Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Nneke, Ngozi, Lloyd, Huw and Crockett, Keeley (2023) Quantification of explainability in black box models using complexity measures. In: 2023 15th International Conference on Innovations in Information Technology (IIT), 14 November 2023 - 15 November 2023, College of Information Technology (CIT), United Arab Emirates University.

Publisher: IEEE

Version: Accepted Version

Downloaded from: https://e-space.mmu.ac.uk/633363/

Usage rights: © In Copyright

Additional Information: © 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enquiries:
If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

Quantification of Explainability in Black Box Models using Complexity Measures

*

Ngozi Nneke
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, UK
Orcid: 0000-0003-3859-2542

Huw Lloyd
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, UK
Orcid: 0000-0001-6537-4036

Keeley Crockett
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, UK
Orcid: 0000-0003-1941-6201

***Abstract*—As a result in the rapid growth of explainability methods, there is a significant interest, driven by industry to develop methods for quantitative evaluation of such explanations. The availability of standard explainability evaluation methods would result in the ability to develop models that suit different stakeholders in different use cases. To address this issue, we propose three measures of the complexity of explanations based on Linear correlation, Monotonicity and ϕ_K. We evaluate these measures on three tabular datasets (Ames House Price, Auto Price, and Wind). We investigate how these complexity measures vary with model accuracy. Our results show that model accuracy varies with complexity measures across the datasets. These variations indicate that models can be developed with the same accuracy but with of models less complex explanations as a result of varying the hyperparameters. We observe a trade-off between complexity measures and model accuracy which is evidenced in Pareto-fronts. We suggest that our metrics could be used for the development of multi-objective optimisation methods for machine learning models with tunable accuracy and simplicity of explanation.**

***Index Terms*—Explainable Artificial Intelligence, Model Accuracy, Explainability, Monotonicity, Linear Correlation, ϕ_k, Complexity Measures, Evaluation, Quantification**

I. INTRODUCTION

The rapid penetration and extensive adoption of Artificial Intelligence systems in critical applications such as healthcare, judiciary, employment, and finance has demonstrated how indispensable these systems are becoming in our world. The high predictive power of such systems makes them attractive to increase efficiency, however, a consequence is that they are often "black box" and not explainable to human stakeholders. Examples of such systems include AlphaGo, which outperformed humans in playing online games [1], rapid diagnosis of diseases by AI systems that are typically difficult to detect by health professionals [2], [3] self-driving cars, question–answer systems and medical assistance systems [4]. Although we know that these systems are powered by black boxes, such systems may be deemed harmful rather than beneficial to the public [5], [6] as a result of a lack of transparency on the algorithmic behaviour of the AI systems. Previous studies have shown that decisions from opaque systems (i.e., systems that hide information on how it makes a decision) are sometimes misleading and characterised by trust issues [6], [7]. Consequently, there has been an impetus to make AI systems more explainable. Furthermore, there is an increasing need, driven by legislation and regulatory bodies, to make every AI system transparent and understandable [8], [9] to all stakeholders from within the developer pipeline, operators in the field and the users themselves [4].

To address the need to explain black box models, numerous explainable (XAI) methods and frameworks have been developed and have been categorised into inherent interpretable models and post hoc methods [4]. Typical examples of explanations are permutation feature importance, partial dependence plot, and global and local explanations [10]. However, these explanations have been subject to criticism. Studies have argued that many of the explanations fail to meet users' objectives because they seem to be inaccurate or deceiving [11], [12]. No standardized objective metrics to evaluate explanation methods [13] currently exist. As a result, the need for explainability evaluation escalated [14], [15], [16] and several methods were developed. As a consequence, previous scholars categorised explainability methods into qualitative and quantitative evaluation methods [13], [17]. Evaluation of explanations prior to deployment becomes a necessity to build the trust and confidence of the public in the use of AI systems. Initially, qualitative measures tend to be the preferable option for stakeholders to assess the quality of an explanation however they been criticised in the literature as anecdotal evidence "showing individual convincing examples that pass the first test of having face validity" [18], [19] where the evaluation is based on the "researchers intuition of what constitutes a good explanation" [20]. Other researchers argue that the outcome of such assessments tend to be biased and subjective because humans could change their opinions. In addition, the process can be time-consuming and expensive [21] when compared with quantitative evaluation. On the other hand, quantitative explainability evaluation does not involve users to assess the quality of explanations rather it adopts criteria of "some measurements that serve as proxies" in evaluating explanations [4], [18]. Quantitative measures seem to be faster in evaluation and more effective when compared with qualitative measures. Recently, a handful of quantitative (functionally – grounded i.e. – no human in the loop) metrics have been suggested such as Faithfulness [22], Localization accuracy [22], Completeness [16], Stability [23] and Sensitivity [22] to measure properties of interest of XAI methods [4]. Although these metrics are gradually being tested in other contexts [22], a recent study has argued that these metrics were introduced as a result of assessing properties of interest in XAI methods [4] which serves as a limiting factor because the properties that were assessed are domain dependent and could not be generalised. The XAI community is still far from reaching an agreement for common quantitative metrics to automate the efficacy of XAI methods because most of the existing quantitative metrics are customised (i.e., based on context, e.g., Healthcare) and cannot be generalised to other explainability methods which have been highlighted [24]. Within explainability research, both XAI methods and evaluation metrics are expanding rapidly, and there is a need to arrive at common quantitative measurement scales [17]. Considering the need to have a generalised evaluation metric, SHAP (Shapley Additive Explanations) [25] is to be the most widely used post hoc explainability method to create explanations across domains [13].

In this paper we propose complexity measures based on Linear correlation, Monotonicity, and ϕ_k, to determine the complexity of explanations. We hypothesise that such complexity measures can be used as proxy measures of the explainability of the model. SHAP as an explainability method is selected because of its popularity in high stakes domains, its widespread use in explaining black box models [12], and because of its grounding in theory, in contrast to other explainability methods [24]. Since SHAP is utilised to interpret the inner working of the machine learning models, rather than attempting to measure the complexity of the models themselves, we use measures of the complexity of the SHAP explanation since this is the way the model is presented to the user and can be used as a proxy for the complexity of the model. It is worth noting that this study is not measuring explainability, which is best measured in a human-centred study, rather we are proposing some measures of the complexity of explanations. We used feature importance-weighted average of these measures over the SHAP explanations for each feature to give a single measure of complexity for a model. Using this idea, the study aims to answer the following research questions:

RQ1: How do feature importance-weighted averages of typical correlation measures vary between different instances of a black box model?

RQ2: Can we observe a trade-off between these measures of the complexity of explanations and the model accuracy?

A positive answer to RQ2 opens the possibility of tuning models for explainability; with a quantitative measure of the complexity of the explanation, and a method for determining the Pareto-optimal set in the trade-off between explainability and accuracy. Models could be selected based on objective criteria for different use cases. Using our three-correlation metrics, we will examine across three open-source datasets the trade-off between explainability and accuracy. Furthermore, we introduce some relevant terminologies for uniformity understanding within the scope of this study.

Previous scholars are yet to reach a consensus about definitions of Explainability and interpretability [18], [26], [27]. Some researchers believed both terms differ [15], [28], [29] whilst other scholars used them interchangeably [16], [18], [30], [31]. However, this study aligns with the latter to postulate that the two terms will be used interchangeably since both definitions' goal is to provide transparency in AI systems. In the research presented in this paper, we defined explainability as an act of explaining any AI systems' predictions to real people (i.e., with the purpose of building public trust. Similarly, quantifying explainability means the process of assessing explanation through the lenses of objective or subjective measures. This study contributes to the growing knowledge of explainability evaluation by providing empirical evidence on the quantification of the complexity of explainability in Black box models. To the best of our knowledge, this is the first study that employs objective metrics such as Linear Correlation, Monotonicity and ϕ_k are used to measure the explainability complexity of the SHAP method.

This paper is structured as follows: section II briefly reviews the related work on model accuracy and explainability. Section III describes the experimental methodology of the empirical study using three measures of the complexity of explainability across three open-source datasets. Section IV discusses the experimental results and finally, section V presents the conclusion and directions for further work.

II. RELATED WORK

A. *Trade-off Between Model Accuracy and Explainability*

In recent years, there has been a growing number of publications focusing on the trade-off between model accuracy and explainability in diverse contexts which is traceable to the theoretical assumption that the higher the predictive power of machine models, the lower the explainability and vice versa [12], [15], [32]. In the existing literature, model accuracy has been assumed as a good indicator of high predictive performances of black box models which makes such models attractive, yet there are cases where explainability is more desired over model accuracy [15]. Sometimes, the decision to select a model based on its accuracy on unseen data over a high explainability model seems to be dependent on the type of data (context), the context of the decision and the background knowledge of the user [18]. Miller [26] argues that explainability is contextualised meaning that what users want in an individual case differs across domains [26]. In critical domains like healthcare, machine models with high predictive power have been in demand because some researchers believed that not all AI systems need to be transparent [11], [12]. Molnar [33], in their groundbreaking paper, proposed three measures: number of features, interaction strength and main effect complexity to demonstrate a trade-off between performance and post-hoc interpretability. This could only be achieved by minimising these measures which tend to improve the interpretability of machine learning models. While Molnar's paper focuses on model complexities using the mentioned measures, this paper measures the complexity of the explanations. Research has supported the opinion that high predictive models are sought after in high-stakes domains like healthcare, finance, and judiciary where the accuracy of decision-making is preferred [34]. Findings from a study in healthcare with citizen juries indicate that high-performing machine models' accuracy was preferred over model transparency [34] as a result of the need for increased efficiency of healthcare services. In contrast, the citizens' juries voted for explainable machine learning models in low-stakes domains [34]. In a related study with physicians, an explainable machine model decision was preferred on the ground that trust was important to them [35]. Building trust with end-users could be achieved when the underlying mechanisms of any AI systems decisions are comprehensible to the users [35], [36]. Aside from healthcare, a recent study by Bell [37] used two real-world policy datasets to demonstrate a trade-off between model accuracy and explainability and the finding indicate no "direct trade-off between accuracy and explainability nor found interpretable models to be superior in terms of explainability" [37]. This study has been the closest to our study to date, but they focus on binary classification tasks rather than regression. In a similar study, Herm [38] empirically explored this trade-off in a user-centred study where the outcome indicated that the theoretical assumptions or recommendation cannot be generalised across contexts rather it seems to be a situational concept, meaning that end-users' perception and context under study tend to affect the trade-off. As a result, this theoretical assumption needs more empirical work to widen the discourse about model accuracy versus explainability trade-off. Overall, our study indicates the importance of further examining the trade-off in different contexts.

III. EXPERIMENTAL METHODOLOGY

In this study described in this paper, we adopt an empirical experimentation approach to investigating complexity measures of explanations. The experimental methodology of this study is largely categorised into four stages which will be explained accordingly. Figure 1 depicts the strategy adopted in this study. Each stage is briefly defined as defined as follows:

Stage 1: Datasets selection, pre-processing of the datasets and modelling of the Random Forest Regressors (black box models).

Stage 2: Model Prediction explanations using SHAP Global explanation method.

Stage 3: Quantification of complexity measures of SHAP explanations.

Stage 4: Evaluation of the trade-off between explainability and model accuracy.

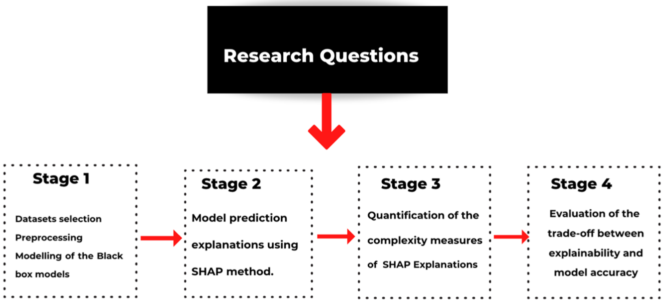
Research Questions
Stage 1
Datasets selection
Preprocessing
Modelling of the Black box models
Stage 2
Model prediction explanations using SHAP method.
Stage 3
Quantification of the complexity measures of SHAP Explanations
Stage 4
Evaluation of the trade-off between explainability and model accuracy

Fig. 1. Overview of the experimental methodology

A. Datasets Selection, Pre-processing, and modelling of the Random Forest Regressor

This section discusses how the datasets have been selected, and pre-processed and Random Forest Regressor models have been trained. In this paper, we selected the Random Forest Regressor because it is a widely used black box model.

1) Dataset Selection and Pre-processing: We employed datasets taken from the widely used repositories Kaggle and Penn machine learning benchmark namely Ames house price, Auto price and Wind [39] which are widely used and well-known to researchers working on tabular regression. Table II summarises the characteristics of the datasets used in this paper. Standard pre-processing techniques were applied to clean the data and perform feature selection. For each dataset, the correlation coefficient [40] with a threshold of 0.5 was used to select the best-correlated features with the target variable. Previous researchers have argued that the selection of the correlated features with the target variable improves the model predictions [41], [42].

2) Modelling of the Random Forest Regressor: The selected features, which are those with a correction coefficient with the target > 0.5, were used in the next phase of training the Random Forest (RF) models. The train test split method was used to evaluate the models, with the data randomly split into train and test set in a ratio of 80 / 20. We trained 1000 models with hyperparameters randomly selected from the ranges given in Table II. For each model, we calculated R^2 and the feature importance-weighted average of the correlation measures for the test data, and plotted these as scatter plots.

B. Model Predictions' explanations using SHAP Method

Predictions from the trained models were explained using the SHAP method. The SHAP method is classified into two types: Local explanations and Global explanations [10]. Local explainability is interested in explaining a specific instance of model predictions whereas Global explanation support how Machine learning models make predictions from a global point of view.

1) SHAP Global explanation Method: In this study, we take SHAP an acronym for Shapley Additive exPlanations is a typical explainability method that is rooted in the famous Shapley values which was introduced by Lord Shapley in 1951

TABLE I
OVERVIEW OF DATASETS

Dataset	Description	Instances	Instances used	Features	Features used
Ames House Price [43]	Ames house price is a well-known dataset in the community. This dataset describes the sale of properties in Ames, Iowa from 2006 to 2010.	1460	1460	81	10
Auto Price [39]	Auto Price is a small dataset from a curated set of benchmarking datasets from the PMLB repository.	159	159	16	9
Wind [39]	Wind Dataset is from the public repository of Penn Machine Learning Benchmark (PMLB)	6574	1500	15	10

TABLE II
RANDOM HYPERPARAMETERS TUNING FOR RANDOM FOREST MODELS

Parameters	Settings
N-estimators	start = 10, stop = 300
Max-features	auto, sqrt, log2
Max-depth	2, 4, 6, 8, 12, 16
Min-samples-split	2, 5, 10, 20
Min-samples-leaf	2, 5, 10, 20
Booststrap	True, False

[25], [10]. Shapley values is a theoretical framework in which players' collective payoff or contribution were strategically shared according to how each of the players contributed to the game [44]. Furthermore, SHAP explanations could be used for a single prediction as well as the global prediction of any ML model predictions.

2) SHAP Feature importance weighted averages: The SHAP values for a given feature, plotted against the feature value for each instance in a dataset is referred to as a SHAP Dependence Plot. We compute a measure of the complexity of this distribution of SHAP values versus feature values for each feature, and then combine these into a single value by weighting with the global feature importance and averaging. Feature importance is defined as a process of determining which feature is more important than another in a model prediction. Similarly, the feature importance depends on how "important" the feature is in the model prediction. In this study, feature importances are used as the weights in the average when combining complexity measures for the different features. Figure 2 shows the SHAP feature importance for the random forest trained on Ames House Price dataset [43]. For each dataset, feature importance-weighted averages of the complexity measures over the SHAP explanations for each feature were calculated to give a single measure of complexity for a model. The values from the feature importance-weighted average scores are then compared with the model accuracy for each dataset.

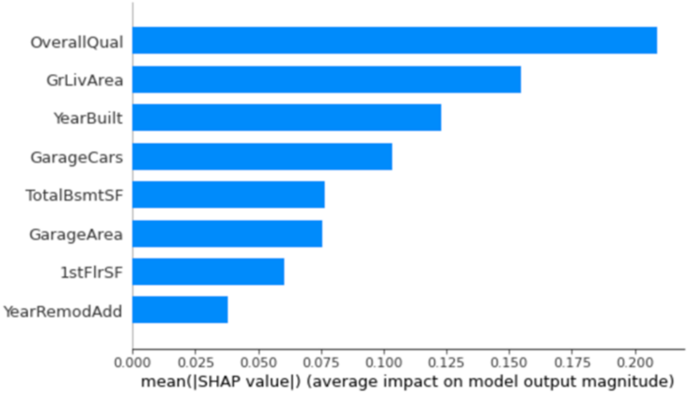

Fig. 2. A typical SHAP feature importance plot

C. Quantification of complexity measures of SHAP explanations

In this section, the implementation of the complexity measures (Linear Correlation, Monotonicity and ϕ_k measures) of SHAP explanations adopted in this study will be defined.

1) Linear Correlation: Linear correlation as should be is a measure to determine correlation of association in two or more random numbers. In this study, we adopted this measure along with two correlation measures (Monotonicity and ϕ_k) to measure complexity of explainability in black box models. The linear correlation coefficient $\rho_{x,y}$ between two variables x and y is given in equation 1

$$\rho(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \tag{1}$$

where $\text{Cov}(x, y)$ is the covariance between the two variables, and σ_x, σ_y are the standard deviations of each variable. The correlation coefficient takes values in the range $\rho(x, y) \in [-1, 1]$, with negative values representing anti-correlation. We disregard the sense of the correlation and henceforth use the absolute value $|\rho_{x,y}|$.

2) Monotonicity: Monotonicity is a metric to either validate or test the monotonic relationship between predictor and target variables of machine learning model predictions [45], [46], [47], [48]. To date, several studies have tested the efficacy of Monotonicity to determine the degree of dependence between features. Kachapova [45] shed light on light on the application of the Monotonicity coefficient to ascertain the dependence of random variables. The coefficient properties were found to be similar to the Pearson correlation which is in the range (-1 to 1). Adopting a similar position, findings from other authors have used insight from the study to reduce social harm and violate ethical principles [47]. Drawing from an extensive range of sources, the Monotonicity measure has been applied to solve classification problems specifically in medical diagnosis and credit scoring by confirming the machine models' predictions through the use of Monotonicity measure [49], [50], although these authors use a different measure to the one due to Kachapova's, which we use in this paper. Taken together, these studies provide converging evidence for Monotonicity as a measurement to determine the degree of association of random variables, but the metric has not been explored to measure explainability [51].

We use Kachapova's monotonicity coefficient [45] $\rho_m(x, y)$ is given by

$$\rho_m(x,y) = \begin{cases} \frac{\text{Cov}(x,y)}{\text{Cov}(x^*,y^*)} & \text{Cov}(x,y) > 0 \\ 0 & \text{Cov}(x,y) = 0 \\ -\frac{\text{Cov}(x,y)}{\text{Cov}(x^*,y')} & \text{Cov}(x,y) < 0 \end{cases} \tag{2}$$

where the superscripts $*$ and $'$ refer to the samples of x or y with their values sorted into ascending and descending order respectively. As with the linear correlation coefficient, negative values indicate anticorrelation, and we use $|\rho_m(x, y)| \in [0, 1]$.

3) ϕ_k: This recently proposed correlation coefficient [52] was based on several refinements of the Pearson correlation of random variables which has been a de facto standard in diverse contexts [52]. While the Pearson coefficient is used to detect a linear association between two random variables, ϕ_k tends to be used to determine non-linear dependence for more than two dichotomous variables. ϕ_k was selected because of its ability to detect nonlinearity unlike the other measures (Linear Correlation and Monotonicity). According to [52], [53], the ϕ_k coefficient tends to work well with categorical, ordinal and interval variables which ϕ_k as the best preferable correlation coefficient. ϕ_k values range from -1 to +1 or 0 to +1 where zero (0) means no correlation, +1 means strongest possible correlation and -1 means negative as relation. Although ϕ_k is a relatively new measure,we will apply the measurement to measure the complexity of explainability and compare it with other established metrics such as Monotonicity and Linear correlation.

IV. EXPERIMENTAL RESULTS

This section summarises the experimental results in figure 3 and 4 of this study. As previously stated, the aim of our experiments is to observe the variation of the complexity measures with the accuracy across the datasets and determine if there is a trade-off (Pareto-Front) between complexity measures and model accuracy as a result of the hyperparameters tuning of 1000 models. We demonstrate how the feature importance-weighted averages of complexity measures vary with the model accuracy.

For the Ames dataset, the models produce values of R^2 between -0.4 and 0.8 (figure 3). Varying the hyperparameters resulted in a wide range of models. With such low values R^2, some of the models are poor predictors. With higher values of R^2, we see wide variations in complexity measures, even at the same accuracy. For instance, at the upper end of accuracy, the Monotonicity varies from around 0.8 to close to 1, with similar ranges in the other measures. In some cases, the linear correlation is close to 1, suggesting that plots for these models will be approximately linear.

For the Auto dataset, R^2 varies between 0.55 and 0.85 (figure 3), meaning that the range of model accuracy in this dataset is better than the former. However, at higher model accuracy, complexity measures seem to vary. For instance, as accuracy increases from 0.70, values of linear correlation coefficient tend to cluster between 0.2 and 0.5 on the y-axis, while values of Monotonicity and ϕ_k coefficients vary towards 1.

In the Wind dataset, the varying hyperparameters still produce wide range of model accuracy of R^2 between 0.20 to 0.55 (figure 3). At the same accuracy, values of complexity measures are steadily on the rise. For example, At model accuracy between R^2 0.20 and 0.30 (figure 3), values of complexity measures increase from 0.80 towards 1.

With these variations in both model accuracy and complexity measures across the cases, we see that it is possible to produce models with the same accuracy but with more or less complex explanations by varying the hyperparameters of a black box regressor.

Further, the Pareto optimal sets shown in Figures 3 and 4 clearly indicate a trade-off between the complexity of explanations and model accuracy. These envelopes represent the Pareto front in a multi-objective optimization problem in which the two objectives are to maximise explainability and accuracy.

A. Discussion

Based on the experimental results in figure 3 and 4, we discovered variations in model accuracy and complexity measures and a trade-off was clearly observed across the cases.

As the model accuracy and complexity measures vary in all the use cases, it appears that as the model accuracy increases, the complexity measures decrease meaning less correlated in

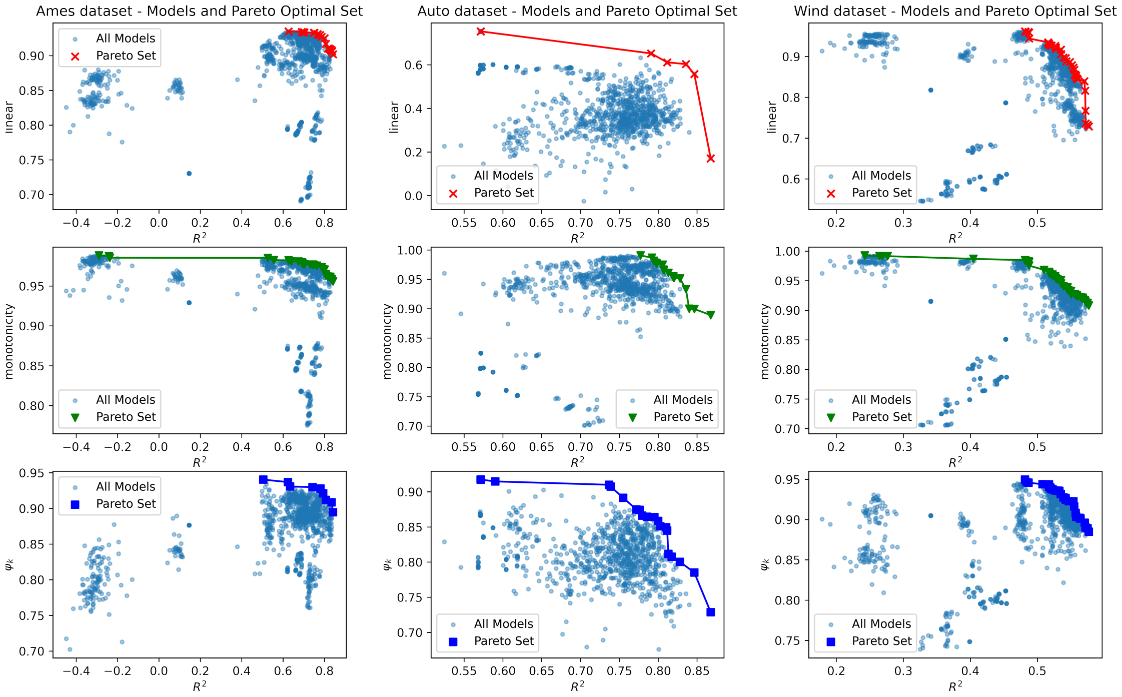

Fig. 3. Scatter plots of Complexity measures against R^2 for 1000 random forest models trained on the three datasets, with Pareto optimal sets highlighted

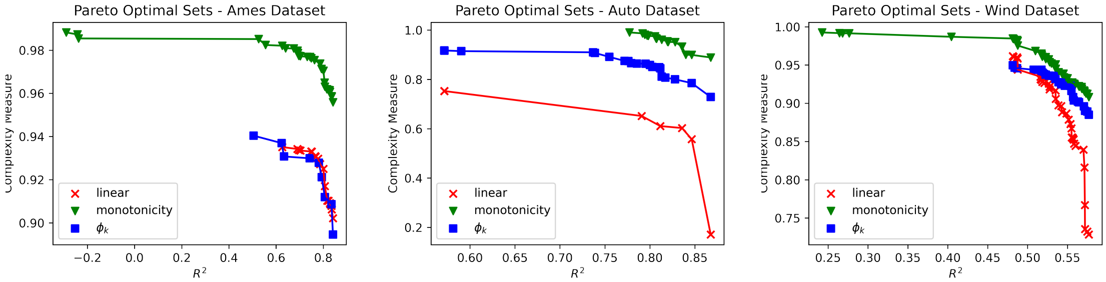

Fig. 4. Pareto optimal sets for the three complexity measures vs R^2 for each dataset, demonstrating the complexity-accuracy trade-off.

explanation. In addition, the variation of complexity measures across datasets indicates that Monotonicity was high when compared with others, meaning that this criterion is easier to satisfy. In general, there is a trade-off between complexity measures and model accuracy which is exemplified in the Pareto-fronts in figure 4.

Interestingly, there is a difference in the Auto Price dataset which seems to have the lowest values (figure 3) of Linear correlation coefficient which is clearly seen in both scatterplots and Pareto fronts. This suggests that it is producing explanations which are monotonic and correlated, but the relationship is not a simple linear one. These sorts of differences between datasets will be investigated further in a human-centred study and underlining the importance of our further work in which we will investigate which of these metrics provides the best measure of the usefulness of SHAP explanations to practitioners.

V. FURTHER WORK AND CONCLUSION

In this paper, we proposed three complexity measures of feature importance-weighted averages of Linear correlation, Monotonicity, ϕ_k to quantitatively measure the complexity of SHAP explanations. The complexity measures and model accuracy have been investigated to determine how the feature importance-weighted average of the measures vary with the Random Forest model accuracy across the three datasets. Also, we investigate the trade-off between the complexity of explainability and model accuracy as a result of models hyperparameters tuning. We found that all three of our proposed measures behaved in accordance with the hypothesis that more accurate models have more complex explanations. The relative values of the three measures differed over the three datasets we used. We were able to observe the complexity-accuracy trade-off by plotting the Pareto front.

As a consequence, further work is needed to determine which of these three measures aligns with how useful post-hoc explanations are to practitioners. It may be different for different groups, and it may be a mixture of different measures. In addition, this paves the way for a multi-objective approach to hyperparameter tuning which balances explainability (measured through the proxy of the complexity of the post-hoc explanations) and accuracy. Further, it is possible that this tuning could be tailored to individual use cases and stakeholder needs.

REFERENCES

[1] Anthony,Alford, "DeepMind's Agent57 Outperforms Humans on All Atari 2600 Games," 2020. [Online]. Available: https://www.infoq.com/news/2020/05/deepmind-ai-atari/

[2] T. Folke, S. C.-H. Yang, S. Anderson, and P. Shafto, "Explainable AI for medical imaging: Explaining pneumothorax diagnoses with Bayesian Teaching," *arXiv:2106.04684 [cs]*, Jun. 2021. [Online]. Available: http://arxiv.org/abs/2106.04684

[3] M. Y. Shaheen, "Adoption of machine learning for medical diagnosis," *ScienceOpen Preprints*, Sep. 2021. [Online]. Available: https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-.PPHMKA6.v1

[4] G. Schwalbe and B. Finzel, "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts," *arXiv:2105.07190 [cs]*, Dec. 2021. [Online]. Available: http://arxiv.org/abs/2105.07190

[5] L. K. S. J. L. Mattu, Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," 2016, publication Title: ProPublica.

[6] D. Hardawar, "Staples, Home Depot, and other online stores change prices based on your location," Dec. 2012, publication Title: VentureBeat. [Online]. Available: https://venturebeat.com/2012/12/24/staples-online-stores-price-changes/

[7] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris, "Detecting price and search discrimination on the internet," in *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI*. Redmond, Washington: ACM Press, 2012, pp. 79–84. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2390231.2390245

[8] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, Oct. 2017. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2741

[9] C. Chance, "Moving Forward on Explainable AI - nEW Guidance From the UK ICO and Turing Institute," p. 8, 2020.

[10] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, "General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models," *arXiv:2007.04131 [cs, stat]*, Aug. 2021. [Online]. Available: http://arxiv.org/abs/2007.04131

[11] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589750021002089

[12] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *arXiv:1811.10154 [cs, stat]*, Sep. 2019. [Online]. Available: http://arxiv.org/abs/1811.10154

[13] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics," *Electronics*, vol. 10, no. 5, p. 593, Jan. 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/5/593

[14] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[15] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103

[16] N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, Jan. 2021. [Online]. Available: http://arxiv.org/abs/2011.07876

[17] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions," in *Machine Learning and Knowledge Extraction*, ser. Lecture Notes in Computer Science, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2020, pp. 1–16.

[18] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," May 2022. [Online]. Available: http://arxiv.org/abs/2201.08164

[19] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Mar. 2017. [Online]. Available: http://arxiv.org/abs/1702.08608

[20] T. Miller, "Explanation in Artificial Intelligence: Insights From the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[21] N. A. M. Ahmed and A. Alpkoçak, "A quantitative evaluation of explainable AI methods using the depth of decision tree," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 6, pp. 2054–2072, Jan. 2022. [Online]. Available: https://journals.tubitak.gov.tr/elektrik/vol30/iss6/4

[22] X.-H. Li, Y. Shi, H. Li, W. Bai, Y. Song, C. C. Cao, and L. Chen, "Quantitative Evaluations on Saliency Methods: An Experimental Study," Dec. 2020. [Online]. Available: http://arxiv.org/abs/2012.15616

[23] R. Calegari, G. Ciatto, and A. Omicini, "On the integration of symbolic and sub-symbolic techniques for XAI: A survey," *Intelligenza Artificiale*, vol. 14, no. 1, pp. 7–32, Sep. 2020.

[24] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective," Feb. 2022. [Online]. Available: http://arxiv.org/abs/2202.01602

[25] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv:1705.07874 [cs, stat]*, Nov. 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

[26] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," Aug. 2018. [Online]. Available: http://arxiv.org/abs/1706.07269

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," Feb. 2019. [Online]. Available: http://arxiv.org/abs/1806.00069

[28] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, Oct. 2019. [Online]. Available: http://www.pnas.org/doi/full/10.1073/pnas.1900654116

[29] G. Montavon, W. Samek, and K.-R. Müller, "Methods for Interpreting and Understanding Deep Neural Networks," *Digital Signal Processing*, vol. 73, pp. 1–15, Feb. 2018. [Online]. Available: http://arxiv.org/abs/1706.07979

[30] P. Hase and M. Bansal, "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" *arXiv:2005.01831 [cs]*, May 2020. [Online]. Available: http://arxiv.org/abs/2005.01831

[31] D. Carvalho, E. Pereira, and J. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, p. 832, Jul. 2019.

[32] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," *Applied AI Letters*, vol. 2, no. 4, p. e61, 2021. [Online]. Available: http://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61

[33] C. Molnar, G. Casalicchio, and B. Bischl, *Quantifying Model Complexity via Functional Decomposition for Better Post-Hoc Interpretability*. [Online]. Available: http://arxiv.org/abs/1904.03867

[34] S. N. van der Veer, L. Riste, S. Cheraghi-Sohi, D. L. Phipps, M. P. Tully, K. Bozentko, S. Atwood, A. Hubbard, C. Wiper, M. Oswald, and N. Peek, "Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2128–2138, Sep. 2021. [Online]. Available: https://academic.oup.com/jamia/article/28/10/2128/6333351

[35] W. K. Diprose, N. Buist, N. Hua, Q. Thurier, G. Shand, and R. Robinson, "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator," *Journal of the American Medical Informatics Association: JAMIA*, vol. 27, no. 4, pp. 592–600, Apr. 2020.

[36] W. Swartout, "XPLAIN: A System for Creating and Explaining Expert Consulting Programs," *Artif. Intell.*, 1983.

[37] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy," in *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM, Jun. 2022, pp. 248–266. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3533090

[38] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability," *International Journal of Information Management*, p. 102538, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026840122200072X

[39] R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, and J. H. Moore, "PMLB: a large benchmark suite for machine learning evaluation and comparison," *BioData Mining*, vol. 10, no. 1, p. 36, Dec. 2017. [Online]. Available: https://doi.org/10.1186/s13040-017-0154-4

[40] S. L. Schober Patrick, Boer Christa, "Correlation coefficients: Appropriate use and interpretation," *Journal of Software*, vol. 126, p. 1763, 2018.

[41] H.-H. Hsu and C.-W. Hsieh, "Feature selection via correlation coefficient clustering," *Journal of Software*, vol. 5, p. 1371–1377, 2010. [Online]. Available: http://ojs.academypublisher.com/index.php/jsw/article/view/3713

[42] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Applied Intelligence*, vol. 52, no. 5, pp. 5457–5474, Mar. 2022. [Online]. Available: https://link.springer.com/10.1007/s10489-021-02524-x

[43] D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," *Journal of Statistics Education*, vol. 19, no. 3, p. 8, Nov. 2011. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10691898.2011.11889627

[44] L. Gianfagna and A. D. Cecco, *Explainable AI with Python*. Springer International Publishing, 2021. [Online]. Available: https://www.springer.com/gp/book/9783030686390

[45] F. Kachapova and I. Kachapov, "A Measure of Monotonicity of Two Random Variables," p. 8, 2012.

[46] M. Hu, X. Deng, and Y. Yao, "On the properties of subsethood measures," *Information Sciences*, vol. 494, pp. 208–232, Aug. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519303524

[47] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. d. Feitas, "Bayesian Optimization in a Billion Dimensions via Random Embeddings," *Journal of Artificial Intelligence Research*, vol. 55, pp. 361–387, Feb. 2016. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10983

[48] F. Sovrano and F. Vitali, "An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability," Sep. 2021. [Online]. Available: http://arxiv.org/abs/2109.05327

[49] J. Sill and Y. Abu-Mostafa, "Monotonicity Hints for Credit Screening," in *Advances in Neural Information Processing Systems*, vol. 9. MIT Press, 1996.

[50] P. Lory and D. Gietl, "Neural Networks for Two-Group Classification Problems with Monotonicity Hints," in *Classification and Information Processing at the Turn of the Millennium*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, R. Decker and W. Gaul, Eds. Berlin, Heidelberg: Springer, 2000, pp. 113–118.

[51] A. G. Asuero, A. Sayago, and A. G. González, "The Correlation Coefficient: An Overview," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 1, pp. 41–59, Jan. 2006. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10408340500526766

[52] M. Baak, R. Koopman, H. Snoek, and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics," Mar. 2019. [Online]. Available: http://arxiv.org/abs/1811.11440

[53] S. Malik and R. Singh, "A Family Of Estimators Of Population Mean Using Information On Point Bi-Serial and Phi Correlation Coefficient," Feb. 2013. [Online]. Available: http://arxiv.org/abs/1302.1658