Towards knowledge-based explainability for deep neural networks

Rim El Cheikh, Issam Falih, Engelbert Mephu Nguifo

Towards knowledge-based explainability for deep neural networks

Résumé. As deep learning (DL) models gain traction in real world applications, there is a growing demand for transparency in their results. The field of explainable AI (XAI) has swiftly responded to this challenge with notable advancements, such as feature-based methods like SHAP and LIME. However, a distinct category of XAI approaches is emerging, which, instead of raw input features, leverages explicit knowledge representations to produce explanations. By incorporating domain-specific knowledge either before, during or after training the model, these methods aim to provide interpretable insights into specific outcomes or the overall functioning of the explained model. This paper reviews these approaches from the perspective of the level at which the knowledge is accounted for into the DL/XAI pipeline, comparing methods and discussing the accompanying challenges and opportunities in enhancing model interpretability.

1 Introduction

Artificial intelligence (AI) based systems, particularly deep learning (DL) models, are increasingly infiltrating aspects of everyday life, delivering impressive results across various applications with varying degrees of stakes. For DL systems to be practically useful, especially in critical scenarios, individuals affected by their decisions must understand why a particular outcome occurred (Hasan et al., 2024; Payrovnaziri et al., 2020). These DL systems are usually described as blackbox models, which contrary to transparent models have a complex and opaque learning and decision process that hinders its comprehensibility. The need for an explanation for the blackbox results has made explainable AI (XAI) a crucial area of research, driven by ethical concerns, user needs, and regulatory requirements (Ali et al., 2023; Saeed et Omlin, 2023). Essentially, XAI methods aim to translate highly-dimensional numeric information that guides the decisions of DL models into human-understandable elements, allowing to extract meaningful insights and interpretations.

Due to growing interest in this field, numerous XAI approaches have emerged (Dwivedi et al., 2023; Hassija et al., 2024), covering various types of DL models, including convolutional neural networks (Kim et al., 2018; Montavon et al., 2017), natural language processing (Bouchacourt et Denoyer, 2019; Lei et al., 2016), graph neural networks (Funke et al., 2022; Ying et al., 2019), and more. The most well-known examples, such as SHAP (Lundberg et Lee,

2017) and LIME (Ribeiro et al., 2016), fall under the category of feature-based explanations, which relate the output to the domain of the input features. In image classification, for example, these methods would highlight pixel blobs as explanatory to the outcome of the model. However, such explanations often lack semantic meaning and may not be fully useful to the user. In this work, we shift away from feature-based explanations and focus on XAI methods that use knowledge in some form to generate explanations, a category we refer to as knowledge-based XAI (K-XAI). Instead of relying on raw input features, K-XAI approaches consider interpretable elements from domain knowledge to support their explanations. This strives to achieve alignment of explanations with the user's realm of intelligibility.

This work sheds light on the landscape of existing knowledge-based XAI methods by proposing a categorization scheme based on the stage at which knowledge is introduced into the DL/XAI pipeline (Fig. 1). We adopt a broad definition of knowledge, encompassing any structured information relevant to the application domain, such as concepts, ontologies, and knowledge graphs. It can range from domain-specific scientific knowledge like medical ontologies, to common knowledge like visual attributes used for object identification. The knowledge can also exhibit varying levels of expressiveness in terms of semantic meaningfulness to the user. Moreover, the focus is on explainability methods for deep neural networks (DNNs), encompassing architectures with multiple hidden layers, such as convolutional architectures, object detection and image segmentation modules. The objective is to provide an overview of existing knowledge-based XAI methods, identifying trends, gaps, and challenges within the field.

It's important to distinguish these methods from knowledge-informed machine learning (Von Rueden et al., 2021), which integrates prior knowledge into the training process to address data insufficiency. Instead, this paper focuses on utilizing interpretable elements from domain knowledge to enhance the explainability of DNNs.

The rest of the paper is structured as follows. Section 2 examines existing work on taxonomy for XAI methods, alongside the presentation of a new framework for K-XAI. Section 3 provides a detailed discussion of existing K-XAI approaches according to our framework. Section 4 focuses on the comparison of the methods, highlighting their strengths and weaknesses. Finally, section 5 discusses the challenges and opportunities in the realm of K-XAI.

2 Categorization of K-XAI methods

The field of XAI research is relatively new and rapidly expanding. Many efforts have been made in order to organize the discussion around XAI, focusing on definitions, standards and taxonomy, which is essential in order to describe and contrast existing methods. In this section, an overview of existing taxonomies is first presented, followed by a new categorization approach, specifically designed to address the knowledge component within K-XAI methods.

2.1 Overview of existing taxonomies

A descriptive comparison of characteristics and properties helps situate existing methods with respect to the user expectations and the intended application. Some works in this vein offer formal frameworks for describing explanations (Amgoud, 2023; Marques-Silva et Ignatiev, 2022), but are often tailored for feature-based explainability. Others aim to catalog properties that can be used to qualify explainability techniques (Sokol et Flach, 2020).

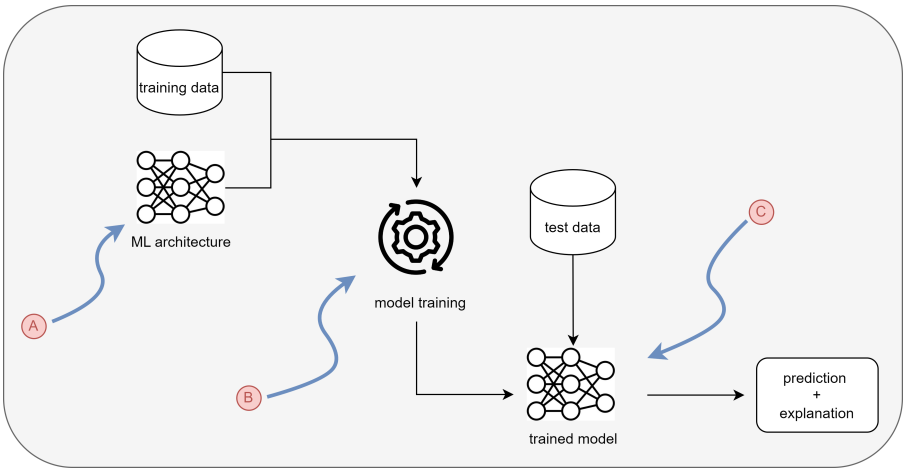

FIG. 1 – *DL/XAI pipeline and the different levels of knowledge integration for explanations : (A) Design level, (B) Training level, and (C) Post-hoc level.*

In terms of surveys, they are plentiful in the XAI field. Some works like Guidotti et al. (2018), cover a wide range of XAI methods by providing a classification based on the problem at hand (model explanation, outcome explanation, model inspection or transparent box design problem), the type of explanator (decision rules, saliency maps, activation maximization, etc.), the type of explained blackbox model (deep neural networks, tree ensemble, support vector machine, non-linear models), and the type of data used as input by the blackbox model (tabular, image, text). The work by Arrieta et al. (2020) establishes a primary distinction between models that are inherently transparent and those that require post-hoc explanations, which is followed by the common categories of model-agnostic and model-specific methods. Similar to Guidotti et al. (2018), model-specific methods can then be classified depending on the ML models they can explain. Arrieta et al. (2020) further classifies methods based on whether they explain the model's processing (execution trace) or its internal data representation (data structure). The paper acknowledges that some methods can fit into different categories depending on their application. Another review work by Samek et al. (2021) starts by distinguishing XAI techniques applicable to DNNs and to models beyond deep networks (clustering and anomaly detection). It then offers a more focused review of specialized methods tailored for specific neural network architectures (graph neural networks and recurrent neural networks). Finally, Vilone et Longo (2021) focuses on the XAI method's output formats (numerical, rules, textual, visual or mixed). It additionally discusses common taxonomies for explanations such as scope (global/local), stage (ante hoc/post hoc) and the problem type (classification/regression).

While these surveys are vital for the organisation and advancement of XAI research, they typically either omit or place less emphasis on the knowledge aspect of explainability methods that include it. It is worth noting that, while efforts have been made to explore the use of knowledge graphs in XAI (Rajabi et Etminani, 2024; Tiddi et Schlobach, 2022), this paper broadens the definition of knowledge to include other representations, such as ontologies and

concepts, both widely used to provide explainability.

2.2 Knowledge-based categorization scheme

This paper advocates for more targeted reviews that specifically address knowledge-driven explainability approaches. In particular, this section discusses existing K-XAI methods in terms of the level at which the knowledge elements are taken into account in the DL/XAI process. Additionally, the form and origin of the knowledge are emphasized in the context of this review, which helps contrast and characterize the methods with respect to the knowledge aspect. Following is a description of the properties that are used to categorize K-XAI methods in Table 1 :

- **Data** type : Some methods work for specific types of data like images **(I)**, text **(T)**, tabular data **(Ta)** or other domain-specific types **(O)**.
- **Target** to explain : A method might aim to explain a single prediction **(P)**, a group of predictions **(G)** like a specific class, or the whole model **(M)**. This criterion matches local, cohort and global explanations respectively.
- Explanation Output **Format** : While all reviewed methods have knowledge elements as support for the explanations, their output can be presented in different formats, namely visual **(V)**, scores **(S)**, rules **(R)**, same as features **(F)** and ontological **(O)**.
- **Level** at which knowledge is introduced in the DL/XAI pipeline : post-hoc level **(P)**, during training **(T)** or during the design of the architecture **(D)**.
- Knowledge **Form** : Knowledge providing explanatory elements for the method can be knowledge graphs **(G)**, ontologies **(O)**, libraries of concepts **(C)**, etc.
- Knowledge **Origin** : Knowledge elements can be pre-defined by the user or the developer of the XAI module **(PD)**, or are automatically extracted by the XAI method **(AE)**.

3 Detailed discussion of K-XAI methods

In this section, existing methods that use knowledge to generate explanations for the neural network outcome are presented. These approaches are categorized in the following subsections based on the level of the DL/XAI pipeline in which knowledge is incorporated by the explainability algorithm. The three levels are : (A) during the design of the model architecture, (B) throughout the model training procedure, or (C) at the post-hoc stage, after training model. This categorization is proposed as the primary classification level for knowledge-based XAI methods, which can be expanded on based on the form and the origin of the used knowledge.

3.1 Design level

Prior knowledge can be integrated directly during the design of the neural topology of the model. By having representations of relevant knowledge embedded within the model architecture and influencing its behavior and decision-making process, the same knowledge elements can be leveraged to provide explanations for the model decisions, enhancing its transparency, and consequently its interpretability.

Methods	Data	Target	Format	Level	Form	Origin
CBM (Koh et al., 2020)	I	P	S	D	C	PD
Deep GONet (Bourgeais et al., 2021)	O	M	S	D	O	PD
SENN (Alvarez Melis et Jaakkola, 2018)	A	P	S	D	C	AE
X-NeSyL (Díaz-Rodríguez et al., 2022)	I	P	V	T	KG	PD
EDUCE (Bouchacourt et Denoyer, 2019)	I+T	P	F	T	C	AE
EDNN-EKG (Daniels et al., 2020)	I	P	F+S	T	KG	PD
TCAV (Kim et al., 2018)	I	G	S	P	C	PD
ACE (Ghorbani et al., 2019)	I	G	F+S	P	C	AE
CCE (Abid et al., 2022)	I	P	S	P	C	PD
CACE (Yeh et al., 2020)	I+T	G	F	P	C	AE
CoCoX (Akula et al., 2020)	I	P	S	P	C	AE
TREPAN Reloaded (Confalonieri et al., 2021)	Ta	M	R	P	O	PD
RevelioNN (Agafonov et Ponomarev, 2023)	I	P	O+V	P	O	PD
Doctor XAI (Panigutti et al., 2020)	O	P	R	P	O	PD

TAB. 1 – *K-XAI methods classified according to the data type (I : images ; T : text ; Ta : tabular ; O : other ; A : any), scope (P : prediction ; G : group ; M : model), explanation format (V : visual ; S : scores ; R : rules ; F : same as features ; O : ontological), level of knowledge integration (P : post-hoc ; T : training time ; D : design), form of knowledge (C : concepts ; O : ontology ; G : Knowledge graphs), origin of knowledge (PD : pre-defined ; AE : automatically extracted).*

Concept Bottleneck Models (CBM) (Koh et al., 2020) are neural network architectures that incorporate an intermediate layer whose neurons represent a set of semantic concepts relevant for the prediction task. By providing concept annotations for the data, which represent the knowledge in this case, the model learns to predict this set of concepts. This first prediction is then propagated to the final task layer to compute the final output. The scores computed in the intermediate concept layer can be considered as inherent concept-based explanations for the final classification outcome.

Similarly, Self-Explaining Neural Networks (SENN) (Alvarez Melis et Jaakkola, 2018) is a class of neural architectures that inherently provide local interpretations for their results. They are designed to be linear in automatically extracted concepts rather than the raw features, allowing the explanations to be importance scores that are associated to these concepts and making of them self-explainable models. High-level concepts are extracted during training with a concept encoder component integrated into the neural design.

Deep GONet (Bourgeais et al., 2021) is another example of this category. It proposes the conception of a new neural architecture, based on a multi-layer perceptron where the learning process is guided by a domain-specific ontology. In that work, the neurons of the network are explainable elements by themselves as they contain domain significance inherited from the ontology. By constraining the learning process with the ontology, Deep GONet ensures that its decision process is grounded in domain-specific knowledge.

3.2 Training level

This category of approaches incorporates knowledge during the training of the neural network. Like the previous category, these approaches fall into the ad-hoc paradigm of explainability methods, where an algorithmic component is anticipated before training the model in order to address its explainability.

One such approach is Explainable Neural-Symbolic Learning (X-NeSyL) (Díaz-Rodríguez et al., 2022), where knowledge is conveyed in the form of a graph and inserts it in the training loop of the neural model. X-NeSyL introduces a novel "XAI-informed training procedure" that incorporates the coherence of feature attribution scores from the neural component with the knowledge graph at the loss function level. This ensures that the predictive model can provide an explanation alongside each prediction.

Another approach that couples explanations along with predictions is Explaining model Decisions through Unsupervised Concepts Extraction (EDUCE) (Bouchacourt et Denoyer, 2019). This method extracts the explanatory elements from the training data using unsupervised learning, which involves identifying relevant concepts without relying on concept-labeled examples. The process of concept extraction is integrated within the training procedure for classification, enabling the model to simultaneously learn both the underlying concepts in the data and the task-specific classification.

Finally, Explainable Deep Neural Models Using External Knowledge Graphs (EDNN-EKG) (Daniels et al., 2020) aligns concepts from a knowledge graph with annotations from the input dataset to create a hierarchy that represents subclass-superclass relationships. This expanded set of object probabilities serves as training input for a self-explainable model, enabling classification with interpretable, human-understandable explanations.

3.3 Post-hoc level

In this category, methods assume access only to an already trained predictive model, treating it as an oracle to explain outcomes or overall functioning.

Testing with Concept Activation Vectors (TCAV) (Kim et al., 2018) is a concept-based explainability method that evaluates how user-defined concepts influence the predictions of a trained neural network. TCAV represents these concepts as vectors in the model's latent space and then measures their impact on a target class by calculating directional derivatives. The given explanation is a percentage, indicating how often the concept positively influences the model's prediction for the target class. TCAV has inspired many other concept-based methods, particularly through its use of Concept Activation Vectors (CAVs) as a way to represent human-understandable concepts for explanation generation.

Conceptual Counterfactual Explanations (CCE) (Abid et al., 2022) provides a score that reflects the effect of adding or removing a concept, represented by a CAV, on the probability of correctly classifying an instance. This approach falls under the category of counterfactual explanations, which explore "what if" scenarios, offering rationales for the model's behavior.

Another notable approach is Conceptual and Counterfactual Explanations via Fault-Lines (CoCoX) (Akula et al., 2020), which uses TCAV to extract class-specific concepts from the training data. Then, CoCoX generates fault-line explanations that specify the minimal set of concepts needed to alter the classification of an instance.

Completeness Aware Concept-Based Explanations (CACE) (Yeh et al., 2020) introduces a completeness score, which reflects whether a set of explanatory concepts is sufficient to fully explain a model's predictions. CACE automatically extracts concepts from the training data and guarantees that they meet the completeness condition. Then, ConceptSHAP is used to quantify the importance of each concept in contributing to the completeness score.

The final approach inspired by TCAV is Automatic Concept-based Explanations (ACE) (Ghorbani et al., 2019). ACE introduces a module for automatic extraction of visual concepts from the data. It then adopts the sensitivity score defined in TCAV to quantify the importance of the extracted concepts to the model's behavior.

Beyond concept-based explanations, TREPAN Reloaded (Confalonieri et al., 2021) utilizes ontologies. This method extracts decision trees that approximate the decision process of a neural model. This is done by strengthening the impact of understandable features when extracting the explainable approximation of the model. The understandability of a feature is judged by its connection to more general concepts in an ontology.

Retrospective Extraction of Visual and Logical Insights for Ontology-based Interpretation of Neural Networks (RevelioNN) (Agafonov et Ponomarev, 2023) also uses ontologies by mapping the network's intermediate outputs to a set of concepts defined in it. This set coincides with the class labels. The concept representations, together with the ontology, are then passed to a reasoning module to generate logical explanations grounded in the defined concepts.

Finally, still involving ontologies, Doctor XAI (Panigutti et al., 2020) is notable for incorporating temporal encoding to account for the sequential nature of the data. It identifies similar neighbors of the instance needing explanation and creates a synthetic neighborhood by using domain knowledge from an ontology to ensure the generation of meaningful instances through perturbation. An interpretable model is trained on this synthetic neighborhood to mimic the original blackbox behavior. From this model, symbolic rules are extracted, providing human-understandable explanations for the blackbox predictions.

4 Comparison of K-XAI methods

This section compares the methods presented in the previous section, highlighting the distinctions across three dimensions : the level of knowledge integration, the origin of the knowledge, and the form it takes within the models.

At the design level, methods like CBMs and Deep GONet provide built-in interpretability by embedding knowledge elements directly within the network, influencing model outputs. These models are especially suitable for users who need explanations grounded in a set with user-defined semantic expressiveness. This is particularly true for Deep GONet, where embedded specialized ontologies ensure that the model decisions are both contextually meaningful and understandable. However, these models require access to a pre-defined set of concepts, which can be limiting in cases where knowledge is not available or its quality not guaranteed. To overcome this limitation, SENN offers to discover concepts automatically. This method yields flexibility by capturing knowledge from the data, allowing users to engage with and refine emerging concepts. CBM and Deep GONet provide a limited degree of human interaction, as users have the opportunity to engage with the knowledge elements, but only before the training of the model. Deep GONet is more rigid due to their use of pre-defined ontologies This rigidness, while it might over-constrain and limit the model's ability to generalize, is ad-

vantageous in critical application, granting the user control over the quality and usefulness of knowledge elements.

Training-level methods, such as X-NeSyL, EDNN-EKG and EDUCE, ensure that explanations are produced alongside predictions, enhancing transparency of the model. The trade-off lies in increased computational overhead and complexity during training. Nonetheless, the advantage is a model that retains robustness while offering predictions and their explanations rooted in knowledge. By tightly coupling the knowledge with the learning objectives, X-NeSyL ensures coherence between feature scores and predefined knowledge, allowing the model to produce semantically meaningful explanations that align closely with known relationships. EDNN-EKG also uses KGs to infuse the learning with the prior knowledge. However, it does so by splitting the overall model into two : the first part is a task-specific blackbox whose predictions are training inputs to the second part, a self-explainable model. This approach may lead to concerns similar to post-hoc methods, where the module responsible for the explainability is a separate entity from the blackbox. Concerning EDUCE, the unsupervised learning of concepts also offers a flexible but less specialized form of semantic meaning.

Post-hoc methods are crucial for explaining blackbox systems already in use, offering explainability to trained models without requiring modifications. TCAV, CCE, CoCoX, ACE and CACE are particularly effective for capturing conceptual meaning by representing concepts within the latent space of the network. These methods offer explainability by associating concepts with model behavior, allowing for explanations that range from specific concept influences to counterfactual scenarios. However, contrary to methods using pre-defined knowledge, ACE and CACE's automatic extraction of visual concepts calls into question the meaningfulness of concepts. On the other side, ontology-based approaches like TREPAN Reloaded, RevelioNN and Doctor XAI provide structured explanations, enhancing transparency of the model's overall decision process. The use of structured knowledge in these methods ensures the semantic meaningfulness of the insights they deliver. TCAV and related approaches allow users to experiment with different sets of concepts and explore their value as explanations, which facilitates user intervention in the generation of explanations. However, the quality of the CAVs, representing concepts in most of these methods, depends on the quality of the linear classifier. Kim et al. (2018) have shown that while CAVs yield good accuracies for low-level concepts like colors or textures, they struggle to accurately represent higher-level concepts such as those relating to people (ethnicity, gender, age range).

Overall, post-hoc methods generally feature shallow knowledge integration, which limits their alignment with the original learning process while allowing for flexible explainability. In contrast, for design or training level methods knowledge is tightly embedded within the model, representing an ad-hoc integration that directly shapes the internal workings of the network and ensures that every decision made by the model is directly influenced by the knowledge.

5 Challenges and opportunities

The progress of K-XAI systems is accompanied by a range of challenges that unveil opportunities to shape the potential of research in this field and its impact on real life applications.

Unlike ad-hoc approaches, post-hoc XAI provides users with an explainability tool without necessitating alterations to the model's architecture or training, thereby averting potential harm to its predictive performance. However, post-hoc methods rely on an external module that re-

presents the knowledge in the internal space of the blackbox model, generating explanations from a separate entity from the trained model. This introduces an additional layer of ambiguity as it becomes necessary to assess whether both the explained model's decision and the interpretation provided by the post-hoc method can be trusted. This is illustrated in the previous section, where it was established that CAVs (Kim et al., 2018) may fall short in capturing complex concepts effectively. In contrast, ad-hoc approaches could ensure closer alignment with the model's prediction process, potentially enhancing the fidelity of explanations to the internal workings of the blackbox. To validate this, a classifier inherently explainable in terms of knowledge elements is needed for direct comparison of post-hoc and ad-hoc fidelity.

Regarding the origin of knowledge in K-XAI methods, those relying on automatic extraction may encounter limitations in efficiently reflecting the user needs. Extracted concepts can sometimes be overly generic, associated with confounding variables, or difficult for human users to discern. While similar concerns may apply to pre-defined knowledge, the user theoretically has control over its quality which is one convenient way to overcome the challenge of subjectivity and usability that are inherent to the explanation process. This question is discussed in Buschmeier et al. (2023), which highlights that the cognitive act of understanding is not static but varies with context and goals, making it a dynamic process requiring the consideration of the user's prior knowledge. Nonetheless, depending on the use case, K-XAI algorithms might need to account for uncertainties or potential biases when generating their explanations, even when they reflect the user's expectations and assumptions on the knowledge.

Another challenging topic in XAI research in general is establishing standardized objective metrics. While evaluation is a less stable aspect for XAI than for other AI fields, some methods allow to inspect XAI approaches in terms of performance and quality of explanations. Current evaluation methods may need refinement to account for the specificities presented by knowledge-based explainability, namely the support of the explanations being knowledge elements that are often extrinsic to the input data. Knowledge-based explanations could offer a more direct comparison with the ground truth, facilitating the validation of K-XAI methods. Existing metrics, such as plausibility – assessing the coherence of explanations with the user's mental model of the task – could be adapted to take into account the knowledge aspect into its computation. As a consequence of the explanation using a shared vocabulary with the domain expert, the evaluation of its coherence with the ground truth defined by these experts might be rendered more straightforward.

Additionally, the use of a shared vocabulary in knowledge-based explanations is likely to demonstrate a higher consistency with the user expectations, reducing the potential for confirmation bias and ambiguity in interpretation of the predictions and their explanations. However, this assumption requires further exploration of the attitude of users towards knowledge-based explanations versus other categories, like feature-based XAI.

More generally, K-XAI is challenged by issues related to the acquisition, maintenance, scaling and complexity of knowledge base. Fortunately, advances in knowledge representation might enhance knowledge-based explainability.

6 Conclusion

This paper has delved into knowledge-based explainability methods for deep neural networks. Methods in this category utilize human-understandable elements from domain know-

ledge to generate explanations for predictions produced by blackbox classifiers. Integration of the knowledge elements can be done at different levels of the DL/XAI pipeline, which we use as a framework to describe and understand how knowledge is leveraged to reach interpretable DL-based systems. This framework showcases the versatility of knowledge-based XAI in addressing diverse scenarios and needs. Anticipating explanatory elements and basing the explanation on them is a way of involving the user, demonstrating consideration for both the vocabulary they would use and their mental model that the DL system is approximating. This contributes to the overall user-friendliness and effectiveness of the explainability process. Additionally, this paper compared the methods based on different properties with respect to the knowledge aspect such as the form of the knowledge and its origin.

Challenges in the field of K-XAI were also identified and discussed, namely in terms of knowledge quality, post-hoc vs ad-hoc paradigms and evaluation methods for K-XAI. These challenges provide occasions for further research in many aspects, such as the task of accounting for knowledge quality or mistakes in the generation of explanations based on this knowledge, or the advancement of evaluation frameworks explainability, particularly ones that leverage the knowledge elements for a more effective assessment of the explainability method.

Références

Rajabi, E. et K. Etminani (2024). Knowledge-graph-based explainable AI : A systematic review. *Journal of Information Science 50*(4), 1019–1029.

Hassija, V., V. Chamola, A. Mahapatra, et al. (2024). Interpreting black-box models : A review on explainable artificial intelligence. *Cognitive Computation 16*(1), 45–74.

Hasan, M. A., F. Haque, S. R. Sabuj, et al. (2024). An End-to-End Lightweight Multi-Scale CNN for the Classification of Lung and Colon Cancer with XAI Integration. *Technologies 12*(4), 56.

Agafonov, A. et A. Ponomarev (2023). RevelioNN : Retrospective extraction of visual and logical insights for ontology-based interpretation of neural networks. In *Conference of Open Innovations Association (FRUCT)*, pp. 3–9. IEEE.

Buschmeier, H., H. M. Buhl, F. Kern, et al. (2023). Forms of Understanding of XAI-Explanations. *arXiv preprint arXiv :2311.08760*.

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

Ali, S., T. Abuhmed, S. El-Sappagh, K. Muhammad, et al. (2023). Explainable Artificial Intelligence (XAI) : What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion 99*, 101805.

Dwivedi, R., D. Dave, H. Naik, et al. (2023). Explainable AI (XAI) : Core ideas, techniques, and solutions. *ACM Computing Surveys 55*(9), 1–33.

Saeed, W. et C. Omlin (2023). Explainable AI (XAI) : A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems 263*, 110273.

Funke, T., M. Khosla, M. Rathee, et A. Anand (2022). Zorro : Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge and Data Engineering 35*, 8687–8698.

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

Tiddi, I. et S. Schlobach (2022). Knowledge graphs as tools for explainable machine learning : A survey. *Artificial Intelligence 302*, 103627.

Díaz-Rodríguez, N., A. Lamas, J. Sanchez, et al. (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.

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.

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.

Von Rueden, L., S. Mayer, et K. s. o. Beckh (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.

Samek, W., G. Montavon, S. Lapuschkin, C. J. Anders, et K.-R. Müller (2021). Explaining deep neural networks and beyond : A review of methods and applications. *Proceedings of the IEEE 109*(3), 247–278.

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.

Vilone, G. et L. Longo (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction 3*(3), 615–661.

Yeh, C.-K., B. Kim, S. Arik, et al. (2020). On completeness-aware concept-based explanations in deep neural networks. *Advances in NeurIPS 33*, 20554–20565.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, et al. (2020). Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion 58*, 82–115.

Payrovnaziri, S. N., Z. Chen, et al. (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.

Sokol, K. et P. Flach (2020). Explainability fact sheets : A framework for systematic assessment of explainable approaches. In *ACM FaccT*, pp. 56–67.

Panigutti, C., A. Perotti, et D. Pedreschi (2020). Doctor XAI : An ontology-based approach to black-box sequential data classification explanations. In *ACM FaccT*, pp. 629–639.

Koh, P. W., T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, et P. Liang (2020). Concept bottleneck models. In *ICML*, pp. 5338–5348. PMLR.

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.

Daniels, Z. A., L. D. Frank, C. J. Menart, et al. (2020). A framework for explainable deep neural models using external knowledge graphs. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, Volume 11413, pp. 480–499. SPIE.

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

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

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

Alvarez Melis, D. et T. Jaakkola (2018). Towards robust interpretability with self-explaining neural networks. *Advances in NeurIPS 31*.

Kim, B., M. Wattenberg, J. Gilmer, et al. (2018). Interpretability beyond feature attribution : Quantitative testing with concept activation vectors (tcav). In *ICML*, pp. 2668–2677. PMLR.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, et D. Pedreschi (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR) 51*(5), 1–42.

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

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

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.

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

Summary

Les modèles d'apprentissage profond (DL) gagnant en popularité dans les applications réelles, la demande de transparence dans leurs résultats augmente. Le domaine de l'intelligence artificielle explicable (XAI) répond à ce défi avec des avancées significatives, telles que des méthodes basées sur les attributs comme SHAP et LIME. Toutefois, une catégorie distincte d'approches XAI est entrain d'émerger, qui, au lieu de s'appuyer sur des attributs brutes, utilise des représentations de connaissances explicites pour produire des explications. En incorporant des connaissances spécifiques au domaine avant, pendant ou après l'entraînement du modèle, ces méthodes XAI visent à fournir des perspectives interprétables sur des résultats ou sur le fonctionnement global du modèle expliqué. Cet article examine ces approches du point de vue du niveau auquel la connaissance est prise en compte dans le pipeline DL/XAI, en comparant les méthodes et en discutant les défis et les opportunités associés à l'amélioration de l'interprétabilité des modèles.