Comparing explanations from glass-box and black-box machine-learning models

Michał Kuk[1][0000−0002−6270−3938], Szymon Bobek[2][0000−0002−6350−8405], and Grzegorz J. Nalepa[2][0000−0002−8182−4225]

[1] AGH University of Science and Technology
[2] Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and Institute of Applied Computer Science, Jagiellonian University

Abstract. Explainable Artificial Intelligence (XAI) aims at introducing transparency and intelligibility into the decision-making process of AI systems. In recent years, most efforts were made to build XAI algorithms that are able to explain black-box models. However, in many cases, including medical and industrial applications, the explanation of a decision may be worth equally or even more than the decision itself. This imposes a question about the quality of explanations. In this work, we aim at investigating how the explanations derived from black-box models combined with XAI algorithms differ from those obtained from inherently interpretable glass-box models. We also aim at answering the question whether there are justified cases to use less accurate glass-box models instead of complex black-box approaches. We perform our study on publicly available datasets.

Keywords: explainable AI · machine learning · artificial intelligence · data mining

1 Introduction

In recent years, the impact of machine learning on our daily life increased significantly, providing invaluable support to the decision making process in many domains. In insensitive areas, such as healthcare, industry, and law, where every decision may have serious consequences, the adoption of AI systems that cannot justify or explain their decisions is difficult and in many cases not desired. Such an observation stays in contradiction to the trend in the AI world, where the most progress is observed in the area of black-box models such as deep neural networks, random forests, etc. This duality led to the development of explainable AI methods, which allows introducing *transparency* and *intelligibility* to the decisions made by not interpretable black-box models. However, this transparency and intelligibility may serve different purposes, depending on the application area and the task that is to be solved with the AI method. In particular, we can define two main goals of XAI methods:

1) understand the mechanics of the ML model in order to debug the model, and possibly the dataset (i.e., what input drives the *model* to classify instance *A* as class *C*,

2) understand the phenomenon that is being modelled with AI methods and to build trust (i.e., what input makes the *instance A* to be classified as *C*).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55

While these goals might be indistinguishable at first glance, there is a fundamental difference in the assumptions that need to be fulfilled in both cases. In the first case, we assume that the model might be wrong, and we want to fix it, hence the information about the *model* is the most important. The model performance is an objective. In the second case, we assume that the model is correct and we want to use it to obtain information about the *class* or *instance* itself. The explanation itself is the main objective. In this paper, we focus on the second case. We provide a discussion on the performance of the glass-box models and black-box models explained with the use of XAI methods, in the situation when the objective is not to learn about the AI model, but about the phenomenon the model captures. We focused on rule-based explanations as one of the most understandable and widely applicable methods in industrial and medical cases. We performed a comparison of these two approaches on datasets from selected scikit-learn datasets and UCI Machine Learning Repository to see if a simple glass-box model can outperform complex XAI algorithms.

The rest of the paper is organised as follows: In Section 2 we describe a few papers which concern the explainable methods and we introduce our motivations. In Section 3 we present our approach to the performance comparison mentioned above. Next, Section 4 presents and discusses the results we obtained. Finally, in Section 5 we summary our work.

2 Related works and motivation

In this paper, we focused to evaluate what is the difference between the explanations obtained from black-box models combined with XAI algorithms and from those obtained based on interpretable glass-box models. To get such evaluation, firstly we verified the existing researches which concern glass-box and black-box models in the application of Explainable Artificial Intelligence.

In [10] the author makes a comparison of white and black box models. The author outlines that in some cases glass-box models could give as accurate results as the black box models. However, it strongly depends on the application domain and the data delivered. In the case of the black-box models, the author highlights that the experts do not need to understand the mathematical transformations behind them, but they proposed to deliver the output data in a similar form as input.

In [1] the authors pay attention to the fact that nowadays there is the need for XAI application due to commercial benefits, regulatory considerations, or in cases when the users have to effectively manage AI results. They outline that the black-box models do not disclose anything about internal design, structure, or implementation. On the other hand, the glass-box is completely exposed to the user.

In [6] the authors used glass-box and black-box models to predict the ambient black carbon concentration. They used several methods, whereas a neural network with LSTM layers gave the best results. However, they highlight that using black-box models like neural network or random forest complicates explanations.

In [9] the authors used the Anchor algorithm to obtain rules which could be explanations of each cluster of data. As a result, the proposed methodology is able to

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55

generate human-understandable rules which could be passed to the experts to support in the explainability process.

In [14] the authors used the black-box model to develop the structured attack (StrAttack), which is able to explore group sparsity in adversarial perturbation by sliding a mask through images. They demonstrate the developed method on datasets consisting of images. Furthermore, they outline that thanks to the sliding masks, they increase the interpretability of the model.

In [15] the authors also concentrate on image classification. They used a deep neural network to assign input to predefined classes. To interpret the models, they considered post-hoc interpretations. More specifically, they focus on the impact of the feature on the predicted result – they tried to uncover the casual relations between input and output.

In [11] the authors created an open-source Python package called InterpretML. They focused in most cases on the feature importance explanations, not on the human-readable rules.

I our work we focus on rule-based explanations, which according to our previous research [3] proves to be one of the most intelligible way of providing explanations to experts. Therefore, we mainly concentrate on the algorithms which can generate explanations which are represented as a logic implication (IF-THEN) by using a conjunction of relational statements. Such explanations can be executed with rule-based engines and verified according to selected metrics such as accuracy, precision, or recall. Having that, a research question arises: If the explanation is as much valuable as the model decisions, what kind of model should be applied to assure good interpretability along with high accuracy of explanations? Should it be 1) a glass-box model to directly generate explainable results, or 2) a complex black-box model and explain its results with Explainable Artificial Intelligence methods? To solve this research problem, we aim to apply XAI methods that are able to generate human-readable rules for complex black-box models and verify if these methods are needed to be applied in the case of simple tabular data or if we should make explanations directly with the use of the glass-box models.

3 Experimental comparison

The scope of this work concentrates on the comparison of using simple glass-box models with complex black-box models explained with the XAI algorithm. In this work, we use a classification task as an exemplary problem to be solved. We considered the most popular classifiers available in the scikit-learn package [12].

For glass-box models, we selected: Decision Tree, Nearest Neighbors. For black-box models, we chose RBF SVM, Gauss Process, Random Forest, Neural Network, AdaBoost, Naive Bayes, QDA. We used a default models' settings as hyperparameter tuning was not the main goal of this paper. Our experiment considers two approaches for solving the classification task: 1) use directly explainable glass-box model, 2) use the black-box model, explain the predictions with XAI method, and solve the main problem based on the explanations obtained.

In the second approach, we took into consideration the XAI algorithms which generate explanations in the form of human-readable rules based on the trained classifier

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55

model. In this work, we considered three XAI methods: Anchor [13], Lux [4] and Lore [7]. Each of them generates instance-based explanations (rules), which subsequently were converted into XTT2 format [8]. To allow the results to be compared with glass-box models we used HEARTDROID inference engine [5] to predict the classification target (label) based on the obtained rules and data instances.

The schematic illustration of the considered approaches is presented in the Fig. 1.

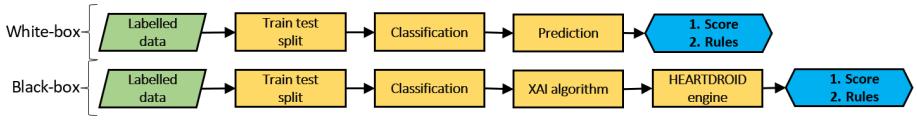

Fig. 1. Glass-box and black-box models approaches comparison.

To test the considered approaches and draw reliable conclusions, we chose data from different sources as an input to the experiments. In the work we used the following datasets: banknote and glass[3], cancer and iris[4], and titanic[5].

Each dataset has been divided into train and test datasets. For each considered classifier, we applied the same train instances to train the model and test instances to make predictions to maximize the reliability of the comparison results. For both considered approaches, we computed the accuracy, recall, and precision scores for each considered classifier to compare these two approaches.

4 Results and discussion

In this section, we present the results of our experiments. Firstly, we compared the performance of all considered classifiers used directly to solve the classification problem. Then, we compared scores for the classification problem solved based on the rules generated for the black-box models with XAI methods (only the best results for each XAI method) vs glass-box model. Finally, we also considered the variance of scores that can be obtained for a selected XAI method depending on the black-box model explained. These results are presented in the following figures.

Fig. 2 shows the results (scores) which were calculated for all classifiers used directly to predict the classification target (label). As can be seen, there are some datasets for which all classifiers perform with a high score (close to 1.0) such as iris and banknote that suggest that the problem to be solved in their case is relatively simple. For cancer and wine datasets, most classifiers also give high scores, but others (e.g. QDA, RBF, SVM) perform much worse. The most difficult problem to be solved is contained in the glass dataset for which the best score is lower than 0.7. Other difficult dataset is titanic for which scores are not greater than 0.8. Comparing different classifiers across all considered datasets, it can be noticed that the glass-box Decision Tree model gives comparable results to black-box models. For the most difficult dataset (glass), it gives the highest precision, keeping recall and accuracy at a competitive level.

[3] See: https://archive.ics.uci.edu/ml.
[4] See: https://scikit-learn.org/stable/datasets.html.
[5] See: https://www.kaggle.com/datasets.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55

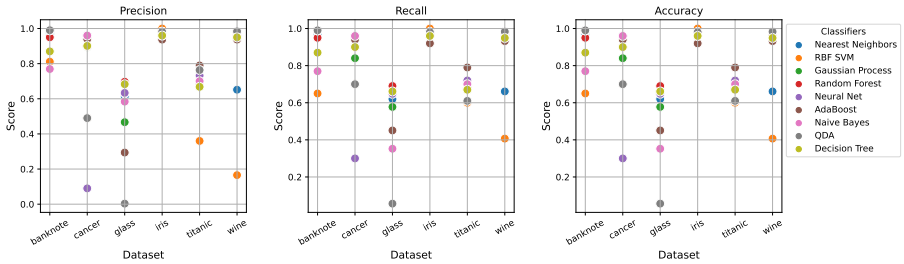

Fig. 2. Classifiers score comparison.

Figure 3 presents the comparison of the glass-box model results with the black-box models explanations obtained with XAI methods and executed with the use of HEARTDROID engine. In this figure, only the best scores for each dataset for each of the XAI methods are presented to compare the best possible results that can be obtained with a particular XAI method. In the case of using of XAI methods to generate rules, in Fig. 3 we can observe that the Anchor algorithm gives the best results in most datasets, but the Lux and Lore algorithms give noticeably worse results. In the case of the Lore algorithm, we noticed some of the bugs which resulted in scores equal to 0 which were marked on the charts. Only for the iris dataset (probably the easiest one), the Lux XAI method gives better results than the Anchor algorithm. Comparing the results from the exampled black-box models with Decision Tree, it can be noticed that for simple datasets like banknote, iris, or glass, the glass-box model gives better results. For slightly more difficult but still simple datasets (cancer and wine), slightly better results are obtained with the Anchor algorithm than the Decision Tree. However, the difference is not significant. In the case of more complex datasets like Titanic, glass-box model gives considerably worse results than the black-box model explained with the Anchor method.

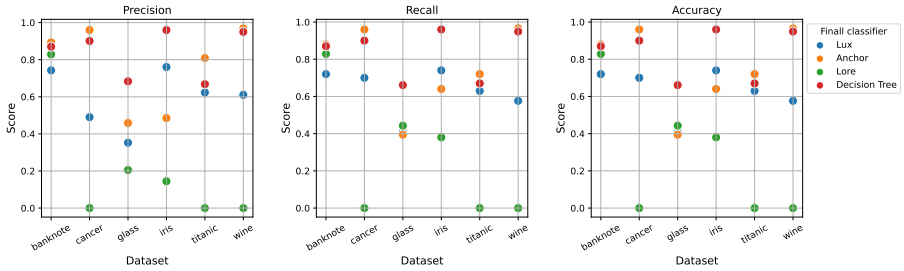

Fig. 3. XAI methods score comparison to glass-box model.

Fig. 4 shows the variance of the performance obtained with the Anchor algorithm, as it gives the best results from the considered XAI methods, depending on the classification model used. We can observe the biggest score variance for cancer and wine datasets (relatively simple) for which the results strongly depend on the classifier model explained. The lowest variance is observed for glass and iris datasets, so the most and the least difficult datasets.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55

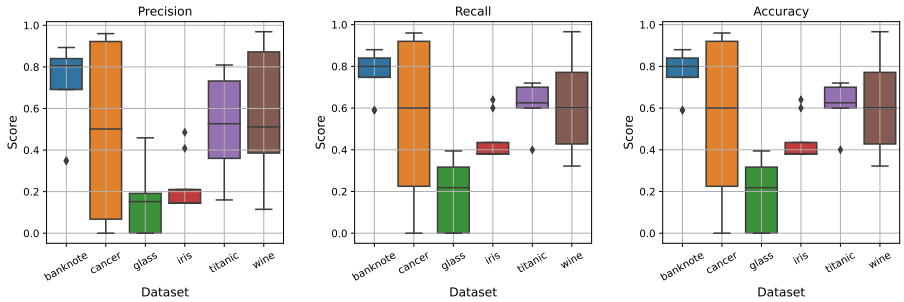

Fig. 4. XAI methods dependency on classifiers.

The obtained results allow us to compare how the explanations derived from black-box models combined with XAI algorithms differ from those obtained by interpretable glass-box models. Executed experiment proves that despite the fact that the black-box models are more complex and universal than the simple glass-box models, there is no need to apply them, especially for simple datasets. We found out that Decision Tree classifier gives competitive results and provides the model in an easily understandable format. However, in some examples, even in relatively simple datasets, it can be beneficial to apply more complex explanation methods (black-box model linked with XAI method) than simple glass-box models. Obtained results suggest also that when we need to consider more complex cases, better results can be obtained using the black-box models explained with XAI methods with human-readable rules. However, the final results strongly depend on the selected classifier. Hence, the properly chosen model which is treated as an input to the XAI method is important and has a significant impact on the final result.

5 Summary

In this paper, we made a glass-box and black-box classifiers comparison with the application in the explainable artificial intelligence area. The main goal of the work was to investigate if we should use the glass-box models to directly generate explanations or rather use a complex black-box model linked with XAI methods? We compared the classification scores for several classification methods and then we used the same trained models to obtain results with the use of XAI algorithm methods. We conducted our experiments based on the publicly available datasets. The results suggest that especially in the case of tabular data, it is worth investing resources into research on inherently explainable models, instead of relying on a combination of black-box and XAI algorithms. However, taking into account more complex analyses that concern e.g. embeddings or latent semantic analysis uses of glass-box models could be insufficient and then, black-box models with XAI methods could be applied. However, the choice should be made carefully, with additional evaluation of XAI results to select the most suitable approach. In future work, we plan to extend this analysis, taking into account different types of data, including time series and images and combine it with explanation evaluation methods [2] to provide a comprehensive study on XAI and glass-box models applicability.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55

Acknowledgements This paper is funded from the XPM (Explainable Predictive Maintenance) project funded by the National Science Center, Poland under CHIST-ERA programme Grant Agreement No. 857925 (NCN UMO-2020/02/Y/ST6/00070)

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
2. Bobek, S., Bałaga, P., Nalepa, G.J.: Towards model-agnostic ensemble explanations. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M. (eds.) Computational Science – ICCS 2021. pp. 39–51. Springer International Publishing, Cham (2021)
3. Bobek, S., Kuk, M., Brzegowski, J., Brzychczy, E., Nalepa, G.J.: Knac: an approach for enhancing cluster analysis with background knowledge and explanations. CoRR **abs/2112.08759** (2021), https://arxiv.org/abs/2112.08759
4. Bobek, S., Nalepa, G.J.: Introducing uncertainty into explainable ai methods. In: Paszynski, M., Dieter, K., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds.) Computational Science – ICCS 2021. Springer International Publishing (2021), accepted
5. Bobek, S., Nalepa, G.J., Ślażyński, M.: HeaRTDroid – rule engine for mobile and context-aware expert systems. Expert Systems **36**(1), e12328 (2019). https://doi.org/10.1111/exsy.12328, https://doi.org/10.1111/exsy.12328
6. Fung, P.L., Zaidan, M.A., Timonen, H., Niemi, J.V., Kousa, A., Kuula, J., Luoma, K., Tarkoma, S., Petäjä, T., Kulmala, M., Hussein, T.: Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration. Journal of Aerosol Science **152**, 105694 (2021). https://doi.org/https://doi.org/10.1016/j.jaerosci.2020.105694, https://www.sciencedirect.com/science/article/pii/S0021850220301798
7. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. ArXiv **abs/1805.10820** (2018)
8. Kaczor, K., Nalepa, G.J.: Critical evaluation of the xtt2 rule representation through comparison with clips. In: KESE@ECAI (2012)
9. Kuk, M., Bobek, S., Nalepa, G.J.: Explainable clustering with multidimensional bounding boxes. pp. 1–10 (2021). https://doi.org/10.1109/DSAA53316.2021.9564220
10. Loyola-González, O.: Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access **7**, 154096–154113 (2019). https://doi.org/10.1109/ACCESS.2019.2949286
11. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability (2019)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
13. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI (2018)
14. Xu, K., Liu, S., Zhao, P., Chen, P.Y., Zhang, H., Fan, Q., Erdogmus, D., Wang, Y., Lin, X.: Structured adversarial attack: Towards general implementation and better interpretability (2019)
15. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_55