2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)

Interpretability in HealthCare: A Comparative Study of Local Machine Learning Interpretability Techniques

Radwa El Shawi
Tartu University
Tartu, Estonia
radwa.elshawi@ut.ee

Youssef Sherif
Zewail City
Cairo, Egypt
S-youssef.mansour@zewailcity.edu.eg

Mouaz Al-Mallah
Houston Methodist Center
Houston, USA
mal-mallah@houstonmethodist.org

Sherif Sakr
Tartu University
Tartu, Estonia
sherif.sakr@ut.ee

***Abstract*—Although complex machine learning models (e.g., Random Forest, Neural Networks) are commonly outperforming the traditional simple interpretable models (e.g., Linear Regression, Decision Tree), in the healthcare domain, clinicians find it hard to understand and trust these complex models due to the lack of intuition and explanation of their predictions. With the new General Data Protection Regulation (GDPR), the importance for plausibility and verifiability of the predictions made by machine learning models has become essential. To tackle this challenge, recently, several machine learning interpretability techniques have been developed and introduced. In general, the main aim of these interpretability techniques is to shed light and provide insights into the predictions process of the machine learning models and explain how the model predictions have been resulted. However, in practice, assessing the quality of the explanations provided by the various interpretability techniques is still questionable. In this paper, we present a comprehensive experimental evaluation of three recent and popular *local* model agnostic interpretability techniques, namely, `LIME`, `SHAP` and `Anchors` on different types of real-world healthcare data. Our experimental evaluation covers different aspects for its comparison including `identity`, `stability`, `separability`, `similarity`, `execution time` and `bias detection`. The results of our experiments show that LIME achieves the lowest performance for the identity metric and the highest performance for the separability metric across all datasets included in this study. On average, SHAP has the smallest average time to output explanation across all datasets included in this study. For detecting the bias, SHAP enables the participants to better detect the bias.**

***Keywords*-Machine Learning; Black-Box Model; Machine Learning Interpretability; Model-Agnostic Interpretability;**

I. Introduction

Machine learning models have been proven to be successful in many application domains such as financial systems, advertising, marketing, criminal justice system, and medicine. Despite the growing use of machine learning-based prediction models in the medical domains [5], clinicians still find it hard to rely on these models in practice. In practice, most of the models developed by data scientists mainly focus on prediction accuracy as a performance metric but rarely explain their prediction in a meaningful way [3], [7]. Generally speaking, there is a trade off between the performance of machine learning models and their interpretability. That is the more interpretable the model such as linear models and decision trees, the lower their performance would be compared to complex models such as deep learning models. Thus, there have been some criticism for using complex machine learning models in the medical domain even with their promise of high accuracy.

In general, the interpretability of machine learning predictions is defined as the degree to which users can understand and comprehend the predictions made by the machine learning models [14]. In particular, the main aim of interpretability is to assist in understanding the most influential features that lead the model to a specific prediction which would significantly help the clinicians to understand the reasoning behind a specific prediction and hence they will be able to accept or reject the prediction. In addition, understating the prediction process is always useful for getting some insights of how this model is working and can help in improving the prediction process and model performance for future predictions.

Recently, interpretability has been receiving a notable attention especially after the new GDPR imposed by the European Parliament in May 2018 that forces industries to 'explain' any decision taken when automated decision making takes place: "*a right of explanation for all individuals to obtain meaningful explanations of the logic involved*" [8]. In general, machine learning interpretability methods can be broadly classified into two categories: *global* or *local* [9]. In principle, global techniques enable clinicians to understand the entire conditional distribution modeled by the trained response function and obtained based on average values. On contrast, local interpretations promote the understanding of small parts of the conditional distribution for specific instances. Since conditional distribution decomposes of small parts that are more likely to be linear or well-behaved, hence they can be explained by interpretable models such as linear regression and decision trees.

With the increasing attention for the interpretability challenge, several techniques have been developed, however, assessing the quality of these interpretability techniques is still an open challenge [6]. This is mainly because there is

2372-9198/19/$31.00 ©2019 IEEE
DOI 10.1109/CBMS.2019.00065

no adequate well defined concise metrics for measuring the quality of the explanations of interpretability frameworks. In practice, one of the main challenges of measuring the quality of interpretability is that it is subjective; good explanation is different from one user to another [4]. Nevertheless, there are few studies focus on the evaluation of interpretability techniques [16]. In this work, we present a detailed experimental evaluation of three recent and popular *local* model agnostic interpretability techniques, namely, `LIME` [17], `Anchors` [18] and `SHAP` [21] on different types (tabular and text data) of real-world healthcare data using five different metrics, namely, `identity`, `stability`, `separability`, `similarity`, `execution time` and `bias detection` (used to evaluate tabular data only). For ensuring repeatability as one of the main targets of this work, we provide access to the source codes and the detailed results for the experiments of our study[1]. The remainder of this paper is organized as follows. Section II provides an overview of the different local interpretability techniques that have been considered in this study. Section III describes the details of our experimental setup in terms of used datasets and evaluation metrics. The detailed results of our experiments is presented in Section IV before we conclude the paper in Section V.

II. Overview on Interpretability Frameworks

In this section, we give an overview of the local interpretability techniques which we consider in this study.

A. LIME

The `LIME` technique [17] has been introduced as a local interpretability technique that relies on the assumption that the decision boundary of a complex machine learning model is linear locally around the instance to be explained. It explains the instance of interest by fitting an interpretable model on perturbed sample around the input instance of interest. In particular, LIME generates a perturbed sample around the instance to be explained. For each instance in the perturbed sample, LIME gets the prediction from the model to be explained (the perturbed sample along with the prediction will act as the training dataset for the interpretable model). Then, the technique assign weights to the instances in the new training dataset according to their proximity to the instance to be explained. Finally, LIME fits an interpretable model on the new training dataset.

B. Anchors

`Anchors` is a rule based model-agnostic local explainer technique [18]. In general, LIME explanations do not have a clear coverage; it is unclear whether the explanation given for a specific instance x is applicable in the region where x is located. Anchors guarantee that the predictions of instances in the same anchor is almost the same. In other words, Anchors find the features that are enough to fix the prediction such that changing the other features has no impact on the prediction. One way to construct anchors is the bottom-up approach [18] in which anchor is constructed incrementally. In particular, Anchors starts with an empty rule and in each iteration the rule is extended with one feature such that the new rule has the highest estimated precision. To select the best rule in each iteration, KL-LUCB algorithm is used [12].

C. SHAP

An idea from game theory was applied to measure the role of each feature on the prediction process. The Shapley value [20] is a method from coalitional game theory that can fairly distribute the gain among players (features), where contributions of players are unequal. In particular, Shapley value fairly distributes the difference between the prediction and the average prediction among the feature values of the instance to be explained. In practice, one of the main challenges in the Shapely value approach is the computation time. In particular, for exact computation of Shapley value, all possible sets (coalitions) of features need to be evaluated (with and without the feature of interest). The exact value calculation becomes hard to compute for large number of features as the number of sets increases exponentially with the number of features. To avoid this issue, sampling techniques were introduced to sample coalitions with fixed number of samples [13]. While other attempts proposed different computation method for Shapely value that comprises weight kernels and regularized linear regression [15] which will be evaluated in this work (SHAP).

III. Experimental Setup

A. Datasets

In our experiments, we used two types of datasets: *Tabular Datasets* and *Text Datasets*. The tabular datasets of this study have been collected from patients who underwent treadmill stress testing by physician referrals at Henry Ford Affiliated Hospitals in metropolitan Detroit, MI in the U.S, FIT Project [1]. In particular, the data has been obtained from the electronic medical records, administrative databases, and the linked claim files and death registry of the hospital between January 1st, 1991 and May 28th, 2009 [1]. The data set includes 43 attributes containing information on vital signs, diagnosis and clinical laboratory measurements. Examples of these attributes include *sex*, *age*, *race*, *%heart rate achieved*, *resting systolic blood pressure*, *resting diastolic blood pressure*, *obesity*, *hypertension*, *history of smoking*, and *METS*. In our previous work, we have developed two machine learning models for predicting the risk of mortality [19] and diabetes [2]. In the experiments of this work, we have used the two datasets of our previous studies: `mortality` dataset and the `diabetes` dataset. The cohort of the mortality dataset includes 34,212 patients

[1] https://github.com/DataSystemsGroupUT/Interpretability-comparison

who completed 10-year follow-up, while the cohort of the diabetes dataset includes 32,555 patients who completed 5-year follow-up. Random forest is the model that have shown the best performances in predicting the risk of mortality and diabetes over these datasets. For more information about the details of the dataset and modeling process of this study, we refer the interested readers to [19], [2].

The text datasets used in this study are the `Drug Review (Drugs.com)` and `Side Effects (Druglib.com)` datasets form the UCI Repository[2]. For these datasets, the random forest model has been also used for predicting whether the review is positive (more than 5 stars) or negative (less than or equal 5 stars) in the drug review dataset and whether the side effect is mild, moderate, severe, extreme or extremely severe in the side effects dataset.

B. Evaluation metrics

We used the following five metrics to compare between different the interpretability techniques:

1) *Identity*: this metric states that if there are two identical instances, then they must have identical explanations [10].
2) *Stability*: this metric states that instances belong to the same class must have comparable explanations [10].
3) *Separability*: this metric states that if there are two dissimilar instances, then they must have dissimilar explanations [10]. This metric holds assuming that the model does not have degree of freedom [22]; means that all the features used in the model are relevant to the prediction.
4) *Similarity*: this metric states that the more similar the instances to be explained, the closer their explanations should be and vice versa.
5) *Time*: this metric represents the average time used by the interpretability framework to output an explanation across all the instances in the testing dataset. The time evaluated on a standard machine with Intel Core i7 6500U and 8 GB RAM.
6) *Bias detection*: the ability to detect bias in training data (used to train a machine learning model) from the explanations of instances in the testing dataset.

IV. Experimental Results

A. Tabular Data

In our experiments, we follow the same pipeline on mortality and diabetes datasets: (i) partition the dataset into 80% for training and 20% for testing, (ii) train a random forest model on the training dataset, (iii) for each instance in the testing dataset, use LIME, SHAP and Anchors to explain the prediction of the model. In principle, measuring the `identity` of the explanations provided by the different techniques on the tabular datasets is straightforward. In particular, for every two instances in the testing dataset with, if the distance between the two instances equals to zero (identical), then the distance between their explanations should be equal to zero. To measure the `separability` metric, we choose a subset S of the testing dataset that has no duplicates and get their explanations. Then for every instance s in S, we compare its explanation with all other explanations of instances in S and if such explanation has no duplicate then it satisfies the separability metric. To measure the `similarity` metric, we cluster instances in the testing dataset, after normalization using DBSCAN algorithm. For each framework, we normalize the explanations and calculate the mean pairwise Euclidean distances between explanations of testing instances in the same cluster. The framework with the smallest mean pairwise Euclidean distances across its clusters is the best reflecting the similarity metric. Measuring the `stability` metric is done by clustering the explanations of all instances in the testing dataset using `K-means` clustering algorithm such that the number of clusters equals to the number of labels of the dataset. For each instance in the testing dataset, we compare the cluster label assigned to its explanation after clustering with the instance's predicted class label and if they match then this explanation satisfies the stability metric.

Table I shows the experimental results on the `mortality` and `diabetes` datasets, respectively. The numbers in this table represent the percentage of instances that satisfy the defined metrics. For each row metric, we highlighted the highest performance in bold font and underlined the lowest performance. For the `identity metric`, the LIME technique has shown the worst performance on the two datasets. The reason behind that is the sampling technique used to generate the dataset that acts as the training dataset for the linear model used to approximate the behaviour of the complex black-box model. More specifically, such training dataset is generated by sampling instances around the instance to be explained uniformly at random. On the other side, the SHAP technique satisfies the identity metric for all tested instances. For the mortality dataset, the LIME technique achieves the highest performance for the stability metric (83%) followed by the Anchors technique (80%), while SHAP comes in the last place (75%). For the diabetes dataset, Anchors achieves the highest performance for the stability metric (74%) followed by SHAP (60%), while LIME comes in the last place (52%). For the separability metric on the diabetes dataset, all the interpretability frameworks achieves performance of 100%. For the separability metric on the mortality dataset, all the frameworks achieve comparable performance between 98% and 100%. For the similarity metric on both datasets, SHAP achieves the highest performance for the similarity metric (3.23 on mortality dataset and 6.06 on diabetes dataset), followed by Anchors, while LIME comes in the

[2]https://archive.ics.uci.edu/ml/

	Mortality			Diabetes		
	LIME	**Anchors**	**SHAP**	**LIME**	**Anchors**	**SHAP**
Identity	0%	23%	**100%**	0%	11%	**100%**
Stability	**83%**	80%	75%	52%	**74%**	60%
Separability	**100%**	99%	98%	**100%**	**100%**	**100%**
Similarity	7.3	6.33	**3.23**	13.13	9.96	**6.06**
Time (Sec.)	**0.23**	9.1	**0.23**	**0.21**	10.38	0.22

Table I
EVALUATION OF INTERPRETABILITY FRAMEWORKS ON TABULAR DATASETS

last place (7.3 on mortality dataset and 13.13 on diabetes dataset). For the diabetes dataset, LIME and SHAP achieve comparable average processing time of 0.21 seconds and 0.22 seconds, respectively while the average processing time of the SHAP framework is 10 seconds on average.

For `Bias detection`, we used visual analytics methods to detect the bias on the mortality dataset inspired by [11]. In particular, we compare between two different user interfaces (tabular interface and aggregate interface) for detecting the bias in each of the studied frameworks. We created a biased dataset from the mortality dataset such that the bias is detectable in both interfaces. The biased dataset is created such that the smoking feature is inversely related to the risk of mortality; if the patient is a smoker then the patient is at low risk of mortality which is counterintuitive. The bias is created such that the biased model achieves higher testing accuracy than the unbiased model. The bias is created with the same degree in both training and testing datasets. We trained a random forest model on both datasets. The testing accuracies on the unbiased and biased datasets are 86% and 94%, respectively. The users evaluated the bias were people with basic knowledge in medical domain.

In each of the biased and unbiased testing datasets, we get the explanation of each instance from the different interpretability frameworks based on all 14 features. To make the interface simple, we consider features that contribute to the prediction without specifying whether the contribution is towards or against the prediction. For comparing the interfaces for bias evaluation, we compare patients who are at high risk mortality and patients of low risk of mortality. In particular, we have implemented and use the following two user interfaces:

Tabular user interface: the tabular user interface for the mortality model is shown as a tabular view such that each row represents the explanation features used (Figure 1). The user can navigate between the biased model and the unbiased one across the three interpretability frameworks. The accuracy of each of the biased and unbiased model is shown in the top of the interface. For LIME and SHAP, the columns in the table are sorted according to the average contribution of features on the testing dataset in each of explanation framework such that the leftmost column has the highest average contribution in the explanations. For Anchors, the columns are sorted according to the number of occurrences of features in the explanations such that the leftmost column has the highest occurrence across the testing dataset. To make comparison between patients who are at high risk and who are at low risk of mortality easy, we use different colour for each group.

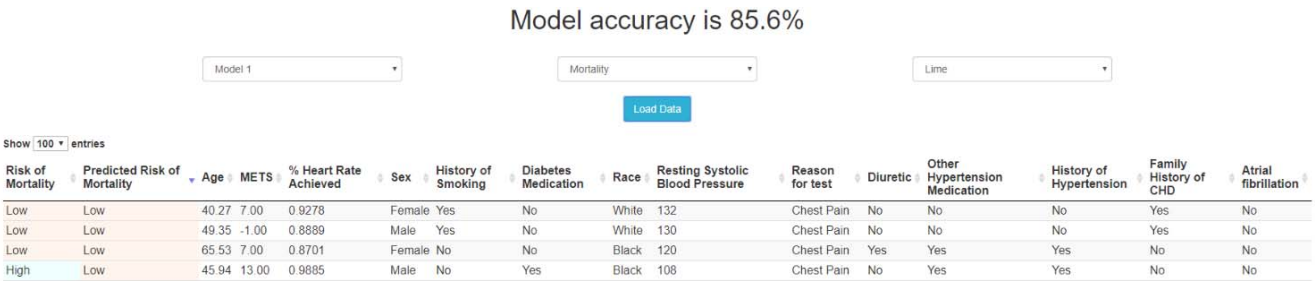
Aggregated user interface: the aggregate user interface shows the distribution of features' values as histograms sorted such that the top-left histogram in each interpretability framework is for the feature with the highest average contribution and the bottom-right histogram is for the feature with the lowest average weight. The aggregated user interface for LIME framework, shown in Figure 2, illustrates only six features, due to space limitations. For each histogram, the height of the bars represent the percentage of instances in each group. The aggregated user interfaces for all the interpretability frameworks using all features are available in the project repository.

We conducted a user study to evaluate the ability to detect the bias by comparing the explanations of the biased and unbiased testing datasets using the tabular and aggregated user interfaces. This study involved 23 fresh graduate students. We divided the participants into 2 groups, one group to evaluate the tabular user interface and the other to evaluate the aggregated user interface. For both groups, we introduced the meaning of accuracy and how it is used to evaluate the model's performance. We then explained the mortality dataset by informing the participants with the meaning of the features and how they logically affect the output class i.e. as age increases, there is a higher chance of mortality. Finally, we explained to the participants how to use the evaluation interface. Out of the 23 participants, only 20 responses were valid. We evaluated validity by asking the participants which model has better accuracy which was a pretty obvious question. All the participants identified the bias from the tabular interface while only 80% identified the bias from the aggregated user interface. It is clear from the results that the tabular user interface enables participants to better detect the bias. Based on these results, we ranked the interpretability frameworks that enabled the participants to correctly detect the bias using the tabular and aggregated user interfaces as SHAP, followed by Anchors and then LIME.

B. Text Data

In our experiments, we preprocessed the text datasets as follows. First, we removed stop words that do not have a sentiment value such as "the" using stopwords corpus from *nltk* package[3]. Second, we removed the `HTML` tags for the text to prevent them from taking place in our dictionary. Finally, we used snowball stemmer in *nltk* to remove morphological affixes from words. This is done to

[3]https://www.nltk.org/



Risk of Mortality	Predicted Risk of Mortality	Age	METS	% Heart Rate Achieved	Sex	History of Smoking	Diabetes Medication	Race	Resting Systolic Blood Pressure	Reason for test	Diuretic	Other Hypertension Medication	History of Hypertension	Family History of CHD	Atrial fibrillation
Low	Low	40.27	7.00	0.9278	Female	Yes	No	White	132	Chest Pain	No	No	No	Yes	No
Low	Low	49.35	-1.00	0.8889	Male	Yes	No	White	130	Chest Pain	No	No	No	Yes	No
Low	Low	65.53	7.00	0.8701	Female	No	No	Black	120	Chest Pain	Yes	Yes	Yes	No	No
High	Low	45.94	13.00	0.9885	Male	No	Yes	Black	108	Chest Pain	No	Yes	Yes	No	No

Figure 1. Tabular user interface for the mortality model using unbiased data on LIME interpretability framework

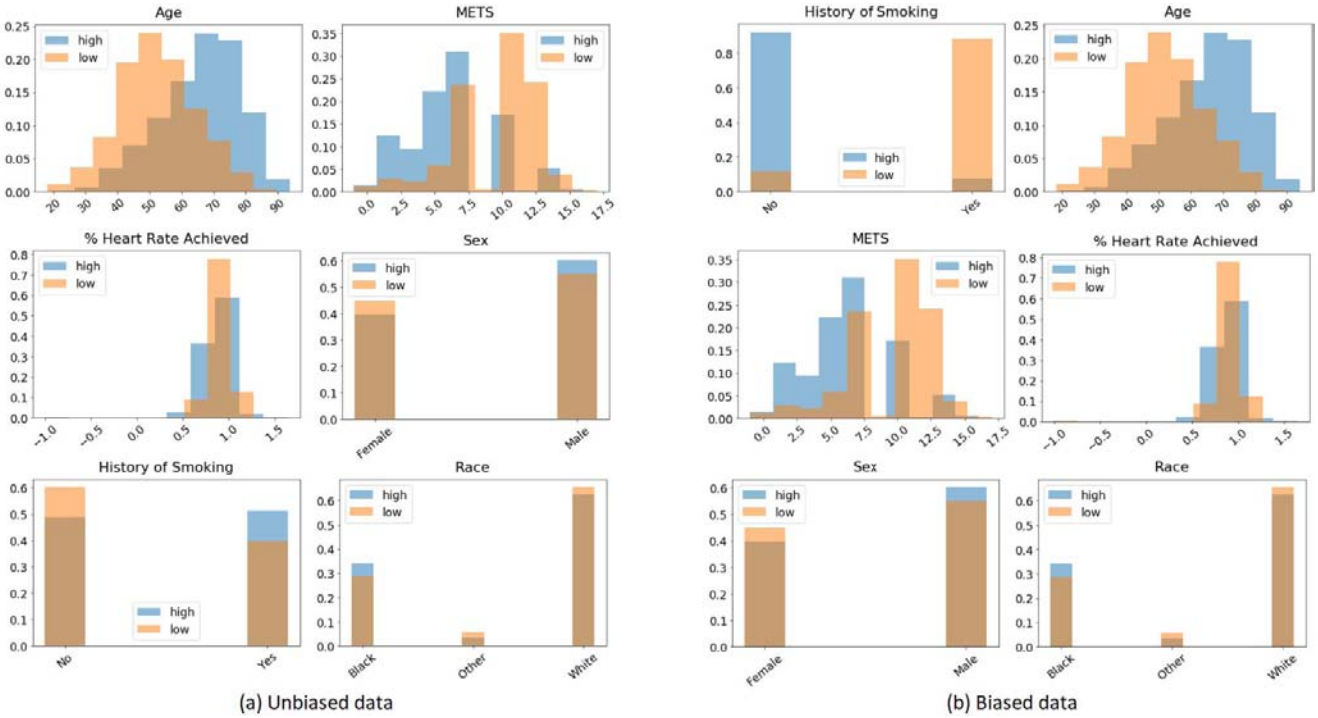


Figure 2. Aggregate explanation user interface for the mortality model on LIME interpretability framework

prevent frequent words from taking more than one place in our dictionary.

For each of the `Drug Review` dataset and the `Side Effects` dataset, we follow the same pipeline used on tabular data, except that the dataset has been partitioned into 70% for training the model and 30% for testing the model. The `identity`, `separability`, `similarity`, and `stability` metrics are measured on text data in the same ways as of tabular data.

Table II shows the results of evaluating the three different interpretability techniques on the `Drug Review` dataset and the `Side Effects` dataset respectively. For each row metric, we highlighted the highest performance in bold font and underlined the lowest performance. The results show that Anchors achieves the highest identity performance of 66% and 78% on the drug review and side effects datasets respectively, followed by SHAP (12% on drug review dataset and 50% on the side effects dataset). LIME achieves the lowest identity performance of 0% and 0.5% on the drug review and side effects datasets respectively. For the stability metric, SHAP achieves the highest performance on the drug review dataset (85%), while Anchors achieves the highest performance on the side effects dataset (70%). LIME and SHAP achieve the highest comparable performance for the separability metric on both datasets. On both datasets, LIME achieves the highest performance for the similarity metric (2.1 on drug review dataset and 8.61 on side effects dataset), followed by Anchors, while SHAP comes in the last place (6.47 on drug review dataset and 21.3 on side effects dataset). Anchors has the highest average time to output explanation while SHAP has the lowest on both datasets.

	Drug Review			Side Effects		
	LIME	**Anchors**	**SHAP**	**LIME**	**Anchors**	**SHAP**
Identity	0%	**66%**	12%	0.5%	**78%**	50%
Stability	74%	82%	**85%**	51%	**70%**	53%
Separability	**100%**	83%	99%	**100%**	96%	99%
Similarity	**2.1**	2.28	6.47	**8.61**	17.4	21.3
Time (Sec.)	1.13	10.3	**0.8**	0.93	3.4	**0.46**

Table II
EVALUATION OF INTERPRETABILITY FRAMEWORKS ON TEXT DATASETS

V. CONCLUSION

In this work, we evaluated three different frameworks namely, `LIME`, `SHAP` and `Anchors` on different types of data (tabular and text). The results show that there is no clear winner. In other words, there is no single interpretability technique that can achieve the best performance for all metrics across different types of data. Thus, it is crucial for the users of the interpretability techniques to clearly specify their interpretability focus (metric) and understand the strength and weakness of each interpretability techniques so that they can achieve their goal for getting reasonable and effective explanations for their used complex machine learning model. As a future work, we plan to extend our work to evaluate the performance of different interpretability frameworks on medical image datasets.

ACKNOWLEDGMENT

The work of Radwa Elshawi is funded by the European Regional Development Funds via the Mobilitas Plus programme (MOBJD341). The work of Sherif Sakr is funded by the European Regional Development Funds via the Mobilitas Plus programme (grant MOBTT75).

REFERENCES

[1] M. H. Al-Mallah et al. Rationale and design of the Henry Ford Exercise Testing Project (the FIT project). *Clinical cardiology*, 37(8), 2014.

[2] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PloS one*, 12(7), 2017.

[3] S. Basu Roy et al. Dynamic hierarchical classification for patient risk-of-readmission. In *KDD*, 2015.

[4] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, 2016.

[5] A. M. Darcy, A. K. Louie, and L. W. Roberts. Machine learning and the profession of medicine. *Jama*, 315(6), 2016.

[6] Doshi-Velez et al. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[7] J. Futoma, J. Morris, and J. Lucas. A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, 56:229–238, 2015.

[8] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93, 2018.

[10] M. Honegger. Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions. *arXiv preprint arXiv:1808.05054*, 2018.

[11] E. B. Josua Krause, Adam Perer. A user study on the effect of aggregating explanations for interpreting machine learning models. *KDD Workshops*, 2018.

[12] E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013.

[13] I. Kononenko et al. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010.

[14] B. Y. Lim, A. K. Dey, and D. Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *SIGCHI*, 2009.

[15] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[16] S. Mohseni, N. Zarei, and E. D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*, 2018.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 2016.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

[19] S. Sakr, R. Elshawi, A. M. Ahmed, W. T. Qureshi, C. A. Brawner, S. J. Keteyian, M. J. Blaha, and M. H. Al-Mallah. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the henry ford exercise testing (fit) project. *BMC medical informatics and decision making*, 17(1):174, 2017.

[20] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[21] E. Štrumbelj and I. Kononenko. A general method for visualizing and explaining black-box regression models. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 21–30. Springer, 2011.

[22] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.