A Review of Explainable Artificial Intelligence

Kuo-Yi Lin, Yuguang Liu, Li Li, Runliang Dou

A Review of Explainable Artificial Intelligence

Kuo-Yi Lin[1,2][0000-0002-7461-8116], Yuguang Liu[1], Li Li [1,2] and Runliang Dou[3,*]

[1] College of Electronics and Information Engineering Tongji University, Shanghai 201804, China.
[2] Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, 201804, China
[3] School of Management, Tianjin University, Tianjin 300072, China
drl@tju.edu.cn

Abstract. Artificial intelligence developed rapidly, while people are increasingly concerned about internal structure in machine learning models. Starting from the definition of interpretability and historical process of interpretability model, this paper summarizes and analyzes the existing interpretability methods according to the two dimensions of model type and model time based on the objectives of interpretability model and different categories. With the help of the existing interpretable methods, this paper summarizes and analyzes its application value to the society analyzes the reasons why its application is hindered. This paper concretely analyzes and summarizes the applications in industrial fields, including model debugging, feature engineering and data collection. This paper aims to summarizes the shortcomings of the existing interpretability model, and proposes some suggestions based on them. Starting from the nature of interpretability model, this paper analyzes and summarizes the disadvantages of the existing model evaluation index, and puts forward the quantitative evaluation index of the model from the definition of interpretability. Finally, this paper summarizes the above and looks forward to the development direction of interpretability models.

Keywords: Explainable, Machine Learning, Classification, Application.

1 Introduction

The deep learning model fits well in the era of big data, but its accuracy and efficiency are based on the improvement of algorithm efficiency and the combination of huge parameter space. This also means that efficient machine learning algorithms are difficult to directly understand or explain. At present, the existing interpretable machine learning methods are mostly classified according to the original model of the interpretation model. Therefore, this paper summarizes the definition, scope, nature and categorizes the interpretation model more comprehensively and discusses the interpretation issues faced in the digital manufacturing process and proposes some suggestions.
The paper is organized into the following sections. Section II provides a review of the related research pertaining to the definition of interpretability and explainable model.

Section III describes the explainable model from the aspects: aims, scope, implementation and classification, especially in the last part, the classification of explainable models has been reorganized in a clearer way. Section IV discusses the application and problems of explainable models, as well as proposes the feasible solutions. In the Conclusion, the paper looks forward to the development prospects the research trends of explainable models.

2 Related works

Historically speaking, since the 1970s, there has been sporadic interest in explanations itself, which began with attention on expert systems. In the following thirty years, the attention of related research shifted to neural networks and recommendation systems, as shown in Fig.1. Nonetheless, progress on these issues slowed about a decade ago. This is because the focus of AI research has shifted to implementing algorithms and models that focus on predictive power, while the ability to interpret decision processes has taken a back seat.

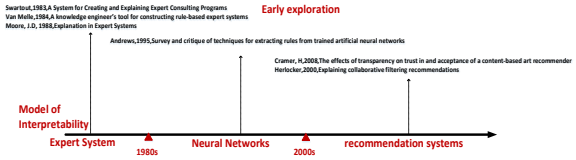

Fig. 1. Early exploration of explainability model and definition

The interpretive definition has improved over time, for example, Gunning [1] adopts that explainable artificial intelligence enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners. Molnar [2] considers that interpretable machine learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans. Miller [3] regards interpretability as the degree to which a human can understand the cause of a decision. The details and the reasons used to explain or even whether the explanation is easy to understand are completely dependent of the audience while these definition neglects the role of the audience. So the definition must reflect the dependence of the explainable model on audience.

Aimed at above problems, [4] gives the definition of explanation that given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand. At this point, we can define the interpretability model as given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

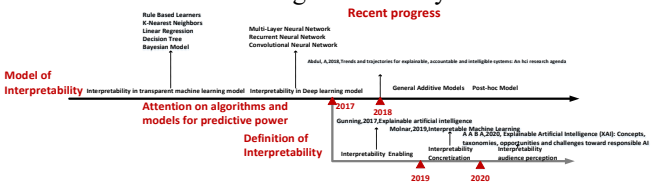

Fig. 2. Recent progress of explainability model and explainability definition

3 Explainable model

3.1 Scope of explainable model

Algorithm transparency is only responsible for answering the question of creating a model without data or learning models [2], but the focus of interpretable machine learning is prediction rather than algorithm [5]. Decomposability refers to the ability of the model to explain each part, which can ensure that the model obtains the interpretation of the input or parameters from the existing conditions. Simulability is the ability to simulate in a more complex environment, within which the model can obtain simulation equations from known conditions.

The global interpretability model starts from the training data and the entirety of each part of the model. The interpretability of the global model can be divided into the overall level and the modular level. The overall level of global interpretability can be realized conceptually [6], but this requires a trained model and corresponding algorithms and data [5], which is difficult to complete in practice [2]. Although the global interpretability at the overall level is difficult to achieve, the interpretability at the module level can be achieved [5]: For example, the interpretable part of a linear model is its weight; the interpretable part of a decision tree is its node splitting and analysis [5] [2]. It is worth noting that the various parameters of the linear model are related to the whole, and feature decoupling may be involved when considering feature interaction [7].

The idea of explaining the prediction of a single sample is to enlarge a single sample and try to understand how the model achieves the prediction [5]. For the black box model, using a simpler interpretability model to approximate a smaller target area can maintain interpretability while maintaining high accuracy. This is because under the assumption that local predictions are linearly or monotonically dependent on certain features, local explanations may be more accurate than global explanations [2]. For the interpretation of a certain type of sample prediction, there are essentially two methods: treating the target sample group as the entire data set for prediction and interpretation; or summarizing the above-mentioned partial interpretation of the prediction of a single sample to achieve the prediction of the sample group [2].

3.2 Implementation of explainable model

The interpretability of the model can be divided into the interpretation model and the interaction with the audience. As stated in the previous definition of interpretability, interpretability is not limited to models, it is for specific audiences. For interaction with users, it may involve prototype interpretation interfaces and visualization of models; or the use of psychological knowledge to explain machine learning principles, etc.

The interpretation model can be realized by simplifying or imposing constraints on transparent or other specific models; it can also be realized by establishing post interpretable models for existing models. Specifically, for a deep interpretation model that aims to explain deep learning, improved deep learning techniques can be used to learn

its interpretable functions. For example, Cheng [8], etc. look for evidence of scene judgment in pictures and learn semantic association. Cheng [8] trains the network to associate semantic attributes with hidden layer nodes and associates labeled nodes with known ontology, which makes it have explanatory power. Researchers at the University of California, Berkeley [9] use the idea of generating image captions and generating visual interpretations to generate visual interpretations, which associate image descriptions with class definitions, identify objects in the image through the CNN model, and use the RNN model to convert the features in CNN model into words and titles.
Interaction modules that enhance audience understanding can be divided into psychology (or humanities) and human-computer interaction (HCI) parts. Specifically, we should first study and model how humans produce and understand explanations, and what attributes can make humans perceive explanations to achieve human interpretability, which belongs to the category of psychology or humanities. In addition, the way and process of interaction between users and entities greatly affects users' understanding and trust in them.

4 Classification of explainable model

The interpretability model can be summarized into three categories according to its training sequence with the original model: pre-model, in-model, and post-model, which is shown in Fig.3.

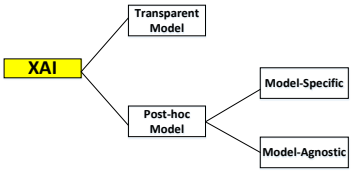

Fig. 3. Current Status of Interpretable Model Research

Solvability models can be divided into specific model-specific interpretation models and model-independent interpretation models according to whether they are restricted to specific models, which is shown in Fig.3. Because model-specific interpretation methods are based on the internal characteristics of certain specific models, they are limited to specific model classes; relatively, interpretation models that have nothing to do with the original model can be applied to a wider range of model interpretation.

4.1 Scope of explainable model

The transparent model refers to a model that is understandable by itself. The range of interpretability can be described by the transparency, decomposability and emulation of the above algorithm. The interpretability of a transparent model can be achieved by imposing constraints on the model, such as sparsity, monotonicity, causality, or physical constraints in the professional field. The current status of interpretable transparent model is shown in Fig.4.

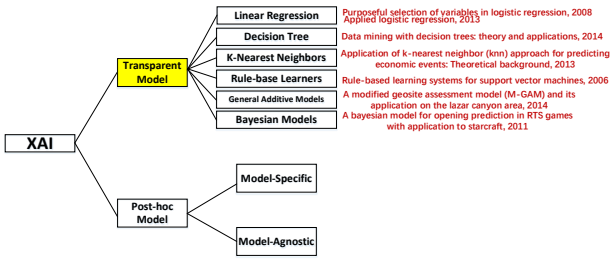

Fig. 4. Current Status of Interpretable Transparent Model

Logistic regression model is essentially a classification model, and a logistic regression model with continuous dependent variables is a linear model. There is a linear dependence between the predictors and predictors of this type of model, so it reduces the flexibility of data fitting while giving the model interpretability. Decision tree is a decision-making hierarchy used to solve classification and regression problems, and it satisfies the constraints of model transparency. K-Nearest Neighbors (K-Nearest Neighbors) makes decisions on test samples by coordinating the prediction results of K-nearest neighbors. The key to the interpretability of this model is the distance and similarity between K neighbors, which is similar to the human decision-making process. Enhancing the interaction between the user and the model [4]. Rule learning is a general term for a certain type of model, which refers to a model that can obtain corresponding rules from training data. The rules can be simple if-then conditional rules or more complex rules combined by simple rules. The number and nature of the rules in the model have a greater impact on the performance and interpretability of the model. The generalized additive model is a linear model that makes predictions or decisions through a combination of smooth functions defined by predictor variables. The Bayesian model adopts the form of probability directed acyclic graph to express the conditional dependence between a group of variables. These models above can be applied to production scenarios with small data but high requirements for real-time and model transparency

4.2 Post-hoc Explainable Models

The post-hoc interpretable model refers to the model that explains the existing initial model. Its purpose is to explain the opaque black box machine learning model. The post-hoc interpretability model can be divided into an explanation model for a specific machine learning model and an explanation model independent of the model.

4.2.1 Model-Specific methods in Shallow ML

Machine learning models include a variety of supervised learning models, many of which are complex algorithms. For example, extracting interpretation rules from LSTM or distilling neural networks into soft decision trees belong to this type of interpretation model method. The current status of interpretable shallow ML model is shown in Fig.5.

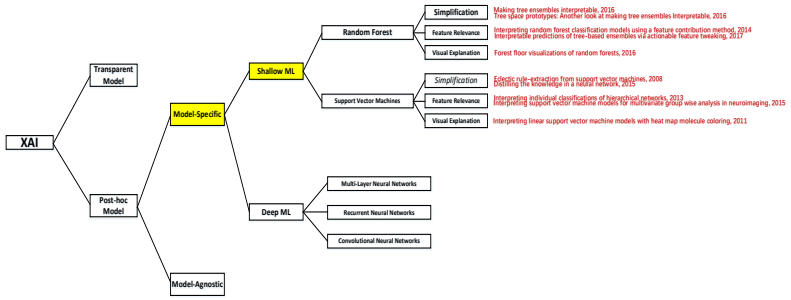

Fig. 5. Current Status of Interpretable Shallow ML Model

The interpretation method for this model can be summarized as model simplification and variable importance analysis. Subsequently, Deng [10] created a simplified tree set learning model (STEL). But the simplification of the model inevitably means the loss of efficiency. S. Hara et al. [11] proposed to use the simplified model for model interpretation while maintaining the original complex model for decision-making. Based on the above problems, Breiman [12] first proposed to use the importance of features in random forests as an explanation basis. Subsequently, Tolomei et al. [13] proposed a method to measure the importance of features by changing the prediction category.

The support vector machine model completes tasks such as classification by constructing one or more hyperplanes in a high-dimensional space. Support vector machines have high generalization performance, and their commonly used model visualization methods include: model simplification, visualization, and local interpretation. In terms of model simplification, [4] divided it into the following four categories according to the depth of the model: 1) Extract a more explanatory rule learning model from the support vector machine; 2) [14] established additional hyperplanes for the support vector machine and constructed a rule learning model to complete the model simplification; 3) trained with the original interpretation model; 4) There are still studies [15] that have created statistics that explicitly explain the margins of support vector machines and specifically explain the multi-factor patterns shown in neuroimaging. The above models are suitable for production scenarios with more data and more complex models required. In this scenario, the interpretability of the model can be enhanced by imposing limiting factors on it.

4.2.2 Model-Specific methods in Deep Learning Models

Although deep neural networks are black box models, their good accuracy makes them widely used in various industries. Explanatory research has always been a hot issue in the industry. As shown in Fig.6, common interpretation methods include, simplified models, feature correlation models, visual interpretation models, text interpretation models, sample interpretation models, partial interpretation models, etc.

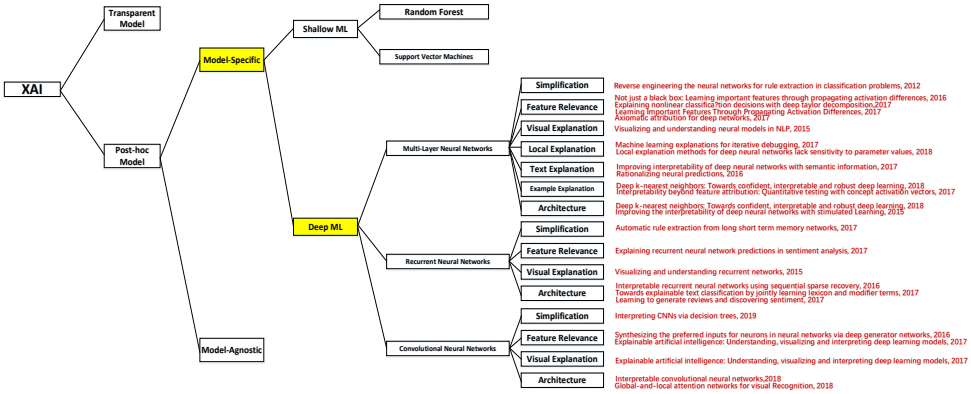

Fig. 6. Current Status of Interpretable Deep ML Model

Multi-layer neural network is one of the most common black box models, and the explanation of the black box model is a necessary condition to realize its application value. There are few studies on the simplified interpretation of multi-layer networks directly, such as the Deep-RED algorithm [16]. [17] et al. treat model simplification as a Post-hoc interpretability method independent of the model. [16] interprets the original model through a gradient boost tree, and [17] reduces the multi-layer model to a model. However, for complex multilayer neural networks with deep layers, the feature correlation model is more suitable. [17] proposed on the basis of the Deep-LIFT algorithm of [18] to explain the original model by correlating the output result with the contribution of the input.

RNN network is widely used in original data with inherent order such as natural language processing and time series because of its ability to store data relationships. [19] et al. proposed a specific propagation rule that works with RNN multiplicative connection to explain the knowledge of RNN neural network. [20] combined the hidden Markov model with the RNN model to ensure the interpretability and accuracy of the model, [21] et al. changed the model itself to understand the process of model decision-making.

The CNN network is widely used in computer vision, and its ability to visualize data essentially strengthens human understanding of the model. By constructing a global average pooling layer between the last convolutional layer of CNN and the fully connected layer of the predicted object, the weight projection of the output layer in the convolutional feature is realized, and the important image for a specific object class is identified by this Area and improve the mapping quality [22]. [23] uses the method of feature correlation analysis, using the loss of each filter in the high-level convolutional layer to force it to learn a specific object. [24] adopted a visual interpretation method and proposed the Grad-CAM algorithm to highlight the interpretation area of the predicted target in the image. [25] combined the CNN network with the RNN network and the LSTM model, and explained the pictures by analyzing the text information. In addition to the above method of explaining the decision-making process by mapping the output in the input space, [26] explained how the CNN network makes decisions by going deep into the network and through the middle layer. [27] combined the above-

mentioned deep model internal method with the input-output mapping method to explore the decision-making process of CNN. In addition, [28] and [29] used LIME algorithm and bio-confrontation detection interpretation method to explain CNN network respectively. The above models are suitable for production scenarios with extremely large data and high accuracy requirements. This scenario usually trains the model in an offline state, but simply imposing constraints on the model cannot balance the performance of the model and its interpretability. Therefore, for such production scenarios, a separate interpretable model is usually trained to explain the original model.

4.2.3 Model-Agnostic methods divided by explainable scope

The model diagnosis interpretation method does not depend on the original model to be explained. Under certain circumstances, model-specific interpretation methods are better than model-independent interpretation methods, but the advantage of the latter is that it is completely independent of the original model, which guarantees the reusability of the model to a certain extent.

This paper will classify and summarize this type of interpretation model within the scope of its interpretation. Specifically, it can be divided into a local interpretability model, a global interpretability model, and a local/global interpretability model, which is shown in Fig.7.

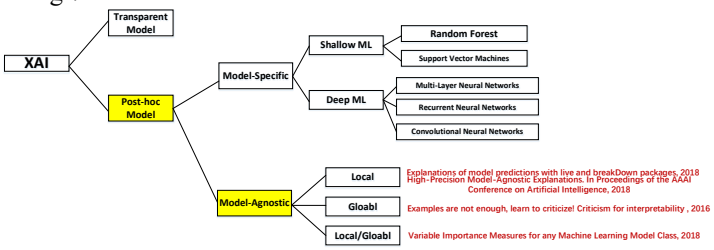

Fig. 7. Current Status of Interpretable Deep ML Model

5 Problems and development

The problems of interpretability models include the trade-off between model interpretability and performance, the evaluation indicators of interpretability models, and other problems or other unresolved problems in existing interpretability models. However, the emergence of ex post interpretability models and mixed interpretation models may provide interpretability while maintaining the performance of the original model. The evaluation index of an interpretability model refers to a measurement system that evaluates the pros and cons of the model from the definition, nature and goals of the model. Starting from the definition of the interpretability model, comprehensibility needs to consider the understanding ability of the audience, so it is a relatively subjective nature. But from an objective point of view, a relatively simple model is easier to understand. Based on the expansion of this definition, Doshi-Velez and Kim [30] proposed three levels of interpretability of evaluation models according to the interpretation cost and the effectiveness of the interpretation results: application-based evaluation, model audience-based evaluation and application evaluation method. Interpretability is essential to promote learning. Humans do not need to explain everything, but when dealing with

a specific event, human curiosity will drive them to explore the reasons for the event and update their environmental thinking models in the process of exploring the reasons. Therefore, the interpretability model does not have a mandatory goal. Its main goal is to develop the original artificial intelligence model toward responsibility, fairness, privacy and data fusion while maintaining a high level of predictive performance.

6 Conclusions

This paper defines the interpretability model and describes the aims, scope, implementation and classification of the explainable model. The paper discusses the application and problems of explainable models, as well as proposes the feasible solutions. Aiming at the contradiction between the interpretability of the interpretability model and the accuracy of the original model, the ex post interpretability model can be used to solve the above contradiction; and for the problem of the evaluation index of the interpretability model, this article considers the nature of interpretability and starting from the definition, put forward a method route that can quantitatively evaluate the model. Follow-up research can do more in-depth exploration of this method.

Progress	Details
Model	Summarize the definition, scope and nature for interpretability model and categorize the interpretability model.
Applications	Analyze the applications in industrial fields and discuss the interpretation issues faced in the digital manufacturing process.
Development	Propose suggestions based on the existing interpretability model.

References

1. D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).
2. Molnar, C. Interpretable Machine Learning. 2019. Available online: https://christophm.github.io/ interpretable-ml-book/ (accessed on 22 January 2019).
3. Miller, T. Explanation in Artificial Intelligence: Insights from the social sciences. Artif. Intell. 2018, 267, 1–38. [CrossRef]
4. A A B A , Natalia Díaz-Rodríguez b, D J D S A C , et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion, 2020, 58:82-115.
5. Carvalho D V , Pereira E M , Cardoso J S . Machine Learning Interpretability: A Survey on Methods and Metrics[J]. 2019.
6. Lipton, Z.C. The mythos of model interpretability. arXiv 2016, arXiv:1606.03490.
7. Bengio, Yoshua, Courville, et al. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-1828.
8. Cheng, H., et al. (2014) SRI-Sarnoff AURORA at TRECVID 2014: Multimedia Event Detection and Recounting.
9. Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations, arXiv:1603.08507v1 [cs.CV] 28 Mar 2016

10. H. Deng, Interpreting tree ensembles with intrees (2014). arXiv:1408.5456.
11. S. Hara, K. Hayashi, Making tree ensembles interpretable (2016). arXiv:1606.05390.
12. L. Breiman, Classification and regression trees, Routledge, 2017.
13. G. Tolomei, F. Silvestri, A. Haines, M. Lalmas, Interpretable predictions of tree-based ensembles via actionable feature tweaking, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 465–474.
14. X. Fu, C. Ong, S. Keerthi, G. G. Hung, L. Goh, Extracting the knowledge embedded in support vector machines, in: IEEE International Joint Conference on Neural Networks, Vol. 1, IEEE, 2004, pp. 291–296.
15. B. Gaonkar, R. T. Shinohara, C. Davatzikos, A. D. N. Initiative, et al., Interpreting support vector machine models for multivariate group wise analysis in neuroimaging, Medical image analysis 24 (1) (2015) 190–204.
16. J. R. Zilke, E. L. Menc´ıa, F. Janssen, Deepred–rule extraction from deep neural networks, in: International Conference on Discovery Science, Springer, 2016, pp. 457–473.
17. R. Traor´e, H. Caselles-Dupr´e, T. Lesort, T. Sun, G. Cai, N. D. Rodr´ıguez, D. Filliat, DisCoRL: Continual reinforcement learning via policy distillation (2019).
18. A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences (2016).
19. L. Arras, G. Montavon, K.-R. M¨uller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis (2017).
20. V. Krakovna, F. Doshi-Velez, Increasing the interpretability of recurrent neural networks using hidden markov models (2016).
21. E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, in: Advances in Neural Information Processing Systems, 2016, pp. 3504–3512.
22. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
23. Q. Zhang, Y. Nian Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8827–8836.
24. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
25. Y. Dong, H. Su, J. Zhu, B. Zhang, Improving interpretability of deep neural networks with semantic information, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4306–4314.
26. D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6541–6549.
27. C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, the building blocks of interpretability, Distill (2018).
28. M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning (2016).
29. N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning (2018).
30. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. arXiv 2017, arXiv:1702.08608.