Towards Polynomial Adaptive Local Explanations for Healthcare Classifiers

Jamie Duell⋆1[0000−0002−8837−7843], Xiuyi Fan[2], and Monika Seisenberger1[0000−0002−2226−386X]

[1] School of Mathematics and Computer Science, Swansea University, Wales
{853435, m.seisenberger}@swansea.ac.uk
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore
xyfan@ntu.edu.sg

Abstract. Local explanations aim to provide transparency for individual instances and their associated predictions. The need for local explanations is prominent for high-risk domains such as finance, law and health care. We propose a new model-agnostic framework for local explanations "Polynomial Adaptive Local Explanations (PALE)", to combat the lack of transparency of predictions through adaptive local models. We aim to explore explanations of predictions by assessing the impact of instantaneous rate of change in each feature and the association with the resulting prediction of the local model. PALE optimises a complex black-box model and the local explanation models for each instance, providing two forms of explanations, one provided by a localised derivative of an adapting polynomial, thus emphasising instance specificity, and the latter a core interpretable logistic regression model.

1 Introduction

The use of eXplainable Artificial Intelligence (XAI) methods enable clarity for the communication of black-box model predictions enabling a person's rights for a '*right for explanation*' in Europe's General Data Protection Regulation (GDPR) [14]. As of 2016, there exist variations of XAI surrogate models that explore different approaches to localised explanations, though the premise of XAI greatly predeceased the recent influx [5]. Perturbation methods have seen success and wide application in the medical domain [3] [9] [13] [17], popular examples being Local Interpretable Model-Agnostic Explanations (LIME) [11], SHapley Additive exPlanations (SHAP) [8] and Scoped Rules (Anchors) [12], where SHAP explores feature summary through additive marginal contribution evaluation and Anchors and LIME explore local surrogate models from a set of readily interpretable models e.g. linear regression.

In this work, we specifically aim to approach local explanations for tabular data with Electronic Health Records (EHR) being a fundamental asset to population and precision health research. In exploration of clinical care, it has been a standing point that local explanations hold high importance to promote trust [16]. For example, being in the field of precision medicine, explanations would naturally need to contain patient specificity,

⋆ This work is supported by the UKRI AIMLAC CDT, funded by grant EP/S023992/1.

to deal with a case-by-case basis of care. We've seen the development of tools utilising pre-existing XAI methods in addendum to data exploration and analytic techniques [6].

With the lack of consistency across explanations [4] they can prove to be untrustworthy. In order to better adapt local explanations on an instance level we need to provide optimized scale-ability, highlighting patient specificity. In an attempt to create clear, efficient and patient specific local explanations, we propose the Polynomial Adaptive Local Explanations (PALE) framework, an end-to-end model aiming to mutually optimize both a complex model and each local explanation with a focus on tabular data. This should enable the transparency of patient predictions in a local domain, by producing explanations on how each patient and each feature can impact the outcome through local surrogate models that adapt to patient specific cases, as such in this work we

1. Produce an end-to-end framework that optimises both the complex model and the local model for each instance;
2. Produce explanations based on the derived scaling polynomial models to understand uni-variate feature impact for local instances;
3. Produce explanations based on a logistic regression model to understand uni-variate feature impact for local instances;
4. Compare local explanations and local explanation performance across the different XAI methods.

2 Related Work

Exploration of local surrogate model explanations saw an effective rise posterior to the efforts of LIME. LIME is a model-agnostic method with a primary focus on local explanation where a local linear model is used on a perturbed set around the instance $\mathbf{x}$. An explanation $\mathcal{E}$ for local point $\mathbf{x}$ is defined

$$\mathcal{E}(\mathbf{x}^{(j)}) = \arg\min_{g \in G} L(f, g, \pi_{\mathbf{x}^{(j)}}) + \Omega(g),$$

where we have a local linear model g from a set of linear models G, aiming to minimise the error of the local linear model, where perturbations around instance $\mathbf{x}^{(j)}$ are subject to a neighbourhood π, where the fidelity of the local model is measured against the complex model f, through L. The Ω term is used to reduce the complexity of the local model g. Perturbations are created around the mean of the data set within one standard deviation following a Gaussian distribution. See [11] for details.

There are various branches of LIME, to which end, the original framework has been adapted and extended in various cases. The authors of deterministic-LIME (DLIME) [18] extend the LIME framework by producing an adaptive neighbourhood using k-nearest neighbours and hierarchical clustering in an attempt to provide consistent explanations. In [20] the authors introduce Stabilized-LIME (S-LIME) which also surrounds the improvement of perturbation points for better local explainability, stability in the former DLIME and S-LIME are measured using the Jaccard similarity coefficient. [10] introduces local explanations and example-based local explanations, where weighting is carried out using random forests for supervised neighbourhood selection.

In [1] the authors propose a ensemble approach to LIME, namely LimeOut in order to reduce the reliance of sensitive features, in order to achieve this the authors replicate a similar idea to drop out techniques that are used in neural networks, aiming to maintain model performance. The authors of [19] introduce Bayesian LIME (BayLIME), in efforts to obtain consistency in explanations and maintain model robustness through integration of prior knowledge and the adaptation of Bayesian reasoning.

Extrapolating to local model fits [15] introduces Tree-LIME, this replacing the local linear model with a decision tree based approach for local interpretability. The authors of [2] draw more comparable intentions, as the authors aimed to fit a quadratic model to extend the LIME local model, the intent to analyse the performance improvement against the linear model. Therefore, the development of this inspired the intent for creating a framework with instance specific explainability to any polynomial degree that fits best for a given case. Feature attribution methods have explored specific feature-types, where we see focus on continuous features, enhancing the idea for the selective perturbation strategy [7].

3 Method

3.1 PALE Framework

We propose a complete framework to optimise the complex model f over all data X, therefore, $f(X)$ denotes our black-box model, where we minimise the residual loss $\mathcal{L}_f$ of the complex model. Our model uses the same neighbourhood setting that is used in the LIME framework. We optimise the local explainer loss for each j^{th} instance, where $X = [\ldots, \mathbf{x}^{(j)}, \ldots]$. We search for the optimal local models $g_m \in G$, where G is a set of polynomial models, for an instance in the local neighbourhood $\pi_{\mathbf{x}^{(j)}}$. Local model error is minimised through $\mathcal{L}_{g_m^{(j)}}$, where the optimal m polynomial degree for each instance is obtained. The framework aims to produce local explanations over classification problems, therefore we assume the complex model f to be some classifier.

Adaptive Model Introducing PALE, the generated surrogate data set $\mathcal{Z}^{(j)}$ is weighted by some neighbourhood $\pi_{\mathbf{x}^{(j)}}$,for an instance of interest $\mathbf{x}^{(j)}$. The surrogate set can be represented by $\{z', \mathbf{y}\} = \mathcal{Z}^{(j)}$, where an instance $\mathbf{z}'_s$ in the surrogate set is defined by $\mathbf{z}'_s \in \mathbb{R}^{1\times N}$, the surrogate data is given by $z' \in \mathbb{R}^{M\times N}$ and labels $\mathbf{y} \in [0, 1]$. We let $f(\mathbf{z}'_s)$ for each instance $\mathbf{z}'_s$ be the labels of the surrogate set using the prediction probability as the target for $g_m(z')$.

We first aim to have a scaling polynomial fit for instance adaptation in order to both provide better localised model performance as well as to provide insight into feature attribution and the affect of feature alteration in the local domain. This is carried out through the optimisation of the objective function $\mathcal{L}(X; \Phi, \cdot, \Psi)$, to obtain the optimal parameter set for both the local and complex model. We optimise our complete objective function in one function to avoid inconsistencies in local explanations, ensuring we obtain the same random seed for perturbation strategies and data split.

$$\mathcal{L}(X;\Phi,\cdot,\Psi) = \underbrace{\mathcal{L}_f(X;\cdot)}_{\text{Complex model loss}} +\lambda_p(\cdot)$$
$$+\underbrace{\sum_{j=1}^{M}\mathcal{L}_{g_m^{(j)}}(\mathcal{Z}^{(j)};\Phi)}_{\text{Explainer loss}}+\lambda_p(\Phi)+\underbrace{\sum_{j=1}^{M}\mathcal{L}_{u^{(j)}}(\mathcal{Z}'^{(j)};\Psi)}_{\text{Logistic Loss}}+\lambda_p(\Psi)$$

$\mathcal{L}_{g_m^{(j)}}$ is used to minimize the loss where we use the root mean-squared error ($RMSE$) to determine localised model performance of some surrogate set, namely $\mathcal{Z}^{(j)}$ in the neighbourhood $\pi_{\mathbf{x}}^{(j)}$ determining error in each model to the m^{th} degree polynomial for a prediction $g_m(\mathbf{z}'_s)$ for the instance of the surrogate set, and the fidelity to the labels y_s assigned by $f(\mathbf{z}'_s)$. We carry this out for the number of instances in each surrogate set and minimise the loss, we do this for every instance $\mathbf{x}^{(j)}$.

$\mathcal{L}_u^{(j)}$ defines the loss function for the logistic regression function $u^{(j)}$ in the surrogate set $\mathcal{Z}'^{(j)}$, by default this is given by the uni-variate binary cross-entropy loss function for each feature z'_i of each instance z'_s in the surrogate set with respect to the true label for the instance in the surrogate set y'_s. The regularization parameter $\lambda_p(\Phi)$ of our local model in given example to be λ_2 ridge regression, in an attempt to avoid over fitting of the local models whilst keeping all features as non-zero weights.

We let $\lambda_p(\cdot)$ be a placeholder for the parameters regularized in the complex model to obtain a best fit e.g. coefficients in regression. Both Φ and Ψ concretely represent coefficients of the local regression model and logistic regression model. From the integration of this objective function, we can then obtain the optimal set of parameters returned for the ideal polynomial for the local model and using a select complex model and associated loss function, we then obtain the matrices of optimal coefficients Φ' and Ψ', where each row is a vector corresponding to coefficients for the $\mathbf{x}^{(j)}$ instance to be explained.

$$\{\Phi', \cdot', \Psi'\} := arg\,min_{\{\Phi,\cdot,\Psi\}}[\mathcal{L}(X;\Phi,\cdot,\Psi)]. \tag{1}$$

Once obtaining the optimal fit for the local model, we extract the best performing m^{th} degree polynomial as the model to explain. After determining the best fit for each instance and extracting the coefficient matrix Φ', we do this to obtain the optimal local models $g_m^{(j)}$ for each $\mathbf{x}^{(j)}$.

Adaptive Local Explanations With models obtained from equation 1 we generate explanations, we produce an ordered absolute value where the associative value corresponds to the feature importance ranked by its value $|\frac{\partial g_m^{(j)}}{\partial x_i}|$ for each feature i to gauge a descending order of feature importance. Generalising to a scaling polynomial fit, we can observe the partial derivative for the m^{th} polynomial degree, such that for each feature x_i we observe the affect of change, where every other feature is kept static $\mathbf{x}_{/i}^{(j)}$, therefore,

$$g_m^{(j)}(x_i + \Delta x_i, \mathbf{x}_{/i}^{(j)}) = g_m^{(j)}(x_i, \mathbf{x}_{/i}^{(j)}) + (\Delta x_i) \cdot \frac{\partial g_m^{(j)}}{\partial x_i}(x_i, \mathbf{x}_{/i}^{(j)}),$$

as such we obtain a complete set of polynomial model partial derivative based explanations over the given data set X. We refer to this set of polynomial explanations as $\mathcal{E}_p(X)$, where each row corresponds to a instance $\mathbf{x}^{(j)}$, and each column corresponds to the features,

$$\mathcal{E}_p(X) = \begin{bmatrix} \frac{\partial g_m^{(1)}}{\partial x_1} & \frac{\partial g_m^{(1)}}{\partial x_2} & \cdots & \frac{\partial g_m^{(1)}}{\partial x_k} \\ \frac{\partial g_m^{(2)}}{\partial x_1} & \frac{\partial g_m^{(2)}}{\partial x_2} & \cdots & \frac{\partial g_m^{(2)}}{\partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m^{(r)}}{\partial x_1} & \frac{\partial g_m^{(r)}}{\partial x_2} & \cdots & \frac{\partial g_m^{(r)}}{\partial x_k} \end{bmatrix}. \tag{2}$$

Precision We introduce a form of local precision, this is a user defined level of precision which is in the range [0,1]. The term γ, is a flexible user influenced term that binds whether an instance explanation is returned, to a given precision of local fidelity where a returned explanation given the value for $\gamma = 1$ would determine $|(f(\mathbf{x}^{(j)}) - g_m^{(j)}(\mathbf{x}^{(j)}))| = 0$. This meaning that the prediction of the local model g accurately represents the point of interest predicted from our complex model f, meaning $g_m^{(j)}(\mathbf{x}^{(j)}) = f(\mathbf{x}^{(j)})$. This is determined through a term given the complex and local model for an instance of interest and a measure of precision γ, such that,

$$\begin{aligned} Precision(g_m^{(j)}, f, \mathbf{x}^{(j)}; \mathcal{T}, \gamma) = |(f(\mathbf{x}^{(j)}) - g_m^{(j)}(\mathbf{x}^{(j)}))|, \\ \text{s.t. } Precision \leq 1 - \gamma. \end{aligned}$$

We also allow the user to select a target value, $\mathcal{T} \in \{0, 1\}$ (*1 by default in the binary case*), this will allow for the partial derivative of the local regression to be associated with some user defined $\mathcal{T}$ for an explanation. If the local model does not meet the precision requirements, the instance explanation will not be returned. Therefore, the purpose of this in the applied case is to return only locally precise explanations.

3.2 Logistic Explanation

In addition to the prior, we provide explanations with respect to the odds ratios (OR), through uni-variate logistic regression analysis on each feature in the perturbed set $\mathbf{z}_i'$. We introduce the logistic model as the function $u^{(j)}$, where $u^{(j)}$ is the local logistic regression model over a surrogate set for instance $\mathbf{x}^{(j)}$. The localised model is a univariate model to explore individual feature importance. To achieve this, we introduce a secondary surrogate set $\mathcal{Z}'$ where, $\{z', \mathbf{y}'\} = \mathcal{Z}'$. A feature vector is denoted by $\mathbf{z}_i' \in \mathbb{R}^{m \times 1}$ and associated label is a binary case $\mathbf{y}' \in \{0, 1\}$, therefore,

$$u^{(j)}(z_i') = P(\mathbf{y}' \mid z_i') = \frac{1}{1 + (exp(-(\Psi_i \times z_i')))}. \tag{3}$$

We introduce a modified version of OR to center odds at the value 0 for ease of interpretation, the logistic explanation $\mathcal{E}_l$ where Ψ_i is the returned log odds, can be represented by,

$$\mathcal{E}_l(\mathbf{x}_i^{(j)}) = exp(\Psi_i) - 1. \tag{4}$$

4 Comparative Methods

We introduce a comparisons of explanations returned by XAI methods. We include SHAP, a linear model, higher degree polynomials and logistic explanations as the XAI methods.

Jaccard Index We can explore the Jaccard similarity index for v features, for this paper we explore $v = 5$. The Jaccard index can be defined by $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, to compare returned sets of feature names between two XAI methods.

Pearson Correlation Coefficient We also compare the Pearson r correlation coefficient for the sets of explanations, given the absolute values returned from the XAI methods.

Logistic Comparison We can use the shift in odds ratio in either $\mathbb{R}^+$ or $\mathbb{R}^-$ of non-absolute value explanations, for each feature i of an instance, to determine similarity between the derived explanation and odds ratio explanation. We determine the ratio of shared explanation shift $LogCompare$ for any $\mathbf{x}^{(j)}$ over N features as,

$$LogCompare(\mathbf{x}^{(j)}) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{[x_i]}, & \text{if } sgn\left(\frac{\partial g_m^{(j)}}{\partial x_i}\right) = sgn\left(\mathcal{E}_l(\mathbf{x}_i^{(j)})\right), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

5 Results

Data for this study uses artificial data from the Simulacrum[3], a synthetic data set developed by Health Data Insight CiC derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service (NCRAS)[4], which is part of Public Health England. We extract a subset of lung cancer patients from the Simulacrum to demonstrate the proposed method. We focus on binary classification problems for the

[3] https://simulacrum.healthdatainsight.org.uk/
[4] http://www.ncin.org.uk/about_ncin/

demonstration of this framework. The binary classes we aim to predict are $<$ *6 Months* and $>$ *6 Months* survival time.

We use an XGBoost model with a 70% train and 30% testing data split as our complex model to demonstrate the explanatory model. The model performance is evaluated using the Root Mean Squared Error (RMSE), obtaining the following,

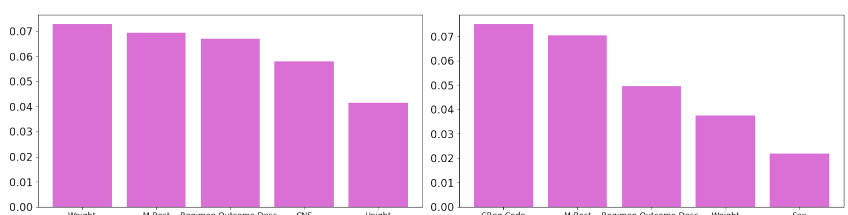
Class	Precision	Recall	F1-Score
$<$ 6 Months	0.97	0.97	0.97
$>$ 6 Months	0.98	0.98	0.98

Posterior to this, we determine a local patient instance of interest to explain.

- Age 66, Sex 0, Morph 8140, Weight 85.90, Height 1.67, Dose Administration 8, Chemo Radiation 0.0, Regimen Outcome Description 0.0, Admin Route 1.0, Regimen Time Delay 0.0, Regimen Stopped Early 1.0, Cycle Number 1.0, Grade 1.0, Cancer Plan 0.0, Cancer Registration Code 301.0, T Best 4.0, N Best 2.0, M Best 0.0, Laterality 2.0, CNS 1.0, ACE 9.0, Performance 0.0, Clinical Trial 2.0.

Prediction: $>$ *6 Months*,
Actual: $>$ *6 Months*.

We explore how higher degree polynomial functions can inform feature attribution on a local level. We use the partial derivative for the 2^{nd} and 3^{rd} degree polynomials, to determine how each feature i interacts with the output for our local model.



(a) Derivation of the quadratic polynomial term - Simulacrum patient instance.

(b) Derivation of the cubic polynomial term - Simulacrum patient instance.

Fig. 1: The explanation determines how an instantaneous increase in each feature value x_i influences the local polynomial function $g_m^{(j)}$, where we have $g_2^{(j)}$ for figure 3a, and $g_3^{(j)}$ for figure 3b.

Evaluating the explanations for the first 5 feature, we observe that the quadratic derivative determine *Weight*, *M Best* and the *Regimen Outcome Description* to have a high attribution in the local model. Conversely, when observing the 3^{rd} degree polynomial, we see *Cancer Registration code* followed by *M Best* and *Regimen Outcome Description* as the highest attribution in the local model.

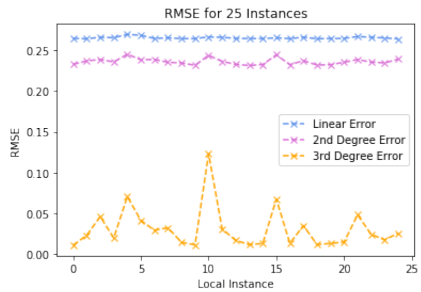
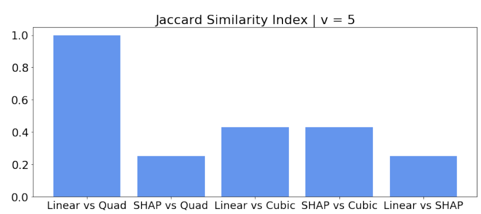


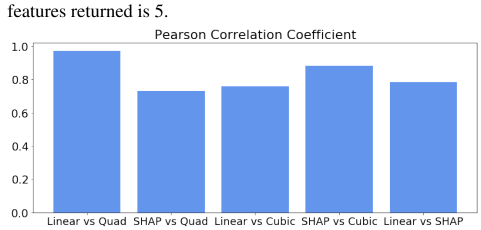
Fig. 2: RMSE measurements for a subset of 25 Simulacrum patient instances. We can observe how the increase in polynomial degree improves the local model accuracy.

We explore the performance of each model to the m^{th} degree polynomial, looking at the $RMSE$ returned for the local model $g_m^{(j)}$ for 25 instances $\mathbf{x}^{(j)} : j = \{1, 2, ..., 25\}$. From this, we determine that an increase in polynomial degree has significant impact on the local model performance over each surrogate set $\mathcal{Z}^{(j)}$.

XAI Models - Similarity Measures For the comparison of XAI models, we determine the Jaccard similarity index between the sets of g_m and the response given by SHAP. Although the PALE framework extracts the ideal polynomial degree and produces an explanation for each instance, we instead manually extract explanations for each degree and compare the similarities amongst each degree polynomial and SHAP.



(a) The Jaccard similarity index where the number of returned features returned is 5.



(b) Pearson correlation coefficient between XAI methods.

Fig. 3: A comparison of explanations given by the linear model, quadratic model, cubic model and the SHAP model for a patient instance.

We observe there exists the greatest Jaccard similarity between that of the 3^{rd} degree polynomial fit and SHAP. We also explore the pearson r correlation coefficient between each model and identify that the 3^{rd} degree polynomial holds a greater correlation with SHAP than other models for the given instance.

Interpretable Odds Ratio Similarity Exploring the agreement between both the quadratic and cubic explanations for the signed floating point values, as opposed to absolute values, so we can determine the amount of shared attribution between the logistic model and local polynomial derivations. From this, we obtain $LogCompare(\mathbf{x}^{(j)}) = 0.48$ for the quadratic model explanation and $LogCompare(\mathbf{x}^{(j)}) = 0.65$ for the cubic model explanation. Therefore, we observe in the given case, the cubic explanation has a greater similarity in explanation with the logistic model than that of the quadratic model.

6 Conclusion and Future Work

We use a similar classification problem as seen in [4], [7], where under similar predictions surrounding survival we see great influence from the likes of *M Best*, *Weight*, amongst other features. Therefore, we observe the selection of important features hold a degree of accuracy with clinical knowledge of cancer survival. The contribution of this work is an end-to-end framework that optimizes both the local and complex model to provide an explanation of how change to a feature will influence the outcome of the model prediction in the local setting. We emphasise the need for patient specificity, thus we produce an adaptive framework at the local level through adaptive polynomials.

We identify that the uni-variate approach shows single feature interaction with the local model, and although predictions are reliant on the kernel and localised feature perturbations which can lead to explanation instability, with ongoing research being focused in this area for the extension of LIME, we instead focus on improving the interpretable local model by adapting explanations to each local instance to increase local specificity. Extending upon this, the interpretable comparison with the logistic regression model poses questions towards the disagreement of explanations, to further analyse this, we will consider statistical significance against the explanations given. We acknowledge the problem of potential polynomial overfitting despite regularization. Further research will be carried out in order to approach the addressed issues and expand upon the framework.

References

1. Bhargava, V., Couceiro, M., Napoli, A.: Limeout: An ensemble approach to improve process fairness. In: ECML PKDD 2020 Workshops. pp. 475–491. Springer International Publishing, Cham (2020)
2. Bramhall, S., Horn, H., Tieu, M., Lohia, N.: Qlime-a quadratic local interpretable model-agnostic explanation approach. In: SMU Data Science Review: No. 1 , Article 4. vol. 3 (2020)
3. Dindorf, C., Konradi, J., Wolf, C., Taetz, B., Bleser, G., Huthwelker, J., Werthmann, F., Bartaguiz, E., Kniepert, J., Drees, P., Betz, U., Fröhlich, M.: Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (XAI). Sensors (Basel) **21**(18), 6323 (Sep 2021)

4. Duell, J., Fan, X., Burnett, B., Aarts, G., Zhou, S.: A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (IEEE BHI 2021). Athens, Greece (Jul 2021)
5. Främling, K.: Decision Theory Meets Explainable AI. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems. vol. 12175, pp. 57–74. Springer International Publishing, Cham (2020)
6. Kapcia, M., Eshkiki, H., Duell, J., Fan, X., Zhou, S., Mora, B.: ExMed: An AI tool for experimenting explainable ai techniques on medical data analytics. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 841–845 (2021)
7. Kovvuri, V.R.R., Liu, S., Seisenberger, M., Müller, B., Fan, X.: On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study. arXiv:2203.12701 (2022)
8. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Advances in NeurIPS 30: Annual Conference on NeurIPS. pp. 4765–4774 (2017)
9. Peng, J., Zou, K., Zhou, M., Teng, Y., Zhu, X., Zhang, F., Xu, J.: An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. Journal of Medical Systems **45**(5) (Apr 2021)
10. Plumb, G., Molitor, D., Talwalkar, A.: Model agnostic supervised local explanations. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 2520–2529. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
11. Ribeiro, M., Singh, S., Guestrin, C.: "Why Should I Trust You?" explaining the predictions of any classifier. arXiv:1602.04938 (2016)
12. Ribeiro, M., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,. pp. 1527–1535. AAAI Press (2018)
13. Sarp, S., Kuzlu, M., Wilson, E., Cali, U., Guler, O.: The enlightening role of explainable artificial intelligence in chronic wound classification. Electronics **10**(12) (2021)
14. Selbst, A.D., Powles, J.: Meaningful information and the right to explanation. International Data Privacy Law **7**(4), 233–242 (12 2017)
15. Shi, S., Zhang, X., Li, H., Fan, W.: Explaining the predictions of any image classifier via decision trees. ArXiv **abs/1911.01058** (2019)
16. Tonekaboni, S., Joshi, S., McCradden, M., Goldenberg, A.: What clinicians want: Contextualizing explainable machine learning for clinical end use. In: MLHC (2019)
17. Yoo, T.K., Ryu, I.H., Choi, H., Kim, J.K., Lee, I.S., Kim, J.S., Lee, G., Rim, T.H.: Explainable Machine Learning Approach as a Tool to Understand Factors Used to Select the Refractive Surgery Technique on the Expert Level. Translational Vision Science Technology **9**(2), 8–8 (02 2020)
18. Zafar, M.R., Khan, N.: Deterministic local interpretable model-agnostic explanations for stable explainability. Machine Learning and Knowledge Extraction **3**(3), 525–541 (2021)
19. Zhao, X., Huang, W., Huang, X., Robu, V., Flynn, D.: BayLIME: Bayesian local interpretable model-agnostic explanations. In: de Campos, C., Maathuis, M.H. (eds.) Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. Proceedings of Machine Learning Research, vol. 161, pp. 887–896. PMLR (27–30 Jul 2021)
20. Zhou, Z., Hooker, G., Wang, F.: S-LIME: Stabilized-LIME for Model Explanation. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining. p. 2429–2438. KDD '21, Association for Computing Machinery, New York, NY, USA (2021)