Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models

Nesaretnam Barr Kumarakulasinghe[†], Tobias Blomberg[†], Jintai Liu[†*], Alexandra Saraiva Leao[†]
Institution for Learning, Informatics, Management and Ethics
Karolinska Institutet
Stockholm, Sweden
tobias.blomberg@stud.ki.se

Panagiotis Papapetrou
Department of Computer and Systems Sciences
Stockholm University
Stockholm, Sweden

***Abstract*—The usage of black-box classification models within the healthcare field is highly dependent on being interpretable by the receiver. Local Interpretable Model-Agnostic Explanation (LIME) provides a patient-specific explanation for a given classification, thus enhancing the possibility for any complex classifier to serve as a safety aid within a clinical setting. However, research on if the explanation provided by LIME is relevant to clinicians is limited and there is no current framework for how an evaluation of LIME is to be performed.**

To evaluate the clinical relevance of the explanations provided by LIME, this study has investigated how physician's independent explanations for classified observations compare with explanations provided by LIME. Additionally, the clinical relevance and the experienced reliance on the explanations provided by LIME have been evaluated by field experts.

The results indicate that the explanation provided by LIME is clinically relevant and has a very high concordance with the explanations provided by physicians. Furthermore, trust and reliance on LIME are fairly high amongst clinicians. The study proposes a framework for further research within the area.

***Keywords*—machine learning, classification model, Local Interpretable Model-Agnostic Explanation, clinical decision support system, patient-specific explanation.**

I. INTRODUCTION

The practice of recording patient data in electronic health record systems (EHR) has resulted in large datasets containing an abundance of clinical data [1,2]. This has allowed for the development of black-box, predictive, and diagnostic models with the use of machine learning methods [2]. Such models have the potential to serve as an intelligent clinical decision support tool, thus contributing to safer and more efficient health care [2,3]. Even though machine learning models have been developed in accordance with data science theory and standards of performance, the adoption of these models by healthcare professionals in clinical practice is low [3].The reasoning behind this could be in regard to the "black-box" nature of machine learning algorithms. Classification models with high performance, generally have low transparency, thus the receivers of the outcome have no insight in the logic behind the specific classification [4]. This causes impediments within the field of healthcare for several reasons. Seeing that a prediction may be the cause for action, incorrect predictions may have severe consequences [5]. Explaining the classification outcome allows the healthcare professional to identify incorrect reasoning, thus prevent negative consequences for the patient [5,6]. Furthermore, recognizing the reasoning behind a prediction allows for more specific actions in order to change the classification outcome if undesirable [7]. Finally, the European Union's General Data Protection Regulation (GDPR) specifies that the usage of machine learning methods for prediction models is approved only when the outcome is interpretable making research in the interpretability of "black-box" machine learning models of significant interest [8].

There are several methods one could use in order to make the outcome of a black-box classification model more interpretable. One approach is to make a global explanation by listing what features are generally more important while making the prediction [6]. In healthcare, it is, however, desirable to obtain an instance-specific explanation. This allows for more individualized decision making, thus providing the patient with more personalized care [5].

Local Interpretable Model-Agnostic Explanation (LIME) is a recently developed framework that can be used with any black-box classification model in order to obtain an explanation for one specific instance [5]. It works by giving a local explanation of the classification and provides the minimum number of features that contribute to the maximum likeliness of the specific class outcome for one observation [5]. Even though LIME has previously been applied to clinical classification models, there is a lack of research on the acceptability and interpretability of LIME explanations among clinical healthcare professionals. This study aims to evaluate the quality of LIME explanations and if these explanations increase health care professionals' trust and reliance on black-box classification models.

A. Contributions

In this paper, we aim to explore the degree to which healthcare professionals agree with the instance-specific prediction explanations provided by LIME for black-box classification models. More specifically, we study to what degree do the prediction explanations provided by LIME overlap with independent explanations for the same instance provided by a clinician. Moreover, we investigate whether healthcare professionals are satisfied with the explanations provided by

[†]Authors contributed equally
[*]Sponsored by KI-CSC scholarship

2372-9198/20/$31.00 ©2020 IEEE
DOI 10.1109/CBMS49503.2020.00009

LIME and assess whether the representation of prediction explanations provided by LIME is adequate for clinicians. Finally, we study the level of trust by clinicians towards LIME's instance-specific explanations

II. METHODS

To accomplish the objectives at hand, the study was conducted in three phases: first, a sepsis classification model was trained; next, the LIME algorithm was applied on selected observations; finally, the explanations provided by LIME were evaluated by physicians.

A. Sepsis Classification Model

Data used for this study originated from ICU patients in the United States of America and were extracted from the 2019 PhysioNet Challenge [9]. The dataset consists of data from 40,336 patients and each patient has one or more observations per hour after being admitted into the ICU, and each observation contains 40 different variables, including demographic data, vital signs, and laboratory tests. The label of each observation is defined as the onset of sepsis or not according to the Sepsis-3 criteria [10]. A total of 27,916 septic patient observations which accounted for 1.8% of the total of 1,552,210 patient observations. The remaining 1,524,294 observations were non-septic patient observations. Since vital signs are commonly used for evaluating sepsis in the clinic setting, all observations with missing values of vital signs were removed. Platelets, creatine, bilirubin, and the fraction of inspired oxygen (FiO2) were included based on the sequential organ failure assessment (SOFA) score, which is well known within the critical care community [12]. Missing values were imputed with Multivariate Imputation by Chained Equation (MICE) algorithm. After that, an under-sampling technique was employed to construct a balanced dataset with an equal class distribution of septic and non-septic patients. This resulted in 13,713 patient observations with 14 variables (age, gender, heart rate, respiration rate, pulse oximetry, temperature, systolic blood pressure, diastolic blood pressure, mean arterial pressure, platelets, creatine, bilirubin, FiO2, and sepsis class label) were finalized for model construction and evaluation. The performance of various black-box classifiers such as Adaboost, Random Forest, and SVM were evaluated with 10-fold cross-validation and their average accuracy, precision, recall, and F1-score were calculated. Various kernel types and number of trees were evaluated for the SVM and Random Forest classifiers respectively in order to optimize the predictive performance. The best performing model with regards to these prespecified metrics was then selected for the implementation of LIME.

B. LIME Implementation

1) Instances for LIME implementation. As to select observations to implement LIME on, all observations from the test set with a prediction probability of more than 80% for both classes i.e., "septic" or "no-septic" were shortlisted as potential cases to be evaluated by physicians. Five septic and five non-septic observations were randomly selected from these shortlisted cases and manually reviewed to ensure their medical relevance. In total, ten observations were selected as the LIME evaluation cases, and one of the examples of a "non-septic" patient is shown in Fig. 1. The number of 10 patient cases to be evaluated by each physician was selected as it represented a reasonable demand on already busy clinicians' participants.

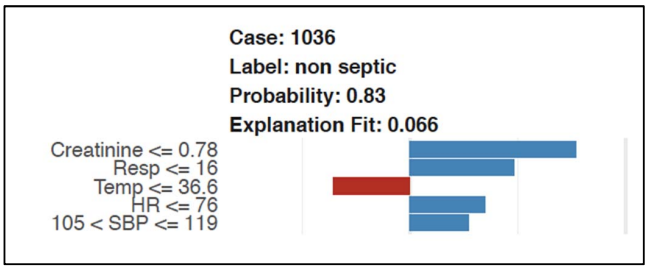

Figure 1. Example of graphical LIME output. The raw output of LIME explanation of a patient predicted as non-septic.

2) LIME parameters. After the sepsis classification model deployment, the LIME algorithm was implemented on selected instances using the LIME package version 0.5.1 in RStudio. For continuous values, the number of bins was set to four. The algorithm used for selecting important features (feature_select) was set to "highest_weights" which fits a ridge regression to select the features with the highest absolute weight. The number of features used for each explanation was set to five to obtain at least three contributory factors for each patient instance. "Three contributory factors" were chosen as it was thought to represent a reasonable balance between the number of available variables(n=14), demands placed on the participating physicians and clinical significance of the obtained result i.e., demonstrating a high concordance between the "top 3" LIME and physician explanations was felt to be more meaningful than a high concordance between the "top 5 (or higher)" explanations. Conversely, a concordance of the "top 2" LIME and physician explanations was felt to be too restrictive. For all other arguments, the default setting was used.

C. LIME Evaluation

1) Evaluators. Registered physicians with prior experience in ICU patient care were contacted via email for participation in this study. These physicians were professional contacts of a team member from two tertiary hospitals in Singapore. A total of 20 physicians were contacted for participation in this study. Physicians with incorrectly filled or incomplete surveys were excluded. Ten physicians completed the survey and fulfilled the inclusion and exclusion criteria at the point of data analysis.

2) Evaluating tools. A three-part survey was created in order to gather data to achieve the objectives of the study. The study questionnaires were developed by the study team based on the study objectives and adapted from existing literature[13], where relevant.

Part 1. The physicians' acceptance of the model prediction. In this part, each physician was first requested to review the demographic information, clinical vital signs, and laboratory measurements of each of the ten patient instances. Next, they were requested to decide if he/she agreed or disagreed with the

model's prediction of "sepsis" or "no sepsis" for each of the ten patient instances without having access to the prediction explanations from LIME. It was deemed that LIME explanations for model predictions that were rejected by physicians would negatively bias the subsequent physicians' evaluation of LIME. Hence, the instances where the model's predictions were rejected by physicians were removed from further evaluation in Part 2. This aims to provide a physician evaluation of LIME that was objective and as independent as possible to the performance of the prediction model.

Part 2. The agreement between the explanations provided by LIME and the explanations provided by physicians of the model's predictions.

In this part, the physicians were initially blinded to LIME explanations and requested to indicate and rank the importance of the top three contributory factors that they felt the model had utilized to make the predictions for specific patient instances. Next, the physicians were unblinded to LIME explanations for the same patient instances and were requested to indicate their satisfaction with the LIME explanations on a Likert Scale (from 1 = completely dissatisfied to 5 = completely satisfied).

Part 3. Physicians' trust and reliance on LIME

In this part, physicians were requested to complete a trust and reliance survey that was adapted from a questionnaire described by Jian et al [11] to evaluate user trust and reliance on Artificial Intelligence. This part consisted of nine questions to evaluate physicians' trust and reliance on a 5-point Likert scale.

Additionally, physicians were also asked to compare the graphical output and a text-based output of LIME explanations and to indicate their preference based on a 5- point Likert Scale.

3) Data analysis. A table with a more detailed description of the data obtained from the three-part surveys and intended performance measure for data analysis is presented in Table

Table 1. Description of analyzed data. The column marked ID will be relevant for the result.

Part	Analyzed data		ID	Performance Metric
1	Physician evaluation of the model's prediction		1.1	Percentage of the model's predictions accepted
2	Agreement between explanations provided by LIME and explanations provided by respondents	General match: The top three variables selected by LIME overlap with the top three variables selected by the respondent.	2.1	Percentage of overlapping
		Exact ranking match: variables selected by LIME fall within physicians' top three and have the match the physicians' importance rank.	2.2	
2	Respondent's satisfaction with explanations provided by LIME	Overall satisfaction score regarding explanations provided by LIME	3.1	Average satisfaction score with a ranking of 1-5, where 1 = completely dissatisfied and 5 = completely satisfied
		Satisfaction score regarding explanations provided by LIME when chosen variables do not overlap with the physician's top three variables	3.2	
		Satisfaction score regarding explanations provided by LIME when chosen variables do overlap with the physician's top three variables	3.3	
3	Physicians trust and reliance on LIME.		4.1	
3	Physician evaluation of LIME output	Satisfaction with the visual representation of LIME explanation	5.1	Average satisfaction score with a ranking of 1-5, where 1 = completely dissatisfied and 5 = completely satisfied Qualitative inductive content analysis
		Satisfaction with the textual representation of LIME explanation	5.2	

III. RESULT

A. Model Performance

The performance of all the classification models that were evaluated after optimization are detailed in Table 2. The best performing model was the Random Forest classification model utilizing ntree=300. It achieved a performance of, (1) accuracy 0.66, (2) precision 0.67, (3) recall 0.64, and (4) F1-score 0.65 and was selected for application in the remaining aspects of this study.

1) Physicians' acceptance of model performance. Ten physicians completed the survey and fulfilled the inclusion and exclusion criteria. Each respondent had ten cases to evaluate. Out of the 100 given cases, respondents agreed with the model prediction in 87 cases, which translates to 87%. These 87 patient cases, with predictions approved by the respondents, were used to evaluate the satisfaction with the LIME explanations.

Table 2. Total result of evaluated classification models. ntree – number of trees, RF – Random forest, SVM – Support vector machine

Model	Accuracy	Precision	Recall	F1-score
AdaBoost	0.655071	0.667456	0.614435	0.639584
RF (ntree = 20)	0.643331	0.649180	0.619645	0.63383
RF (ntree = 50)	0.655436	0.661675	0.631791	0.646193
RF (ntree = 100)	0.657406	0.663753	0.634193	0.648326
RF (ntree = 300)	0.664553	0.672163	0.638966	0.654906
RF (ntree = 500)	0.662510	0.670486	0.635586	0.652340
SVM (linear)	0.612121	0.632632	0.529121	0.576115

SVM (radial)	0.655948	0.681666	0.581665	0.627492
SVM (polynomial)	0.638154	0.712038	0.460663	0.559153

B. LIME Evaluation

1) Agreement between LIME and Physicians. The ratio of the overlap between the top three variables selected by LIME and the top three variables selected by the physician can be found in Table 3 i.e., the top 3 LIME explanations that also fall within the top 3 physicians' explanations.

Furthermore, in 25.29 % of the rankings, not only was there an overlap but there was an exact match between the importance ranking of explanations provided by LIME and the explanation provided by the physicians. Examples of this are cases where both the physician and LIME ranked a particular variable with the same importance rank i.e., most important (#1 rank) or (#2 rank), etc.

Table 3. General agreement between LIME and physicians.

The overlap between the variables selected by LIME and physicians	Number of matches out of 87 cases	Ratio
3 of 3	13	14.94 %
2 of 3	47	54.02 %
1 of 3	27	31.03 %

2) Physicians' satisfaction, trust, and reliance on LIME. Table 4 presents an overview of the physicians' satisfaction with the explanations provided by LIME, trust, and reliance on LIME and the evaluation of LIME output. It is noteworthy to mention that, out of the LIME explanations that did not overlap with the physician explanations, 27.59% received a high physician satisfaction score of at least 4 out of 5(80%).

In addition to the results presented in Table 4, a qualitative analysis regarding the visual and textual output of LIME was done. The results of this indicated that the physicians preferred the visual output from LIME over the textual. Several of the respondents mentioned that the output was "easily interpreted" and "...informative. Easy to get used to…". One informant brought up that the chart could be organized in a different manner as to be easier understood and mentioned that "The 'probability' is probably more helpful for the layman".

Regarding the usefulness of a textual representation of the LIME explanation, the opinions were scattered. Some respondents thought that the text would be abundant, due to the hectic work environment of physicians. Others responded that "the narrative is easier to digest". Additionally, several respondents brought forth that they would have appreciated information on what the cut-offs were for the different groupings and the reasoning behind this.

Table 4. Physicians' satisfaction, trust, and reliance on LIME. The column ID corresponds to the analyzed data variables in Table 1

ID	Physician	A	B	C	D	E	F	G	H	I	J	Mean
3.1.	*Overall satisfaction score regarding explanations provided by LIME*	4.1	4.2	3.2	2.9	3.8	3.6	3.7	4.7	4.2	4.6	3.9
3.2.	*Satisfaction score regarding explanations provided by LIME when chosen variables do not overlap with the physician's top three variables*	3.1	3.2	2.4	2.6	3.4	2.7	2.0	3.5	2.6	4.1	3.0
3.3.	*Satisfaction score regarding explanations provided by LIME when chosen variables overlap with the physician's top three variables*	4.8	4.9	3.6	3.0	4.4	4.4	4.5	4.2	4.7	4.8	4.3
4.1.	*Trust and reliance score*	3.1	4.1	3.3	3.3	1.7	3.3	3.9	3.1	3.4	4.2	3.4
5.1.	*Satisfaction with the visual representation of LIME explanations*	3.0	5.0	3.0	4.0	5.0	-	-	2	4	4	3.8
5.2.	*Satisfaction with the textual representation of LIME explanations*	2.0	2.0	4.0	2.0	1.0	-	-	4	2	4	2.6

IV. Discussion

The usage of black-box classification models within the healthcare field is highly dependent on being interpretable by the receiver [6]. LIME provides a patient-specific explanation for a given prediction by a black-box classification model, by ranking the importance of each variable that contributes to the prediction. [5].

However, research on if the explanations provided by LIME are relevant to clinicians is very limited and there is no current framework for how an evaluation of LIME should be performed. To our knowledge, there has been only one previous study [13] by Tajgardoon et al. that attempted to evaluate the explanations provided by LIME with physicians.

In our study, we performed an "Application-level evaluation" as described by Doshi-Velez et al. [12] It refers to evaluations where the end-users are engaged in an evaluation of the interpretability of a machine learning model Specifically, in this study, we performed a blinded comparison between LIME and physician explanations of a black-box classification model for the first time. Additionally, we built upon the research of Tajgardoon et al. [13] by studying physician's satisfaction with LIME explanations with finer granularity and formally studied physicians' perceptions of LIME output representation. Finally, we explored the level of trust and reliance of physicians with regards to the application of LIME to explain the predictions of black-box machine learning models which has not been previously performed. Below the study findings are discussed in the context of the research aims stated previously.

1) To what degree do the prediction explanations provided by LIME overlap with independent explanations for the same instance provided by a clinician? According to the results, LIME and blinded physicians achieved a general match of at least one variable in 100% of the patient cases, at least two variables in 68.97% of the patient cases and in all three variables in 14.94% of the patient cases. Remarkably, 25.29% of the variables selected by LIME not only overlapped with the top three variables selected by the physicians but also had the same importance rank. This suggests that the explanations provided by LIME are comparable to the explanations provided by the physician when analyzing the predictions provided by a black-box machine learning model, thus making them medically relevant.

2) To what degree are clinicians satisfied with the prediction explanations provided by LIME? When asked to evaluate the explanations provided by LIME against their own explanations, physicians were generally satisfied, with a 78% satisfaction score. This suggests that physicians generally agree with the LIME explanation which, in turn, reinforces the medical relevance of the LIME explanations. The relationship was stronger, with a satisfaction score of 86% when the LIME explanations overlapped with the physicians' top three explanations. In the cases where the explanation provided by LIME did not overlap with the physician's top three explanations, a lower, 60% overall physician satisfaction score was reported. However, as described above, 29.23% of the top three variables selected by LIME but that did not overlap with the physicians' top 3 variables still received a high physician satisfaction score of at least 80%. Hence it is reasonable to assume that these LIME explanations were still of significant clinical value even though they did not match the physician's initial explanations for the patient case.

3) To what degree are clinicians satisfied with the representation of prediction explanations provided by LIME? When asked to weigh in on their preference for how the LIME explanations should be communicated to the physicians, it was clear that the surveyed physicians preferred a diagrammatic approach. They reported that the raw chart output (Fig. 1) was satisfactory with a mean agreement score of 3.8 (out of 5), much higher than the score that a narrative text output received. This is in line with previous knowledge on that information in the form of images is more easily understood than that in the form of text in the healthcare setting [14].

4) To what degree do healthcare professionals trust the instance-specific explanation provided by LIME? The results provide a realistic reflection of the level of trust physicians have for LIME. A trust score of 68% suggests a fairly positive attitude towards LIME which is encouraging. One physician had very low trust and reliance scores despite being fairly satisfied with the explanations provided by LIME. While this relationship should be studied formally with a larger sample of physician reviewers it does suggest that the satisfaction with the LIME explanations does not automatically translate to increased trust and reliance of LIME and that there are other factors at play. We hypothesize as with the social-technical aspect of information technology systems, some physicians may be cautious in making clinical decisions based on the outcome of a machine learning model [15]. It is, however, important to note, that such hesitance may more likely be related to the use of the machine-learning classification model itself, rather than the explanations provided by LIME.

B. Proposed Evaluation Framework

With this study, we propose a framework for the objective evaluation of LIME on black-box machine learning models within the context of healthcare.

One key issue to be addressed when evaluating LIME or other model agnostic interpretability methods is the interdependence of the performance of LIME with the performance of the machine learning model. This is important since the end-user may view both the machine learning model and LIME as one system. In other words, LIME will be evaluated unfairly in the presence of a poorly performing model. To overcome this issue, the most apparent solution is to optimize the performance of the model in order to reduce its negative influence on the evaluation of LIME.

Furthermore, we propose and used two additional methods to overcome this bias. Firstly, we selected cases to be explained by LIME, for both "positive"(septic) and "negative"(non-septic) classes based on a probability of outcome predicted by the model. A cut off of 0.8 was selected for this study based on the study by Tajgardoon et al [13]. Next, we proposed that the physician first evaluate the model's prediction for the selected cases to evaluate its clinical relevance prior to evaluating the explanations provided by LIME. Cases that the model's prediction was deemed to be medically unsound were not subjected to further physician assessments of LIME explanations.

In addition, the initial blinding of the physicians to the LIME explanations provided our study several advantages. Firstly, with blinding, we were able to obtain an unbiased comparison between LIME and the physicians' own explanation of the model's prediction. This also helped to avoid reviewers from readily agreeing with the LIME's explanations without careful review. This is in contrast to the fabricated explanations that were used in the study by Tajgardoon et al [13] to achieve the same goal. Furthermore, some of the explanations that were offered by LIME, but not selected by the physician, represented additional variables that were later deemed to be important by the physician after unblinding.

C. Limitations and future directions

1) Physician reviewers and patient cases. In order to increase the validity of this study, a higher number of physician reviewers and patient cases would be beneficial. especially with regards to the trust and reliance questionnaire. Besides this, one also has to consider that the selection of physicians for this study was rather homogenous regarding years and field of practice which is not representative of all healthcare professionals. Furthermore, in this study physicians were asked to evaluate the machine learning models predictions and LIME explanations based only on the 14 patient variables. This is in

contrast to a realistic clinical scenario where physicians have access to the entire patient health record..

2) Dataset. The dataset utilized for this study was initially highly unbalanced with regards to the class labels. It also contained a significant number of missing values. While data processing techniques such as random under-sampling and multiple random imputations were used to overcome this, it is undeniable that these limitations in the dataset may introduce bias.

3) Machine learning model. One can argue that the machine learning model used in this study could have a better performance of precision, recall, and F1-score. However, due to limitations in the dataset and given that the primary aim was to evaluate of LIME, not to create the best possible predictive model, we chose to overcome the limitations in the predictive ability of the model with the two methods described earlier instead of further refinement of the model itself. Future research would certainly include using more optimal datasets as well as further optimization of machine learning models alongside a physician evaluation of LIME explanations.

4) Exclusion of "erroneous" model predictions. In this study, 13 patient instances with model predictions that were rejected by physicians were considered to be "erroneous" and excluded from further physician.This was done to provide an objective physician assessment of LIME performance that was independent of the machine learning model's performance. However, realistically speaking, a machine learning model's predictions will never be accepted by all physicians for all patient cases. Hence including these "erroneous" predictions as part of the LIME evaluation may have provided a more realistic physician evaluation. Furthermore, the LIME explanations of these "erroneous" predictions may provide insight to the physicians about the rationale and reason behind these "erroneous" predictions. Future research, building on the current work will include these "erroneous" machine learning model predictions and provide a "misclassification" analysis.

5) LIME method. LIME has many advantages. It can be used in a wide range of machine learning models, it is easy to deploy, has a fast computation time [16], and provides short, selective explanations which seem to suit busy physicians. However, there are some limitations that form the bulk of the ongoing research into LIME. Firstly, explanations provided by LIME can sometimes be inconsistent or unstable [17], due to variations in sampling or delineation of the extent of local data points to be included in the local model. In addition, the explanations of a simplistic, local linear model used by LIME may not always be consistent with the global logic of the model for complex datasets [16]. Furthermore, there is no clear guidance on how to select the number of features or kernel width for the local model which results in a trade-off between the local model's complexity and its simplicity i.e interpretability [16].

V. CONCLUSION

In conclusion, this study performed an evaluation of LIME against physician respondents to explain the predictions of sepsis or lack thereof that were produced by a "black-box" machine learning model. The results show an encouraging degree of physician consensus and satisfaction with the explanations provided by LIME. Furthermore, a cautiously positive degree of physician trust and reliance on LIME was observed. Further research to improve LIME as well as to study the factors that would promote its trust amongst physicians will be vital to the widespread adoption of "black-box" machine learning predictive tools in healthcare.

REFERENCES

1. Williams DC, Warren RW, Ebeling M, Andrews AL, Teufel Ii RJ. Physician Use of Electronic Health Records: Survey Study Assessing Factors Associated With Provider Reported Satisfaction and Perceived Patient Impact. JMIR Med informatics [Internet]. 2019 Apr 4;7(2):e10949–e10949. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30946023
2. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA - J Am Med Assoc. 2013;309(13):1351–2.
3. Jabbar MA, Samreen S, Aluvalu R. The future of health care: Machine learning. Int J Eng Technol. 2018;7(4):23–5.
4. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare. 2015;1721–30.
5. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 1135–44. (KDD '16). Available from: http://doi.acm.org/10.1145/2939672.2939778
6. Ahmad MA, Teredesai A, Eckert C. Interpretable machine learning in healthcare. Proc - 2018 IEEE Int Conf Healthc Informatics, ICHI 2018. 2018;447.
7. Tolomei G, Silvestri F, Haines A, Lalmas M. Interpretable predictions of tree-based ensembles via actionable feature tweaking. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2017;Part F1296:465–74.
8. Tesfay WB, Hofmann P, Nakamura T, Kiyomoto S, Serna J. I Read but Don't Agree. 2018;2:163–6.
9. Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019 [Internet]. [cited 2020 Jan 16]. Available from: https://archive.physionet.org/challenge/2019
10. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Jama. 2016;315(8):801–10.
11. Jian J-Y, Bisantz AM, Drury CG. Towards an Empirically Determined Scale of Trust in Computerized Systems: Distinguishing Concepts and Types of Trust. Proc Hum Factors Ergon Soc Annu Meet [Internet]. 1998 Oct 5;42(5):501–5. Available from: http://journals.sagepub.com/doi/10.1177/154193129804200512
12. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017;(Ml):1–13. Available from: http://arxiv.org/abs/1702.08608
13. Tajgardoon M, Samayamuthu MJ, Calzoni L, Visweswaran S. Patient-Specific Explanations for Predictions of Clinical Outcomes. 2019;88–97. Available from: https://doi.org/
14. Balkac M, Ergun E. Role of Infographics in Healthcare. Chin Med J (Engl). 2018;131(20):2514–7.
15. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, Desroches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res. 2019;21(3).
16. Molnar C. Interpretable machine learning. A Guide for Making Black Box Models Explainable. 2019.
17. Alvarez-Melis D, Jaakkola TS. On the Robustness of Interpretability Methods. 2018;(Whi). Available from: http://arxiv.org/abs/1806.08049