# Explainable artificial intelligence (XAI): a trend in biomedical data science

**Nguyen Quoc Khanh Le**

Over the past decade, artificial intelligence (AI) models have been drawn a lot of attention because of its efficiency in different fields. Many diverse studies have proven the effectiveness of AI models compared to traditional statistical ones. However, most current AI applications are built into the states called “black box AI” where people insert the data into machine to get the prediction outcomes without any further explanation. More specifically, current efforts aim to improve the predictive ability of algorithms by introducing more complicated systems that make it difficult for human to understand and, ultimately, what aspect of evidence these models rely on to make decisions. Despite the numerous benefits of widespread industrial implementation of machine learning models in a vast array of applications, a critical moral-issues-related domain such as medicine should be given much attention due to the enormous value it offers to humans. Furthermore, from the standpoint of sentient research, the uncertainty of sophisticated systems in making predictive decisions impedes their adoption rate in medical situations, as uninterpretable systems are reluctant to convince. Explainable artificial intelligence (XAI) models are essential for clinicians and patients to comprehend and convince their assumptions [1]. As a result, the XAI, which focuses on methods for interpreting machine learning models, has been rekindled in recent years. XAI can help practical significance of computer models by providing robust yet human-readable systems that provide straightforward explanations and improve safety, reliability, and transparency.

There are 2 ways to interpret a model as follows:

- Global interpretation: to understand how a model makes decisions for the overall structure. For example: we would like to predict the risk of disease in patients.
- Local interpretation: to understand how a model makes decisions for a single instance. For example, we would like to understand why a people has high risk of the disease.

In this article, we would like to discuss two well-known methods for local interpretation including SHapley Additive exPlanations (SHAP) [2] and local interpretable model-agnostic explanations (LIME) [3]. Both explanations are done on a public dataset related to biomedicine (Pima Indian diabetes dataset) from Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). This dataset contained lab data of both diabetes and non-diabetes samples; and could be used for conducting machine learning models to predict whether a subject who has diabetes or not. The data structure and variables are as follows:

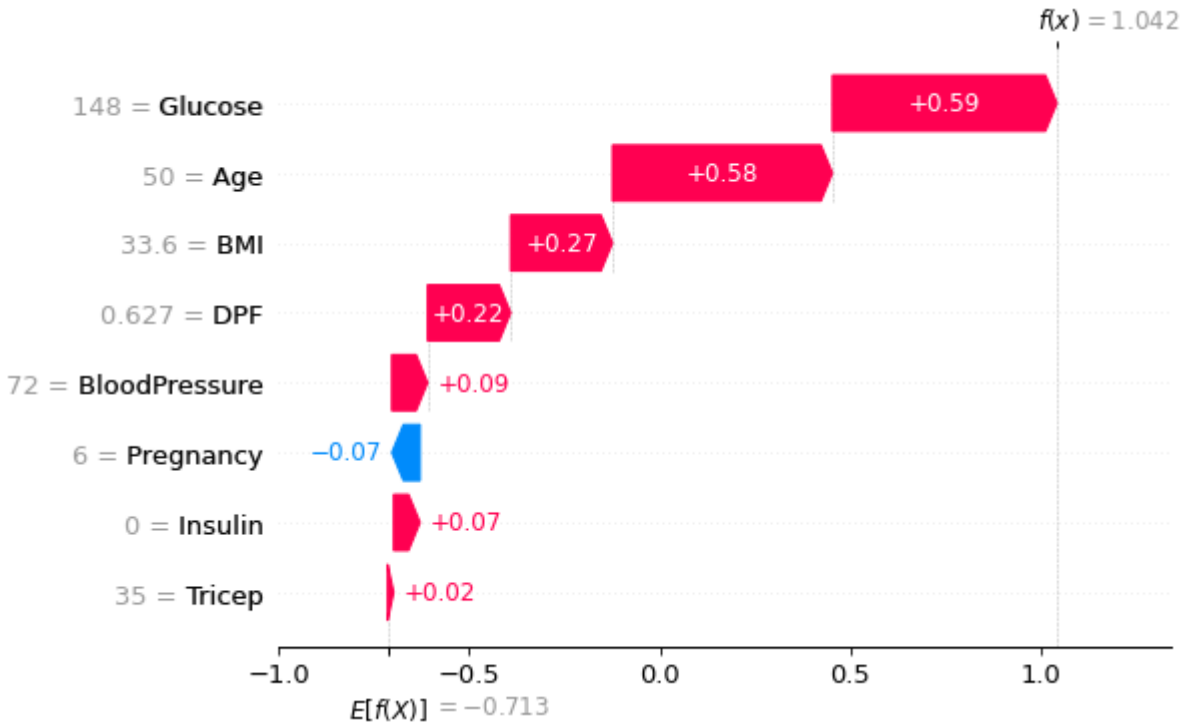
| Pregnancy | Glucose | BloodPressure | Tricep | Insulin | BMI | DPF | Age | Diabetes |
|---|---|---|---|---|---|---|---|---|
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |

Where “Diabetes” can be used as outcome variable, and the other variables are personal data or lab data that can be used to predict the diabetes outcome. Assuming that we try to solve the same machine learning problem to predict diabetes patients, now let see how we can interpret such kinds of model.

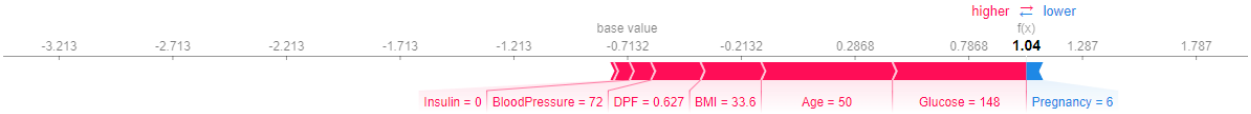
**(1) SHAP** is a game-theoretic technique that can be used to explain the output of any AI algorithms. It integrates optimal credit allocation to local explanations by employing traditional game-theory's Shapley values and their associated extensions. Detail information can be shown in the representative publication [2]. Released in 2017, it has been used in a lot of publications and this paper’s citations are more than 6,797 times recently (data from Google Scholar – May 2022). SHAP can be implemented using Python programming language. The tutorials for installing and using can be accessed at its homepage https://github.com/slundberg/shap. In details, SHAP can be installed freely using this command line syntax:

- ***pip install shap***

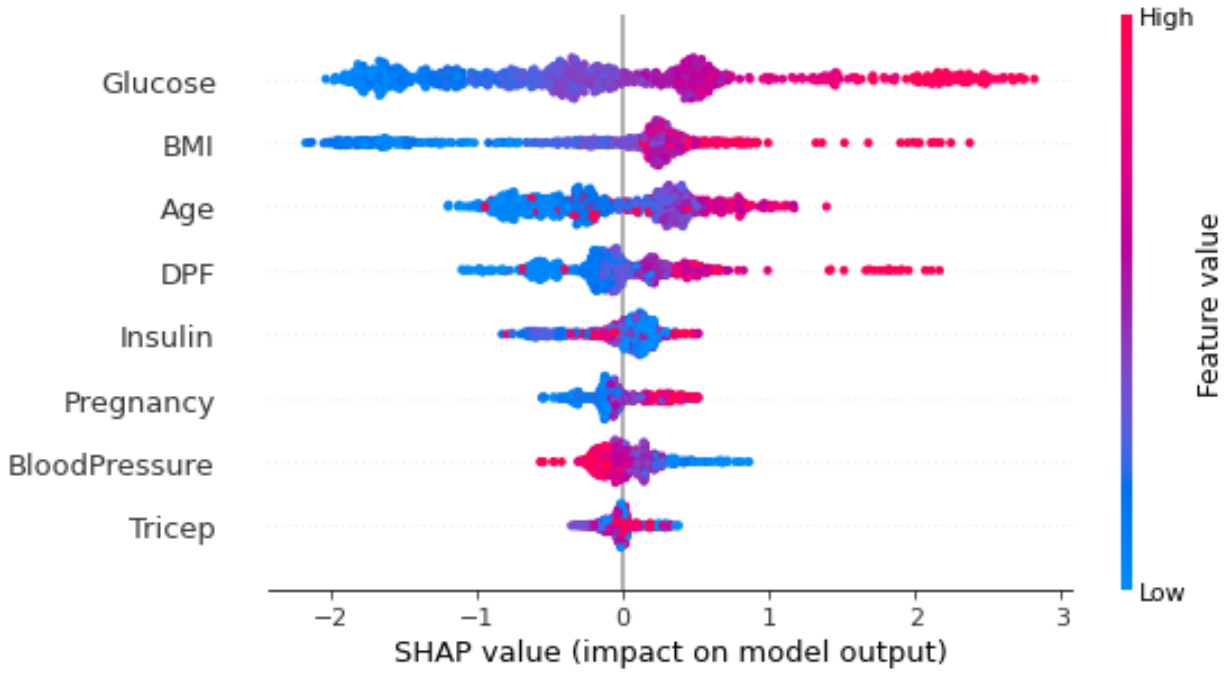
Now we try to explain our model using SHAP analysis on our sample dataset. For the first explanation, we would like to see the contribution of features when our model predicts an individual sample. Taking a sample for explanation, we plot the waterfall chart from SHAP analysis to see the feature contribution. As shown in the below figure, the *Glucose* feature pushes the prediction higher (with red color) than the other features, followed by the *Age* feature. On the other hand, the *Pregnancy* feature pushes the prediction lower (with blue color). Via this way, we can explain any samples and to see which features play an important role to push the performance of the model.




SHAP also provides another way to explain these features individually via a force plot (as shown in the below figure). From this plot, we can see that the most important features are ranked from *Glucose*, *Age*, and so on; while the bad one is *Pregnancy*.




The idea of SHAP interpretation based on the SHAP values and feature values. We can get an overview of which features are important in our model (on the whole dataset) via a bee swarm plot. This plot summarizes the effects of all the features that contribute to the model via their SHAP values. From the below figure, we may see that in overall, the most important features are *Glucose*, and followed by the *BMI* and *Age*. Taking the *Glucose* feature to have more detailed information of this plot, we can see that the high values of *Glucose*, the high impact (high SHAP values) on the model output. On the other hand, the low value of *BloodPressure*, the high impact on the model output.

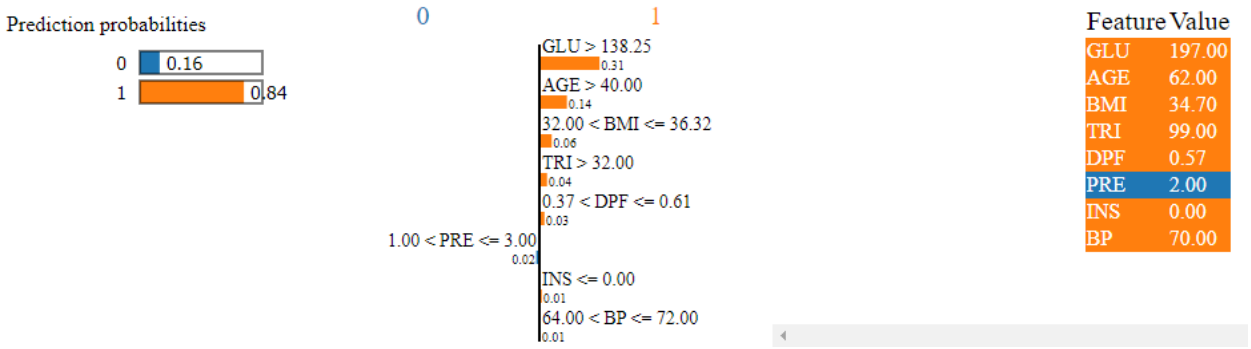



The idea of SHAP approach is popular recently and attracts many researchers to interpret any AI model. It has been used successfully in a variety of applications related to medical informatics [4] or bioinformatics [5], thus it is still growing in the near future.

**(2) LIME** [3] is a relatively new explainable technique that learns an interpretable model locally around the prediction to describe any classifier's predictions in an interpretable and faithful manner. LIME focuses on training local surrogate models to explain individual predictions rather than training a global surrogate model. Detail information can be shown in the representative publication [3]. Released in 2016, it has been used in a lot of publications and its citations are more than 8,646 times recently (data from Google Scholar – May 2022). LIME can be implemented using Python programming language. The tutorials for installing and using can be accessed at its homepage https://github.com/marcotcr/lime. In details, LIME can be installed freely using this command line syntax:

- ***pip install lime***

Now we try to explain our model using LIME analysis on our sample dataset. Since our problem is to create a machine learning model to predict whether a subject who has diagnosed with diabetes or not. We conducted LIME analysis to have an insight into the samples and features of validation dataset. Here, after creating a machine learning model to predict diabetes patients, taking a random sample for running LIME analysis and it shows as follows:




For positive prediction (diabetes – class 1), "*Glucose* > 138.25", "*Age* > 40", "32 < *BMI* < 36.32", "*Tricep* > 32", "0.37 < *DPF* <= 0.61", "*Insulin* <= 0", and "64 < *BloodPressure* <= 72" are the conditions to identify diabetes patients with a probability of 0.84. On the other hand, "1 < *Pregnancy* <=3" is the condition to reach the probability of 0.16 in non-diabetes patients. Therefore, this sample is classified as the diabetes patients since it reaches higher probabilities in positive prediction. This LIME analysis can be used to explain any sample in our dataset and they support different types such as tabular data, image data, or even text data. Recently, we have successfully applied LIME to explain our machine learning model on central precocious puberty prediction [6], and we think it holds potential for further applications in biomedical informatics.

In overall, we discuss the establishment of XAI, especially local interpretation models in a specific data related to biomedical informatics. Given the remarkable advances in biomedicine-based computational models over the last few years, model interpretability flaws exposed significant limitations. As a result, we genuinely think that XAI in biomedical data science still has a lot of untapped potential for future research.

**Data availability**

Source codes of this article can be accessed freely at https://colab.research.google.com/drive/15TPQJNsCEFIa-ESBOiq7nO1lYHRQYQ0J?usp=sharing.

**References**

1. Vo, T. H.; Nguyen, N. T. K.; Kha, Q. H.; Le, N. Q. K., On the road to explainable AI in drug-drug interactions prediction: a systematic review. *Computational and Structural Biotechnology Journal* **2022**.
2. Lundberg, S. M.; Lee, S.-I., A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **2017,** *30*.
3. Ribeiro, M. T.; Singh, S.; Guestrin, C. In *" Why should i trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; pp 1135-1144.
4. Hung, T. N. K.; Le, N. Q. K.; Le, N. H.; Tuan, L. V.; Nguyen, T. P.; Thi, C.; Kang, J.-H., An AI-based Prediction Model for Drug-drug Interactions in Osteoporosis and Paget's Diseases from SMILES. *Molecular Informatics* **2022,** *n/a* (n/a), 2100264.

5. Le, N. Q. K.; Ho, Q.-T., Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods* **2021**.
6. Huynh, Q. T. V.; Le, N. Q. K.; Huang, S.-Y.; Ho, B. T.; Vu, T. H.; Pham, H. T. M.; Pham, A. L.; Hou, J.-W.; Nguyen, N. T. K.; Chen, Y. C., Development and Validation of Clinical Diagnostic Model for Girls with Central Precocious Puberty: Machine-learning Approaches. *PLOS ONE* **2022,** *17* (1), e0261965.