Information Fusion 71 (2021) 28–37

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI

Andreas Holzinger [a,b,*], Bernd Malle [a], Anna Saranti [a], Bastian Pfeifer [a]

[a] Medical University Graz, Austria
[b] Alberta Machine Intelligence Institute, Canada

ARTICLE INFO

Keywords:
Information fusion
Explainable AI
xAI
Graph Neural Networks
Multi-modal causability
Knowledge graphs
Counterfactuals

ABSTRACT

AI is remarkably successful and outperforms human experts in certain tasks, even in complex domains such as medicine. Humans on the other hand are experts at multi-modal thinking and can embed new inputs almost instantly into a conceptual knowledge space shaped by experience. In many fields the aim is to build systems capable of explaining themselves, engaging in interactive what-if questions. Such questions, called counterfactuals, are becoming important in the rising field of explainable AI (xAI). Our central hypothesis is that using conceptual knowledge as a guiding model of reality will help to train more explainable, more robust and less biased machine learning models, ideally able to learn from fewer data. One important aspect in the medical domain is that various modalities contribute to one single result. Our main question is "How can we construct a *multi-modal* feature representation space (spanning images, text, genomics data) using knowledge bases as an initial connector for the development of novel explanation interface techniques?". In this paper we argue for using Graph Neural Networks as a method-of-choice, enabling information fusion for multi-modal causability (causability – not to confuse with causality – is the measurable extent to which an explanation to a human expert achieves a specified level of causal understanding). The aim of this paper is to motivate the international xAI community to further work into the fields of multi-modal embeddings and interactive explainability, to lay the foundations for effective future human–AI interfaces. We emphasize that Graph Neural Networks play a major role for multi-modal causability, since causal links between features can be defined directly using graph structures.

1. Introduction

Current medical AI is very successful in certain tasks due to the great advances in statistical machine learning. One of the most cited examples is the work of Esteva et al. (2017) [1], where classification of skin cancer via convolutional neural networks achieved a performance on par with human experts, demonstrating that AI is capable of classifying skin cancer with a level of competence comparable to dermatologists. Another success story is the work of De Fauw et al. (2018) [2]: they achieved human expert performance on the classification of optical coherence tomography (OCT) scans to detect choroidal neovascularization (CNV) which is commonly known as age-related macular degeneration (AMD), the major cause of blindness. A very recent work of Faust et al. (2019) [3], on mapping brain tumour histomorphologies shows that deep learning provides a highly dynamic data-driven approach that can help to automate traditionally laborious and qualitatively difficult image-based analyses in pathology. Moreover, they showed that machine-engineered features correlate with salient human-derived morphological constructs. This is an important step in achieving an overlap between human and AI, helping in eliminating bias and improving the accountability for future AI assisted medicine. They also emphasized that the currently best performing methods, apart from requiring a lot of top-quality data, are highly opaque, so even narrow classification tasks lack interpretability as well as user-defined feature selection. This opaqueness (commonly called "Black-Box" behaviour) of statistical machine learning models shown in the best-practice examples mentioned above is inherent in model free stochastic approaches such as deep neural network machine learning [4,5].

The paradigms that underlie these problems fall into the growing field of explainable AI (xAI). Here, methods for the implementation of transparency and *explainability* of such Black-Box methods are developed, motivated often by legal issues [6]. However, in the medical domain we are facing another complex challenge, which lies in the integration, fusion and mapping of various distributed and heterogeneous

* Corresponding author at: Medical University Graz, Austria.
E-mail address: andreas.holzinger@medunigraz.at (A. Holzinger).

https://doi.org/10.1016/j.inffus.2021.01.008
Received 21 October 2020; Received in revised form 20 January 2021; Accepted 22 January 2021
Available online 27 January 2021

data in arbitrarily high dimensional spaces in a multi-modal (MM) manner, i.e. we must always consider that diverse data and different features contribute to a result [7,8]. A good example is cancer research [9], or radiomics where multi-faceted data from diverse sources contribute to a decision [10].

Therefore, Arrieta et al. (2020) [11] emphasize, that both *explainability* and *information fusion* are important, most importantly as *different* users (laymen, physicians, computer scientists, ...) need *different* explanations. Via intra-modal feature extraction and MM embedding, a *low-dimensional representation space* comprising *relevant* data to assist the medical decision making process is necessary. We describe the challenge of constructing such a MM embedding space in Section 2.

In certain domains, especially in the medical field, there is a need for causability, introduced by Holzinger et al. (2019) [12]. *Causability* is not a synonym for causality in the sense of Judea Pearl [13]; the term causa-bil-ity was introduced in reference to usa-bil-ity. Whilst explainability (represented by the field of xAI) is about the technical implementation of transparency and traceability in AI approaches, causability is about measuring and ensuring the *quality of explanations* [14]. That means the measurable extent to which an (xAI) explanation to a (human) user achieves a specified level of *causal understanding*, measured with effectiveness, efficiency and satisfaction in a specified context of use — similar as usability [15]. To promote a better understanding, we summarize the definitions here:

- (A) Explainability := technically highlights decision relevant parts of machine representations and machine models i.e., parts which contributed to model accuracy in training, or to a specific prediction for a particular observation. Here the xAI community has already developed a variety of successful methods. Explainability does *not* refer to a human model.
- (B) Causality := the relationship between cause and effect in the sense of Judea Pearl [13].
- (C) Causability := the measurable extent to which an explanation – resulting from (A) – to a human expert achieves a specified level of causal understanding. This can be measured e.g. with the System Causability Scale [14]. Causability refers to a human model.

Understanding can be ensured when we can map explainability (the "technical explanation") with causability ("the human understanding"). Successful mapping between explainability and causability requires new human–AI interfaces which allow domain experts to interactively ask questions and counterfactual questions to gain deep insight into the underlying *independent* and *disentangled* explanatory factors of a result, adapted to the needs of the respective end user [16].

In an ideal world both human and AI statements would be identical and congruent with the *ground truth,* which is defined for both humans and AI equally [14]. However, in the medical domain we are in the real world, thus we face two problems: (i) ground truth cannot always be well defined, especially when making a medical diagnosis; and (ii) human (scientific) models are often based on causality in the sense of Judea Pearl as ultimate aim for understanding the underlying explanatory mechanisms.

While correlation is accepted as a basis for decisions, it can only be an intermediate step, due to the importance of validity in medicine [17]. Moreover, it is necessary (a) to build human trust, and (b) to build a kind of "AI experience" among the medical professionals, according to (Cabitza, Campagner & Balsano, 2020) [18].

Currently there is a huge debate in the AI community about the avoidance of bias and how to ensure fairness in AI decisions [19]. Bias is a core topic in causality, and causability is a possible measure. Validation of causal effects under determined causal structures is especially needed if and when such effects are estimated in limited settings. In the medical domain a good use case for such a limited setting are randomized controlled trials. Such trials permit to test for causal hypotheses, because a randomization-by-design is guaranteed, even with limited domain knowledge. A particular problem of generalizability has been described by (Bareinboim & Pearl, 2013) [20], which is called *transportability* and can be seen as a "data fusion framework" for the external validation of intervention models and counterfactual queries. Transportability enables to transfer causal effects learned in experimental studies to a new setting, in which only observational studies can be conducted. Transportable models can be integrated into clinical guidelines to augment domain experts with "action-savvy" predictions, in pursuit of better precision medicine [21].

The field of xAI generally has huge potential to contribute towards a better understanding of diseases, which can furthermore lead to more accurate diagnoses, more rational disease prevention strategies, better treatment selection, and the development of novel therapies. Moreover, a better understanding of diseases can contribute to the long term goal of predictive preventive personalized participatory (P4) *precision* medicine which genuinely seeks to redefine the understanding of disease onset and progression, treatment response, and health outcomes through the most precise measurement of molecular, genetic, environmental, and behavioural individual factors that contribute to health and disease. In this case it is imperative that AI decisions are fully re-traceable across all modalities involved, giving the medical domain expert the power to (i) understand, (ii) confirm or (iii) overrule them (see Fig. 1). Whatever future human–AI interfaces may look like, they must enable a medical expert to understand the causal pathways to produce meaningful counterfactuals [22]. Here, the use of graphs and graph representation learning can be beneficial and therefore we describe some possible approaches in Section 3.

In the following sections we propose three core challenges and a series of experiments carried out in succession as part of an overall pipeline to establish *a multi-modal, decentralized, explainable machine learning infrastructure for the medical domain.* Each stage in the pipeline has a different focus and can theoretically work on its own, meaning experiments could be conducted in isolation. However, building upon and mutually extending their individual advancements, our *integrated* approach will yield its maximum benefit.

2. Challenge I: Constructing a multi-modal (MM) embedding space

Learning on fused data from different sources and modalities can substantially outperform traditional methods on just one type of data structure. Jointly learning on input data of different modalities is a standard routine [23,24], with a constant stream of innovation in recent years [25,26].

The fundamental challenge in fusing disparate modalities lies in bridging the *semantic gap* between them and handling potential disagreements thereof — this problem is identified within the literature as *aligning local geometries* of subjects across feature spaces [24]. In earlier works multi-modality was understood to be limited to different image taking techniques (e.g. CT, MRI, PET, etc.) or different resolutions. Simple methods just concatenate feature vectors to fuse different domains, not considering varying distance or neighbourhood metrics across domain boundaries.

Wei et al. (2019) [27] proposed a MM Graph Convolutional Network in order to consider the different modalities contained in videos — visual, acoustic, textual — by constructing a user–item interaction graph within each modality, thus learning different user preference functions which eventually fuse into a unified representation in a specifically designed *combination layer*. Dourado, Tabbone & Torres (2019) [28] deal with multi-source and MM features in the problem domain of information retrieval and rank aggregation, where they use graphs to encapsulate and correlate ranks as well as graph embeddings to reduce these graphs to vectors which are eventually fused into a response. Mai, Hu & Xing (2020) [29] observe a *modality gap* in cross-modal fusion techniques due to a failure to learn joint embedding spaces. They propose an encoder–decoder architecture translating

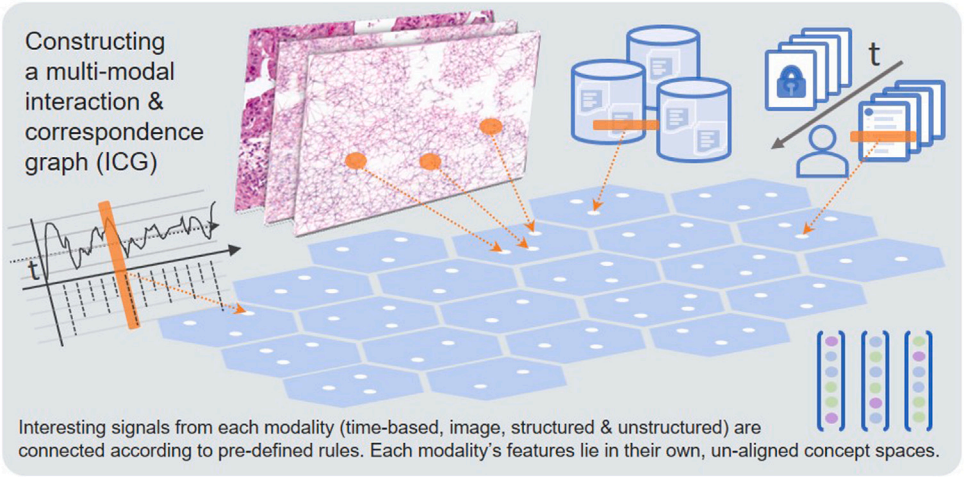

Fig. 1. Graph fusion. Data points from four different input modalities – time-series, histopathological images, knowledge databases as well as patient histories in the form of unstructured text – are mapped into an *interaction & correspondence graph (ICG)*. Combined with their intra-modal geometry (network or similarity structure), we can generate positive/negative samples from this graph (akin to word-coocurrences in a text corpus) and embed them into low-dimensional, dense representations in a modality-aligned concept space.

each source modality into a *modality-invariant embedding space* via adversarial training, including a *Modality Attention Network* since not all modalities contribute equally. In multi-similarity metric fusion for semi-supervised Label Propagation, [30] propose to extend the *Flexible Manifold Embedding* technique to consider both label and feature spaces, constructing a label-based *Correlation Graph* to interact with other similarity graphs. Vivar et al. [26] utilize MM Graph Attention Networks to deal with missing features, stemming from different sources of medical data.

In summary, it is interesting to note that although multi-modality is an important issue across the literature, and data/graph fusion is a method of choice for many, it seems that the concept is understood in different ways by researchers in different areas applying it to a varying range of objectives, from label propagation, to recommendations, to fusing data of different dimensionality, or compensating for missing features. Therefore, it is important to define clear goals, anticipated challenges and potential methods to overcome them, which we contribute to in this work.

It is important to initially capture each source domain's intrinsic *ontology* — e.g. relations, hierarchies, partitions in a graph, analogies, co-occurrence, and other forms of semantics within texts, or pathways on an *omics level. Concurrently, it is necessary to define *cross-links* – interactions and correspondences – between entities of different domains. For instance, a superpixel (which has perceptual meaning) in an image may correspond to an entry in a controlled vocabulary, or a mutation within a gene causes a different behaviour on the protein level. For a lack of pre-defined *cross-domain semantics* most of this work will have to be done either manually or by deriving connection rules from knowledge provided by human experts, which requires intensive inter-disciplinary and cross-domain effort. Sampling this ontology-enriched cross-domain graph produces positive/negative instances w.r.t. a potential decision boundary, which are subsequently fed into an embedding algorithm (Fig. 1):

Two possible approaches to this end are *joint embeddings* [31] and *Graph Representation Learning (GRL)* [32,33].

Generally, *embeddings* are low-dimensional vector representations of entities which are usually learned from large corpora in an unsupervised fashion, i.e. by forcing an algorithm to approximate pair-wise distance in the embedding space to a given similarity function in the original, higher-dimensional input space. There is a wealth of literature on so-called *neural embeddings*, with the most prominent work presented by Bengio et al. (2003) [34].

In order to connect information from images, text & *omics data we need to compute and learn representations for nodes/subgraphs in this low-dimensional space, considering node-level features as well as their structural surroundings (e.g. *graph neighbourhoods*). Within this embedding space, we can subsequently compare items in diverse ways (e.g. hierarchies, analogies, clustering, etc.) across different input modalities, whose intrinsic geometries have been aligned during the joint embedding procedure.

The most important aspect in this phase is generating positive/negative samples within and across modalities as input features – for example, *Word2Vec* [35], as an unsupervised representation learning algorithm, assumes word co-occurrence within sentences as positive samples – this is highly domain specific and dependent on human assumptions, which constitutes a *bias trap* that must be carefully avoided through repeated experimentation with diverse sets of assumptions and automated as well as human control (objective performance measures and inter-expert validation combined).

Recent research has also successfully demonstrated a combination of different methods in a two-step process: (i) pre-processing intra-modal data using traditional assumptions about their respective domains, and (ii) feeding the resulting, normalized feature vectors into a graph representation learner (neighbourhood aggregation, GAN, etc.). It is a challenge to extend this approach to MM data sets. Others utilized traditional machine learning techniques (e.g. Random Forest) to build intra-modal similarity matrices, which are subsequently fused [25]. However, this requires the full matrix to exist at computation time, which is unrealistic for larger graphs or distributed graphs. Depending on the approach, it might be necessary to utilize a pre-instantiated target dictionary of features curated by human experts (e.g. cell types visible in an image).

Frequent problems in machine learning (ML) and statistics concern missing data or data of different feature dimensionality. While the latter can be easily dealt with the application of dimensionality reduction techniques, there are no standard solutions for dealing with non-existent information. However, especially in the field of graph based ML, the structure of a network alone often carries sufficient information in order to obtain respectable results. Our own experiments on small, well-structured graphs showed that – depending on learning task and target attributes – randomly initialized representations can yield results on par with carefully designed and pre-processed feature vectors.

In this phase it is necessary to produce a corpus of MM feature representations, originating in e.g. histopathological images, patient case files (text), *omics data as well as medical knowledge bases, where a *knowledge graph* can be utilized as an initial connector. Such pre-trained concept embeddings would constitute an advanced pendant to

word & document embeddings, inlcuding the well-known *GloVe* [36] or *fasttext* corpora. This could help institutions worldwide to elevate the utility of their already existing, yet unstructured and unconnected databases.

Beyond establishing broadly applicable embeddings, GRL can also be used in a supervised fashion by integrating a label-based term into the loss function. This would enable end-to-end learning on specific tasks, where one network architecture comprises all processing steps from handling raw inputs up to the final prediction. However, this comes with the downside of decreased generality in the learned representations (since those are now task-specific and thus not reusable across applications), unless informed data augmentation [37] techniques are used. Moreover, embeddings learned as an intermediate step within a neural network will generally not correspond to concepts of human understanding and will therefore lack causability by nature.

3. Challenge II: Distributed GRL

Distributed learning has become a trending topic in recent years, either for reasons of limited central memory & processing power, legal restrictions on data transmission, non-identically independend distributed data (non i.i.d.) over local sub-populations, or for the potential of hyper-scalable systems operating at lower costs. This is especially true for the medical domain, where locally generated data is sensitive, practically "stationary" (are not allowed to be transferred between institutions), and of huge volume (on the order of many Terabytes per day).

There are three main versions of distributed learning: (i) purely decentralized, where local models do not automatically contribute to each other above manual sampling of the models and update of hyperparameters; (ii) federated learning schemes [38,39], where a global model is constructed considering update signals of all local clients which are merged into a global model and distributed back to the clients; (iii) collaborative learning in various forms, where the goal is to exchange information about internal model formation among the parties involved in a peer-to-peer fashion, yet keep the local training data confidential (a variant could also train on de-centralized features supposedly modelling the same underlying instances [40]). The great challenge for distributed learning of embeddings is to keep the representation space aligned across location boundaries — akin to the feature alignment problem in the MM setting.

Considering local pockets of data as natural clusters whose representative *supernodes* can be used as inputs to a more abstract graph embedding step. The literature contains several approaches to achieve such coarsened embeddings: while (i) simple aggregation of node features can already suffice [41], (ii) introducing *virtual nodes* and learning their embeddings together with the rest of the subgraph [42] promises a very flexible clustering strategy, e.g. one could run a series of local predictions and record which nodes in the graph had the greatest influence on a desired outcome, then connect this set to a virtual node whose embedding can be expected to be particularly helpful in distributed predictions of the same kind. It is worth to mention that virtual nodes do not exist in the original graph and can therefore not be sampled, they are rather *representatives* of interesting clusters w.r.t. solving specific tasks; (iii) sampling strategies such as Random Walks (RW) or Anonymous RWs [43], where the local graph structure itself is translated to fixed-size vectors which are subsequently used as inputs to an embedder.

Lastly, Ying et al. (2018) [44] have proposed an extension of their earlier GraphSAGE (Graph Sampling & Aggregation) approach to take hierarchical feature levels into account; they invented graph coarsening *DiffPool* layers akin to pooling layers in traditional CNNs which learn a node assignment matrix, thus performing an embedding-level clustering step. Although this approach is innovative, it is unclear whether and to what extent it can handle distributed data.

Consequently, our proposed pipeline broadly consist of the following steps (and refer to Fig. 2):

- Once a repertoire of GRL/embedding methods has been established, a decentralized scenario can be simulated. We understand *distributed* with the additional challenge that we have no throughput/latency guarantees concerning the connections between local subgraphs; in the extreme case we even need to consider common internet & mobile connections. Thus, the principle challenge lies in propagating aggregated information per subgraph in such a way that GRL is still feasible from a performance & accuracy standpoint.
- Therefore it is necessary to extend existing *neighbourhood aggregation* techniques, selecting & experimenting with different *aggregation schemes*, some of which are rather trivial (mean, max, sum) but nevertheless perform well on certain tasks according to literature; others can be complete neural networks in themselves, such as Long Short-Term Memory (LSTM) [45], which were successfully used in GraphSAGE (Graph Sampling & Aggregation).
- Depending on its *size, domain, or shape*, experiments with different aggregators per local subgraph to obtain fitting representations are necessary. In practice, aggregation functions can have significant impact on results — it is therefore desirable to train a network so it is capable of establishing heuristics of when to use which aggregator, enabling the network to optimize its aggregation strategy autonomously. The heuristics can incorporate domain knowledge or be derived from user interpretation in a deep reinforcement learning setting [46], thus getting incorporated to define the reward function and maximize the accumulated reward of achieving this goal. The need for an explicit definition of predefined and rigid logic rules is thereby avoided, whereas the emergence of new strategies is enabled.
- Regarding LSTMs, a complementary challenge lies in explaining their good performance on what are intuitively *permutation invariant* data, like the ordering of node neighbourhoods in a graph. We conjecture that many graph-related learning problems could be modelled under the aspect of *information flow*, where the sequence of nodes propagating signals is decisive. However, this assumption would need to be tested in various real-world scenarios.

4. Challenge III: Explainable AI/interpretable ML

Graph Neural Networks (GNN) have been known for quite a while and are a well-established method to learn from graph structures [47, 48]. GNN can be used for encoding representations of data that do not have an ordered grid structure, contain different types of relationships and have obtained state of the art performance on different tasks including graph classification [49], node classification [50] and link prediction [51].

There are different types of GNN architectures [52] and different graph embeddings [53].

The need for dynamic GNN has led to newly invented architectures such as Pointer Graph Networks (PGN) [54], to enable the processing of adaptive graphs (also explained in detail in a recent survey [55]).

The Open Graph Benchmark [56] contains representative datasets for GNN benchmarking and is continuously updated. However, GNNs can be treated as black box models, lacking interpretability. This lead to an increase in development of interpretation techniques for all those different GNN architectures.

Duvenaud et al. (2015) [41], first introduced a convolutional neural network architecture for graphs which generalizes standard molecular feature extraction methods based on circular fingerprints, and demonstrated interpretability and predictive performance on fingerprint data.

Baldassarre et al. (2019) [57] studied the explainability of GNN output using gradient-based and decomposition-based techniques on two tasks and implemented Sensitivity analysis (SA) [58], Guided backpropagation (GB) [59], and Layer wise relevance propagation (LRP) to explain node prediction (we briefly discuss LRP below).

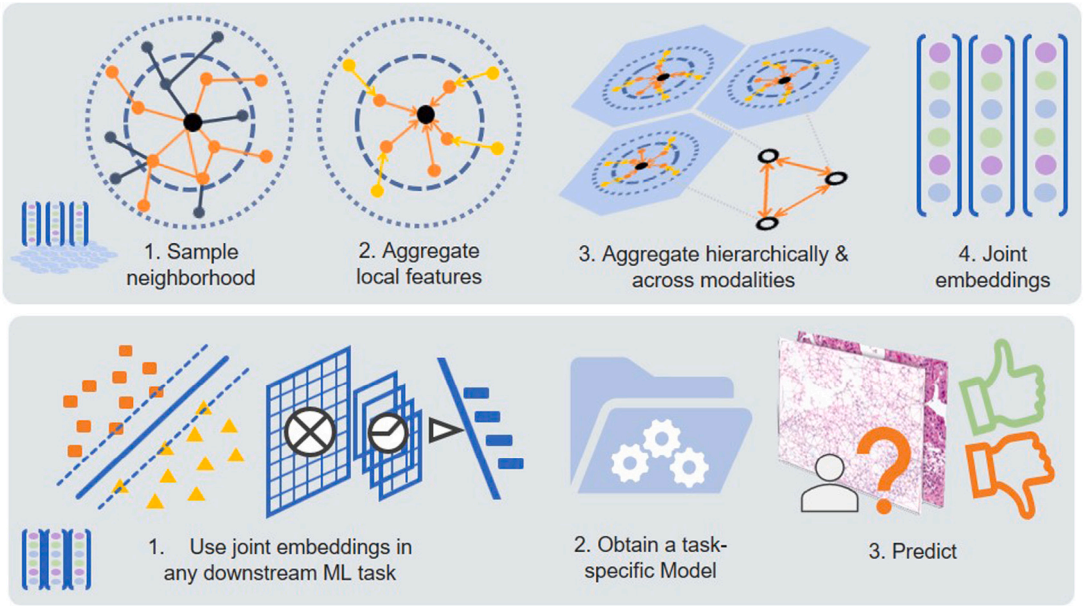

Fig. 2. Learning hierarchies & logical clusterings of nodes in a graph. This approach extends the original GraphSAGE (Graph Sampling & Aggregation) as well as [44] to a decentralized setting where node clusters are generated locally & their aggregated feature vectors propagated across the network in order to connect them in higher-order clusters based on some similarity metrics..

Recent work [60] focuses on explaining comprehensive features such as node- and edge features as well as connecting patterns in a weighted graph for node classification. In addition, simulation on synthetic data was performed to compare results with human interpretation.

Prediction on graphs is usually influenced by complex combinations of nodes and edges between them; accurate prediction of node labels can only be achieved when they are considered together [61]. Such a joint contribution cannot be modelled as simple linear combinations of individual contributions. Some research work on GNNs uses *attention mechanisms* for interpretability [50,62,63].

The learned edge attention values indicate relevant graph structures, however, the values are the same across the nodes connected by that edge; this does not hold in many applications where the edge is the most important element for the label prediction of one particular node, but not the labels of other nodes. Furthermore, these approaches cannot explain predictions by considering both graph and node features, and are thus limited to specific GNN architectures. Motivated by these problems, Ying et al. (2019) [64] proposed GNNExplainer, a model-agnostic approach that can provide interpretable explanations for predictions of any graph based model. The advantage of GNNExplainer is that the generated explanation is a rich subgraph (part of the entire graph on which GNN was trained), such that the subgraph maximizes mutual information with GNN prediction. In addition, this approach is capable of handling both single- and multi-instance explanations. In single instance, GNNExplainer explains predictions of one instance, such as a node, label or new link, whereas in multi-instance explanations it provides explanations that consistently explain a set of instances, such as nodes of a given class.

Another new approach for explanations has been introduced recently by Yuan et al. (2020) [65], called XGNN, for interpreting deep graph models at model level. XGNN introduced a technique of finding graph patterns that maximize a certain prediction through graph generation. It is formulated as a reinforcement learning (RL) problem and can generate graph patterns repetitively.

Certain graph rules are incorporated to make the generated graphs human-intelligible and valid. GraphLIME [66] is a local interpretable model explanation framework that finds the most representative features as explanations in a non-linear manner. GraphLIME is a nonlinear graph variant of the Local Interpretable Model-Agnostic Explanation (LIME) method [67], which considers perturbation near the node being explained and applies a linear interpretable model. There exist some promising approaches which aim at capturing better model hierarchical relationships or causality mechanisms [68,69].

With the increased adoption of graph convolutional neural networks (GCNNs) [33], explainability methods for GCNNs have been also introduced recently [70]. Three prominent explainability methods of convolutional neural networks include: Contrastive Gradient-based (CG) saliency maps, Class Activation Mapping (CAM), and Excitation Backpropagation (EB) as well as their variants: Gradient-weighted CAM (Grad-CAM) and contrastive EB (c-EB) are extended to GCNNs. The explanations of each method are categorized into three metrics: fidelity, contrastivity and sparsity. The results from two application domains demonstrated that Grad-CAM is currently the most suitable among the studied methods for explanations on graphs of moderate size. Another work on local fidelity for explanation for GNNs is introduced in [71]. A post-hoc framework known as TraP2 is proposed, which is based on local fidelity of any GNN model and generates high fidelity explanations. Furthermore, different gradient attribution analysis approaches for GCNNs have been proposed [72], known as Node Attribution Method (NAM), which can get the model contribution from central node as well as its neighbouring nodes to the model output. In addition, Node Importance Visualization (NIV) visualizes the central node and its neighbour nodes based on the value of the contribution. Afterwards, perturbation analysis is utilized to verify the efficiency of the NAM based on citation network datasets. With the use of this method as well as visualization of node contribution in the decision making, more comprehensive contributions of each node are obtained.

Schnake et al. (2020) [73], proposed the GNN-LRP approach which is derived from higher-order Taylor expansions based on Layer-wise Relevance Propagation (LRP). The LRP algorithm for pixel-based images explains a classifier's prediction specific to a given data point by attributing relevance scores to important components of the input by using the topology of the learned model itself.

This method is originally based on two methodological principles: (i) the network propagation technique via max-pooling with rectified linear units by Zeiler & Fergus (2014) [74], and (ii) Taylor decomposition [75]. The overall idea is to identify patterns in the input

data of a network that are linked to a particular activation. As a first step, Bach et al. (2015) [76] have used this for images as pixel-wise decomposition and demonstrated how this decomposition can be used as an xAI method in combination with a pixel-wise relevance propagation algorithm. For this purpose they distinguished between (i) viewing the neural network as a function whilst disregarding the network topology, and (ii) message passing approaches (similar to learning representations by back-propagating errors [77] and learning parameters of Conditional Random Fields [78]). That way one can do a "pixel-wise decomposition of a function:"

Let f be a positive-valued function $f : \mathbb{R}^d \to \mathbb{R}^+$. The image can be decomposed as a set of pixel values $\boldsymbol{x} = \{x_p\}$ where p denotes a particular pixel. The function $f(\boldsymbol{x})$ quantifies the presence (or amount) of a certain type of object(s) in the image. This quantity can e.g. be a probability or the number of occurrences of the object. A function value $f(\boldsymbol{x}) = 0$ indicates the *absence* of such object(s) in the image; a function value $f(\boldsymbol{x}) > 0$ expresses the *presence* of the object(s) with a certain probability.

The goal of this algorithm is that each pixel p in the image is associated to a so-called *relevance score* $R_p(\boldsymbol{x})$. This relevance score indicates to what extent a certain pixel contributes to the classification result. The relevance of each pixel can eventually be stored in a heatmap [79] denoted by $\boldsymbol{R}(\boldsymbol{x}) = \{R_p(\boldsymbol{x})\}$, which is defined as the sum of the relevances in the pixel space corresponding to the total relevance detected by the model, i.e.

$$\forall \boldsymbol{x} : \; f(\boldsymbol{x}) = \sum_p R_p(\boldsymbol{x}) \tag{1}$$

Montavon et al. (2017) [80] observed that the application of deep Taylor decomposition to neural networks used for image classification yields rules, that are similar to those proposed by Bach et al. (2015) [76] for pixel images. Consequently, they presented a heatmapping method for explaining the classification $f(\boldsymbol{x})$ of a data point $\boldsymbol{x}$, that is based on the Taylor expansion of the function f at some well-chosen (the selection of which is difficult) *root point* $\widetilde{\boldsymbol{x}}$, where $f(\widetilde{\boldsymbol{x}}) = 0$. As we know from standard math literature [75] the first-order Taylor expansion of a function is given as

$$\begin{aligned} f(\boldsymbol{x}) &= f(\widetilde{\boldsymbol{x}}) + \left(\frac{\partial f}{\partial \boldsymbol{x}}\Big|_{x=\widetilde{x}}\right)^{\top} \cdot (\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \varepsilon \\ &= 0 + \sum_p \underbrace{\frac{\partial f}{\partial x_p}\Big|_{x=\widetilde{x}} \cdot (x_p - \widetilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon \end{aligned} \tag{2}$$

The sum $\sum_p$ is calculated over all pixels in the image, and $\{\widetilde{x}_p\}$ are the pixel values of the root point $\widetilde{\boldsymbol{x}}$. The added-up elements are the relevances $R_p(\boldsymbol{x})$. The ε denotes second-order and higher-order terms. Most of the terms in the higher-order expansion involve several pixels at the same time and are therefore difficult to redistribute, therefore Montavon et al. (2017) proposed to consider *only the first-order terms for the heatmapping*. Consequently, the heatmap can be expressed as the element-wise Hadamard product $x \odot y$ between the gradient of the function $\partial f / \partial \boldsymbol{x}$ at the root point $\widetilde{\boldsymbol{x}}$ and the difference between the image and the root point $(\boldsymbol{x} - \widetilde{\boldsymbol{x}})$:

$$\boldsymbol{R}(\boldsymbol{x}) = \frac{\partial f}{\partial \boldsymbol{x}}\Big|_{x=\widetilde{x}} \odot (\boldsymbol{x} - \widetilde{\boldsymbol{x}}). \tag{3}$$

For more details please directly refer to the work of Montavon et al. (2017). Images consist of grid-shaped inputs; non-grid-shaped inputs like graphs are processed by GNNs. The approach of Schnake et al. (2020) is able to generate a decomposition of the GNN prediction as a collection of relevant walks. The higher-order Taylor expansions are computed using multiple backpropagation passes from the top layer to the first layer of the GNN. LRP applied in image processing neural networks only needs one backpropagation of relevance. Those backpropagations are not to be confused with the backpropagation training procedure; LRP is applied after the training is done and depends on its performance.

The graph is defined as an ordered pair $G = (\mathcal{V}, \mathcal{E})$ by its node set of objects $\mathcal{V} = \{v_1, \ldots, v_n\}$ and edge (often called link) set of objects $\mathcal{E} \subseteq \{(v_i, v_j) | v_i, v_j \in \mathcal{V}\}$. Each node v can be represented by one or more features, which are in general tensors. The same applies for edges; if all nodes and all edges of the graph have the same type of features, then the graph is homogeneous.

As input, the GNN receives the structure of the graph, which is expressed by the adjacency matrix enhanced by self-connections $\boldsymbol{\Lambda}$ and an initial state $\boldsymbol{H}_0$ which corresponds to the initial representations of the node- and edge features. The GNN computes the following function (4):

$$f(\boldsymbol{\Lambda}; \boldsymbol{H}_0) = g(\boldsymbol{H}_t(\boldsymbol{\Lambda}, \boldsymbol{H}_{t-1}(\boldsymbol{\Lambda}, \ldots \boldsymbol{H}_1(\boldsymbol{\Lambda}, \boldsymbol{H_0})))) \tag{4}$$

where $t = 1 \ldots T$ the number of layers, and g a readout function. The computation of the feature representations at each intermediate layer is based on the representation of the features of the previous layer, as expressed by Eqs. (5) and (6):

$$Z_t = \boldsymbol{\Lambda} \boldsymbol{H}_{t-1} \tag{5}$$

$$\boldsymbol{H}_t = (C_t(Z_{t,K}))_K \tag{6}$$

The GNN-LRP method explains the prediction by attributing relevance to sequences of edges that connect nodes from the input to the output layer of the GNN. These paths now correspond to *walks* on the input graph $\mathcal{G}$. Thereby, the attribution of relevance of a node or edge is not made because it is important on its own, but also because of its connection to other relevant nodes or edges which they are connected to.

The prediction results from applying the transition function in Eq. (4) iteratively from layer 0 to layer T followed by the top-level readout function f. This function receives as input the final state $\boldsymbol{H}_T$ which itself depends on the input graph through its representation $\boldsymbol{\Lambda}$. Consequently, to extract walks in the input graph which are relevant for the GNN prediction, (Schnake et al. 2020) proposed to use the higher-order Taylor expansion of $f(\boldsymbol{\Lambda})$. They considered a T-order Taylor expansion and looked at terms of this expansion that match particular paths in the GNN, i.e. particular walks in the input graph. That means to let $\mathcal{W} = (\ldots, J, K, L \ldots)$ be *one such walk* and view $|\mathcal{W}|$ as the number of edges in the walk $\mathcal{W}$. The $|\mathcal{W}|$th order terms of the Taylor expansion of $f(\boldsymbol{\Lambda})$ at some *root point* $\widetilde{\boldsymbol{\Lambda}}$ can then be expressed as follows:

$$f(\boldsymbol{\Lambda}) = \sum_{k=0}^{\infty} \sum_{\mathcal{B} \in \mathbb{B}_k} \frac{1}{\alpha_{\mathcal{B}}!} \frac{\partial^k f}{\partial \lambda_{\mathcal{E}_1} \ldots \partial \lambda_{\mathcal{E}_k}}\Big|_{\widetilde{\boldsymbol{\Lambda}}} \cdot \prod_{i=1}^{k} (\lambda_{\mathcal{E}_i} - \widetilde{\lambda}_{\mathcal{E}_i}) \tag{7}$$

An appropriate root point $\widetilde{\boldsymbol{\Lambda}}$ is difficult to find, but under the constraint that C_t and g are piecewise linear and positively homogeneous, $\widetilde{\boldsymbol{\Lambda}} = s\boldsymbol{\Lambda}$ with $s \to 0$ is a mathematically sound choice. $\alpha_{\mathcal{B}}$ denotes the multiindex of a bag of edges $\mathcal{B}$, $\lambda_{\mathcal{E}_i}$ refers to the element in the adjacency matrix $\boldsymbol{\Lambda}$ corresponding to edge $\mathcal{E}_i$. It is shown by [73], that all terms where $k \neq T$ vanish, so that Eq. (7) has the following form:

$$f(\boldsymbol{\Lambda}) = \sum_{\mathcal{B} \in \mathbb{B}_T} \frac{1}{\alpha_{\mathcal{B}}!} \frac{\partial^T f}{\partial \lambda_{\mathcal{E}_1} \ldots \partial \lambda_{\mathcal{E}_T}} \cdot \prod_{i=1}^{T} \lambda_{\mathcal{E}_i} \tag{8}$$

This reduced formulation still admits any graph $\mathcal{G}$ as input and many of the established GNN models: GCNN, Graph Isomorphism Network (GIN) [49], and SchNet [81] are contained in this formulation.

The relevance of the walk can be computed in two ways; the first corresponds to the "Gradient × Input" (GI) attribution, which is expressed by Eq. (9):

$$R_{\mathcal{W}} = \frac{\partial}{\partial \ldots} \left(\frac{\partial}{\partial \lambda^{\star}_{JK}} \left(\frac{\partial \ldots}{\partial \lambda^{\star}_{KL}} \cdot \lambda^{\star}_{KL} \right) \cdot \lambda^{\star}_{JK} \right) \cdot \ldots \tag{9}$$

The "Hessian × Product" propagation rule is more robust towards shattering gradient problems that occur in deeper neural networks:

$$R_{\mathcal{W}} = \underbrace{\mathrm{LRP}\Big(\underbrace{\mathrm{LRP}\Big(\underbrace{\mathrm{LRP}\Big(\ldots, \lambda^{\star}_{KL}\Big)}_{R_{KL\ldots}}, \lambda^{\star}_{JK}\Big)}_{R_{JKL\ldots}}, \ldots\Big)}, \tag{10}$$

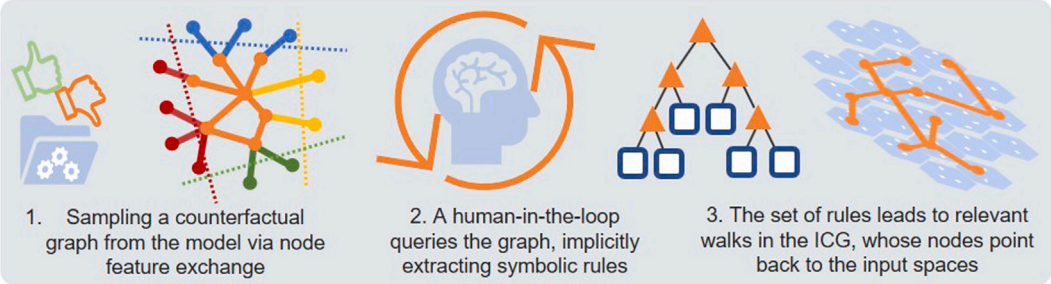

Fig. 3. Generating a counterfactual graph (CG) by sampling from a trained model. A human-in-the-loop interacts with and refines the CG which is subsequently transformed to a decision forest & reduced to an easily-interpretable decision tree. Rules obtained from the decision tree are translated back to relevant paths in the original multi-modal fusion graph/embedding space.

The LRP rule is different according to GNN architecture; for a GCNN in particular it is provided by the following Eq. (11):

$$R_{jKL\ldots} = \sum_{k \in K} \frac{\lambda_{JK} h_j w_{jk}^{\wedge}}{\sum_J \sum_{j \in J} \lambda_{JK} h_j w_{jk}^{\wedge}} R_{kL\ldots} \tag{11}$$

The learned weights w_{jk} on the connection between neuron j and neuron k in the subsequent layers J and K, are used for the backpropagation of relevance. The equation $(\cdot)^{\wedge} = (\cdot) + \gamma\rho(\cdot)$, gives the user the opportunity to experiment with the hyperparameter γ and influence the intensity of the computed relevances. This is common to LRP derivation equations for other neural network architectures.

All terms of the expansion that are not bound to a walk $\mathcal{W}$ in $\mathcal{G}$ converge to zero, implying the conservation property $\sum_{\mathcal{W}} R_{\mathcal{W}} = f(\boldsymbol{\Lambda})$. Conservation is a commonly stated property that explanation techniques should satisfy and consists of a validation method for computational correctness. Despite the simplicity of Eq. (9), computing this quantity directly, e.g. using automatic differentiation, can be extremely difficult; for further details on this approach please refer to the original work of Schnake et al. (2020) [73].

LRP is successfully applied in Graph Convolutional Networks (GCNN) applied in text sentiment analysis by graph classification [82], quantum chemistry, and image processing [73]. The benefit of LRP in this case – although the implementation is quite different and tailored to a particular architecture – is that this method applies to graph classification, whereas most of the non-model agnostic methods explain GNNs that do node classification or link prediction. LRP reveals hidden dynamics over all layers, a property that other XAI methods for graphs [50] do not exhibit. Furthermore, perturbation experiments on node flipping tasks show the monotonic degradation of classification performance, when removing the relevant elements one by one, sorted by decreasing relevance. This fact underlines that the highlighted features are important for the prediction in a proportional manner; this phenomenon is not evident in other XAI methods.

5. Towards an integrated vision

We envision an integrated medical ML pipeline starting at the input data level, building directly on image/signal taking processes. Via intra-modal feature extraction and multi-modal embedding alignment, we arrive at a *low-dimensional representation space* comprising *relevant* data to assist the medical decision making process. Based on current state-of-the-art deep learning models we end up at classification/prediction results that can be presented interactively to medical professionals for inspection. The interactive procedure needs to be in place, facilitating re-integration of human feedback into the whole algorithmic loop. Although there is already a cornucopia of xAI methods to choose from, we find the expressiveness of counterfactuals particularly appealing for an interactive, *interview-style exploration process* due to their contrastive nature.

Initially, there is an urgent need to extend current state-of-the-art methods in graph explainability to the MM case, thereby establishing a baseline for evaluation of our own counterfactual-based approach. Thus, we need to describe, develop and evaluate a model to find the underlying explanatory factors already present in the low-level data, i.e. to answer questions of "What is relevant?" to change the graph structure correspondingly. Since automatic extraction will often fail, a human-in-the-algorithmic-loop [83], will be necessary here to enable the interactive learning setting. Due to the fact that humans are unable to directly orient themselves in high dimensional data sets, we need to design, develop and evaluate subspace visualization methods [84,85] to let the human expert interactively manipulate automatically generated samples, thereby iteratively assisting in forming causal and conceptual models.

In order to utilize the insights gained in this procedure, a crucial step is to formalize, develop & test techniques to channel human-originated feedback back into the automated decision-making process, either via gradual model parameter updates or directly via efficient re-training of input feature representations.

Complex diseases such as cancer need to be studied on a systems level, because the interplay of highly diverse modalities (DNA mutations, Gene expression, Methylation, etc.) substantially contributes to disease progression [86]. Here, graphs provide a natural way to efficiently model this phenomenon; however, semantic links between biological and disease-relevant features across modalities are largely unknown to this date.

We believe that learning these *semantic links* can again be facilitated by a human-in-the-loop as a regulator who is sometimes able to comprehend the context, and based on her/his conceptual understanding may judge the underlying ML decision paths [87]. In Section 4 we have reviewed a range of XAI methods for the detection of walks within the input graph which are relevant for a prediction (e.g a prediction of disease relapse in biomedical applications). The degree of causability obtained from these paths, however, highly depends on the structure of the input graph defining the causal links. Our vision is the conversion of detected *relevant paths* to *disease causing paths* by incorporating human knowledge into the algorithmic loop.

Therefore, human–AI interactions should be based on the low-level (unfolded) input features in order to efficiently discover, reject and confirm causal links between biomedical modalities (using e.g disease ontology databases). Once these links are computed and the structure of the input graphs is updated accordingly, methods for explainable GNN (see Section 4) are consequently regularized towards decision paths with informed disease causing effects. The aforementioned procedure will certainly not converge within a single run, it requires multiple iterations with human domain knowledge in the loop promoting the model training process. Human interactions could be realized by *"what-if"* requests (counterfactuals) to the system resulting in a *counterfactual graph* (CG), where features are defined as nodes and the edges point to combinations of features, which we call *counterfactual paths*. Initially, the CG can be generated in a purely data-driven manner: Based on a test

set comprising a sufficient number of samples, an algorithm will walk through the feature space swapping feature values between nearest neighbours of a different outcome class until the class of the instance itself changes. The nearest neighbour based sampling will result in adversarial examples of realistic patient profiles and thus are built upon plausibly counterfactuals. Furnishing such counterfactuals based roughly on the internals of a model does not suffice for explainability. The plausibility of the adversarial change is a must, i.e. the "adversarial path" leading to the label change should have a real chance of occurring in practice for the counterfactual to be realistic. In this regard, extensions of recent attempts at plausible counterfactuals for image classification [88] should be extended to models for graph data.

The sampled feature path leading to the class change will be stored and forms a *counterfactual decision path*. Repeating this procedure results in a graph comprising *multiple decision paths,* which could be utilized as a communication channel back into the Black-Box model.

We suggest to transform the CG to a Decision Forest (DF) [89] classifier comprising multiple trees derived from semantically enriched counterfactual subgraphs. The predictive power of this DF classifier could be compared with the classification outcomes of the Black-Box model. Decision Trees strongly supporting the outcome of the Black-Box model are directly linked to connected nodes within the counterfactual graph and thus may serve as the *local* explanatory factors of a decision. Moreover, it would be interesting to study whether the decision paths within the CG can be mapped back to the explained decision paths inferred by the methods in Section 4.

Recent work has shown how to efficiently reduce a DF to a single decision tree [90,91] from which counterfactuals can be easily observed by the leaf nodes, so it could be used as a model for *global* explanations (see Fig. 3). In this approach the *human-in-the-loop* will have the opportunity to study this consensus decision tree, thereby adopting the modifications to the counterfactual graph. Exploring the effects on the decision trees caused by modifications of the counterfactual graph may facilitate the definition of symbolic rules in order to revise the internal structure of the input graph (see Fig. 3). Possible modifications include adding or deleting semantic links between modalities, but also adjustments of their edge weights. A big advantage of our visionary approach is that the impact of regulation could be explored efficiently without the need for compute-intense re-training of the model after each modification. Accumulated modifications to the CG can be back-propagated as knowledge-based constraints and overall will regularize the training process of the Black-Box Model.

6. Conclusion

In this paper we motivated the need for a novel, holistic approach to an automated medical decision pipeline building on state-of-the-art Machine Learning research, yet integrating the human-in-the-loop via an innovative, *interactive & exploration-based* explainability technique called *counterfactual graphs*. We emphasized the need for multi-modality in every stage of this integrated approach, since medical decisions are mostly directed by various influence factors stemming from a multitude of underlying signals and knowledge bases. Towards this end, we outlined potential solutions to the challenge of *aligning local geometries* across different input feature spaces. Last but not least, we pointed out the necessity of computing a joint multi-modal representation space in a *decentralized* fashion, for the reasons of scalability & performance as well as ever-evolving data protection regulations.

This effort is indented as a motivation for the international research community and a launchpad for further work in the fields of *multi-modal embeddings*, *interactive explainability*, *counterfactuals*, *causability*, as well as necessary foundations for effective future *human–AI interfaces*.

Abbreviations

- AI = Artificial Intelligence
- CAM = Class Activation Mapping
- c-EB = contrastive Excitation Backpropagation
- CG = Counterfactual Graph
- CRF = Conditional Random Fields
- CT = Computational Tomography
- DF = Decision Forest
- DGNN = Dynamic Graph Neural Network
- EB = Excitation Backpropagation
- GAN = Generative Adversarial Network
- GB = Guided Backpropagation
- GCNN = Graph Convolutional (Neural) Network
- GIN = Graph Isomorphism Network
- GloVe = Global Vectors for Word Representation
- GNN = Graph Neural Network
- Grad-CAM = gradient-weighted Class Activation Mapping
- GraphSAGE = Graph Sampling & Aggregation
- GRL = Graph Representation Learning
- ICG = Interaction & Correspondence Graph
- LIME = Local Interpretable Model-Agnostic Explanations
- LSTM = Long Short-Term Memory
- LRP = Layer Wise Relevance Propagation
- MM = Multi-Modal
- MRI = Magnetic Resonance Imaging
- NAM = Node Attribution Method
- NIV = Node Importance Visualization
- OCT = Optical Coherence Tomography
- OGB = Open Graph Benchmark
- PGN = Pointer Graph Network
- PET = Positron Emission Tomography
- RW = Random Walks
- ReLU = Rectified Linear Unit
- SA = Sensitivity Analysis
- xAI = explainable Artificial Intelligence
- XGNN = Explanations of Graph Neural Networks

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful for the valuable comments of the anonymous reviewers. Parts of this work have received funding from the EU Project FeatureCloud. The FeatureCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826078. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 "explainable Artificial Intelligence".

References

[1] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118, http://dx.doi.org/10.1038/nature21056.

[2] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature Med. 24 (9) (2018) 1342–1350, http://dx.doi.org/10.1038/s41591-018-0107-6.

[3] K. Faust, S. Bala, R. van Ommeren, A. Portante, R. Al Qawahmed, U. Djuric, P. Diamandis, Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning, Nature Mach. Intell. 1 (7) (2019) 316–321, http://dx.doi.org/10.1038/s42256-019-0068-6.

[4] J. Pearl, The limitations of opaque learning machines, in: J. Brockman (Ed.), Possible Minds: 25 Ways of Looking at AI, Penguin, New York, 2019, pp. 13–19.

[5] J. Pearl, The seven tools of causal inference, with reflections on machine learning, Commun. ACM 62 (3) (2019) 54–60.

[6] D. Schneeberger, K. Stoeger, A. Holzinger, The european legal framework for medical ai, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 2020, pp. 209–226, http://dx.doi.org/10.1007/978-3-030-57321-8-12.

[7] A. Holzinger, M. Dehmer, I. Jurisica, Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions, BMC Bioinformatics 15 (S6) (2014) I1, http://dx.doi.org/10.1186/1471-2105-15-S6-I1.

[8] A. Holzinger, B. Haibe-Kains, I. Jurisica, Why imaging data alone is not enough: Ai-based integration of imaging, omics, and clinical data, Eur. J. Nucl. Med. Mol. Imaging 46 (13) (2019) 2722–2730, http://dx.doi.org/10.1007/s00259-019-04382-9.

[9] C. Jean-Quartier, F. Jeanquartier, I. Jurisica, A. Holzinger, In silico cancer research towards 3r, BMC Cancer 18 (1) (2018) 408, http://dx.doi.org/10.1186/s12885-018-4302-0.

[10] Q. He, X. Li, D.N. Kim, X. Jia, X. Gu, X. Zhen, L. Zhou, Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: Applications in medical prognosis prediction, Inf. Fusion 55 (3) (2020) 207–219, http://dx.doi.org/10.1016/j.inffus.2019.09.001.

[11] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (xAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[12] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 9 (4) (2019) 1–13, http://dx.doi.org/10.1002/widm.1312.

[13] J. Pearl, Causality: Models, Reasoning, and Inference, second ed., Cambridge University Press, Cambridge, 2009.

[14] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations, KI - Künstliche Intelligenz (German J. Artif. Intell.) 34 (2) (2020) 193–198, http://dx.doi.org/10.1007/s13218-020-00636-z, Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt.

[15] A. Holzinger, Usability engineering methods for software developers, Commun. ACM 48 (1) (2005) 71–74, http://dx.doi.org/10.1145/1039539.1039541.

[16] A. Holzinger, Explainable ai and multi-modal causability in medicine, J. Interact. Media 19 (3) (2020) 171–179, http://dx.doi.org/10.1515/icom-2020-0024.

[17] F. Cabitza, D. Ciucci, R. Rasoini, A giant with feet of clay: On the validity of the data that feed machine learning in medicine, in: Organizing for the Digital World, Springer, Cham, 2019, pp. 121–136.

[18] F. Cabitza, A. Campagner, C. Balsano, Bridging the last mile gap between ai implementation and operation:data awareness that matters, Ann. Transl. Med. 8 (7) (2020) 501, http://dx.doi.org/10.21037/atm.2020.03.63.

[19] M.J. Kusner, J.R. Loftus, The long road to fairer algorithms, Nature 578 (2020) 34–36, http://dx.doi.org/10.1038/d41586-020-00274-3.

[20] E. Bareinboim, J. Pearl, A general algorithm for deciding transportability of experimental results, 2013, pp. 1–28, arXiv:1312.7485.

[21] M. Prosperi, Y. Guo, M. Sperrin, J.S. Koopman, J.S. Min, X. He, S. Rich, M. Wang, I.E. Buchan, J. Bian, Causal inference and counterfactual prediction in machine learning for actionable healthcare, Nature Mach. Intell. 2 (7) (2020) 369–375, http://dx.doi.org/10.1038/s42256-020-0197-y.

[22] D. Kahneman, Varieties of counterfactual thinking, in: N.J. Roese, J.M. Olson (Eds.), What Might Have Been: The Social Psychology of Counterfactual Thinking, Taylor and Francis, New York, 1995, pp. 375–396.

[23] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature Methods 11 (3) (2014) 333–340, http://dx.doi.org/10.1038/nMeth.2810.

[24] S. Liu, S. Liu, S. Pujol, R. Kikinis, D. Feng, W. Cai, Propagation graph fusion for multi-modal medical content-based retrieval, in: 13th International Conference on Control Automation Robotics & Vision (ICARCV), IEEE, 2014, pp. 849–854, http://dx.doi.org/10.1109/ICARCV.2014.7064415.

[25] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, Multi-modal classification of alzheimer's disease using nonlinear graph fusion, Pattern Recognit. 63 (3) (2017) 171–181, http://dx.doi.org/10.1016/j.patcog.2016.10.009.

[26] G. Vivar, H. Burwinkel, A. Kazi, A. Zwergal, N. Navab, S.-A. Ahmadi, Multi-modal graph fusion for inductive disease classification in incomplete datasets, 2019, pp. 1–9, arXiv:1905.03053.

[27] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video, in: L. Amsaleg, B. Huet, M. Larson, G. Gravier, H. Hung, C.-W. Ngo, W.T. Ooi (Eds.), Proceedings of the 27th ACM International Conference on Multimedia, ACM SIGMM, 2019, pp. 1437–1445, http://dx.doi.org/10.1145/3343031.3351034.

[28] I.C. Dourado, S. Tabbone, R. d. S. Torres, Multimodal Prediction based on Graph Representations, 2019, arXiv. http://arxiv.org/abs/1912.10314.

[29] S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, 2020, pp. 164–172, http://dx.doi.org/10.1609/aaai.v34i01.5347.

[30] S. Bahrami, A. Bosaghzadeh, F. Dornaika, Multi similarity metric fusion in graph-based semi-supervised learning, Computation 7 (1) (2019) 15, http://dx.doi.org/10.3390/computation7010015.

[31] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston, Starspace: Embed all the things!, 2017, pp. 1–9, arXiv:1709.03856.

[32] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, NeurIPS, 2017, pp. 1024–1034.

[33] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, pp. 1–14, arXiv:1609.02907.

[34] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. (JMLR) 3 (2) (2003) 1137–1155.

[35] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C.J. Burges, L. Buttou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, NIPS 2013, NIPS foundation, 2013, pp. 3111–3119.

[36] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1532–1543.

[37] M.D. Bloice, P.M. Roth, A. Holzinger, Biomedical image augmentation using augmentor, Oxford Bioinformatics 35 (1) (2019) 4522–4524, http://dx.doi.org/10.1093/bioinformatics/btz259.

[38] B. Malle, N. Giuliani, P. Kieseberg, A. Holzinger, The more the merrier - federated learning from local sphere recommendations, in: Machine Learning and Knowledge Extraction, in: Lecture Notes in Computer Science LNCS, vol. 10410, Springer, Cham, 2017, pp. 367–374, http://dx.doi.org/10.1007/978-3-319-66808-6-24.

[39] F. Sattler, S. Wiedemann, K.-R. Müller, W. Samek, Robust and communication-efficient federated learning from non-iid data, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (2019) 3400–3413, http://dx.doi.org/10.1109/TNNLS.2019.2944481.

[40] Y. Hu, D. Niu, J. Yang, S. Zhou, Stochastic distributed optimization for machine learning from decentralized features, 2018, pp. 1–10, arXiv:1812.06415.

[41] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 2015, pp. 2224–2232.

[42] T. Pham, T. Tran, H. Dam, S. Venkatesh, Graph classification via deep learning with virtual nodes, in: IJCAI Workshop on Learning in Graphs, 2017, pp. 1–5.

[43] S. Ivanov, E. Burnaev, Anonymous walk embeddings, in: International Conference on Machine Learning (ICML 2018), PMLR 80, 2018, pp. 1–10, arXiv:1805.11921.

[44] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, J. Leskovec, Hierarchical graph representation learning with differentiable pooling, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, NeurIPS, 2018, pp. 4800–4810.

[45] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[46] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowl.-Based Syst. http://dx.doi.org/10.1016/j.knosys.2020.106685.

[47] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, IEEE, 2005, pp. 729–734, http://dx.doi.org/10.1109/IJCNN.2005.1555942.

[48] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2008) 61–80, http://dx.doi.org/10.1109/TNN.2008.2005605.

[49] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? 2018, pp. 1–17, arXiv:1810.00826.

[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, 2017, pp. 1–12, arXiv:1710.10903.

[51] M. Zhang, Y. Chen, Link prediction based on graph neural networks, in: Advances in Neural Information Processing Systems, 2018, pp. 5165–5175.

[52] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–21, http://dx.doi.org/10.1109/TNNLS.2020.2978386.

[53] H. Cai, V.W. Zheng, K.C.-C. Chang, A comprehensive survey of graph embedding: Problems, techniques, and applications, IEEE Trans. Knowl. Data Eng. 30 (9) (2018) 1616–1637.

[54] P. Veličković, L. Buesing, M.C. Overlan, R. Pascanu, O. Vinyals, C. Blundell, Pointer graph networks, 2020, pp. 1–19, arXiv:2006.06380.

[55] J. Skarding, B. Gabrys, K. Musial, Foundations and modelling of dynamic networks using dynamic graph neural networks: A survey, 2020, pp. 1–21, arXiv:2005.07496.

[56] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, J. Leskovec, Open graph benchmark: Datasets for machine learning on graphs, 2020, pp. 1–33, arXiv:2005.00687.

[57] F. Baldassarre, H. Azizpour, Explainability techniques for graph convolutional networks, 2019, pp. 1–21, arXiv:1905.13686.

[58] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, pp. 1–14, arXiv:1409.1556.

[59] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: Y. Bengio, Y. LeCun (Eds.), 3rd Internatioanal Coference on Learning Representations ICLR 2015 (Workshop Track), 2015, pp. 1–14.

[60] X. Li, J. Saude, Explain graph neural networks to understand weighted graph features in node classification, arXiv preprint arXiv:2002.00514.

[61] A.K. Debnath, R.L. Lopez de Compadre, G. Debnath, A.J. Shusterman, C. Hansch, Structure–activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity, J. Med. Chem. 34 (2) (1991) 786–797.

[62] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (14) (2018) 145301.

[63] C. Lin, G.J. Sun, K.C. Bulusu, J.R. Dry, M. Hernandez, Graph neural networks including sparse interpretability, 2020, pp. 1–10, arXiv:2007.00119.

[64] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, GNNExplainer: Generating explanations for graph neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 2019, pp. 9244–9255.

[65] H. Yuan, J. Tang, X. Hu, S. Ji, XGNN: towards model-level explanations of graph neural networks, in: Y. Liu, R. Gupta (Eds.), Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '20, ACM, 2020, pp. 430–438, http://dx.doi.org/10.1145/3394486.3403085.

[66] Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, Y. Chang, Graphlime: Local interpretable model explanations for graph neural networks, 2020, pp. 1–10, arXiv:2001.06216.

[67] M.T. Ribeiro, S. Singh, C. Guestrin, 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, 2016, pp. 1135–1144.

[68] A.M. Saxe, J.L. McClellans, S. Ganguli, Learning hierarchical categories in deep neural networks, in: M. Knauff, M. Pauen, N. Sebanz, I. Wachsmuth (Eds.), Proceedings of the 35th Annual Meeting of the Cognitive Science Society, COGSCI 2013, Cognitive Science Society, Austin, TX, 2013, pp. 1271–1276.

[69] A. Chattopadhyay, P. Manupriya, A. Sarkar, V.N. Balasubramanian, Neural network attributions: A causal perspective, 2019, pp. 1–10, arXiv:1902.02302.

[70] P.E. Pope, S. Kolouri, M. Rostami, C.E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10772–10781.

[71] C. Ji, R. Wang, Y. Li, H. Wu, Perturb. more, Perturb more trap more: Understanding behaviors of graph neural networks, 2020, pp. 1–25, arXiv:2004.09808.

[72] S. Xie, M. Lu, Interpreting and understanding graph convolutional neural network using gradient-based attribution method, 2019, pp. 1–10, arXiv:1903.03768.

[73] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K.T. Schütt, K.-R. Müller, G. Montavon, xAI for graphs: Explaining graph neural network predictions by identifying relevant walks, 2020, pp. 1–24, arXiv:2006.03589.

[74] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Tuytelaars T. Pajdla (Eds.), ECCV, in: Lecture Notes in Computer Science LNCS, vol. 8689, Springer, Cham, 2014, pp. 818–833, http://dx.doi.org/10.1007/978-3-319-10590-1-53.

[75] C.M. Bender, S.A. Orszag, Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory, McGraw-Hill, New York, 1978.

[76] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140, http://dx.doi.org/10.1371/journal.pone.0130140.

[77] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536, http://dx.doi.org/10.1038/323533a0.

[78] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

[79] S. Bach, A. Binder, K.-R. Müller, W. Samek, Controlling explanatory heatmap resolution and semantics via decomposition depth, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 2271–2275, http://dx.doi.org/10.1109/ICIP.2016.7532763.

[80] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognit. 65 (2017) 211–222, http://dx.doi.org/10.1016/j.patcog.2016.11.008.

[81] K. Schütt, P.-J. Kindermans, H.E.S. Felix, S. Chmiela, A. Tkatchenko, K.-R. Müller, Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, in: Advances in Neural Information Processing Systems, 2017, pp. 991–1001.

[82] R. Schwarzenberg, M. Hübner, D. Harbecke, C. Alt, L. Hennig, Layerwise relevance visualization in convolutional text graph classifiers, arXiv preprint arXiv:1909.10911.

[83] A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics 3 (2) (2016) 119–131, http://dx.doi.org/10.1007/s40708-016-0042-6.

[84] F. Jeanquartier, C. Jean-Quartier, A. Holzinger, Integrated web visualizations for protein-protein interaction databases, BMC Bioinformatics 16 (1) (2015) 195, http://dx.doi.org/10.1186/s12859-015-0615-z.

[85] M. Hund, D. Boehm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D.A. Keim, L. Majnaric, A. Holzinger, Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the doctor-in-the-loop, Brain Inform. 3 (4) (2016) 233–247, http://dx.doi.org/10.1007/s40708-016-0043-5.

[86] M. Gustafsson, C.E. Nestor, H. Zhang, A.-L. Barabási, S. Baranzini, S. Brunak, K.F. Chung, H.J. Federoff, A.-C. Gavin, R.R. Meehan, Modules, networks and systems medicine for understanding disease and aiding diagnosis, Genome Med. 6 (10) (2014) 1–11, http://dx.doi.org/10.1186/s13073-014-0082-6.

[87] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G.C. Crişan, C.-M. Pintea, V. Palade, Interactive machine learning: experimental evidence for the human in the algorithmic loop, Appl. Intell. 49 (7) (2019) 2401–2414, http://dx.doi.org/10.1007/s10489-018-1361-5.

[88] A. Barredo-Arrieta, J. Del Ser, Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples, arXiv:2003.11323.

[89] L. Rokach, Decision forest: Twenty years of research, Inf. Fusion 27 (1) (2016) 111–125, http://dx.doi.org/10.1016/j.inffus.2015.06.005.

[90] R.R. Fernández, I.M. De Diego, V. Aceña, A. Fernández-Isabel, J.M. Moguerza, Random forest explainability using counterfactual sets, Inf. Fusion 63 (11) (2020) 196–207, http://dx.doi.org/10.1016/j.inffus.2020.07.001.

[91] O. Sagi, L. Rokach, Explainable decision forest: Transforming a decision forest into an interpretable tree, Inf. Fusion 61 (2020) 124–138, http://dx.doi.org/10.1016/j.inffus.2020.03.013.