifying meaningful patterns or insights from the explanations may become more challenging.

Model Sensitivity

SHAP values can be sensitive to the choice of model architecture, training data, and hyperparameters. Variations in these factors can lead to different SHAP values, impacting the consistency and reliability of the explanations provided by SHAP.

Assumption of Feature Independence

SHAP values assume feature independence, meaning that the contribution of each feature to the model's prediction is considered in isolation. In reality, features may be correlated or interact with each other, leading to potentially misleading or incomplete explanations from SHAP.

Potential for Misinterpretation

There is a risk of misinterpretation or misrepresentation of SHAP values, particularly when communicating explanations to end-users. Without proper context or guidance, users may draw incorrect conclusions or make decisions based on incomplete or misleading information derived from SHAP explanations.

Limited Support for Non-Tabular Data

While SHAP has been primarily applied to tabular data, its applicability to other data types, such as text, images, or time-series data, may be limited. Adapting SHAP to handle these data types effectively remains an area of ongoing research and development.

Materials and Methods

Based on the literature review, we summarize the challenges of using SHAP and LIME metrics:

Inadequacy and Refutation

Some researchers argue about the inadequacy and refutation of Shapley values for explainability, highlighting potential limitations or drawbacks in its application. (Huang and Marques-Silva 2023) presented papers critiquing the suitability of Shapley values for explainability.

Assumptions of Feature Independence

The existing methods in explainable machine learning are based several assumptions, which can impact the reliability and consistency of SHAP metrics (Krishna *et al*., 2022).

Adversarial Attacks

SHAP, along with LIME, is susceptible to adversarial attacks, which can undermine the trustworthiness of the explanations generated by these methods (Slack *et al*., 2020).

Feature Relevance Quantification

(Janzing *et al*., 2020) point out that quantifying feature relevance in explainable AI poses a causal problem, suggesting potential challenges or limitations in accurately interpreting and utilizing SHAP values for this purpose.

Critical Review

There are critical reviews of SHAP values in the literature, indicating that despite their popularity, there may be aspects of SHAP metrics that require further scrutiny or refinement (van der Horst *et al*., 2020).

These challenges collectively underscore the need for careful consideration and ongoing research to address limitations and enhance the effectiveness of SHAP metrics in explaining machine learning models. SHAP is model agnostic only in theory but requires separate implementation for each algorithm. For example, SHAP implementation built for XGBoost will not work for the random forest algorithm. SHAP implementation built for neural networks will not work for non-neural network-based models. SHAP replaces the features with random values to compute the deviations from the mean predictions. So there is an unknown effect of the random values impacting the contributions in features (both prediction and correlation).

Experiments on the New Approach

We propose a new method to replace a feature and compute the deviations of prediction to the original prediction. We can try this approach in the following two ways:

- Physically remove the feature in the dataset and retrain the model

- Neutralize the feature by setting a uniform value in the dataset and run the prediction (without retraining the model)

The first method seems to be complex and has technical implications, as retraining the model may alter prediction patterns (model parameters) from the original. The second method seems to be simpler and we can compare the two performances of the same model with the observed value and constant value of the features. This means model parameters are not altered as we are not retraining the model.

We start with the following proposition of defining an ML prediction problem having n features and the goal of producing classes of outputs (typical classification problem). It can be stated as follows:

P0 = F(x1,x2,x3,x4,……. Xn)

Where x1, x2, x3, x4 …. are the features P0 is the probability value of the predicted class

F is a prediction function using ML algorithm with the features.

The Figure 1 shows formulation of solution which can be described as follows:

Step 1: Choose features one at time from a list of features x1, x2,x3,x4

Step 2: Identify a method to neutralize the effect of the chosen feature in the prediction model as follows:

P1 = F(x1, 0,x3,x4,x5…..xn)

Where P1 is the probability score of the predicted class.

Here, feature x2 is neutralized with a value (typically a constant value) that will produce a uniform effect of the feature for all the observations. Hence, the neutralization makes the features dummy in the prediction. The objective of neutralization is to measure the deviations in the prediction with a constant value of the feature to the prediction with the observed value of the feature.

Step 3: Identify a measure that quantifies the influence of the features on the outcome. We can define the measure as the deviation of the relative change in the outcome to the original outcome. It can be calculated as follows: P0 – P1, where P0 is the probability score of the original predicted class with observed feature value, P1 is the revised probability score of the predicted class with a constant feature value.

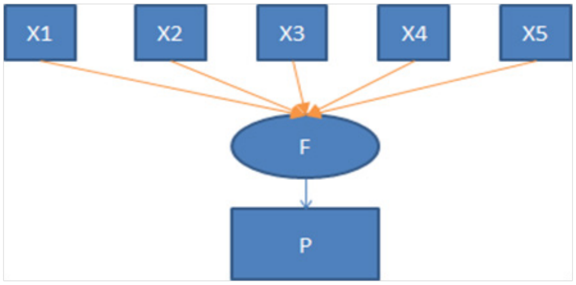


Figure 1: A symbolic formulation of the prediction

We will repeat the calculation for the sample of validation data on the same model F. The deviations of the probability scores are recorded in the dataset for each feature. The deviation can be plotted, showing maximum variations (or swings) in the probability scores due to neutralized features. The longer the swing, the higher is the impact of the feature.

We can formalize the approach as follows:

Feature neutralization impact score is defined as the interquartile range (IQR) value of measures of deviations of revised prediction probability score (with a neutral value of the feature) to the original probability score (with an observed value of the feature).

The high-level algorithm for the calculation of the feature neutralization impact ratio (FNI ratio) for a classification type of ML model as follows:

Algorithm

Feature neutralization impact (F as ML model, X as Sample data):

Let F be the prediction function (typically called as ML model) trained using historical data

Let X be the set of validation data collected from sample real-time data for testing purposes.

Let D be the variable to store the measure of deviation.

Let P0 = probability score of the prediction with original sample data Let FNI is the collection of FNI(f) for all the features in X

For each feature f in the feature set of sample validation data X: For the sample validation data X

Let f1 = neutralization feature with an assigned constant value in the revised sample data

Let X1 be the revised copy of X with neutralization of the feature f1

Let P1 = the probability score of the prediction with revised sample data X1

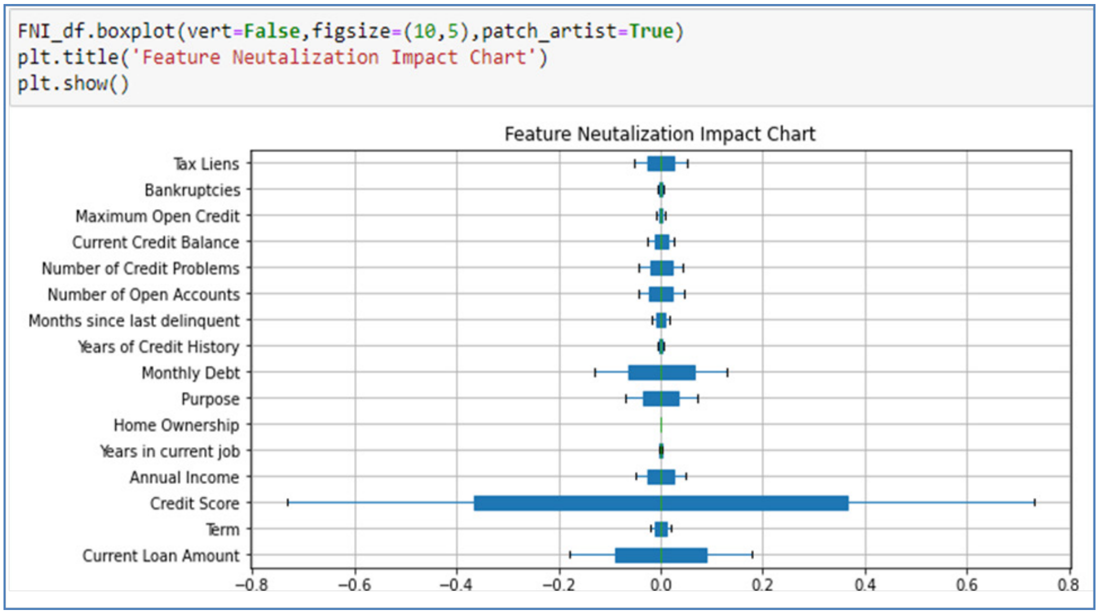
Let D(f) = P0 – P1 Let FNI(f) = IQR(D(f))

Let FNI = Collection of FNI(f) Return FNI

The collection of FNI(f) values can be plotted using a boxplot and we can observe the range of D(f) for each feature.

Results and Discussions

We take a credit approval problem with a set of features to predict whether a credit can be approved or rejected based on the features evaluated by the ML model. We set up an experiment using a credit dataset and evaluated the results using test samples. We will use the probability score instead of the accuracy score for the algorithm. This is because we have observed probability score provides more insight of the variations of the impact of feature neutralization instead of the averaged accuracy score of validation data. We have used credit data to train a model by using a scikit-learn package of random forest algorithm.



```
FNI_df.boxplot(vert=False,figsize=(10,5),patch_artist=True)
plt.title('Feature Neutalization Impact Chart')
plt.show()
```

Figure 2: A box plot of class probabilities for the sample validation data

Here, credit_asmt_model_base is the trained model using credit data. X_test_base is the sample validation data. The probability score for each credit record in X_test_base is calculated using the pred_proba() method of the random forest classifier object in the scikit-learn package (Pedregosa, F. *et al.*, 2011). So, we get the class probabilities for each prediction outcome.

The code snippet for getting class probability values are as follows: *pred_proba_original=credit_asmt_model_base.predict_proba(X_test_base)* Next, we create a copy of the validation data and replace each feature with neutralizing values one by one. The calculated class probabilities for each feature neutralization step as follows:

for idx, feature in enumerate(cols_list):

New_X = X_test_base.copy()

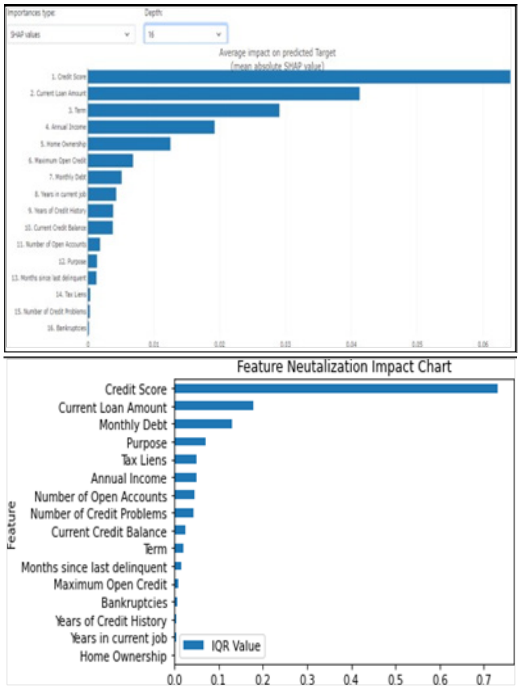
New_X[feature] = 0

pred_proba_revised= credit_asmt_model_base.predict_proba(New_X) FNI_df[feature] = pred_proba_original[1] - pred_proba_revised[1]

The FNI_df contains the deviations in class probabilities for each neutralized feature. The collection of FNI(f) values can be plotted using boxplot as shown in Figure 2.

Figure 2 shows the inter quartile range (IQR) that measures the deviation from the original probability score to the revised score after feature neutralization. The wide band (high IQR) indicates the variation (or change) is significant due to the feature neutralization. Hence, we can consider the larger the IQR, the higher will be impact of the feature on the outcome.

We also present here a comparison result for SHAP and the new method for explainable factors on a model (global interpretations) depicted in Figure 3 as follows (Table 1):



Importances type: SHAP values

Depth: 16

Figure 3: Comparison of feature importance depicted in SHAP and new method

Table 1: Explainable factors on the model

Explainable factors	*SHAP method*	*New method (Feature neutralization impact)*
Feature importance of the model score	Shap values plotted with bar chart on sample set	The Feature Neutralization Impact score on a box plot on the sample set

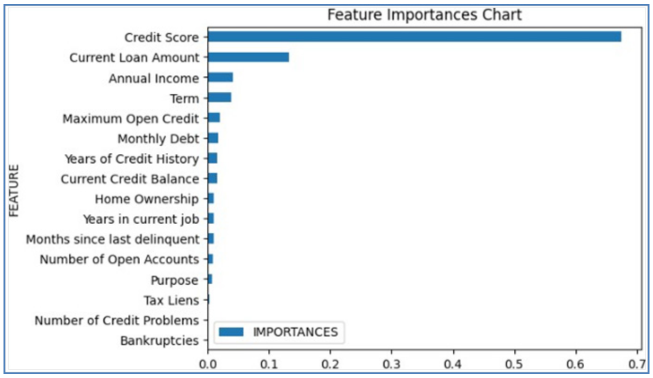


Figure 4: Feature importance depicted by random forest

To verify the results of the neutralization method, we have generated the feature importance of random forest ML algorithm implementation is shown in Figure 4.

Comparing Figures 3 and 4, we can see that features displayed with bands have shown similar results with top 5 features.

Further, we can use the same approach for regression or forecasting type of ML model by using the error values instead of mean square error methods like RMSE or MSE scores.

Conclusion

So we have studied the various challenges in using SHAP and LIME approaches to XAI and proposed and a new approach to overcome the challenges. The new approach has the following advantages when compared to SHAP and LIME methods:

- No technical knowledge is required to assess the model. The method uses a simple formula of relative change.
- This method does not require separate implementation for each machine learning algorithm.
- The same formula can be applied to any model after the training with data.
- Computational complexity doesn't increase with the increase in number of features.

Further, we would like to propose future development of this approach in the following areas:

- Study the implementation of this approach with text and image data used in AI models.
- Develop applications of this approach for testing and validation of AI models.
- Propose improvement methods to increase the trustworthiness and responsible AI metrics to AI governance.

Acknowledgments

Nil

References

Agarwal, N., & Das, S. (2020, December). Interpretable machine learning tools: A survey. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1528-1534). IEEE.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

Huang, X., & Marques-Silva, J. (2023). The Inadequacy of Shapley Values for Explainability. arXiv preprint arXiv:2302.08160. https://arxiv.org/abs/2302.08160

Kalyanathaya, Krishna Prakash, & Krishna Prasad, K., (2022). A Literature Review and Research Agenda on Explainable Artificial Intelligence (XAI). International Journal of Applied Engineering and Management Letters (IJAEML), 6(1), 43-59. DOI: https://doi.org/10.5281/zenodo.5998488

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602. https://arxiv.org/abs/2202.01602

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

Molnar, C. (2022). Interpretable Machine Learning:A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable- ml-book/

Oblizanov, A., Shevskaya, N., Kazak, A., Rudenko, M., & Dorofeeva, A. (2023). Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data. Applied System Innovation, 6(1), 26. https://www.mdpi.com/2571-5577/6/1/26

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830. https://scikit- learn.org/0.16/

Slack, Dylan, 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, pp. 180-186 (2020)

van der Horst, J., Wouters, K., & Gerckens, M. (2020). "A critical review of SHAP values." In Proceedings of the 2020 ACM Southeast Conference (pp. 153-156).