October 2019

Machine Learning Explainability in Breast Cancer Survival

Tom JANSEN [a,b], Gijs GELEIJNSE [b], Marissa VAN MAAREN [b,c], Mathijs P. HENDRIKS [b,d], Annette TEN TEIJE [a], and Arturo MONCADA-TORRES [b,1]

[a] *Dept. of Computer Science, Vrije Universiteit Amsterdam, NL*
[b] *Netherlands Comprehensive Cancer Organization (IKNL), Eindhoven, NL*
[c] *University of Twente, Enschede, NL*
[d] *Dept. of Medical Oncology, Northwest Clinics Alkmaar, NL*

Abstract. Machine Learning (ML) can improve the diagnosis, treatment decisions, and understanding of cancer. However, the low explainability of how "black box" ML methods produce their output hinders their clinical adoption. In this paper, we used data from the Netherlands Cancer Registry to generate a ML-based model to predict 10-year overall survival of breast cancer patients. Then, we used Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to interpret the model's predictions. We found that, overall, LIME and SHAP tend to be consistent when explaining the contribution of different features. Nevertheless, the feature ranges where they have a mismatch can also be of interest, since they can help us identifying "turning points" where features go from favoring *survived* to favoring *deceased* (or vice versa). Explainability techniques can pave the way for better acceptance of ML techniques. However, their evaluation and translation to real-life scenarios need to be researched further.

Keywords. Artificial Intelligence, interpretability, oncology, prediction model

1. Introduction

Although it has been shown that Machine Learning (ML) methods are able to predict oncological outcomes [1], there are still a few factors that hinder their widespread clinical adoption. One of these factors is the lack of trust in the models. Often, ML tools are considered black boxes. If decisions need to be made that are based (at least partially) on predictions made by ML algorithms, users need to be able to understand how and why the algorithm has come up with that decision [2].

In the last couple of years, ML explainability has gained considerable interest. Recently, two techniques have been proposed and developed to make ML models more interpretable: Local Interpretable Model-Agnostic Explanations (LIME) [3] and SHapley Additive exPlanations (SHAP) [4]. However, evaluation of these tools remains relatively unexplored. Although some studies have attempted to evaluate explanations of

[1] Corresponding author: Arturo Moncada-Torres; Zernikestraat 29, 5612 HZ Eindhoven, NL; E-mail: a.moncadatorres@iknl.nl.

model predictions by letting users test which explanations are more understandable or intuitive [5], more analytical, standardized evaluations are needed.

In this paper, we used data from the Netherlands Cancer Registry (NCR) to generate a predictive model of 10-year overall survival (OS) after curative breast cancer surgery. Then, we applied LIME and SHAP to the obtained model to explain its predictions. Finally, we evaluated said interpretability methods by analysing their explanations and the agreement between them, which allowed us to identify features' turning ponts.

2. Materials & Methods

We used NCR data granted under data request K18.999. It consisted of demographic, clinical, and pathological data of patients in the Netherlands diagnosed between 2005 and 2008 with non-metastatic breast cancer who underwent surgery. Features included age, tumor characteristics, hormonal receptor statuses, clinical and pathological TNM-staging, and number of removed and positive lymph nodes. We imputed missing values using Deep Learning and K-Nearest Neighbor (KNN). We defined 10-year OS as the target variable for our model. The final dataset consisted of 46,284 patients and 31 features.

We performed feature selection using a combination of 21 different filter and wrapper methods. Each of them output a ranking ordering feature predictiveness. Then, we computed the median of these rankings and chose the six best ranked features: 1. *age*, 2. ratio between the number of positive and removed lymph nodes (*ratly*), 3. number of removed lymph nodes (*rly*), 4. tumor size in millimeters (*ptmm*), 5. pathological TNM stage (*pts*), and 6. tumor grade (*grd*).

We experimented with a variety of ML tools: Random Forest, Extreme Gradient Boosting (XGB), KNN, Artificial Neural Networks, Naïve Bayes, and Logistic Regression. We performed a randomized grid search to train, test, and optimize each of them. Since our target variable had a class distribution of roughly 75% (*survived*) versus 25% (*deceased*), we used stratified 10-fold cross-validation to optimize the models' hyperparameters. Then, we evaluated their performance using the Area under the Curve (AUC) as a metric, with XGB yielding the highest value (0.78). Therefore, we used XGB for the rest of this study.

In order to better understand the model's predictions, we used LIME and SHAP. On the one hand, LIME approximates *individual predictions* of a (black box) model with a *local* (interpretable) surrogate model that is as close as possible to the original one. Explanations are produced by minimizing a loss function between the predictions of both of them. The complexity of the surrogate model is used to explain the original model [3]. On the other hand, besides offering local interpretability, SHAP allows to explain a model *globally* by expressing it as linear functions of features [4]. In other words, it explains how much the presence of a feature contributes to the model's *overall predictions*.

To assess the consistency of LIME in explaining individual predictions, we applied it to each of the predictions of the test set (20% of the data) 100 times. We defined consistency when LIME assigned values with the same sign in all cases. Then, we evaluated global variable importance using SHAP. Finally, we tested agreement between LIME and SHAP values by comparing their instances (i.e., local) explanations in the test set. We defined agreement when both methods assigned either a positive or a negative contribution to the same feature of the same data instance.

3. Results

Figure 1 shows five representative data instances (i.e., patients) picked at random of the LIME consistency test. The *x*-axis shows the values for a particular feature, while the *y*-axis denotes the feature weight assigned to that value by LIME. A positive weight means it contributed to *survived*, while a negative one contributed to *deceased*. Across all plots, the position of each box is the same for each patient. For instance, the first box of each plot corresponds to the same patient, who was 51 years old, had a *ratly* value of 0, a *rly* value of 1, etc. LIME was consistent in >98% of the cases of *age*, *pts*, and *grd*. However, for *ratly*, *rly*, and *ptmm* LIME was consistent in 74%, 8%, and 55% of the cases, respectively. These cases correspond to the boxes in Figure 1 that cross 0 (dotted line), which means that LIME yielded contradictory weights.

Figure 2 shows the SHAP values of the six features used by the global model. The *x*-axis shows the SHAP values that correspond to the features shown on the *y*-axis. The color scale indicates the feature values, which range from low (blue) to high (red). Similarly to LIME, a positive SHAP value contributes to *survived*, while a negative one contributes to *deceased*.

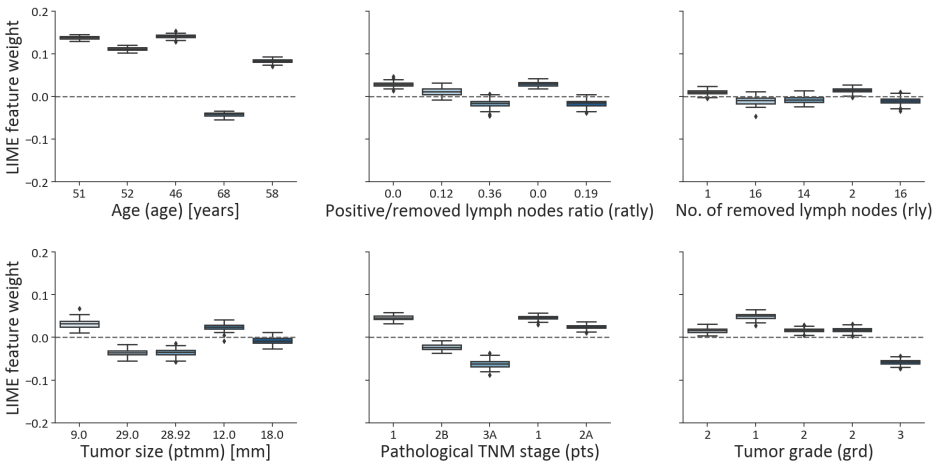

Figure 1. LIME consistency of five representative data instances (i.e., patients) picked at random. Contradictory explanations correspond to boxes that cross the dotted line at 0. Practically, *age*, *pts*, and *grd* had consistent LIME values, while the opposite can be said for *ratly*, *rly*, and *ptmm*.

The percentage of instances where LIME and SHAP agreed on their explanations for each feature was as follows: *age*, 97.5%; *ratly*, 95.9%; *rly*, 87.8%; *ptmm*, 91.9%; *pts*, 99.6%; *grd*, 99.9%. Figure 3 shows the individual feature weights for *age* (since it was the best-ranked feature) and for *rly* (since it was the feature with the largest disagreement). The *x*-axis denotes the feature values, while the *y*-axis denotes the feature weights assigned by the interpretability methods. Circles indicate an agreement between them, while crosses indicate the opposite.

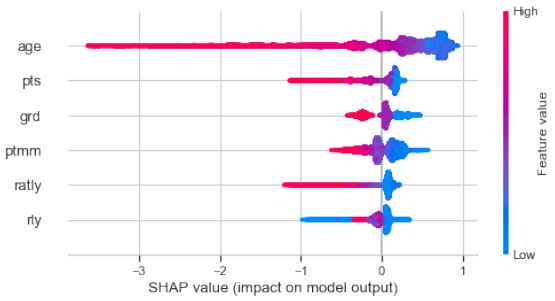

Figure 2. Summary plot of all SHAP values. Globally, *age* has the biggest impact on the model output.

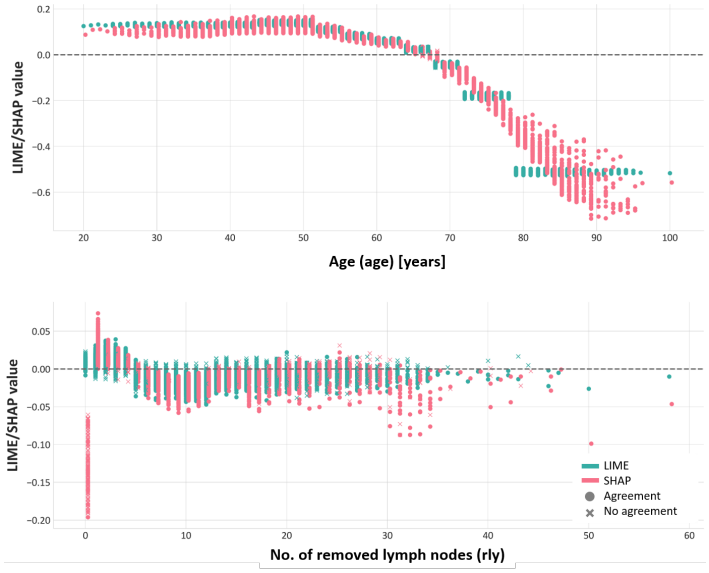

Figure 3. Comparison of the feature weights assigned by LIME and SHAP to all instances in the test set for *age* and *rly*. SHAP markers were shifted slightly along the *x*-axis for clarity.

4. Discussion

Figure 1 shows that LIME tends to assign consistent values to categorical features. For example, a *pts* of 1 is assigned almost identical feature weights in different patients. A similar thing occurs for a *grad* of 2. However, LIME has more difficulties with numerical features. For instance, there is very little difference in the impact that having 1 or 16 lymph nodes removed has on the model predictions. We think this is because LIME discretizes continuous features by binning them and treating them as categorical. Figure 1 also shows that LIME weights can be contradictory. For example, in the *rly* case, LIME values for the same patient are often inconsistent. It has been suggested that LIME's uncertainty can be explained by randomness in the sampling procedure and the variation of interpretation quality across different data instances [6], which is in line with the presented results.

Figure 2 combines the features' effects (x-axis) with their importance (y-axis). At a global level, *age* is the most important feature, while *rly* is the least important. This could be explained by its non-monotonic behaviour (i.e., low *rly* values are assigned positive and negative weights).

Although LIME and SHAP values show a similar trend overall for both *age* and *rly* (Figure 3), we can also distinguish specific regions of mismatch. These are of particular interest, since they can help us identify "turning points" in the features' values. For example, in the case of *age*, mismatches occur approximately between 65 and 68 years. This could explain where the model considers *age* to contribute towards *survived* or towards *deceased*.

5. Conclusion

In this study, we used breast cancer data from the NCR to generate an XGB-based model for predicting 10-year OS. We explained the model's predictions using LIME and SHAP and compared their performance. In few cases, LIME showed inconsistent and contradictory explanations of individual predictions. Furthermore, comparing LIME and SHAP showed agreement between them in 95.4% of the instances. The regions of mismatch allowed us to identify "turning points" in the features' values, which indicate where features go from favoring *survived* to favoring *deceased* (or vice versa).

Methods like LIME and SHAP are a first step to provide a more interpretable way of explaining complex models than what the models are capable of themselves. It is important to keep in mind that perfect explanations are also infeasible, since there is no gold standard to which the explanations can be compared. This also makes the evaluation of these methods a challenge. These type of methods pave the way for larger use and acceptance of ML techniques. However, their evaluation and translation to different fields need to be researched further.

References

[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[2] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" *Preprint arXiv:1712.09923*, 2017.

[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 Int. Conf. on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[5] A. A. Freitas, "Comprehensible classification models: A position paper," *ACM SIGKDD explorations newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[6] H. Fen, K. Song, M. Udell, Y. Sun, Y. Zhang, *et al.*, "Why should you trust my interpretation? Understanding uncertainty in LIME predictions," *Preprint arXiv:1904.12991*, 2019.