Jaccard Index (SHAP & LIME): 0.54
Jaccard Index (SHAP & DICE): 0.45
Jaccard Index (LIME & DICe): 0.45

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Is Explainable AI ready for Digital Health?

Wei Feng Sim[a], Jamie Duell[b] and Xiuyi Fan[a]
[a]Nanyang Technological University, Singapore
[b]Swansea University, United Kingdom

Introduction

Traditional ML models often lack interpretability, posing challenges in critical domains like healthcare where explanations are crucial [1], since majority of ML models are traditional black-box models. Hence, explainable AI (XAI) techniques are developed to address the black-box issue, especially when there is a need for good explanations in predictions for healthcare. This study extends an existing XAI comparison study [2] by analysing a further selection of metrics. We examine state-of-the-art eXplainable AI (XAI) techniques Shapley Additive exPlanations (SHAP) [3], Local Interpretable Model-Agnostic Explanations (LIME) [4] and Diverse Counterfactual Explanations (DiCE) [5].

We evaluate these using the following metrics:

- Spearman's rank correlation;
- Pearson correlation;
- Jaccard similarity index

We aim to determine the consistency in explanations and evaluate the readiness of XAI for application in electronic health records.

Methods

- The dataset contains lung cancer patients filtered and retrieved from synthetic data provided by the Simulacrum version 1.20.
- We use a dataset containing 950 lung-cancer patient data, with (70/30)% train/split.
  - Categorical Features: 16
  - Continuous Features: 6
  - Total Features: 22
- We apply SHAP, LIME and DiCE on MLPRegressor to generate the explanations for comparison.
  MLPRegressor Accuracy: **RMSE** – 65.063273, **R^2 Score** – 0.9909

In Figures 1a, 1b and 1c, we illustrate the top 8 returned explanations produced by each XAI method.

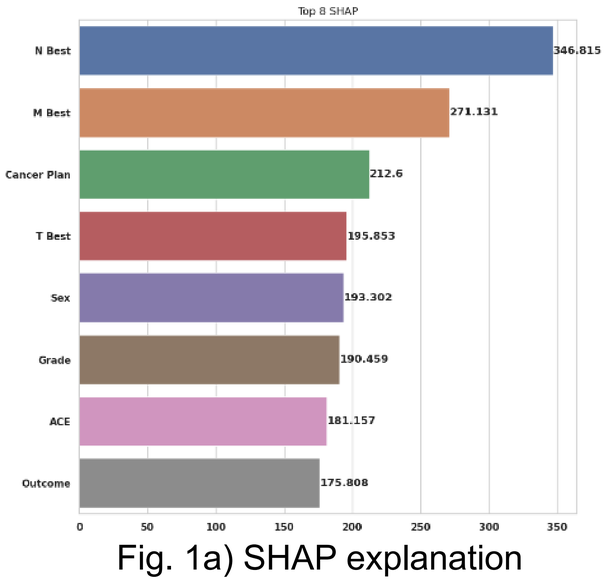


Fig. 1a) SHAP explanation

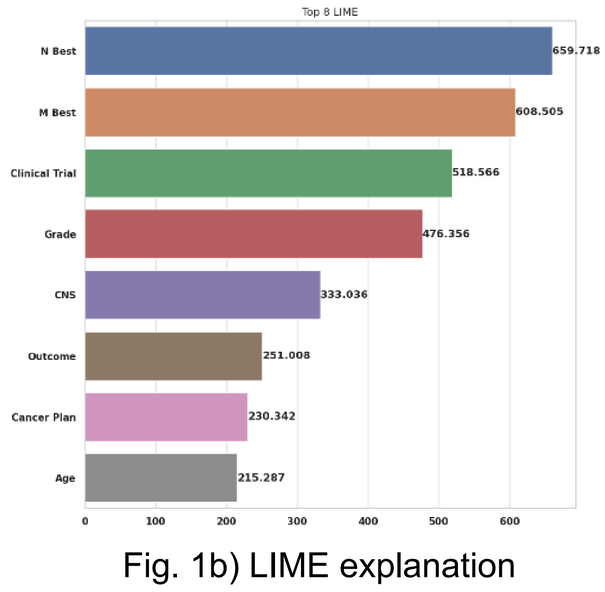


Fig. 1b) LIME explanation

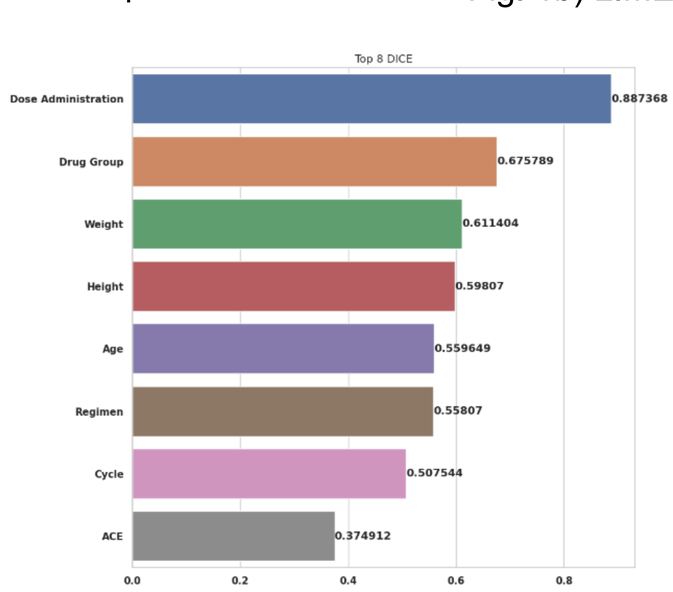


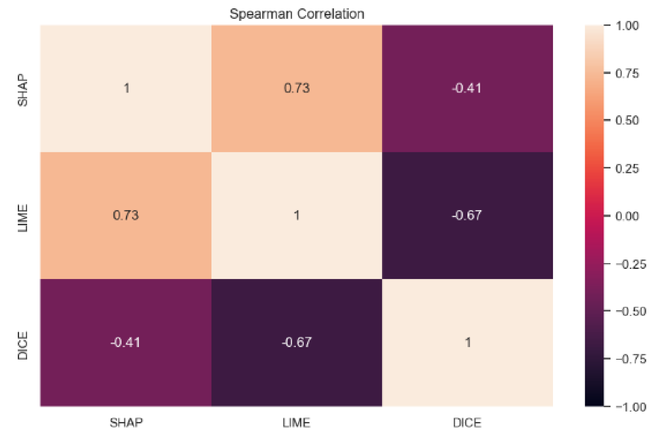
Fig. 1c) DiCE explanation

Metrics

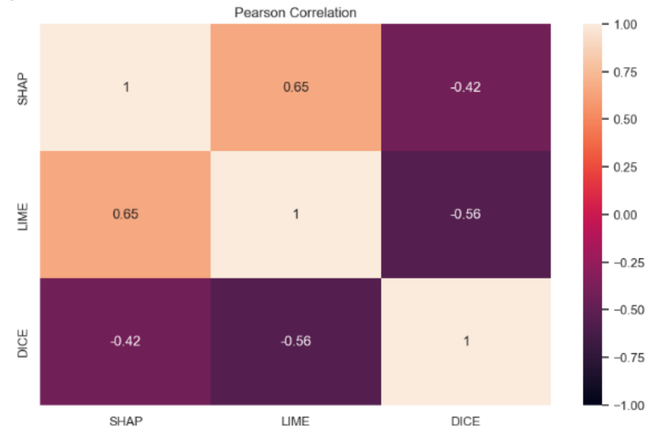
Spearman Rank Correlation	$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$
Pearson r Correlation	$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$
Jaccard Similarity Index	$J(A,B) = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$

Results

Spearman Rank Correlation:



Pearson r Correlation:



From the correlations above, we observed that DiCE have a negative correlation with both SHAP and LIME. For the Jaccard Similarity Index, we flipped the features rankings provided by DiCE and get the Jaccard Similarity Index between the three techniques.

Jaccard Similarity Index (k = 10):

Pairs	Jaccard Similarity Index
SHAP & LIME	0.54
SHAP & DiCE	0.54
LIME & DiCE	0.54

Conclusion

- SHAP and LIME are similar when they are ranking their features, here we see LIME and SHAP have p = 0.73.
- DiCE aberrates from SHAP and LIME in the returned explanation ranking.
- Correlation between all methods is in the range of $-0.56 \leq r \leq 0.65$.
- Jaccard Similarity where k=10 gives a similarity of 0.54 between LIME, SHAP and DiCE. However, DiCE have just as high of the Jaccard similarity due to the flipping of the ranking features. From this observation and the correlations, it can be determined that DiCE does not seem to agree with the other two XAI techniques explanations.

References

[1] A. Pannu, "Artificial intelligence and its application in different areas," Artificial intelligence, vol. 4, no. 10, pp. 79–84, April 2015.
[2] J. Duell et al., "A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records," BHI '21
[3] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in NeurIPS '17.
[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proceedings of KDD '16,
[5] R.K.Mothilal,A.Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proceedings of FAT* '20, Barcelona, Spain.