2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) | 978-1-6654-6710-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/FUZZ-IEEE55066.2022.9882773

Measuring Model Understandability by means of Shapley Additive Explanations

Ettore Mariotti, Jose M. Alonso-Moral
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
{ettore.mariotti, josemaria.alonso.moral}@usc.es

Albert Gatt
Utrecht University
Utrecht, Netherland
a.gatt@uu.nl

Abstract—**In this work we link the understandability of machine learning models to the complexity of their SHapley Additive exPlanations (SHAP). Thanks to this reframing we introduce two novel metrics for understandability: SHAP Length and SHAP Interaction Length. These are model-agnostic, efficient, intuitive and theoretically grounded metrics that are anchored in well-established game-theoretic and psychological principles. We show how these metrics resonate with other model-specific ones and how they can enable a fairer comparison of epistemically different models in the context of Explainable Artificial Intelligence. In particular, we quantitatively explore the understandability-performance tradeoff of different models which are applied to both classification and regression problems. Reported results suggest the value of the new metrics in the context of automated machine learning and multi-objective optimisation.**

I. Introduction

The availability of data and computing power is making machine learning (ML) and Artificial Intelligence (AI) more and more important for extracting value and optimizing processes. In this frame there has been a strong drive for developing very sophisticated models that reach high-performance scores in classification and regression benchmarks. This has though come at the price of having models that are more and more obscure and difficult to inspect and explain. This is an issue in all those applications where an action in the real world has strong consequences and bears responsibilities, such as medicine, justice and finance. The need for having both powerful and *understandable* models is what has driven the development of eXplainable Artificial Intelligence (XAI) [1], [2]. In this context the paradigm is not anymore the automation that enables a machine to solve a task, but rather an integration of predictive tools that give suggestions and insights to a human user who happens to be the one that ultimately is accountable and responsible for high impact decisions (this paradigm is also known as *human-centric computing*).

There are different approaches to XAI. While some researchers try to develop model-agnostic tools that are as widely applicable as possible (post-hoc explanations) [3]–[5], others instead push for the development of intrinsically interpretable models, arguing that white-box models can be as powerful as the so-called black-box models [6]–[8]. We thus have lots of models which are explainable to different degrees, but what we lack is a model-agnostic way of capturing the interplay between (a) the simplicity of the model (regarding Ockham's razor principle) and (b) its empirical performance on one or more tasks. In practice, typically what is done is to measure the complexity of a specific model by looking at some model-specific properties (for example the number of nodes of a decision tree). This is fine when we restrict our explorations to just one class of models, but it is problematic when we want to explore and compare different families of models (e.g., is it fair to compare the number of nodes in a tree to the number of non-zero coefficients in a linear model?).

In this work we present a metric for evaluating model understandability that is model-agnostic, intuitive, computationally efficient and mathematically grounded. This can enable a fairer comparison between different models on specific datasets. In addition, it enables the possibility of multi-objective optimisation guided by automated metrics for both performance and understandability. And last but not least, it can push the development of more explainable models.

The rest of the manuscript is organized as follows. Section II introduces preliminary concepts and reviews related publications. Section III presents two novel metrics for measuring understandability. Section IV uses the metrics of section III to assess the goodness of classifiers and regressors. In addition, we show the utility of the new metrics in the context of an illustrative example of multi-objective optimisation. Finally, Section V concludes the paper and outlines future work.

II. Related Work

There are several approaches to XAI evaluation. Here we report only the papers that are most in line with our approach. For further details, the interested reader is kindly referred to [9].

A. Complexity as a proxy for understandability

Automatic measures of understandability typically consist in measuring some model-specific complexity metric as a proxy for understandability [10], e.g., the number of rules in an expert system, the number of non-zero coefficients in a linear model, the number of nodes in a decision tree, etc. This way of doing might work while dealing within just one class of models but becomes difficult to compare across different classes of models. Molnar et al. [11] proposed

978-1-6654-6710-0/22/$31.00 ©2022 IEEE

to overcome this issue by creating a metric for complexity based on functional decomposition. They count the number of features used by a pipeline with permutation-based evaluation and estimate the interaction effects with accumulated local effects. While being a good approach (for example for multi-objective optimisation) it is not straightforward to see how their metric is related to intelligibility. Furthermore, the way with which their complexity metric is set up is somewhat arbitrary and not intuitive.

Zhou et al. [12] developed a hybrid survey-based evaluation method to find out how a linear model and a tree are intelligible. In addition, they created a meta-model that fits the human evaluation using some attributes of the underlying model. Unfortunately, this approach does not allow the use of the new metric out of the scope of the conducted experiments. Moreover, the reported results (and the generated meta-model) highly depend on the population on which the survey was conducted.

B. *Post-hoc measures*

An interesting approach is to build upon SHapley Additive exPlanations (SHAP) [4]. This is a model-agnostic game-theoretic approach that assigns a relevance score ϕ_i to each feature i using Shapley Values [13]. This allocation is the only additive feature attribution method that satisfies a variety of axioms for an n-person coalitional game v:

- *Null player*: if the presence of player i never adds anything to any coalition, then $\phi_i = 0$;
- *Equal treatments of equals*: if two actors i and j always "bring the same value" to any coalition, then $\phi_i = \phi_j$;
- *Efficiency*: the attributions sums up to the total payoff, that is $\sum_{i \in N} \phi_i = v(N)$;
- *Linearity*: the attributions to player i for a game that is a linear combination of two games $v^{'}$ and $v^{''}$, is the linear combination of the attributions of the sub-games $\phi_i^{'}$ and $\phi_i^{''}$.

It is worth noting that SHAP is becoming more and more popular in practical applications [14], [15]. In this context, Weerts et al. [16] carried out an extrinsic evaluation of the efficacy of SHAP on some specific tasks with humans. They observed that large Shapley values did affect the reasoning applied by their participants. They also highlighted that "*large Shapley values can bring feature values of the instance to attention that are otherwise ignored*". This is a key insight upon which our approach is built, as we will see in the next section.

III. Model-agnostic Automatic Metrics: SHAP Length and SHAP Interaction Length

We are looking for a novel metric with the following desiderata. The new metric must be:

- model agnostic;
- computationally efficient;
- theoretically grounded;
- in agreement with human intuition.

In order to achieve these desired properties of the explainability metric, we will take as starting point SHAP explanations and the principle of minimum cognitive load [17]. The latter states that information before being processed and integrated must first pass through working memory. This buffer, being limited in both capacity and duration [18], limits the overall throughput of information that we can digest. Miller conjectured that this limit was seven plus or minus two concepts to be simultaneously handled by humans [19]. The same limitation was observed when humans did preference judgments in the context of an analytic hierarchy process [20]. We take working memory load to be an estimate of overall cognitive load and formulate our research hypothesis as follows: an explanation with a lower cognitive load has more possibilities to be digested by humans (and can thus be deemed as "understandable") than one with a higher load.

A. *Intuitive idea*

For a user to understand the behaviour of a model by means of SHAP, he/she has to pay attention to every non-zero value of a given explanation. We can further generalise this notion to the user having to look at a subset of explanations whose aggregate contribution sum up to a certain threshold. The idea is that if for a specific prediction we have a few non-zero Shapely values, the user has to look at fewer *details*. The explanation is thus relatively more intelligible than another where instead many features contribute to the output (see Fig. 1). We can also think about this as a lossy compression of the explanation, where the amount of information loss can be controlled precisely. We can then extrapolate from the aggregation of many explanations on the same dataset that if the explanations associated with a model involve on average less features, then the model is less complex than one that involves more of them.

In the following sections we introduce two novel explainability metrics which are built upon SHAP:

- SHAP Length (SL): based on Shapley Values [21].
- SHAP Interaction Length (SIL): based on Shapley Interaction Values [22], where an importance score is given also to pairwise interaction effects.

B. *Formal Definitions*

Lundberg and Lee [21] introduced the use of SHAP as a numerical and visual explanation of the impact that each feature in a given data instance had on the associated prediction. The set $\Phi := \{\phi_i\}$ corresponds to the Shapley allocation of a coalitional game where i features are seen as players and the payoff is the difference between the model output and its average value (also called baseline).

Let's now consider the following useful definitions:

Definition III.1 (Explanation mass)**.** For a local SHAP explanation on an instance with p features $\{\phi_i, i = 1...p\}$, we define explanation mass associated to each feature i to be $|\phi_i|$.

Definition III.2 (Complete explanation)**.** A subset of Shapley values $\Phi_c \subseteq \Phi$ is *complete* if $\Phi_c := \{\phi_i > 0, \forall i\}$. In other

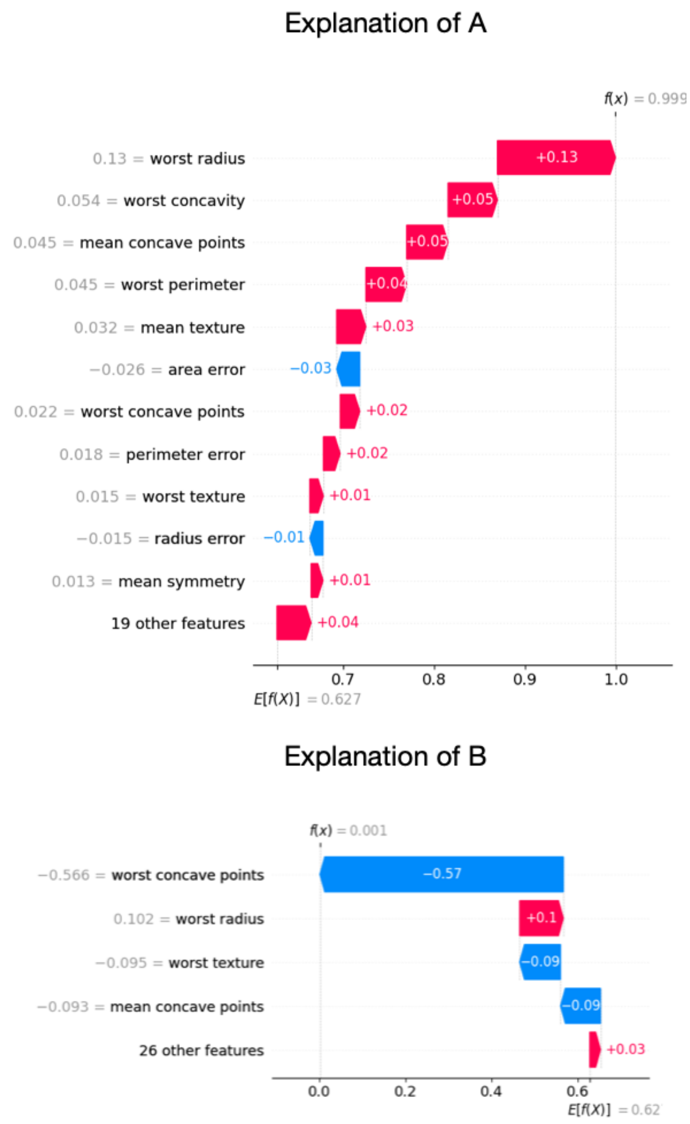

Fig. 1. Two different Breast Cancer instances classified by Random Forest and explained by SHAP whose output is compressed to a 95% mass threshold (see Section III-B for further details). Instance (B) is relatively easier to be understood wrt (A) in virtue of carrying a lower cognitive load (i.e., it involves fewer pieces of information) to a user.

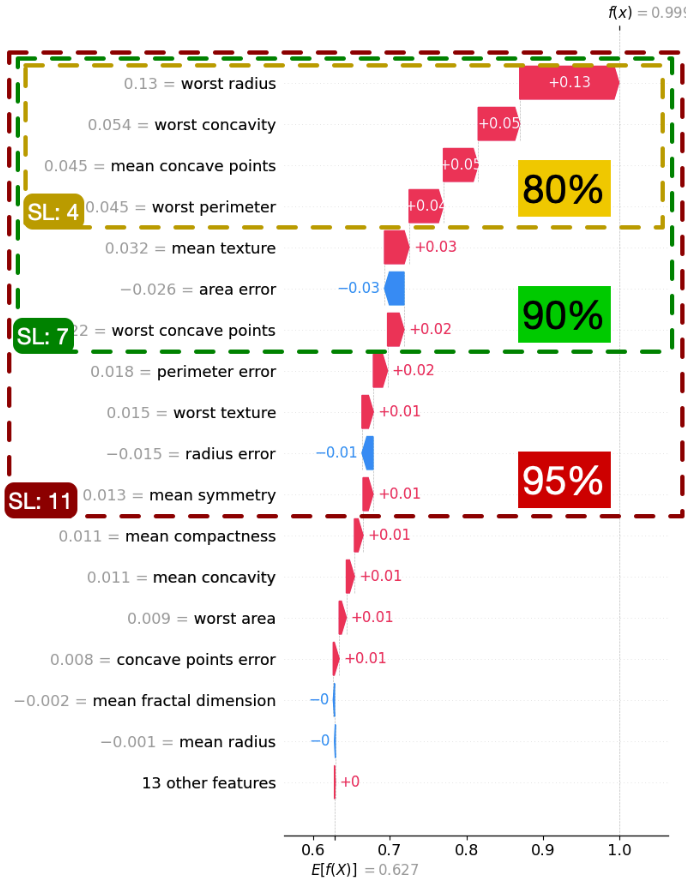

Fig. 2. Examples of different SHAP Length (SL) values for different explanation mass thresholds, respectively of 80%, 90% and 95% for instance #424 of Breast Cancer Dataset. For ease of visualisation, the value of each ϕ_i is reported rounded to the second digit.

words, it is an explanation that describes all the i features that have contributed to the outcome (i.e., $\phi_i > 0$). For example, given a local explanation $\{\phi_1 = 0.4, \phi_2 = 0, \phi_3 = -1.2\}$, we have that the explanation $\{\phi_1 = 0.4, \phi_3 = -1.2\}$ is still complete (that is, it still contains all information).

Definition III.3 (Explanation completeness)**.** For a given subset $\Phi_{subset} \subseteq \Phi_c$, we call *completeness* the scalar

$$\Gamma(\Phi) := \frac{\sum_{i \in \Phi_{subset}} |\phi_i|}{\sum_{i \in \Phi_c} |\phi_i|}.$$

In other words, it is the fraction of explanation mass of a complete explanation that a given subset of explanands captures. Intuitively it represents what is the fraction of information that the subset has compared to a complete one. Notice that $0 \leq \Gamma(\Phi) \leq 1$ and we can refer to it in a percentage form for convenience. As a practical example, in Fig. 2 the subset of explanands in the big red box have a completeness of $\Phi(\{\phi_i \text{ for } i \in \text{red box}\}) = 95\%$.

Definition III.4. $p\%$-complete explanation is the smallest set of Shapely values Φ such that $\Gamma(\Phi) \geq p$.

Based on the previous definitions, it is possible to formulate the following additional property:

Definition III.5. $SL_{p\%} := ||\Phi_p||$ is the number of features in the $p\%$-complete explanation. Similarly, $SIL_{p\%}$ is the cardinality of the $p\%$-complete explanation (i.e., the number of involved features) in a SHAP interaction graph, where one can take only the lower (or upper) triangular matrix of the explanation, due to its symmetric nature, for efficient computing purpose.

C. Implementation Details

The algorithm for computing SL simply orders the absolute Shapley values from the largest to the lowest (normalised by the total). Then the SL value is the first index for which the cumulative sum is greater or equal than the given threshold. A pseudocode version is given in Algorithm 1. SIL is computed with the same Algorithm 1 with the difference that it uses the flattened vector of the lower triangular matrix of Shapley interaction values. In general, it is computationally expensive to compute Shapely interaction values. Fortunately, it is possible to have a very fast exact implementation for tree ensembles that can leverage GPU hardware [23]. For practical purposes, we will compute both SL and SIL with Gradient Boosted Trees from the XGBOOST package (xgb) [24] which acts as a surrogate of the original model. The fidelity of this surrogate is measured as the R^2 score between the prediction of the original model (log odds in case of classification) and the output of the surrogate.

Algorithm 1: SHAP Length (SL)

```
Data: $\{\phi_i\}$ for $i = 1...p$, $th$
Result: $SL_{th\%}$
$ExplanationMass \leftarrow |\phi_i|$
 $TotalMass \leftarrow Total(ExplanationMass)$
 $Ordering \leftarrow ArgSort(ExplanationMass)$
 $CumNormMass \leftarrow$
 $CumSum(ExplanationMass[Ordering])/TotalMass$
 for $i = 0, i < p, i++$ do
     if $CumNormMass[i] > th$ then
         $SL_{th\%} \leftarrow$i
         return $SL_{th\%}$
     end
end
```

IV. Experimental Study

In this section we present some practical use cases to illustrate the validity and utility of SL and SIL. We explore how different models populate the performance-understandability tradeoff and how the proposed metrics relate to other well-known model-specific complexity metrics. In addition, we will compare several families of models, with different degrees of transparency, on a classification dataset and a regression dataset. We are looking for the models that have the best performance and the best understandability (i.e., the smallest complexity in terms of SL and SIL). For this is useful to pay attention to the associated Pareto front (the ranking of models for which there is no other competitor which is better in either performance or understandability) for the datasets under consideration.

A. Balance between performance and understandability

We propose an empirical exploration of the performance-understandability tradeoff in terms of the balance between accuracy and complexity for different algorithms on a classification dataset (Breast Cancer[1]) and a regression dataset (Boston House[2]). On the one hand, the Breast Cancer dataset corresponds to a binary classification problem in the medical domain, with a total of 569 data instances (212 instances associated with Malign class and 357 instances associated with Benign class) which relate 30 real-valued features. On the other hand, the Boston House dataset includes 506 data instances that relate 13 real-valued features with the target of predicting housing prices.

The following algorithms are tested: k-Nearest neighbours (i.e., 10-nearest and 5-nearest, with k=10 and k=5 respectively), Decision Tree (dt), Fuzzy Decision Tree Classifier with s-shaped fuzzy partitions (fuzzy_tree) [25], Decision Tree with a maximum depth of 2 (dt_depth_2), Fuzzy Decision Tree Classifier with a maximum depth of 2 (fuzzy_tree_depth_2), Explainable Boosting Machines (ebm), Explainable Boosting Machines without interaction values (ebm_0) [26], Logistic Regression with L2 regularisation (lr_l2), Logistic Regression with L1 regularisation (lr_l1), Random Forest (rf), Support Vector Machine (svm), Gradient Boosted Trees (xgb). For the regression experiments we used the regression version of the above-mentioned models (except for Fuzzy Decision Trees which are only available in Python for classification) with the addition of Linear Regression with L2 (lr_l2) and L1 (lasso) regularisation. Each model is trained with default parameters and evaluated with 10-fold cross-validation using the Python package scikit-learn [27]. It is worth noting that the xgb-based surrogate reported mean R^2 score of 0.997 and 0.982 on the Breast Cancer [28] and Boston House [29] datasets, respectively.

Reported results for the Breast Cancer dataset are summarised in Fig. 3. It is interesting to appreciate quantitatively the tension between understandability and performance which is reported also elsewhere in the literature [2]. In particular, we can see how decision trees (including in this category fuzzy_tree_depth_2) and nearest neighbour approaches produce by far the simplest models but at the cost of the smallest performance. The Pareto front with the non-dominated models in terms of SL and F1-score (see the plot in the left of Fig. 3) includes dt_depth_2 (the simplest and least accurate model), dt, 10-nearest, rf, fuzzy_tree, xgb, lr_l1 and lr_l2 (the most complex and accurate model). Interestingly, fuzzy_tree turns up as a compromise solution. Notice that fuzzy_tree is close to rf and xgb regarding both performance and complexity while it is a single model versus ensemble models. Moreover, fuzzy_tree can be endowed with linguistic interpretability which is likely to be appreciated in a number of applications.

The Pareto front with the non-dominated models in terms of SIL and F1-score (see the plot in the right of Fig. 3) is slightly different. It includes: dt_depth_2 (the simplest and least accurate model), dt, 10-nearest, xgb, lr_l1 and lr_l2 (the most complex and accurate model). It is worth noting that xgb dominates rf and fuzzy_tree when interactions come into play. Indeed, SHAP interaction graphs are harder to interpret and this is reflected in the fact that SIL values are higher than SL

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[2] http://lib.stat.cmu.edu/datasets/boston

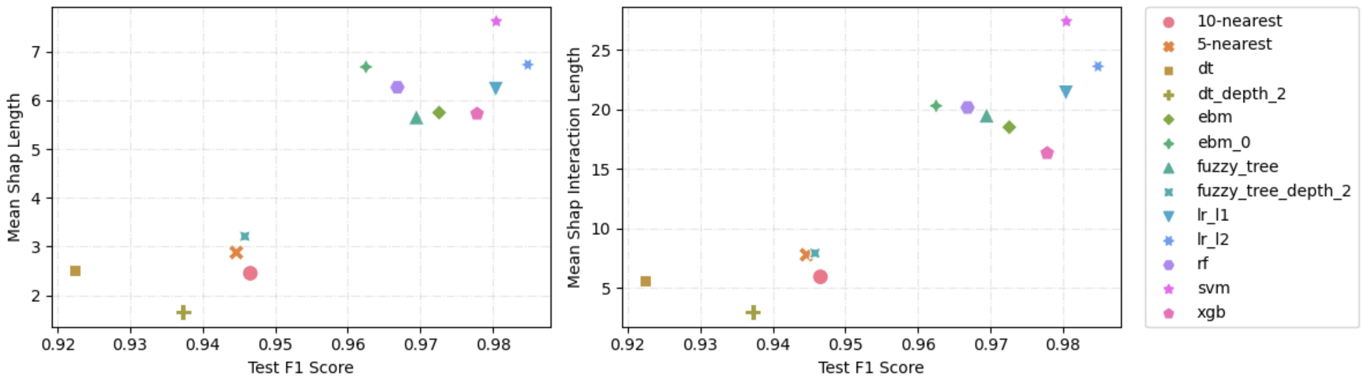

Fig. 3. An empirical exploration of performance-understandability tradeoff of different algorithms on the Breast Cancer dataset. Each metric has been computed with 10-fold Cross-Validation. Performance is measured as the F1 Score. Understandability is associated with the $SL_{90\%}$ and $SIL_{90\%}$. The best models would be in the bottom-right part of each plot.

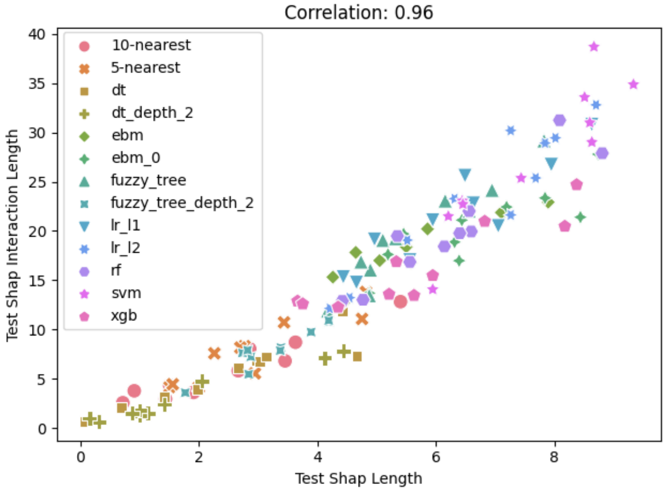

Fig. 4. Correlation graph with SL versus SIL. Here is shown the value for each of the 10 folds for each model trained on the Breast Cancer.

values. F1-score computes the harmonic mean of precision and recall; thus giving the relation between true positives (TP), false positives (FP) and false negatives (FN), which is especially informative in the case of unbalanced classification datasets:

$$F1\!-\!score = 2*\frac{precision * recall}{precision + recall} = \frac{2}{2*TP + FN + FP}$$

As observed in Fig. 4, SL and SIL are highly correlated. This means we can rely on SL (which is easier to compute) as an estimate for SIL. This is the reason why the two Pareto fronts in Fig. 3 are so similar.

The same behaviour is observed in the Boston Dataset dataset (see Fig. 5) but in this case the non-dominated models are only dt_depth_2 (the simplest model) and rf (the most accurate model). In this regression problem we measure performance as the negative mean square error (Negative MSE) as follows:

$$Negative\ MSE = -\frac{1}{N} * \sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2$$

where N is the number of test data instances, $\hat{y}_i$ is the predicted output for instance i and y_i is the actual output for the same instance.

B. Multi-objective optimisation

As a use case we present an exploration of a grid-search for decision regressor trees and linear regression with L1 and L2 regularisation (also known as elastic net) on the Boston House Dataset. We choose these models as they are considered interpretable and we deem interesting to evaluate which part of the complexity-performance space they occupy.

The parameter grid for trees spanned the depth of the trees from 2 to 12 and a regularisation parameter ccp-α from 0 to 1. The parameters of Linear Regression instead were the penalty coefficient α spanning from 0 to 1 and the “l1 amount ratio”, which controlled how much the l1 regularisation was wrt l2, also spanning from 0 to 1. Fig. 6 summarises the reported results. Each point is a different configuration evaluated with 4-fold cross-validation. Interestingly, for this dataset it seems that regression trees can be better both in performance and understandability compared to linear regression. By comparing it to Fig. 5, we can see how in the space of intelligible models we could find solutions that can be competitive to other black-box approaches.

C. Proportionality with respect to popular complexity measures associated with white-box models

For these novel metrics (SL and SIL) to be valid we would expect them to be somehow proportional to some other intuitive, well-established (yet model-specific) metrics commonly used in the literature to measure complexity. In particular, we explored on the Boston House dataset the family of regression trees with the same hyperparameters grid-search used in IV-B, evaluating 1000 configurations with 4-fold cross-validation. In Fig. 7 we compare $SL_{98\%}$ with the number of

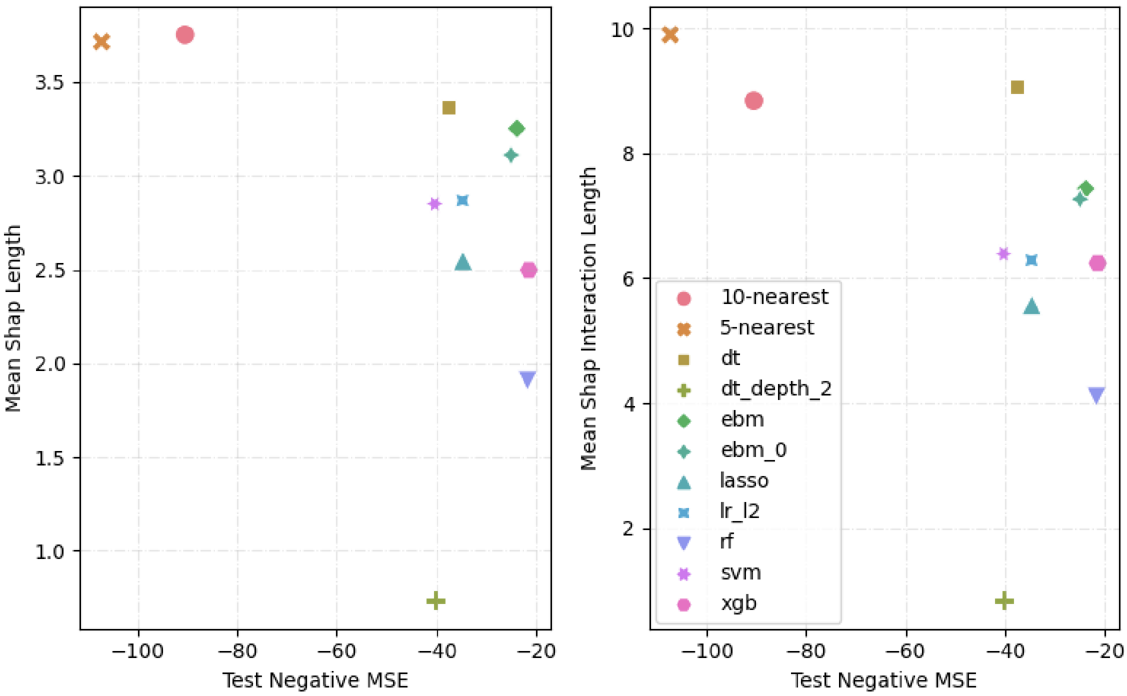

Fig. 5. An empirical exploration of the performance-understandability tradeoff of different algorithms on the Boston House dataset. Each metric has been computed with 10-fold Cross-Validation. Performance is measured as negative Mean Square Error (MSE). Understandability is associated with the $SL_{90\%}$ and $SIL_{90\%}$. The best models would be in the bottom-right part of each plot.

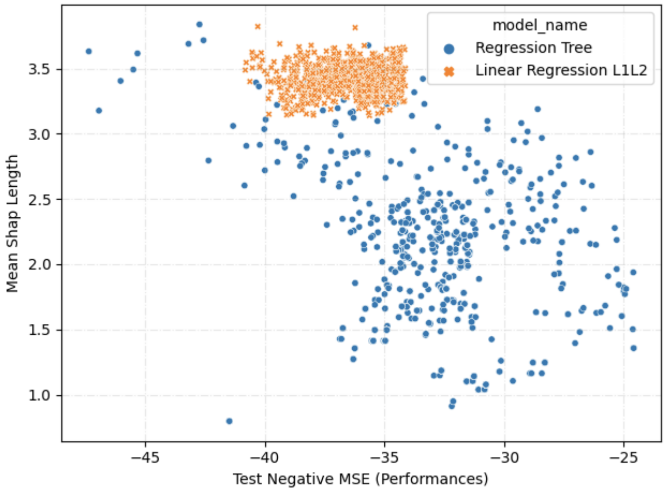

Fig. 6. Complexity (90% SL)/Accuracy (Negative MSE) tradeoff of an exhaustive grid search of decision trees on the Boston House Dataset (4-fold cross validation).

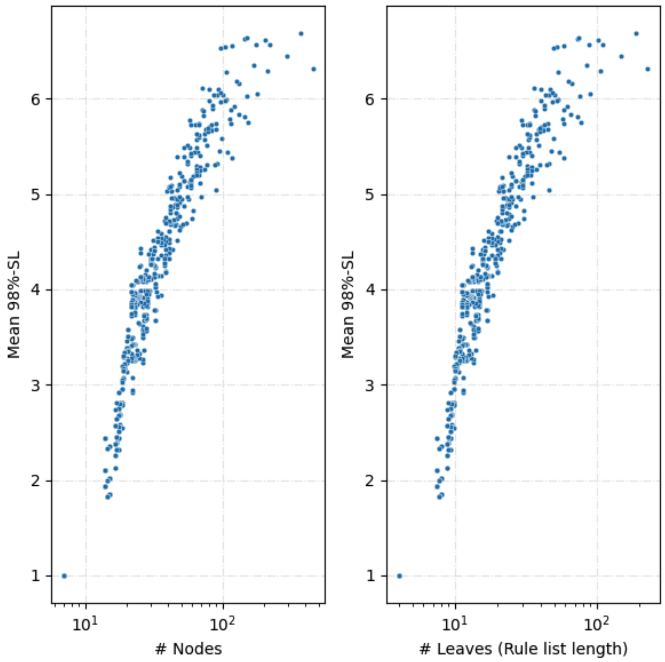

Fig. 7. SL is proportional to the complexity of the tree measured as the number of nodes (roughly equivalent to ”concepts”) and the number of leaves (equivalent to the length of a rule list). Notice that #Nodes and #Leaves are very similar to each other up to a scale factor due to the specific details of the tree construction.

nodes (plot on the left) and with the number of leaves (plot on the right) in a regression tree. The number of nodes is a measure of how many concepts do the tree need for doing well the regression task while the number of leaves translates to how long would be an equivalent list of rules. Both number of nodes and leaves are widely used in the literature to measure the complexity of trees. Notably, SL tends to saturate while the other metrics do not, this is because SL has an upper bound determined by the number of features of the dataset, while the complexity metrics of a tree are unbounded because a tree can in principle become arbitrarily large. We can see that the two metrics are proportional to each other such that a higher SL corresponds to a higher complexity metric. We can thus reproduce results of a model specific metric, but with a model agnostic approach. In this sense with SL we do not lose explanatory power but instead we have a gain in generalisability.

V. Conclusions and Future Work

In this paper we introduced two new metrics (SL and SIL) for evaluating the "understandability" of ML pipelines. The metrics anchor on Shapley values and psychological considerations. These foundations make SL and SIL model-agnostic, efficient, theoretically grounded and more intuitive than other alternatives available in the literature up to now. We have shown how these metrics can be used for comparing how different models behave on some datasets (regarding both classification and regression tasks). As a result, we hope that these proposed metrics can provide a common ground benchmark for pushing the comparison of alternative approaches and the reproducibility of empirical results in the context of XAI.

In addition, the new metrics are ready to be used in the context of automated ML (AutoML). The criterion for selecting the best setup pipeline could be the simplest model within the best performing models. Further research needs to be done in order to establish solid relationships between SL and generalisation power, as it could possibly be a useful indicator of overfitting. As future work, we plan to scale up this benchmark to include more datasets and to try to delineate a ranking of accurate yet understandable models available in the literature.

Acknowledgment

This work is conducted within the NL4XAI project which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. Jose M. Alonso-Moral is a *Ramon y Cajal* Researcher (RYC-2016-19802). This work is funded by the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29, ED431G/08, ED431G2019/04). Some of the previous grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

[1] 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.

[2] 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.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, Jan. 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3236009

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019.

[7] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges," *arXiv:2103.11251 [cs, stat]*, Jul. 2021, arXiv: 2103.11251.

[8] J. M. Alonso, C. Castiello, L. Magdalena, and C. Mencar, *Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems*, 2021. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-71098-9

[9] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3-4, pp. 24:1–24:45, Aug. 2021. [Online]. Available: https://doi.org/10.1145/3387166

[10] J. M. Alonso, L. Magdalena, and G. González-Rodríguez, "Looking for a good fuzzy system interpretability index: An experimental approach," *International Journal of Approximate Reasoning*, vol. 51, pp. 115–134, 2009. [Online]. Available: https://doi.org/10.1016/j.ijar.2009.09.004

[11] C. Molnar, G. Casalicchio, and B. Bischl, "Quantifying Model Complexity via Functional Decomposition for Better Post-hoc Interpretability," in *Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, Sep. 2019, pp. 193–204.

[12] Q. Zhou, F. Liao, C. Mou, and P. Wang, "Measuring Interpretability for Different Types of Machine Learning Models," in *Trends and Applications in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, M. Ganji, L. Rashidi, B. C. M. Fung, and C. Wang, Eds. Cham: Springer International Publishing, 2018, pp. 295–308.

[13] L. S. Shapley, "A Value for n-Person Games," in *Contributions to the Theory of Games (AM-28), Volume II*, H. W. Kuhn and A. W. Tucker, Eds. Princeton University Press, Dec. 1953, pp. 307–318.

[14] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749–760, Oct. 2018.

[15] M. J. Ariza-Garzon, J. Arroyo, A. Caparrini, and M.-J. Segovia-Vargas, "Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending," *IEEE Access*, vol. 8, pp. 64 873–64 890, 2020.

[16] H. J. P. Weerts, W. van Ipenburg, and M. Pechenizkiy, "A Human-Grounded Evaluation of SHAP for Alert Processing," *arXiv:1907.03324 [cs, stat]*, Jul. 2019, arXiv: 1907.03324.

[17] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[18] A. Baddeley, *Working Memory, Thought, and Action*, ser. Oxford Psychology Series. Oxford: Oxford University Press, 2007.

[19] G. A. Miller, "The magical number seven plus or minus two: some limits on our capacity for processing information." *Psychological review*, vol. 63(2), 1956.

[20] T. Saaty and M. Ozdemir, "Why the magic number seven plus or minus two," *Mathematical and Computer Modelling*, vol. 38, pp. 233–244, 2003.

[21] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[22] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," Feb. 2018.

[23] R. Mitchell, E. Frank, and G. Holmes, “GPUTreeShap: Massively Parallel Exact Calculation of SHAP Scores for Tree Ensembles,” *arXiv:2010.13972 [cs]*, Jul. 2021.

[24] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[25] C. Olaru and L. Wehenkel, “A complete fuzzy decision tree technique,” *Fuzzy Sets and Systems*, vol. 138, no. 2, pp. 221–254, Sep. 2003.

[26] H. Nori, S. Jenkins, P. Koch, and R. Caruana, “Interpretml: A unified framework for machine learning interpretability,” *arXiv preprint arXiv:1909.09223*, 2019.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear Feature Extraction For Breast Tumor Diagnosis,” vol. 1905. Proceedings of the Conference on Biomedical Image Processing and Biomedical Visualization, 1993.

[29] D. Harrison and D. L. Rubinfeld, “Hedonic housing prices and the demand for clean air,” *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, Mar. 1978.