Towards knowledge-based explainability for deep neural networks

Rim El Cheikh, Issam Falih, Engelbert Mephu Nguifo

Towards knowledge-based explainability for deep neural networks

Rim El Cheikh*, Issam Falih*
Engelbert Mephu Nguifo*

*Université Clermont Auvergne, Clermont Auvergne INP, ENSMSE, CNRS, LIMOS,
63000 Clermont–Ferrand, France
rim.el_cheikh@doctorant.uca.fr
{issam.falih,engelbert.mephu_nguifo}@uca.fr

Résumé. As machine learning models gain traction in real world applications, user demand for transparent results grows. The field of explainability (XAI) is meeting this challenge with remarkable speed and efficiency. Notable examples include SHAP and LIME, which are feature-based XAI methods. In this work we aim to review a distinct category of XAI approaches, whose support for providing explanations is interpretable explanatory elements representing user knowledge, instead of raw input features. We categorize these methods based on the stage at which the knowledge is integrated to the XAI pipeline. Furthermore, we highlight the literature around the assessment of XAI methods. We emphasize the importance of the metric of faithfulness of knowledge-based explanations, not only to the real world but also to the underlying model.

1 Introduction

AI-based systems are increasingly reaching many aspects of everyday life. Machine learning (ML) models, mainly deep learning models which are called black-boxes given their high complexity, are being used with impressive results in applications of varying degrees of stakes. Examples span from recommendation systems in entertainment services to credit scoring in banking and diagnostic assistance in medicine. In order to reap the benefits of the advancement of ML in applications that touch users, especially where outputs obtained by these black-box systems intervene in critical scenarios, it is imperative that the user or the person affected by the decision are able to clearly understand why the outcome is as it is (Payrovnaziri et al., 2020). Hence, explainability is continuously evolving into an indispensable facet of AI research, further driven by ethical considerations, user-friendly interfaces and regulatory compliance. Also, explainability of black-box models serves as a valuable tool for computer scientists working on improving ML models and debugging, particularly concerning bias mitigation.

XAI approaches of various kinds are becoming abundant. The extensive interest in this field makes it so XAI approaches are covering most types and variations of ML models, from convolutional neural networks (Kim et al., 2018; Montavon et al., 2017) to natural language processing (Lei et al., 2016; Bouchacourt et Denoyer, 2019) to graph neural networks (Ying

et al., 2019; Funke et al., 2022), etc. In this work, we divert the attention from the conventional and most common category of XAI methods which are feature-based approaches and shift the focus to a different category : XAI approaches that use various forms of knowledge to generate explanations. Rather than producing explanations from raw input features, some XAI methods, which we call knowledge-based XAI methods, take into consideration interpretable explanatory elements and use them as support for explanations. Feature-based XAI methods, such as SHAP (Lundberg et Lee, 2017) and LIME (Ribeiro et al., 2016), provide explanations that belong to the domain of the input features. In the example of image classification, explanations that result from these methods will be blobs of pixels that were deemed important for the outcome of the model. This means that the explanation itself is unaware and deprived from any semantic meaning and by results, does not guarantee to be interpretability for the user. In contrast, knowledge-based XAI approaches draw on elements that are either considered as interpretable by the user or extracted in a way that ensure certain interpretable properties. These elements will act as potential explanations that conform to the user's realm of intelligibility.

The objective of this work is to shed light on the existing knowledge-based XAI approaches. We propose to categorize these methods based on the step at which knowledge is introduced into the prediction-explanation pipeline (see figure 1), offering a novel perspective on the process of providing meaningful explanations. Additionally, we delve into the literature dealing with the assessment of XAI methods. We classify the reviewed methods based on that literature while highlighting the lack of consideration of the knowledge aspect, which is significant when describing and evaluating knowledge-based explainability. We also put forward possible ways to close that gap by arguing that faithfulness of explanations to both real world observations and the underlying model is crucial for the effective deployment of XAI techniques.

Section 2 sets the context that is of interest for this review along with definitions of some notions that are repeatedly used in this review. In section 3, existing XAI methods that can be categorized as knowledge-based are exposed. Section 4 explores existing ways to assess explainability methods while drawing attention to missing aspects that are relevant for knowledge-based approaches. Section 5 discusses the limitations and concerns that are revealed for knowledge-based XAI while also emphasizing their advantages and the assets they bring to further progress the XAI field. Finally, section 6 recapitulates the review and lay out the opportunities that knowledge-based XAI research provides.

2 Background

This section sets the context surrounding the subject matter of this review. Then, key notions and terminology relevant to discussions in the field of XAI are defined.

2.1 Context

Within the scope of this review, a deep neural network (DNN) is considered as any neural architecture with more than one hidden layer. DNNs also include convolutional architectures, object detection modules, image segmentation modules and any deep architecture for feature extraction, followed by classification neural layers.

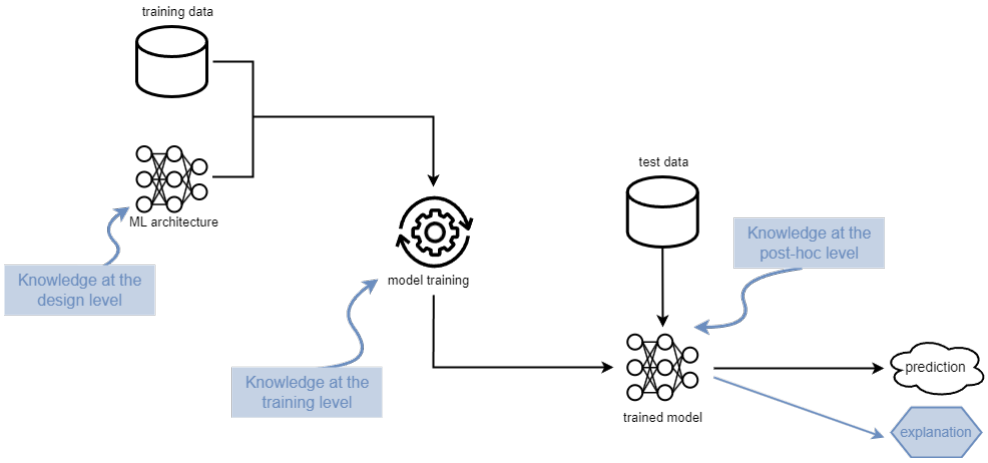

FIG. 1 – *ML pipeline and the different levels of knowledge integration for explanations*

As for knowledge, it is characterized as any information that the user regards as relevant to the learning or application domain, and which could be taken into account to generate explanations. Knowledge can exist at various levels ranging from domain-specific scientific knowledge, such as medical ontologies, to common knowledge, such as visual attributes used for object identification. Knowledge can also exhibit varying levels of meaningfulness and expressiveness in terms of semantics relevance to the user.

We note that the works we review should not be confused with knowledge-informed machine learning Von Rueden et al. (2021). The latter intends to integrate prior knowledge to the predictive system at the training level in order to overcome the challenge of insufficient training data. However, the problem this review tackles is how to utilize interpretable explanatory elements to guarantee the explainability of a DNN. By adopting this approach, explanatory elements can be anticipated and makes it possible to have an assurance of their alignment with the user's vocabulary and expectations of the task at hand.

2.2 Notions related to XAI

The resurgence and fast growth of XAI literature comes with the challenge of reaching a shared and standard terminology. For this reason, it is important to set definitions of XAI-related concepts in order to level the vocabulary of this review.

— **Interpretability.** Interpretability is related to the idea of extracting human-graspable insights from a set of information that accurately reflects the underlying model and its decision process.

— **Explainability.** Explainability is linked to the conversion of the latent information (usually numeric and highly-dimensional) that the DNN is basing its decisions on, into elements that could be digestible by humans. The goal is that these elements would be used to interpret the model and its outcomes.

— **Black-box model.** A black-box model is a predictive system that goes through a complex and non-intelligible process to provide decisions. A black-box, by itself, does not provide the necessary components to guarantee its explainability.
— **Self-explainable model.** A self-explainable model is one that anticipates, inherently or by design, elements that could be used to explain inferences. This notion can be associated with **transparent models**, whose explainability is ensured without an additional module or algorithm.

3 Existing Methods For Knowledge-Based XAI

In this section we present existing methods that use explanatory elements to generate explanations. The state-of-the-art approaches are discussed based on which step of the explanation-prediction generation pipeline the knowledge could be introduced to be accounted for by the explainability algorithm. The three steps are : (1) after training the model and being only able to use it as a blackbox, (2) when training the model and (3) when designing the architecture of the model. We propose this as a first level of categorizing knowledge-based XAI methods.

3.1 Using the model as a black-box

In this category of knowledge-based approaches, only access to an already trained predictive model is assumed. These methods could also be called post-hoc approaches, i.e., can only utilize the pre-trained model as an oracle in order to explain how it comes to a certain result or its overall functioning.

Testing with Concept Vectors (TCAV) Kim et al. (2018) is a concept-based explainability method that exploits the internal high-dimensional state of a convolutional neural network to produce representations of concepts, pre-defined by the user, at the level of some layer l. These representations are the Concept Activation Vectors (CAV). They help explain the internal state at the level of a specific layer. The explanation is a computed percentage reflecting the number of inputs, classified as class k, and whose activation vector at layer l was positively influenced by the concept C. Many other concept-based methods approaches were inspired by TCAV, specifically CAVs as a way to represent human-understandable concepts. The Conceptual and Counterfactual Explanations via Fault-Lines (CoCoX) Akula et al. (2020) approach aims to extract concepts, using TCAV, and provide them as counterfactual explanations that allow the prediction to an alternative desired outcome. Similarly, conceptual Counterfactual Explanations (CCE) Abid et al. (2022) provides a score reflecting the effect that the adding/removing of a concept, represented by a CAV to/from the instance has on the probability of correctly classifying it. Completeness Aware Concept-Based Explanations (CACE) Yeh et al. (2020) is a method that provides concept-based explanations that verify the completeness condition as defined by the authors of the paper. This means that the explanations be a set of concepts that are considered to be complete, i.e., having access to them is enough to fully explain the prediction given by the model. Continuing to consider the influence of TCAV, some approaches make use of the sensitivity score introduced by that method. Automatic Concept-based Explanations (ACE) Ghorbani et al. (2019) introduce a module for automatic extraction of visual concepts which are then given as explanations with a score, computed by TCAV.

Outside of the concept-based explanations, some post-hoc knowledge-based XAI methods use ontologies for explainability. TREPAN Reloaded Confalonieri et al. (2021) allows the extraction of decision trees to approximate the decision process of a blackbox. This is done by strengthening the impact of understandable features while extracting the explainable approximation of the model. The understandability of a feature is judged by its connection to more general concepts present in a ontology.

3.2 Training the model

Some XAI approaches propose to take into account knowledge when training the neural network. Contrary to the previous category, this one (as well as the next one) goes into the ad-hoc approach of explainability. This means that a complementary module is anticipate before training the model in order to address its explainability. This should not be confused with transparent models, where the classification algorithm itself provides explanations for its decision, i.e., no additional method should be implanted or thought in order to understand the results of the model.

Explainable Neural-Symbolic Learning (X-NeSyL) Díaz-Rodríguez et al. (2022) is an XAI method that models knowledge in a graph form and insert it in the training loop of the neural model. The authors propose an "XAI-informed training procedure", called SHAP-Backprop, which takes into account, at the level of the loss function, the coherence of feature attribution scores computed from the neural component with the knowledge graph. The predictive model would therefore be able to give a prediction in tandem with its explanation. Explaining model Decisions through Unsupervised Concepts Extraction (EDUCE) Bouchacourt et Denoyer (2019) also aims to provide explanations alongside the prediction. However, the explanatory elements in this case are automatically extracted concepts.

3.3 Designing the neural network

In this section we review methods that propose to integrate prior knowledge at the level of defining the topology of the neural architecture. Some of these works do not explicitly aim to provide explanations. However, they are included given that they are motivated by making the architecture of neural network based on domain knowledge instead of random, which means that such an architecture would contains elements from prior knowledge. These elements could then be utilized, once the model is trained, to provide explanations to its decisions.

Concept Lattice-based Artificial Neural Network (CLANN) Tsopzé et al. (2007) is an approach that propose to generate the topology of the neural network from prior knowledge, the training data in this case, transformed into a concept lattice. The obtained topology would be justifiable and could facilitate the extraction of rules from the trained network. This motivation could be aligned with the aim to provide a degree of intelligibility to the decision process of the neural network. DeepGONet Bourgeais et al. (2021) suggests the conception of a new neural architecture, based on a multi-layer perceptron, whose learning process is constrained by a domain ontology. In this work, the neurons of the network are explainable elements by themselves as they contains domain significance inherited from the ontology. Another approach, OntoClassifierBourguin et al. (2021) makes use of ontologies to provide explanations for the user. The work uses a semantic bottleneck approach : semantic layers are anticipated in the

architecture of the classifier to allow the extraction of semantic features. This makes the extracted features, which leads to a prediction, semantically meaningful to the user and therefore making the classifier explainable.

4 Assessment Methods

This section presents the existing approaches for assessing and evaluating XAI methods. It also highlights ways of consideration of the knowledge aspects in knowledge-based methods when assessing them. We propose to review the methods based on two approaches. In a first part, a descriptive approach in presented, where XAI methods are compared and judged in terms of their properties and characterization. The second part highlights some existing evaluation techniques for the performance of XAI methods.

4.1 Properties and Characterisation

A descriptive comparison in terms of characteristics and properties allows to situate the existing methods with regards to the expectation of the user of the explanations and its intended application. Many works exist in this vein. Some provide formal frameworks for describing explanations (Amgoud, 2023; Marques-Silva et Ignatiev, 2022). These formalization are mostly restricted for feature-based explainability. Other works aim to provide a sort of grid or catalog of properties to describe and contrast explainability techniques (Sokol et Flach, 2020). Table 1 summarizes the previously presented knowledge-based XAI methods categorized by some of the most useful and common properties.

Here's a description of these properties and their symbols as used in table 1 :

— **Step** at which the knowledge is integrated in the XAI process : As presented in our categorization of knowledge-base XAI literature in section 3, knowledge can be taken into account at the post-hoc level **(P)**, the training level **(T)** or the design of the neural architecture level **(D)**.

— **Data type** : Some XAI methods work for a specific type of data, such as images **(IMG)**, text **(TXT)** or tabular data **(TAB)**. They might also be conceived for a domain-specific type of data **(OTH)**.

— **Target to explain** : An explainability method might aim to explain different elements of the machine learning model : a prediction **(PRED)**, a group of predictions **(GRP)** which usually equates to predictions for a specific class, or the whole model **(MOD)**. This categorization matches local, cohort and global explanations respectively.

— **Family of explanations** : In a general way, the definition of an explanation in not unique. Therefore many different family of explanations exist. Abductive explanations **(ABD)** which are the most plausible or likely explanations. Self-explainable surrogate models **(SUR)** can generated to approximate the behavior of the black-box. Counterfactuals **(CNT)** answer the question *What should be changed in the input in order for it to be classified as a class of interest ?*. Many other families of explanations exist.

— **Portability** : XAI approaches that are conceived in a way that is independent from the specifications of the model they are trying to explain are called model-agnostic (ANY). In contrast, model-specific methods utilize the architecture and the implementation characteristics of the blackbox in order to generate explanations. It can be specific to any

deep neural networks **(DNN)**, convolutional neural networks **(CNN)**, multi-layer perceptron **(MLP)**, etc.

— Knowledge **Form** : Knowledge providing explanatory elements that are comprehensible by humans can take different shapes : knowledge graphs **(G)**, ontologies **(O)**, collections of concepts **(C)**, etc.

— Knowledge **Origin** : Knowledge used to supply explanatory elements can be predefined **(PD)**, guaranteeing their relevance to the user's mental model of the system. They can also be automatically extracted **(AE)** by the XAI module itself. This discovery might lessen the interpretability of the explanatory elements based to the degree to which the extraction is user-attended.

	Step	**Data type**	**Target**	**Family**	**Portability**	**Form**	**Origin**
TCAV (Kim et al., 2018)	P	IMG	GRP	ABD	DNN	C	PD
ACE (Ghorbani et al., 2019)	P	IMG	GRP	ABD	DNN	C	AE
CCE (Abid et al., 2022)	P	IMG	PRED	CNT	DNN	C	PD
CACE (Yeh et al., 2020)	P	IMG+TXT	GRP	ABD	DNN	C	AE
CoCoX (Akula et al., 2020)	P	IMG	PRED	CNT	CNN	C	AE
TREPAN Reloaded (Confalonieri et al., 2021)	P	TAB	MOD	SUR	ANY	O	PD
X-NeSyL (Díaz-Rodríguez et al., 2022)	T	IMG	PRED	ABD	DNN	G	PD
EDUCE (Bouchacourt et Denoyer, 2019)	T	IMG+TXT	PRED	ABD	DNN	C	AE
CLANN (Tsopzé et al., 2007)	D	TAB	-	-	MLP	-	-
DeepGONet (Bourgeais et al., 2021)	D	OTH	MOD	ABD	MLP	O	PD
OntoClassifier (Bourguin et al., 2021)	D	IMG	PRED	ABD	DNN	O	PD

TAB. 1 – *Properties and characteristics of knowledge-based XAI methods*

We note that knowledge form and knowledge origin are not commonly found when describing XAI approaches. However, they are crucial for knowledge-based explainability as they help position on which method is more adequate for the intended application by the user of the explanation.

4.2 Evaluation Techniques

One of the challenges of XAI is the lack of standardized objective performance metrics, making it difficult to compare and contrast the quality and the performance of XAI approaches in a formal and systematic way. Nonetheless, there exist some works on quantitative evaluation that allow to inspect XAI methods in terms of performance metrics. Rahnama et al. (2023) propose to classify evaluation methods for additive XAI methods based on the availability or not of ground-truth explanations. When ground-truth is unavailable, robustness measures (Alvarez-Melis et Jaakkola, 2018; Agarwal et al., 2022) can be used to measure the effect that (un)important features (explanations given by the local additive XAI methods) have on the prediction by nullifying them. When ground truth is on hand, additive explanations can be evaluated by using a synthetic dataset or an interpretable model. In this case, the accuracy of the explainability method will be the similarity between the local explanations it provides and the intrinsic explanation or the weights of the model respectively. These metrics can also be classified by their target : evaluating the explainer VS evaluating the explanation. Although

this categorization is suggested by Rathee et al. (2022) in the context of explainable graph neural networks, it can be extended to evaluating XAI in general. Evaluating the explainer itself goes back to checking the faithfulness, the sparsity and the correctness of the XAI algorithm. While evaluating the explanations that are produced means investigating the plausibility of it in comparison to the human rationale. Table 2 summarize evaluation metrics according to the aforementioned taxonomies.

	Ground-Truth	**No Ground-Truth**
Explainer	- Similarity with interpretable model weights - Correctness - Faithfulness	- Robustness
Explanation	- Similarity with intrinsic explanation from synthetic data - Plausibility	- Sparsity

TAB. 2 – *Evaluation methods for XAI*

It is worth noting that, to our knowledge, these evaluation metrics have not been formalized to take into account the user knowledge aspect of knowledge-based XAI methods. To address this issue we propose a definition of **faithfulness** of knowledge-based XAI methods in two levels. The first we call **plausibility** and it evaluates the alignment of the explanations with the real world (according to the user), i.e., prior knowledge that exists on the domain and that is relevant for the user. The second is **fidelity** and it reflects the alignment of the explanations provided with the underlying model. The choice of these definitions go with our argument that explanations provided by an XAI method are expected to be not only attuned with the real world or the system as seen by the user but should also reliably mirror the underlying model that the method is trying to explain.

The analysis of **plausibility** implies a comparison of the XAI output with an established model of the system that the user has in mind. This definition have a likeness to the correctness metrics as defined by Rathee et al. (2022) for graph neural networks. In order for the XAI method to be useful for the user, provided explanations should fall into the vocabulary that the user uses himself when dealing with the domain in question. The availability of explanatory elements that are used as support for the explanations makes it easier to compare XAI explanations with ground truth since they belong to the same domain. Such analysis comes with difficulties, specially in terms of the subjectivity of the mental model that each user or expert has on the domain in question. Hence, although the actual measure of plausibility may be quantitative, it hides a certain subjectivity. Ideally, an evaluation of plausibility should also be careful to take into account the uncertainty or the potential error in the paradigm that the user is basing its expectations on.

As for **fidelity**, its evaluation could be considered as an objective reflection of the relationship between the predictions and its explanations. At this level, no human-defined elements are required in order to compute the metric. The goal of fidelity is to investigate whether the XAI algorithm is producing explanations that uphold the hypothesis representing the black-box classifier. To quantify fidelity, Sokol et Flach (2020) propose to compute a performance

metric between the output scores of the black-box model and the score of the explanations. Fidelity can be linked to what Vilone et Longo (2021) introduced as the explanation completeness methodology for formal comparison of XAI methods. This approach aims to evaluating which XAI method produces explanations that most comprehensively describe the inferential process of the underlying model.

5 Discussion and Challenges

Most knowledge-based post-hoc approaches heavily relies on the concept activation vectors (CAVs) as defined by the TCAV approach (Kim et al., 2018). However, the quality of the CAVs which represent the concept depends on the quality of the linear classifier. It was shown by its authors that despite obtaining CAVs with good accuracies for low level concepts such as colors or texture, that is not true when dealing with higher level concepts such as characteristics relating to people (ethnicity, gender, age range). Therefore, relying on CAVs to represent the elements the user wishes to use to generate explanations can quickly become problematic when the complexity of the concept can't be sufficiently captured by a linear classifier. Nevertheless, the post-hoc approach to knowledge-based explainability is valuable when the incorporation of domain knowledge before model training is not a viable option due to various reasons. It also offers significant advantages in scenarios where users are already employing a black-box model and now have a need to gain insights into its decision-making process. Post-hoc explainability methods provide users with an explainability tool without the need to modify the model's architecture or retrain it, hence harming its predictive performance.

As for the origin of the knowledge, approaches resting on automatic extraction present some limitations as to whether the explanatory elements reflects efficiently the user's needs. Also, some extracted concepts can be very generic, attributed to a confounding variable or not easily discernible. This reproach could also be true for pre-defined knowledge. Should the XAI algorithm trust the provided elements to generate explanations ? Or is it its responsibility to verify an existing relation between explanatory elements and input data ?

However, despite having some limitations and despite being less discussed as a separate category in the field of XAI, we believe that explainability relying on elements, that are either pre-defined or automatically extracted in a way to be interpretable by the user, presents advantages compared to feature-based approaches. First and foremost, knowledge-based explanations offers a more straightforward comparison with the ground-truth. When defining defining ground-truth and when interpreting explanations, leveraging established knowledge allows for a degree of context and reliability that raw features alone often lack. Furthermore, as a consequence of utilizing a shared vocabulary, it is reasonable that knowledge-based explanations may demonstrate a higher consistency with the user's expectations and might be more likely to resonate with them. As a result, we suspect that this would minimize the potential for confirmation bias and ambiguity in interpretation of the predictions and their explanations.

A concern that is also raised by this review is the question of whether ad-hoc methods are more faithfulness to the real world and to the model compared to post-hoc methods. To answer this question, plausibility and fidelity as defined in section 4 should be more thoroughly investigated. First, we assess plausibility by comparing explanations with the ground-truth. This evaluation method allows to gauge the accuracy and coherence of explanations in relation to the mental model that the user might have on the system. Additionally, the added need to test

the linkage of test data with the knowledge used in post-hoc methods, which seems realistic and cautious in real-life applications, further underscores the complexity of assessing fidelity in a post-hoc context. Second, measuring fidelity involves computing metrics that quantify the alignment between the explanation and prediction. It's worth noting that the nature of post-hoc explanations reduces the interconnectedness of the process of producing a prediction and its explanation. This could hamper fidelity as the detachment from the initial prediction process may lead to explanations that lack cohesion with the model's original intent.

6 Conclusion and Opportunities

In conclusion, this review has delved into knowledge-based explainability methods for deep neural networks. In order to generate explanations for the predictions produced by a black-box classifier, this category of XAI utilizes interpretable elements that are judged to be relevant for the user of the explanations. We identify three stages at which the knowledge could be taken into consideration by the XAI algorithm and we categorize the methods based on this criteria. The three levels are : while designing the neural architecture, while training the model, and after training it. This classification provides a structured framework for understanding how knowledge is leveraged to reach interpretable ML-based systems. It also highlights the versatility of knowledge-based XAI in addressing various scenarios and needs by the user.

In addition to categorizing knowledge-based XAI approaches, this work tackled the existing methods for assessing explainability methods. The review has been conducted at two distinct tiers. The first concerns descriptive comparison approaches which include formalization of explanations and cataloguing of properties and characterizations to compare and contrast XAI methods. The second tier pertains to quantitative evaluation techniques which involve the application of metrics and statistical measures to evaluate the quality and performance of XAI methods. Exploring assessment techniques for XAI helped identify gaps, particularly in the case of knowledge-based XAI. Current evaluation methods may need refinement to account for the specificities presented by knowledge-based XAI.

Essentially, anticipating explanatory elements and basing the explanation on them is a way of involving the user. It shows a certain degree of consideration not only of their mental model of the realm that model is approximating but also of the vocabulary that they would normally use to provide explanations. This contributes to the overall effectiveness and user-friendliness of the explanation process.

The challenges that we present in section 5 provide an opportunity for a deeper experimental investigation to answer two questions. First, does knowledge, represented as explanatory elements, provide better explanations than raw features ? Second, is anticipating the explanatory elements at the level of designing or training the model a way to guarantee faithfulness of the explanations not only to the real world, but also to the underlying black-box ?

Références

Abid, A., M. Yuksekgonul, et J. Zou (2022). Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, pp. 66–88. PMLR.

Agarwal, C., S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, et H. Lakkaraju (2022). Openxai : Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems 35*, 15784–15799.

Akula, A., S. Wang, et S.-C. Zhu (2020). Cocox : Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 34, pp. 2594–2601.

Alvarez-Melis, D. et T. S. Jaakkola (2018). On the robustness of interpretability methods. *arXiv preprint arXiv :1806.08049*.

Amgoud, L. (2023). Explaining black-box classifiers : Properties and functions. *International Journal of Approximate Reasoning 155*, 40–65.

Bouchacourt, D. et L. Denoyer (2019). Educe : Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv :1905.11852*.

Bourgeais, V., F. Zehraoui, M. Ben Hamdoune, et B. Hanczar (2021). Deep gonet : self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. *BMC bioinformatics 22*(10), 1–25.

Bourguin, G., A. Lewandowski, M. Bouneffa, et A. Ahmad (2021). Towards ontologically explainable classifiers. In *Artificial Neural Networks and Machine Learning–ICANN*, pp. 472–484. Springer.

Confalonieri, R., T. Weyde, T. R. Besold, et F. M. del Prado Martín (2021). Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence 296*, 103471.

Díaz-Rodríguez, N., A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, et F. Herrera (2022). Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs : The monumai cultural heritage use case. *Information Fusion 79*, 58–83.

Funke, T., M. Khosla, M. Rathee, et A. Anand (2022). Z orro : Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*.

Ghorbani, A., J. Wexler, J. Y. Zou, et B. Kim (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems 32*.

Kim, B., M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. (2018). Interpretability beyond feature attribution : Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR.

Lei, T., R. Barzilay, et T. Jaakkola (2016). Rationalizing neural predictions. *arXiv preprint arXiv :1606.04155*.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*.

Marques-Silva, J. et A. Ignatiev (2022). Delivering trustworthy ai through formal xai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 36, pp. 12342–12350.

Montavon, G., S. Lapuschkin, A. Binder, W. Samek, et K.-R. Müller (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition 65*, 211–222.

Payrovnaziri, S. N., Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, et Z. He (2020). Explainable artificial intelligence models using real-world electronic health record data : a systematic scoping review. *Journal of the American Medical Informatics Association 27*(7), 1173–1185.

Rahnama, A. H. A., J. Bütepage, P. Geurts, et H. Boström (2023). Can local explanation techniques explain linear additive models ? *Data Mining and Knowledge Discovery*, 1–44.

Rathee, M., T. Funke, A. Anand, et M. Khosla (2022). Bagel : A benchmark for assessing graph neural network explanations. *URL : https ://arxiv. org/abs/2206 13983*.

Ribeiro, M. T., S. Singh, et C. Guestrin (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.

Sokol, K. et P. Flach (2020). Explainability fact sheets : a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 56–67.

Tsopzé, N., E. M. Nguifo, et G. Tindo (2007). Clann : Concept lattice-based artificial neural network for supervised classification. In *CLA*, Volume 331. Citeseer.

Vilone, G. et L. Longo (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion 76*, 89–106.

Von Rueden, L., S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al. (2021). Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering 35*(1), 614–633.

Yeh, C.-K., B. Kim, S. Arik, C.-L. Li, T. Pfister, et P. Ravikumar (2020). On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems 33*, 20554–20565.

Ying, Z., D. Bourgeois, J. You, M. Zitnik, et J. Leskovec (2019). Gnnexplainer : Generating explanations for graph neural networks. *Advances in neural information processing systems 32*.

Summary

Au fur et à mesure que les modèles d'apprentissage automatique gagnent du terrain dans les applications réelles, la demande des utilisateurs pour des résultats transparents augmente. Le domaine de l'explicabilité (XAI) relève ce défi avec une rapidité et une efficacité remarquables. Parmi les exemples notables, on peut citer SHAP et LIME, qui sont des méthodes XAI basées sur les caractéristiques. Dans ce travail, nous visons à examiner une catégorie distincte d'approches XAI, dont le support pour fournir des explications est constitué d'éléments explicatifs interprétables représentant la connaissance de l'utilisateur, au lieu de caractéristiques d'entrée brutes. Nous classons ces méthodes en fonction de l'étape à laquelle la connaissance est intégrée au pipeline XAI. En outre, nous mettons en lumière la littérature relative à l'évaluation des méthodes XAI. Nous soulignons l'importance de la métrique d'exactitude des explications basées sur la connaissance, non seulement par rapport au monde réel mais aussi au modèle sous-jacent.