A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records

Jamie Duell*, Xiuyi Fan*, Bruce Burnett*, Gert Aarts*, Shang-Ming Zhou†
* *Swansea University*, Swansea, United Kingdom, {853435, xiuyi.fan, 989563, g.aarts}@swansea.ac.uk
† *University of Plymouth*, Plymouth, United Kingdom, smzhou@ieee.org; shangming.zhou@plymouth.ac.uk

***Abstract*—eXplainable Artificial Intelligence (XAI) aims to provide intelligible explanations to users. XAI algorithms such as SHAP, LIME and Scoped Rules compute feature importance for machine learning predictions. Although XAI has attracted much research attention, applying XAI techniques in healthcare to inform clinical decision making is challenging. In this paper, we provide a comparison of explanations given by XAI methods as a tertiary extension in analysing complex Electronic Health Records (EHRs). With a large-scale EHR dataset, we compare features of EHRs in terms of their prediction importance estimated by XAI models. Our experimental results show that the studied XAI methods circumstantially generate different top features; their aberrations in shared feature importance merit further exploration from domain-experts to evaluate human trust towards XAI.**

***Index Terms*—Explainable AI, Black-box, Glass-box, Machine Learning, Electronic Health Records**

I. Introduction

Though machines outperform human experts in some applications, one question remains: *how can we assure that the AI solutions are trustworthy?* Commonalities arise in algorithm applications, to where a model can exhibit different behaviours compromising the trust factor, this is where "black-box" models become an adversary to human trust, as understanding the internal mechanisms of such models is difficult, if not impossible [1].

The medical and health sciences have witnessed a growing interest of using Machine Learning (ML). However, in light of ensuring trust in human-AI collaboration, Tonekaboni et al. [2] have looked at the question *"what clinicians want?"* They identify that merely having highly accurate ML models is not sufficient for clinicians; notably a single metric such as classification accuracy does not provide insight to how the solution was obtained or provide depth to the models effectiveness [3]. Medicine needs a requisition of clarity due to the fragile nature of data in the field.

XAI is a sub-field of AI that aims to provide intelligible explanations to the end user [4]. There is a pressing need to develop XAI methods, tools and techniques, as traditional AI approaches suffer from doubts from human experts and general public [5]. This adheres to ethical concern and regulatory considerations that need to be made within the domain, should there be bias or discriminatory results. Healthcare is susceptible to such doubts; and XAI methods have been thus developed and applied in e.g., [6], [7] and [8]. It is believed that XAI will provide the much needed trust in human-AI collaboration for critical applications in healthcare.

In this paper, we apply three XAI methods to demonstrate the usability of an explainable tertiary appendix to ML models and provide data interpretability for large-scale EHR data. In particular, we study *lung-cancer mortality* with multiple ML and XAI models. Lung-cancer is the leading cause of cancer mortality in men and women worldwide [9]. Early predicting the mortality of lung-cancer patients can help identify patients that will benefit from treatment and those at risk of relapse, so help healthcare professionals develop preventative measures and treatment plans.

Data for this study used artificial data from the Simulacrum, a synthetic dataset created by Health Data Insight CiC derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service (NCRAS), which is part of Public Health England. The experimentation of XAI has focused on generating the feature importance regarding a models prediction indicative of contribution. Therefore, the use of XAI allows us to inform domain experts the trends identified through features that can be difficult for a human expert to deduce due to data quantity. This work compares state-of-the-art XAI approaches and serves as a demonstration of what XAI may deliver for healthcare applications.

II. Background

We focus on experimenting three XAI techniques: SHAP, LIME and Scoped Rules. They are *feature attribution* methods which, for a ML prediction, assign "weights" to features used to predict. We briefly review them below.

Shapley Additive exPlanations (SHAP) [10] explains an ML prediction based on feature attribution towards the prediction. The use of Shapley values directs the model on how to fairly distribute the importance of the feature(s), determined by the feature investments. The implementation of Shapley value sees the incorporation of a characteristic function G, for a number of features $F = \{0, 1, 2, \ldots, n\}$ ensuring that coalitions $G \subseteq F$, where the Shapley value

This work is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence, Machine Learning & Advanced Computing, funded by grant EP/S023992/1. Xiuyi Fan is supported with a funding contribution from the Welsh Government Office for Science, Sêr Cymru III programme – Tackling Covid-19.

calculates average contribution over a permutation of features. Additive explanations can be defined as a linear function of binary inputs, defined as

$$g(z') = \phi_0 + \sum_{k=1}^{M} \phi_k z'_k,$$

where g is the given explanation for an original prediction $f(x)$ and where z' is a coalition vector of simplified inputs, such that $z' \subseteq \{0,1\}^M$; M is the maximum number of simplified inputs. Explanatory models use simplified feature inputs due to the complexity of the original input data. Finally, ϕ_k is feature attribution for feature k, and $g(z')$ is the sum of all feature contributions for the linear model g, this is fit for z'.

Local Interpretable Model-agnostic Explanations (LIME) [11] measures whether or not the explanation is close to the prediction of the original model. LIME focuses on local interpretablility, achieved by accessing a single input feature that fits to a line of linearity using a regularization constraint to the linear regression model. To obtain an explanation for a local point x, the faithfulness of the explanation g to the original $f(x)$ is measured for a local prediction, defined as $L(f, g, \pi_x)$ applying the regularization parameters $\Omega(g)$, where L denotes the square loss function that is minimized. Explanation for x is such that:

$$E(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g).$$

Scoped rules (Anchors) [12] explains individual predictions of a classification model by prioritising a search for a decision rule to increase the prediction likelihood. A perturbation-based strategy is used to create an explanatory output of black-box machine learning models. Anchors follow the idea of factual explanations, identifying cases where certain anchors give conditional ”AND” statements that are identified as true.

Scoped rules provides a rule-based format of communication with IF-THEN rules in the explanation structure. The introduction to such extension of LIME provides a better localized understanding. An anchor can be defined as

$$\mathbb{E}_{D_x(z|A)} \left[1_{f(x)=f(z)} \right] \geq \tau, A(x) = 1,$$

where predicate A is an anchor if the expected evaluation of the neighbours of instance x of the distribution D matching A is greater-than or equal-to the precision boundary set on some threshold τ.

III. Data Preparation and ML Prediction

In this work, we compare XAI methods for explaining predictions made towards the likelihood of mortality for lung-cancer patients. Specifically, we pre-process the Simulacrum dataset and solve the following classification problem:

> *Given a lung-cancer patient with features collected from the Simulacrum dataset (see Table I as an example), predicting whether he is likely to survive.*

Table I
A SAMPLE PATIENT RECORD IN THE SIMULACRUM DATASET AFTER PRE-PROCESSING.

Age	75	**Grade**	G3
Sex	Male	**Morph**	8041
Weight	71.2	**Cancer Plan**	Curative
Dose Administration	150	**Outcome**	Treatment completed as prescribed
Drug Group	Etoposide	**Administration Route**	Oral
Behaviour	Malignant	**Regimen Time Delay**	No
T Best	4	**Regimen Stopped Early**	No
N Best	3	**Regimen**	Cisplatin + Gemcitabine
M Best	1	**Clinical Trial**	2
Cycle	1	**Site**	C34
Height	1.57	**CNS**	99
Chemo Radiation	No	**ACE**	9

with a classifier and then generate explanations to the classification with three XAI approaches.

The Simulacrum data set consists of 1,322,100 synthetic cancer patients, allowing for model development and evaluation whilst maintaining patient confidentiality, reflecting a high degree of accuracy to properties found in the NCRAS dataset.[1] We first apply data pre-processing, performed to construct a clean data set to support ML usability. To clean data, we remove null values and obvious errors. For instance, logical inconsistencies in the data, such as patients having weight or height that is unrealistic, or where a patient is listed as undergoing a regimen after death are removed. These conditions are amended based on logical assumption. Following this, the data is then balanced to the lower bounds bias for developing a trustworthy AI model.

We compose the data in a tabular format with each row corresponding to a single patient record. We treat each tabular category as a feature. Then, we obtain the contribution towards one of the two output classes, “Alive” or “Deceased”, representing whether the patient survived or not, respectively. We assume that the model will provide knowledge association in explanations regarding domain specificities, reflective of input features.

Before running XAI methods, we first execute black-box classifiers to generate predictions. Two deterministic algorithms inheriting an explanation are Logistic Regression and XGBoost. We also compare the performance of baseline algorithms against a glass-box method, *Explainable Boosting Machine (EBM)* [13], with all results displayed in Table II. Our testing dataset contains 49,456 Lung Cancer patients, randomly selected from the Simulacrum dataset, with 48.94% decease and 51.06% alive; the models are trained on 70% of the given data and tested on remaining 30%.

Table II
CLASSIFICATION PERFORMANCE OF LR, XGBOOST AND EBM.

	Precision (%)	**Recall**(%)	**Accuracy**(%)
Logistic Regression	68	68	68
XGBoost	78	78	78
EBM	67	67	67

It is evident that XGBoost is the best performing algorithm in terms of classification accuracy. We thus drop both Logistic Regression and EBM approaches and use XGBoost as the baseline algorithm for the explanation extension of XAI.

[1] http://www.ncin.org.uk/about_ncin/

IV. Explaining Classifications with XAI

We apply SHAP, LIME and Scoped Rules to the given lung-cancer mortality problem and compare their explanations. For illustration, we present the *global explanation* obtained with SHAP on our testing dataset in Figure 1. We observe that "M-Best", *Presence or Absence of Distant Metastatic Spread*, "T-Best", *Size and extent of the primary tumor* and "N-Best", *Extent of involvement of regional lymph nodes* provide the most attribution towards model output. Similar figures on global explanations can be produced by LIME and Anchor. They are omitted due to space limit.

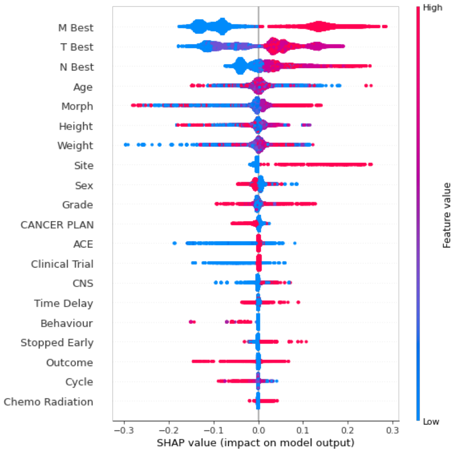


Figure 1. SHAP global explanation. The x-axis provides a weighting, Red being a shift towards death which is countered by blue being alive, with the value "0.0" indicative of minimal impact.

To conduct quantitative comparison between XAI algorithms, we focus on *local (instance) explanation*. Figures 2, 3 and 4 present explanation visualisation from the three methods, respectively, for the patient instance given in Table I. We observe that all three methods identify "M-Best", as the most important feature for the classification (the value "1b" means that cancer has spread to other parts of the body). However, Scoped Rules has identified no other feature being influential to the prediction, and the ranking for the remaining features differs between SHAP and LIME. E.g., SHAP consider "N-Best", as the second most important feature to the prediction whereas LIME considers "Behaviour", *Behaviour of the tumour*, as the second most important.

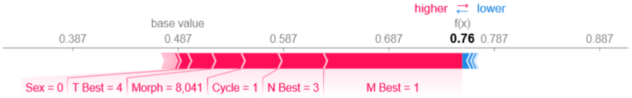


Figure 2. SHAP local explanation. The width of each descriptive block and colour are indicative of the shift in probability for the instance.

Noticing the discrepancy amongst SHAP and LIME, we study the scale of their differences. As top features can be identified by absolute value irrespective of the classification outcome, we count the cases where SHAP and LIME differ on the top k features, for both "Alive" and "Deceased" cases.

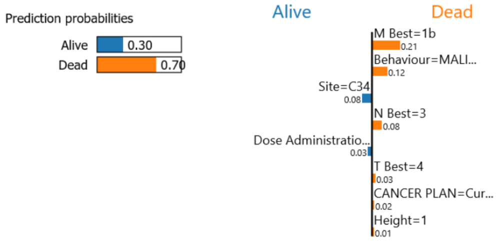
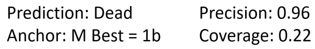


Figure 3. LIME providing an explanation given probability of Alive / Deceased cases. Supported by each feature value and importance towards the corresponding class of feature weighting.



Prediction: Dead Precision: 0.96
Anchor: M Best = 1b Coverage: 0.22

Figure 4. Anchors give a conditional junction of cases. In this case there is a single Anchor.

In other words, we count cases where SHAP and LIME share a given feature F for the kth ranking, written as SHAP(F_k) = LIME(F_k), with $k = 1$, 2 and 3, with results shown in Figure 5. From these shared features, it can be seen that there exist inconsistencies amongst the relationship where the indexed feature is the same across the explanations. But, in differential comparisons, e.g. where SHAP(F_1) = LIME(F_2) there are a high majority of features as priority on the most important feature. Note that Scoped Rules are not included in this analysis as the algorithm does not always find an anchor given τ is set to 0.95, to increase the quality of response, over a collection of instances, this is not applicable to all cases. Scoped Rules and LIME are kindred in calculation; therefore, the results hold little integrity. Conversely, SHAP and LIME differ in calculation providing a more meaningful comparison.

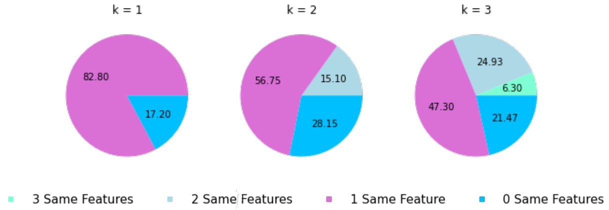


Figure 5. Comparing the shared features using SHAP and LIME.

Given such information, priority features can be determined from the dataset, though they may not be shared for each instance between explainers. It is optimal that the identified important features are relational to a degree, providing a form of validation supporting the consistency of knowledge representation. Therefore, this study provides a demonstration of feature importance, extracted from the first 1000 instances taken from the test data, where we extract the most important feature under the condition that $k = 1$ for each of LIME and SHAP, as well as the first Scoped Rules anchor. In this way, the most influential factors were identified towards a patient either ["Deceased", "Alive"] as shown in Figure 6.

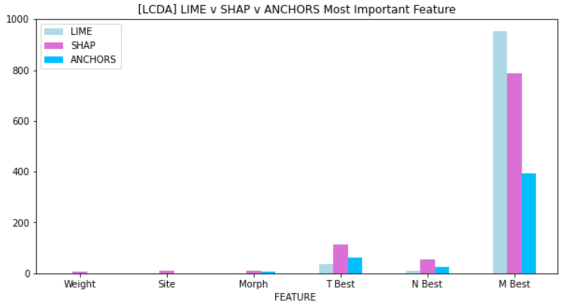


Figure 6. Most important feature returned or the first anchor (scoped rules) for the first 1000 instances on the test data.

It can be seen that the top 3 most important features are M-Best, N-Best, and T-Best. This is a particularly interesting result as it confirms the superiority of TNM based cancer stage classification [14]. Cancer staging is a critical step in the diagnosis process with multifarious objectives [14], such as helping identify treatment plans, providing indication of prognosis, showing the evaluation of the results of treatment, and facilitating the exchange of information of cancer development. The findings of these 3 top features are also consistent with another data-derived cancer prognosis study [15] indicating that TNM stage remains the most important prognostic factors, while being followed by tumor histologic grade, patient sex, age, and performance status. Conversely, note that our study shows that cancer morphology and patient weight are the following important factors for predicting mortality of lung-cancer patients (Figure 1).

V. Conclusion

XAI has been considered as an answer to the ML trust problem in healthcare. In this work, we have compared state-of-the-art XAI techniques on a large-scale EHR dataset in answering the lung-cancer mortality question. We show that the SHAP illustrations bring clarity when communicating both a local and a global explanation to a problem, thus providing more than just a prediction supporting what clinicians want through reasoning. We believe that the tertiary extension of knowledge can improve the rate of case deduction and support human-expert reasoning, and improve trust. Although all three methods, SHAP, LIME and Scoped Rules, have identified M-Best as the single most important feature in deciding a patient's mortality, coinciding with known medical knowledge, they differ on identifying secondary or tertiary features. Thus, although the explanations of XAI models can generate clear feature importance, which help inform clinical decision supports, they cannot work to substitute a human expert. To support the discernible necessity of human-expert trust, a user study will be conducted in future work to determine effectiveness of given explanations.

References

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence Volume 267, February 2019, Pages 1-38*, 2019.

[2] S. T. et al., "What clinicians want: Contextualizing explainable machine learning for clinical end use," *arXiv preprint 1905.05134*, 2019.

[3] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.

[4] W. F. M. Lent and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," *Proc of AAAI*, 2004.

[5] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Open Access Journal*, 2018.

[6] M. Aghamohammadi, M. Madan, J. K. Hong, and I. Watson, "Predicting heart attack through explainable artificial intelligence," in *Computational Science - ICCS - 19th International Conference*, Springer, 2019, pp. 633–645.

[7] S. M. Lauritsen, M. Kristensen, M. V. Olsen, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, *Explainable artificial intelligence model to predict acute critical illness from electronic health records*, 2019. arXiv: 1912.01266.

[8] S. Khedkar, G. Shinde, V. Subramanian, and P. Gandhi, "Explainable ai in healthcare," *Proc. ICAST 2019*, 2019.

[9] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global epidemiology of lung cancer," *Ann Glob Health*, vol. 85, no. 1, 2019.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in NeurIPS 30: Annual Conference on NeurIPS*, 2017, pp. 4765–4774.

[11] S. S. M. Ribeiro and C. Guestrin, "
why should i trust you?" explaining the predictions of any classifier," *arXiv:1602.04938*, 2016.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,*, AAAI Press, 2018, pp. 1527–1535.

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

[14] H. Lemjabbar-Alaoui, O. Hassan, Y.-W. Yang, and P. Buchanana, "Lung cancer: Biology and treatment options," *Biochim Biophys Acta*, vol. 1856, no. 2, pp. 189–210, 2015.

[15] G. A. Woodard, K. D. Jones, and D. M. Jablons, "Lung cancer staging and prognosis," in *Lung Cancer: Treatment and Research*. Cham: Springer International Publishing, 2016, pp. 47–75.