Interpretable Machine Learning: A Case Study of Healthcare

Feyza Yıldırım Okay
Department of Computer Engineering
Gazi University
Ankara, Turkey
feyzaokay@gazi.edu.tr

Mustafa Yıldırım
Department of Computer Engineering
Gazi University
Ankara, Turkey
mustafa.yildirim2@gazi.edu.tr

Suat Özdemir
Department of Computer Engineering
Hacettepe University
Ankara, Turkey
ozdemir@cs.hacettepe.edu.tr

2021 International Symposium on Networks, Computers and Communications (ISNCC) | 978-1-6654-0304-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISNCC52172.2021.9615727

***Abstract*—With the evolution of artificial intelligence, Machine Learning (ML) techniques have become more powerful predictors, and accordingly, the use of ML techniques has become a part of our daily life in different application scenarios such as disease diagnosis, movie recommendation, monitoring system, or detection of malicious attacks. Although ML provides high accurate predictions, it suffers from opacity. By behaving like a black box it excluded users about how to reach particular decisions. Interpretable Machine Learning (IML) is a recent technology that offers a promising solution to the opaqueness problem of complex ML techniques. It provides transparency of how the inner workings of ML lead to certain decisions and allows users to be aware of the decision-making process. Especially, in critical scenarios such as healthcare, it may become extremely important to know the reasons that affect the decision as well as the result. In this study, we aim to show the benefits of IML over a healthcare case study. In experiments, we employ SHAP and LIME IML models for the Random Forest (RF) and Gradient Boosting (GB) algorithms for the problem of diagnosing diabetes and its explanations. Overall results exhibit that applying IML models to complex and hard-to-interpret ML techniques ensures detailed interpretability while maintaining accuracy. We also perform experiments for local interpretability by focusing on an instance, which is another advantage of IML.**

***Index Terms*—IML, interpretability, SHAP, LIME, healthcare**

I. Introduction

Machine Learning (ML) has proved its success in learning complex patterns to make consistent and reliable predictions. From past to present, it has offered great solutions and opportunities to different problems in different fields. Especially with the empowerment of ensemble and deep learning models, it has become preferred by huge companies in various real-world applications, such as movie recommendation, machine translation, speech recognition [1]. ML models have superiority over humans in terms of speed, reproducibility, and scaling in the learning process. After these models are implemented, they can complete a task faster, produce consistent results, and can be replicated to another machine easily and cheaply. On the other hand, the training of a human to complete a task takes a long time, requires extra cost if it is younger and impossible to replicate its knowledge [2].

Most powerful but complex ML models such as random forests, deep learning, and gradient boosting methods have better performance at yielding highly accurate results on various real-world classification, regression, and prediction problems as compared to simple ML models such as linear regression, decision tree, and naive bayes. Unfortunately, they are also opaque to their users about how to decide and which particular decisions are made while predicting. This is a challenging issue in some scenarios where it is important not only the outcome but also how the decisions that affect the outcome are made like healthcare. The lack of interpretability of the ML models limits the widespread adoption of ML models in the field of healthcare. The clinicians hardly trust ML models since the model is designed and evaluated on particular decisions and often relies on the limited statistical and ML knowledge [3]. For example, a patient can be diagnosed with cancer. Accordingly, it may be requested to decide on the treatment method. However, the patient or the doctor cannot access information on which syndromes are more effective or ineffective in diagnosing cancer. There is also the possibility that treatment will do more harm to the patient, especially if misdiagnosed. Where the outcome is highly critical, such as human life, the model needs to be more clear and interpretable.

Interpretable Machine Learning (IML) has emerged to compensate for the opacity problems of powerful but complex ML models by providing transparency about their behaviors in decision-making while maintaining high performance. Although the explainability of intelligent systems approaches are very old, IML models have become very popular especially in recent years. Thus, it sheds light on unknown or hidden facts of ML models to make them understandable by humans [4]. Note that, in some comprehensive literature reviews, IML models classify ML models in different sub-models according to before, during, or after training. In this study, we will consider IML models to cover only IML models applied after training.

This study provides interpretability with well-known IML methods in an area such as healthcare, where not only the results but also the reasons behind the results are important, allowing the users to interpret and understand the results. Hence, it adds transparency, consistency, fairness, and trust to ML-based healthcare systems. It also gives different insights of the understanding the results over an instance-based (local interpretability) and population-based (global interpretability). In order to analyze the advantages of IML models in the

TABLE I. IML models in the field of healthcare

Reference	IML model	ML model	Metric	Healthcare Dataset	Key Finding
ElShawi et al. [5]	LIME, Anchors, SHAP, LORE, ILIME, MAPLE	RF	identity, stability, separability, similarity, time, bias detection, trust	mortality, diabetes, drug review, side effects datasets	The highest performance achieved by each IML models as follows: Identity: MAPLE Stability: MAPLE Separability: LIME, SHAP and MAPLE Similarity: LIME, ILIME, and MAPLE Time: MAPLE Bias detection: SHAP and MAPLE Trust: Anchors
Panigutti et al. [6]	Doctor XAI	RNN with GRU	fidelity, hit, exp.complexity	MIMIC-III dataset	Exploiting ontological information increases the explanation performance.
Zafar et al. [7]	DLIME, LIME	RF, NN	stability	Breast cancer, liver patients, hepatitis patients	DLIME produces stable explanations, while LIME generates unstable generations.
Knapic et al. [8]	SHAP, LIME, CIU	CNN	support, time	Gastral images collected from Video capsule endoscopy	CIU is better than SHAP and LIME in interpreting the decision more rapidly. Also, it has superior support in decision making.
Hakkoum et al. [9]	LIME, FI, PDP	MLP, RBFN	trust	Wisconsin Original dataset for breast cancer diagnosis	The accuracy results of RBFN is higher than MLP, the interpretability of MLP is more consistent with DT classifier. When combining global and local interpretability, the level of understanding increases.

healthcare scenario, two popular IML models, SHAP and LIME, are applied to two of the most well-known complex ML algorithms including Random Forest (RF) and Gradient Boosting (GB).

II. Related Work

With the expansion of traditional ML techniques into the fields of healthcare, the interpretation of ML models has gained great emphasis on healthcare applications [10]. Since a possible wrong decision in prediction may harm the patient's health seriously, which makes it critical to interpreting such results, as well as to predict with high accuracy [11]. Therefore, the use of IML models in healthcare has been recently increased in diagnosing various diseases such as cardiology, neurology, and psychology, determining clinical care pathways, and classifying patients according to their risks. Due to the high capability of interpretability and easiness to apply, transparent approaches such as linear regression, decision trees, and naive bayes are utilized in the healthcare domain commonly [12], [13]. In this study, we rather focus on post-hoc IML methods which are applied after training with ML models in the domain of healthcare, since we aim to show the advantages of IML methods when applied to the complex ML models.

Although the IML is a popular concept and widely studied recently, its applications in the healthcare domain are still limited. ElShawi et al. [5] compare the six popular local post-hoc IML models consisting LIME, Anchors, SHAP, LORE, ILIME, and MAPLE, on different healthcare datasets by considering seven quantitative metrics which are identity, stability, separability, similarity, time, bias and trust detection. Extensive analysis of real-world healthcare datasets is performed by comparing different local IML models in terms of various explanation metrics. According to the results, MAPLE has superiority in most of the evaluation metrics. Panigutti et al. [6] proposed Doctor XAI, which is the first post-hoc IML model addressing multi-labeled sequential and ontology linked data (MIMIC-III) is proposed in the healthcare domain. The explanation results are evaluated in terms of fidelity, hit, and explanation complexity. The results show that exploiting ontological information increases the explanation performance. Zafar et al. [7] proposed deterministic LIME (DLIME) intending to obtain more stable explanations than LIME in determining healthcare status. It leverages hierarchical clustering to partition the training data into different clusters. With DLIME more consistent explanations are achieved compared to LIME. Knapic et al. [8] compares different post-hoc IML models such as SHAP, LIME, and Contextual Value and Utility (CIU) and evaluates their explanation time and support over gastral images collected from video capsule endoscopy. Results clearly illustrate that CIU is better than SHAP and LIME in interpreting the decision more rapidly. Also, it has more support in decision-making. Hakkoum et al. [9] measure the trustworthiness of LIME, Feature Importance (FI), and Partial Dependence Plot (PDP) to diagnose breast cancer. Although the accuracy results of Radial Basis Function Network (RBFN) is higher than Multi-Layer Perceptron (MLP), the interpretability of MLP is more consistent with DT classifier. They also suggest combining global and interpretability to increase the understanding level.

III. Interpretable Machine Learning

Although the term IML is new, the problem of explanation of an expert system is very old, which has been tried to be solved since the mid-1970s. Recently, it has gained renewed attention in literature after 2016. The term IML can use interchangeably with Explainable Artificial Intelligence (XAI) by practitioners and academicians in literature. Most studies focus on the explainability of ML models, and these models are generally named Interpretable Machine Learning (IML). Although these two concepts can be used interchangeably, IML is a substantial part of XAI that focuses on explaining the logic in ML algorithms [14], [15].

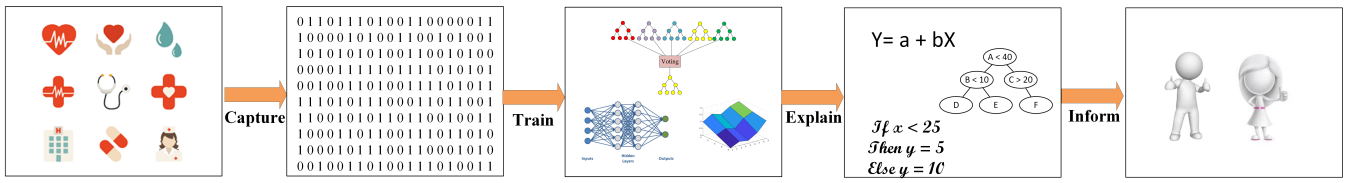


Fig. 1. The general flow of interpretability of healthcare problems

The need for IML models is a debatable issue according to the type or importance of the problem. Bunt et al. [16] discuss the question of 'Are Explanations Always Important?'. The extra cost of performing this analysis should also be considered when assessing the importance of the user's understanding of the models' behavior patterns. For example, for a movie recommending system the cost of interpretability may outweigh the benefits, on the other hand, for critical applications such as disaster management or diagnosis systems, it is also important to find out the reasons behind the decision.

The study of [14] explains the opportunities of using IML by expressing four fundamental reasons: (i) IML aims to justify the outcomes, rather than giving a description of the inner workings of ML models. When an unexpected decision is made, the results are supported with further justifications to the results be more fair and reliable decisions. (ii) From a larger perspective, IML examines and controls the behavioral patterns of ML models. Then, it can quickly identify unforeseen errors and defects and correct them for a seamless decision-making process. (iii) IML enables the improvement of ML models by making them more understandable by humans. With a clearer chain of reasoning by humans, it becomes easier to make smarter decisions. (iv) IML helps to gain new insights into the ML problems, and discover the hidden facts and knowledge.

With the evolution of ML methods, they have become good at predicting many real-world problems with high performance. However, the more the ML model becomes complex, the less interpretability is provided. While methods such as linear regression and decision tree provide predictions with high interpretability but low accuracy, complex algorithms such as random forest and gradient boosting provide high accuracy but low interpretability.

The IML methods in the literature can be classified in different ways according to when the technique is applicable (pre-model, in-model, post-model) and the scope of the instances (local, global) [2], [4], [14].

Depending on whether the applicability of interpretability before, during, or after the building of the model, it can be grouped into three sub-models: (i) Pre-model techniques are applied to the data before model selection, which allows to exploring and understanding of data. It provides interpretability and sparsity by extracting meaningful intuitive features that are used to train the model. Pre-model techniques include feature selection, clustering, and some statistical method such as information gain, k-means, PCA. (ii) In-model, also called intrinsic techniques inherently restrict the complexity of machine learning models during the period of training. (iii) Post-model, also called agnostic, methods provide analysis and interpretation of the model with certain methods after the training process.

Another popular categorization is based on the scope of interpretability, which is global and local interpretability. (i) Local interpretability provides a local understanding of why and how specific predictions can be made [17]. It focuses on providing interpretability by looking at a particular single or multiple instances. It may be preferred in cases where the predictions are linearly dependent on some features rather than complex dependence of all features [18]. Shapley value [19] and LIME (Local Interpretable Model-agnostic Explanations) are some popular methods presented for individual predictions. (ii) Global interpretability aims to describe the model as a whole. More specifically, it requires comprehending decisions, features, structures, and each learned component such as weights and parameters. Especially, it seeks to answer the question of which features are more important and what kind of interactions are realized among them.

IV. Methodologies

In this section, detailed information is given about the ML and IML models used in the experiments. Some ML models like DT, LR have the ability to interpret the model intrinsically. However, as the prediction accuracy of the model increases, its interpretation ability decreases. Therefore, post-hoc IML models have been applied to powerful ML models, allowing for both high accuracy and interpretability. While ML models are used to train the model, IML models are used to interpret the trained model.

In the experiments, we select RF and GB algorithms as our ML models and SHAP and LIME algorithms as our IML models. We select these algorithms because of their popularity in the literature. All of them are well-known algorithms to solve different prediction problems in the literature.

A. ML models

1) RF: RF is a well-known ML model which can be applied to both classification and regression problems. It consists of many individual decision trees. It draws its strength from the fact that ensemble prediction is better than each prediction of individual trees. It employs bagging and feature randomness

to build an uncorrelated forest of trees so as to operate as an ensemble [20]. To overcome the overfitting problem of DT, RF uses feature randomness to split the tree into sub-trees depending on the features. Then, it trains them all separately. After voting the trees among themselves, it chooses the best subtree with the highest prediction performance. Different from DT, it selects a random feature subset that ensures a lower correlation and more diversification [21].

2) GB: GB is a powerful ML model for different regression and classification problems. The basic idea of GB is originated from AdaBoost proposed by Freuman and Schapire [22]. While RF constructs an ensemble of deep individual trees, GB builds an ensemble of shallow trees. GB combines and transforms weak trees into strong trees by controlling loss function. Until the loss function is minimized, GB updates its prediction by adding new models. It generally uses DT as base learners, and constructs new trees until it is overfitting or reaching the optimal results [23]. It uses the loss function to calculate pseudo-residuals.

B. IML Models

1) SHAP: SHapley Additive exPlanations (SHAP) is a popular approach to explain the complex ML models in the literature. It is proposed by Lundberg and Lee [24] as a game-theoretic approach depending on Shapley values. SHAP aims to explain the attributes of an instance to the output. To this end, Shapley values attempt to measure the marginal contribution of a player to the team score in team games. The Shapley values control how the score changes when each player on the team is removed one at a time. Thus, the marginal contribution of each player to the score can be measured. Since the measuring is realized by removing one by one, the computational complexity is increased exponentially with the increase in the number of players. Similarly, SHAP assumes each attribute as a player and explains by measuring the effects of each attribute on the prediction result.

SHAP integrates LIME and Shapley values and can be formalized as follows [24]:

$$g(z') = \Phi_0 + \sum_{j=1}^{M} \Phi_j z'_j \tag{1}$$

where g is an explanatory model, M is the maximum number of attributes, and $z' \in [0,1]^M$ is the simplified new dataset. Lastly, $\Phi_j \in R$ is the contribution of the j attribute, which means that it is the Shapley value. Here, Shapley value, Φ_j can be calculated as Eq. 2 [24]:

$$\Phi_j = \sum S \sqsubseteq F \setminus j \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{j\}) - f(S)] \tag{2}$$

where Φ_j denotes the difference between the prediction with attribute j and the prediction without attribute j. Thus, the marginal contribution of the j attribute is determined. Also, $f(S)$ denotes the subset before including the attribute j. $f(S \cup \{j\})$ is the new set after including attribute j. F shows all the attributes. Lastly, $S \sqsubseteq F \setminus j$ shows possible subsets excluding the attribute j .

2) LIME: Local Interpretable Model-agnostic Explanations (LIME) [25] is used to explain the behavior of machine learning algorithms on a single instance of a data set. It uses an interpretable surrogate model to describe individual predictions. This surrogate model is an interpretable model like LR and DT. It intrinsically operates and considers the model as a black box. LIME aims to understand why ML makes a particular prediction for an instance. LIME checks for changes in the predictions when it perturbs the variations of the data. Accordingly, new instances are generated and respective predictions are performed. Each instance is weighted according to the degree of closeness. LIME is trained with an interpretable surrogate model on a new weighted dataset. Interpretability is ensured by the weights given to the variables by the surrogate model locally. Note that, it does not guarantee that global interpretability.

LIME can be formalized as the following equation [24]:

$$exp(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega_g \tag{3}$$

where x denotes the sample to be explained, g denotes linear regression or surrogate model such as DT, f denotes the actual model that performs predictions such as RF, and π_x denotes the closeness scale of the sample x. Also, L is the function that minimizes the loss function, such as the least mean square error. Ω_g is used to keep the number of attributes of the surrogate model low. However, in practice, the user determines the number of attributes himself.

V. Experimental Results and Discussion

In the aim of providing interpretability, the dataset is first trained with ML algorithms, then IML models ensure further explanations on the trained dataset. The experimental studies are conducted from two perspectives including global interpretability and local interpretability.

A. Dataset

In this study, Sylhet Diabetes dataset, which is available in UCI repository, is used. It consists of 17 attributes with 520 instances. All attributes are categorical and values are complete. By using the dataset, it is aimed to predict the early stage of diabetes risk by ML models. Furthermore, IML models add interpretability and transparency to the ML models.

B. Pre-processing

The dataset is split into training and test sets as 70%-30%. Two popular algorithms, RF and GB, are chosen due to their high prediction power and low explanation power. Also, the prediction results on the test set are evaluated and compared in terms of precision, recall, F1-measure and accuracy as shown in Table II. According to the prediction results, both algorithms are powerful in predicting with high accuracy scores. Also, RF is slightly better than GB. After the training process, each trained model is ready for interpretation by post-hoc IML models.

TABLE II. Pre-processing of ML algorithms

Model	Precision	Recall	F1-score	Accuracy
RF	0.99	0.96	0.98	0.97
GB	0.99	0.95	0.97	0.96

C. The Results with Global Interpretability

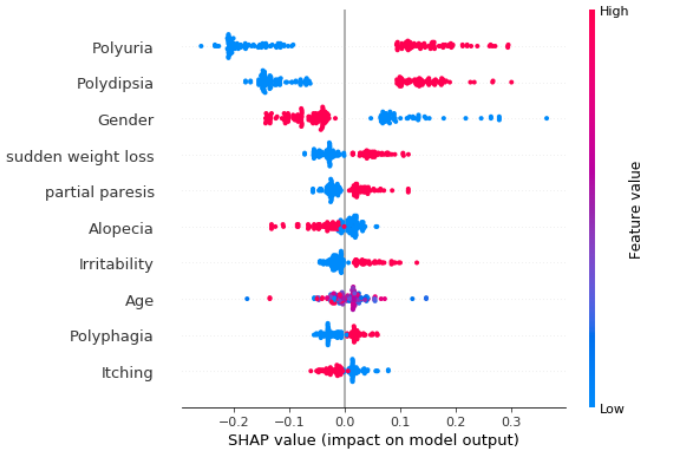
Global interpretability describes the data as a whole. SHAP can support the interpretation of an ML model globally and locally, while LIME only supports local interpretations. For this reason, global interpretability results were obtained only with SHAP in the experiments. SHAP lists the attributes according to their feature importance on the result. Blue and red colors indicate whether the values of features range from low to high. Figure 2 shows the IML results for RF and GB separately. According to the SHAP results on RF, *Polyuria*, *Polydipsia* and *Gender* attributes are more important than the other attributes. Similar results are achieved with GB algorithm. However, after the first three attributes, they differ from each other. As we inferred from the results, the first three attributes are shared by both algorithms, which means these attributes have a key role in reaching the prediction result.

D. The Results with Local Interpretability

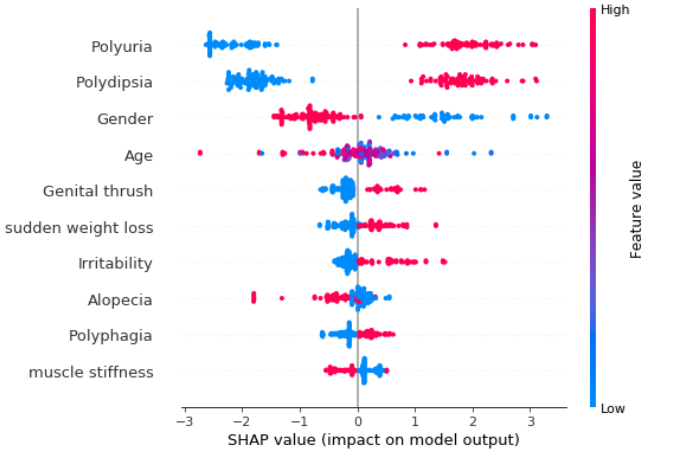
Local interpretability focuses on the explanation of a single instance rather than representing all data. If we give an example from the diabetes dataset, local interpretability provides detailed explanations for a single patient, not all patients. General implications for all patients do not have to be valid for a certain patient. The main factors causing the disease of the patient may be unique and different from the general factors. Local interpretability has the ability to reveal these particular factors.

Since both presented IML algorithms are capable of explaining the ML models locally, we evaluated the interpretability results for both SHAP and LIME. Also, two IML models are applied to the trained data with RF and GB. Figure 3 and Figure 4 give the SHAP and LIME local interpretability results of RF and GB for a patient, respectively. In Figure 3, red-colored attributes push the prediction higher, while green-colored attributes lower. The size of each attribute represents the feature importance. As seen in Figure 3, the results with RF and GB are similar to each other. It can be inferred from these results, the interpretability does not change much according to the models and gives consistent results. In Figure 3(a), the base value is 0.61 The value of the patient is 0.64, which indicates that the estimate is 1, that is, the probability of diabetes is higher. The lines represent the variables that affect the result most. The longer the line, the more the result is affected by these variables. The colors show the direction of the effects. Green-colored attributes support the probability of being healthy and red attributes support the possibility of being diabetes. For example, if the attribute *Polydipsia* is 1 for this patient, this attribute increases the likelihood of diabetes the most.

When we evaluate the difference between global and local interpretability results, they are not consistent for all patients.



(a)



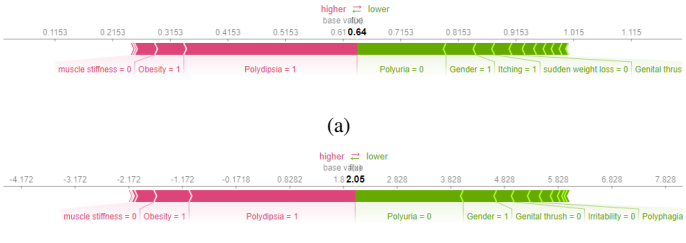
(b)

Fig. 2. SHAP results for global interpretability (a) RF (b) GB

As seen in Figure 2, there is no *Obesity* attribute among the 10 most important attributes of global interpretability. However, when the local interpretability of the selected patient is considered, it is seen that the second most important attribute that increases the probability of diabetes is *Obesity*. With local interpretability, it is possible to obtain more detailed results on a sample basis. Figure 4 shows the local interpretability results of the LIME for the same patient with SHAP. The high similarity of the results with SHAP and LIME strengthens the consistency of the results of both algorithms.

VI. Conclusion

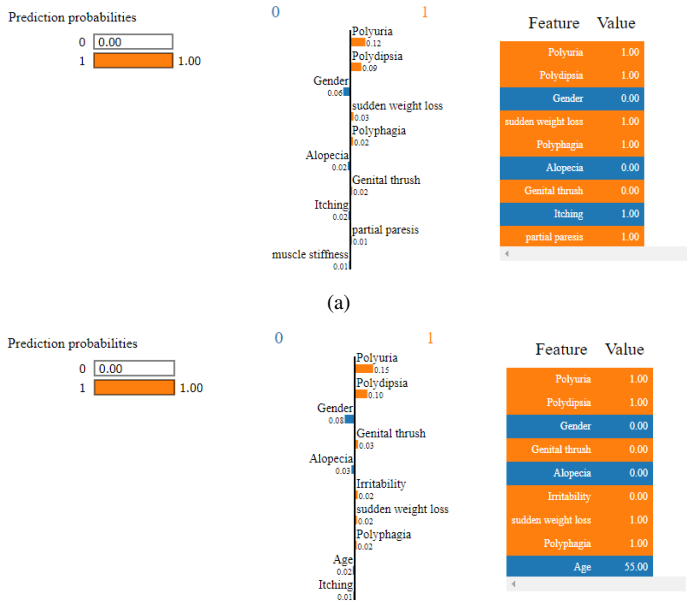
In critical areas such as healthcare, it is not enough to simply make high-accuracy predictions with ML. Interpretability is also critical to understand how to behave ML internally in minimizing the potential risk of error or bias of the results. IML is a promising interpretability approach to allow ML more transparent to its users. With IML, the interpretability of the results ensures that the values that affect the result are understood by ML users (clinicians). In this study, two popular



(a)

(b)

Fig. 3. SHAP results for local interpretability (a) RF (b) GB



(a)

(b)

Fig. 4. LIME results for local interpretability (a) RF (b) GB

IML models, SHAP and LIME, are applied after the data is trained with two complex ML models, which are RF and GB algorithms. The experimental results are evaluated globally and locally. According to the results, interpreting the complex and hard-to-interpret ML models gives insightful information about the most important attributes that affect the result, while maintaining high accuracy. It also enables the interpretability of a local instance (a patient).

REFERENCES

[1] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

[2] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[3] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

[4] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[5] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 2020.

[6] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 629–639, 2020.

[7] Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

[8] Samanta Knapič, Avleen Malhi, Rohit Salujaa, and Kary Främling. Explainable artificial intelligence for human decision-support system in medical domain. *arXiv preprint arXiv:2105.02357*, 2021.

[9] Hajar Hakkoum, Ali Idri, and Ibtissam Abnane. Assessing and comparing interpretability techniques for artificial neural networks breast cancer classication. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–13, 2021.

[10] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

[11] Devam Dave, Het Naik, Smiti Singhal, and Pankesh Patel. Explainable ai meets healthcare: A study on heart disease dataset. *arXiv preprint arXiv:2011.03195*, 2020.

[12] Safae Sossi Alaoui, Brahim Aksasse, and Yousef Farhaoui. Data mining and machine learning approaches and technologies for diagnosing diabetes in women. In *International Conference on Big Data and Networks Technologies*, pages 59–72. Springer, 2019.

[13] Augusto J Guimarães, Vinicius J Silva Araujo, Vanessa S Araujo, Lucas O Batista, and Paulo V de Campos Souza. A hybrid model based on fuzzy rules to act on the diagnosed of autism in adults. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 401–412. Springer, 2019.

[14] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[15] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[16] Andrea Bunt, Matthew Lount, and Catherine Lauzon. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 169–178, 2012.

[17] Lara Marie Demajo, Vince Vella, and Alexiei Dingli. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*, 2020.

[18] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. *arXiv preprint arXiv:2010.09337*, 2020.

[19] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[20] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[21] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272, 2012.

[22] Y Freund and RE Schapire. Experiments with a new boosting algorithm. in proceeding of the thirteen international conference on machine learning: 1996; san francisco edited by: Saitta l, 1996.

[23] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[24] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.