Chapter 1
Prologue: Introduction to Explainable Artificial Intelligence

Moamar Sayed-Mouchaweh

1.1 Explainable Machine Learning

Machine learning methods, especially deep neural networks, are becoming increasingly popular in a large variety of applications. These methods learn from observations in order to build a model that is used to generalize prediction (classification or regression) to unknown data. For instance in the healthcare domain, a model is learned in order to decide whether a patient has a cancer or not by treating its microscopic radio images as input. Other example is credit scoring where a model is used to decide whether a candidate may obtain a loan or not.

Machine learning methods generate or learn a highly effective mapping between input and output. Hence, they behave as "black box" entailing a huge difficulty for humans to understand how and why the prediction (output) was made. However in many applications, it is important to explain to users how the decision (prediction) was made by the model and its meaning using understandable terms. Indeed, explainable models allow users to trust them and to better use them thanks to the detailed information (explanation) on how and why they arrived to the provided prediction. Therefore, making machine learning models transparent to human practitioners or users leads to new types of data-driven insights.

Explainable Artificial Intelligence (XAI) [1] aims at producing explainable models that enable human users to understand and appropriately trust the obtained results. The produced explanations allow to reveal how the model functions, why it behaves that way in the past, present and future, why certain actions were taken or must be taken, how certain goals can be achieved, how the system reacts to certain inputs or actions, what are the causes for the occurrence of a certain fault

M. Sayed-Mouchaweh (✉)
Institute Mines-Telecom Lille Douai, Douai, France
e-mail: moamar.sayed-mouchaweh@mines-douai.fr

M. Sayed-Mouchaweh (ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems*, https://doi.org/10.1007/978-3-030-76409-8_1

and how this occurrence can be avoided in the future, etc. The need for explanations is increasingly becoming necessary in multiple application domains, such as smart grids [2], autonomous cars [3], smart factory or industry 4.0 [4, 5], telemedicine and healthcare [6], etc., in particular within the context of digital transformation and cyber-physical systems [7]. For instance in credit scoring or loan approval system, providing an explanation why a loan was refused to applicants allows bias detection. In cybersecurity [8], explanations may help to ensure the security and safety of a system by determining abnormal changes in its output. These changes were generated by hackers in order to fool the model. Indeed, model transparency allows assessing the quality of output predictions and warding off adversaries [9]. However, explanation is required for critical systems where a mistake may have serious consequences such as for self-drive cars.

Making machine learning-based AI explainable faces several challenges. Firstly, the explanations must be adapted to different stakeholders (end users, policy makers, industries, utilities, etc.) with different levels of technical knowledge (managers, engineers, technicians, etc.) in different application domains. Secondly, it is important to develop an evaluation framework and standards in order to measure the effectiveness of the provided explanations at the human and the technical levels. For instance, this evaluation framework must be able to verify that each explanation is consistent across similar predictions (similar observations) over time, is expressive in order to increase the user confidence (trust) in the decisions made, promote impartial and fair decisions, and improve the user task performance. In order to address these challenges, several questions need to be answered:

- How explanations can be integrated into new and already existing AI systems and models?
- How effective user explanation interfaces can be designed?
- How the social (e.g., user requirements, user confidence) and ethical (e.g., bias detection, fairness) aspects of explainability can be guaranteed and evaluated in the designed XAI models?
- How objective benchmarks, standards, and evaluation strategies can be developed in order to measure the effectiveness of XAI models for different applications, uses, and stakeholders of different expertise levels?

Explanations can be generated either to explain the model output (prediction) or its parameters. For instance for the cancer detection model, explanations can be a map of the microscopic image's pixels that contribute the most to the provided decision or prediction. It can also be the features or activation functions' parameters that contributed or related the most to the provided prediction.

Explanations can be generated either during the system design or during operation in order to ensure its quality and reliability. Explainable systems can be either self-explainable or user-triggered. Self-explainable systems are able to generate explanation whenever something requires to be explained in particular at run-time. User-trigger explanation provides explanation whenever a user requests it. For both cases, a model is required allowing to understand the system's internal dynamics,

its goals and requirements, its surrounding environments (contextual knowledge), and the recipient of the explanation.

In general, XAI approaches can be categorized in terms of their scope, methodology, and usage [9]. The scope indicates the focus of the explanation on a local instance [10] or on the model as a whole [11]. The methodology refers to the focus of the used approach on the input data instance [10] or the model parameters [12]. Finally, the usage concerns how the explainability is integrated to the model either to a specific model (intrinsic) or to any model as a post hoc explanation (model-agnostic) [10]. Each of the previous approaches has its advantages and drawbacks. For instance, local explanation is much easier than global explanation since it is easier to provide an explanation about a local instance than for a whole model. Moreover, post hoc agnostic models have the advantage to be dependent of the machine learning algorithm used to train the model and therefore can be applied to any already existing model.

1.2 Beyond State-of-the-Art: Contents of the Book

The following chapters in this book overview, discuss, and propose different explainable AI structures, tools, and methods in order to address the aforementioned challenges in different applications (healthcare, finance, cybersecurity, document summarization).

1.2.1 Chapter 2: Principles of Explainable Artificial Intelligence

This chapter presents a state of the art around the major explainable machine learning (XAI) methods and techniques. It starts firstly by highlighting the motivation of XAI in particular when it is used in automated decision-making process for critical systems such as healthcare or autonomous vehicles. Then, the chapter presents a set of dimensions that can be used to analyze or evaluate the capacity of XAI to explain in understandable terms to a human, so-called interpretability. The chapter divides XAI methods into categories according to their capacity to provide global or local explanations, if the explanation is related to a specific machine learning model or model-agnostic explainers, to the time that a user is allowed to spend on understanding an explanation as well as their capacity to guarantee the safety desiderata such as privacy, secrecy, security, and fairness. Then, the chapter evaluates some well-known explainers according to the aforementioned dimensions and categories, such as Trepan, Rxren, Corels, Lore, and Anchor, explainers based on the use of a saliency map, prototype-based explanations, and counterfactual explanations.

1.2.2 Chapter 3: Science of Data: A New Ladder for Causation

This chapter proposes a theoretic framework to create an explainable AI tool that is capable of reasoning. The chapter starts by discussing how deep neural networks explain the decision they made by finding the causes generating the current events or situations. To this end, the chapter proposes a cognitive architecture (Learning Intelligent Distributed Agents (LIDA)) equipped with probabilistic fuzzy logic and graphical neural networks for reasoning. Probabilistic fuzzy logic treats three types of uncertainty: randomness, probabilistic uncertainty, and fuzziness. Therefore, it can manage the uncertainty of our knowledge (by the use of probabilities) and the vagueness inherent to the world's complexity (by data fuzzification). It is integrated with graphical neural networks for learning since it cannot learn by itself and needs experts to define intervals before applying fuzzification.

1.2.3 Chapter 4: Explainable Artificial Intelligence for Predictive Analytics on Customer Turnover

This chapter presents an interactive explainable artificial intelligence web interface that integrates and enhances the state-of-the-art techniques in order to produce more understandable and practical explanations to nontechnical end users. It is applied for the prediction of a financial institution's customer churn rate. The Churn is the rate of customers who stopped using a service or product in a given time frame. It is used in business sector such as banking, retail, telecommunications, and education. The proposed explainable web interface combines visualization and verbalization. The visual screens display local and global features importance in order to provide users with the relevance of each feature to the decision made on a certain instance or on the global model. The verbalization is used as an alternative explanation other than the one provided by the visual screens. In addition, it is used as a recommendation to what to do in order to prevent a customer from leaving the company. The proposed explanation web interface is evaluated and compared with some well-known similar explanation tools, such as GAMUT, TELEGAM, and XAI Clinical Diagnostic Tool, using the following criteria: global explanation, local explanation, interactive (contrastive), search table, target instance capability, and the targeted audience.

1.2.4 Chapter 5: An Efficient Explainable Artificial Intelligence Model of Automatically Generated Summaries Evaluation

This chapter handles the problem of the evaluation of automatically generated summaries. The latter is used for facilitating the selection process for a document and index documents more efficiently, in particular when dealing with massive textual datasets. The chapter presents the most important cognitive psychology models for text comprehension such as the Resonance Model, Landscape Model, Langston and Trabasso Model, and the Construction Integration Model. Then, it proposes a cognitive protocol for Automatically Generated Summaries Evaluation (AGSE) based on a cognitive psychology model of reading comprehension. The originality of this protocol is that it takes into consideration the extent to which an abstract is a good abstract by using three criteria: retention, fidelity, and coherence. The retention checks whether the generated output covers all the concepts reported in the source document, the fidelity gives insights into the extent to which the generated summary accurately reflects the author's point of view by focusing on salient concepts conveyed in the source text, and the coherence checks if the automatically generated summary is semantically meaningful. The retention and fidelity scores are modeled using three linguistic variables, "low," "medium," and "high" represented as membership functions. Then the score combining both the retention and fidelity through operators OR and AND is calculated as fuzzy score using three rules. For instance, *If the retention score is low and the fidelity score is also low, then, the R-F score is low.* Finally, text units having the highest *R-F* scores after defuzzification will present candidate sentences of an ideal summary. Three datasets containing articles and extractive summaries about different topics, such as crisis or protest or war in some Middle East countries, are used for the evaluation of the presented protocol.

1.2.5 Chapter 6: On the Transparent Predictive Models for Ecological Momentary Assessment Data

This chapter describes the use of some well-known classification methods (Bayesian Network, Gradient-boosted trees, Naïve Bayes, Multi-Layer Perceptron, Random Forest, and Decision Tree) in order to assess individuals' eating behavior. The goal of this classification is to predict the future eating behavior of an individual regarding eating healthy or unhealthy in order to intervene just in time if the future behavior is eating unhealthy. The chapter explains the different steps used in order to perform this prediction as well as its interpretation. The dataset represents samples collected from 135 overweight participants over 8 weeks. The goal of the interpretation step is to understand the triggers that lead users to make a choice, which is less healthy than the others are. The extracted features are categorical, such

as the location of eating, and continuous, e.g., craving. These features' capacity to discriminate a certain class from the others is used for the interpretation. Each individual is represented by a circle with a size, color, and edges. The color indicates the cluster type, while the size represents how much the corresponding individual shares behavior patterns with the other individuals belonging to the same cluster. The edges correspond to the links between the different individuals belonging to the same clusters.

1.2.6 Chapter 7: Mitigating the Class Overlap Problem in Discriminative Localization: COVID-19 and Pneumonia Case Study

This chapter treats the problem of distinguishing COVID-19 from other pneumonia within a single model trained to detect both diseases using computed tomography (CT) scans, lung ultrasonography (LUS) imagery, and chest X-rays. The challenge is that both diseases are very close entailing to have overlapped classes in the feature space. In addition, there are much fewer COVID-19 labels to learn from entailing class imbalance problems. Therefore, this chapter proposes an approach, called Amplified Directed Divergence, that works with ensembles of models to deal with class overlap and class imbalance while ensuring confident predictive assessments. This approach does not require localized labels, since they can be labor-intensive, but rather exploits the Class Activation Maps (CAMs) computed at the final convolutional layer for each class in a Convolutional Neural Network (CNN) model. The goal is to perform the classification and localization of COVID-19 Regions of Interest (RoI) from CT scans, LUS imagery, and chest X-rays. The salient regions in the COVID-19 CAMs can then be unsampled to the size of the original image in order to localize the features most conducive to the classification of the chosen class. In order to mitigate aleatoric uncertainty, related to overlapped classes, a kernel method is used. It accepts two class activation maps from expert models, each trained on specific overlapped classes, and extracts activations that are relevant to one of them, i.e., target class (COVID-19). The result is a new class activation map that better localizes objects of interest in the presence of class overlap. The proposed approach is validated using COVID-19 and Viral Pneumonia imagery. The obtained results show that the proposed approach enables machine learning practitioners and subject matter experts in various domains in order to increase their confidence in predictions provided by models trained on image-level labels when object-level labels are not available.

1.2.7 Chapter 8: A Critical Study on the Importance of Feature Selection for Diagnosing Cyber-Attacks in Water Critical Infrastructures

This chapter proposes the use of feature selection techniques in order to improve the cyber-attack detection and classification system. This improvement is achieved thanks to the elimination of irrelevant and redundant features from the original data, in particular when they are described by high dimensional and low-quality feature space. In addition, this elimination allows to reduce learning time and prevent overfitting. The chapter proposes the comparison of 12 feature selection techniques in order to effectively select the optimal set of features for detecting intrusion. Then, the selected features are used by two different supervised classification methods k-Nearest Neighbors (kNN) and Decision Trees (DT) in order to perform the intrusion (cyber-attack) classification. The proposed intrusion detection system is applied to a water storage system. The latter is a cyber-physical system vulnerable to seven different types of attacks. The chapter compares the performance of the 12 feature selection techniques and the impact of selected features on the kNN and DT intrusion detection and classification accuracy. The feature selection and analysis can be seen as a way to identify the features that contribute the most to detect a cyber-attack. Then, these features can be used in order to explain the nature, type, and behavior of that detected type of attack.

1.2.8 Chapter 9: A Study on the Effect of Dimensionality Reduction on Cyber-Attack Identification in Water Storage Tank SCADA Systems

This chapter proposes the use of dimensionality reduction techniques (Locally Linear Embedding, Isomap, Linear Discriminant Analysis, Multidimensional Scaling, Principal Component Analysis, etc.) in order to improve the cyber-attack detection and classification system. This chapter is complementary to the previous chapter in the sense that feature selection usually works when at least a number of features possesses very useful information, while dimensionality reduction tries to rectify the feature space and obtain an improved distribution. Indeed, dimensionality reduction can be very helpful in the design of intrusion detection systems (IDS). For instance, if a cyber-attack can be detected by monitoring a large number of features, dimensionality reduction can yield a feature space in which only one or a very small number of features are enough to explain a change that indicates a cyber-attack. In contrast, other techniques such as feature selection may not result in the same efficiency, as the features may not have enough information to only select a small number of them to detect a cyber-threat. The chapter discusses the benefit of using dimensionality reduction to provide an explanation of the detected intrusion and

confidence of the obtained decision (detected intrusion). Indeed, the feature space may contain hidden characteristics that are dormant to human eye. Dimensionality reduction techniques improve the explainability by capturing the complex structure of the original data, and then transforming it into a low-dimensional space, which facilitates visualization, revealing relationships between samples, understanding and monitoring the dynamics of the system.

References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
2. Murray, D., Stankovic, L., & Stankovic, V. (2020, November). Explainable NILM networks. In *Proceedings of the 5th International Workshop on non-intrusive load monitoring* (pp. 64–69).
3. Li, Y., Wang, H., Dang, L. M., Nguyen, T. N., Han, D., Lee, A., & Moon, H. (2020). A deep learning-based hybrid framework for object detection and recognition in autonomous driving. *IEEE Access, 8*, 194228–194239.
4. Lughofer, E., & Sayed-Mouchaweh, M. (Eds.). (2019). *Predictive maintenance in dynamic systems: Advanced methods, decision support tools and real-world applications*. Springer Nature.
5. Christou, I. T., Kefalakis, N., Zalonis, A., & Soldatos, J. (2020, May). Predictive and explainable machine learning for industrial internet of things applications. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)* (pp. 213–218). IEEE.
6. Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., & Zatloukal, K. (2017). Towards the augmented pathologist: Challenges of explainable-AI in digital pathology. *arXiv preprint arXiv:1712.06657*.
7. Sayed-Mouchaweh, M. (Ed.). (2020). *Artificial intelligence techniques for a scalable energy transition: Advanced methods, digital technologies, decision support tools, and applications*. Springer Nature.
8. Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity, 2021*.
9. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371*.
10. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.
11. Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics, 9*(3), 1350–1371.
12. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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* (pp. 1135–1144).