Dec 12th, 12:00 AM

Towards a Theory-Based Evaluation of Explainable Predictions in Healthcare

Suparna Ghanvatkar
National University of Singapore, sups.ghanvatkar@gmail.com

Vaibhav Rajan
National University of Singapore, vaibhav.rajan@nus.edu.sg

Towards a Theory-Based Evaluation of Explainable Predictions in Healthcare

Short Paper

Suparna Ghanvatkar
National University of Singapore
15 Computing Drive, Singapore 117418
suparnag@comp.nus.edu.sg

Vaibhav Rajan
National University of Singapore
15 Computing Drive, Singapore 117418
vaibhav.rajan@nus.edu.sg

Abstract

Modern Artificial Intelligence (AI) models offer high predictive accuracy but often lack interpretability with respect to reasons for predictions. Explanations for predictions are usually necessary in making high-stakes clinical decisions. Hence, many Explainable AI (XAI) techniques have been designed to generate explanations for predictions from black-box models. However, there are no rigorous metrics to evaluate these explanations, especially with respect to their usefulness to clinicians. We develop a principled method to evaluate explanations by drawing on theories from social science and accounting for specific requirements of the clinical context. As a case study, we use our metric to evaluate explanations generated by two popular XAI algorithms in the task of predicting the onset of Alzheimer's disease using genetic data. Our preliminary findings are promising and illustrate the versatility and utility of our metric. Our work contributes to the practical and theoretical development of XAI techniques and Clinical Decision Support Systems.

Keywords: healthcare, clinical decision support systems, explainable artificial intelligence

Introduction

Artifical Intelligence (AI) has reached or surpassed human abilities in many domains such as computer vision and natural language processing. AI-based predictive models, in particular, deep neural networks, also have the potential to provide insights for clinical decision-making beyond what humans can do (Greenspan et al. 2016). However, their practical use in Clinical Decision Support Systems (CDSS) is still limited by their lack of interpretability with respect to the reasons for predictions (Sutton et al. 2020). It is well known that such black-box predictions lead to lower user acceptance, as discussed by Gregor and Benbasat (1999), who argue for the need for explanations in DSS to improve performance and user perception. Further, recent European General Data Protection Regulation (GDPR) requires organizations using electronic personal data in DSS to provide meaningful explanations about how an algorithm reaches its final decisions (Kim and Routledge 2018). Such regulations further emphasize the need for explanations for predictions in CDSS.

To fulfill this need for explanations in AI methods, many Explainable AI (XAI) approaches are being actively developed (Molnar 2022). These approaches use a variety of different techniques to provide explanations for black-box models. Many techniques, in different ways, provide scores for each input feature (or variable), indicating the importance or *relevance* of the feature for the prediction made on a given subject. These scores can serve as reasons for the prediction, i.e., the explanation, which, in turn, can be used in subsequent decision-making. E.g., consider a predictive model that uses an MRI image of the brain as input to predict the presence of a disease. An XAI algorithm can then be used to indicate, for a specific patient, the voxels in the MRI that are most relevant for the prediction. This information can be used in a CDSS to highlight the MRI portions most important for the patient's prediction to the clinician.

A large number of XAI algorithms have been developed and the number is rapidly increasing with the pervasive growth of explainable AI. For example, Hakkoum et al. (2022) review *174* different techniques that have been used for clinical tasks. Thus, CDSS designers face the problem of choosing the best possible explanation, possibly from different XAI algorithms, for a prediction in a given clinical context. Our aim is

to develop a metric to evaluate explanations from XAI algorithms. Designers can use such a metric when designing and implementing a CDSS.

In most previous studies, for clinical contexts, explanations are usually not evaluated in a rigorous manner (Hakkoum et al. 2022). One of the main reasons for this is the lack of empirical evaluation metrics for the explanations generated (Miller 2019). Some metrics have been proposed to evaluate the ability of explanations to correctly represent the reason for prediction. These evaluation metrics have been developed with little involvement and consideration of the end-users of the explanations, who in this case are the clinicians (Jin et al. 2022). To evaluate the usefulness and usability of the explanations in a CDSS, we need to consider how clinicians understand and use explanations. This calls for an approach that accounts for both the clinical context and how humans receive and process explanations.

To develop such an approach, we turn to theories of explanation from the social sciences that have studied how people generate and evaluate explanations in general, i.e., outside the machine learning context. They provide clear requirements for a good explanation suitably accounting for human information processing and behavior. In addition, we factor in the unique challenges and requirements of the clinical context. Clinicians require explanations using relatively few domain-specific concepts at a granularity that supports their time-critical decision making. Such explanations are more likely to fit into their existing workflows and lead to higher adoption of XAI in CDSS (Sutton et al. 2020). From these theory-based and context-specific requirements, we formalize the specific requirements for a metric and design our evaluation metric.

Our metric has several novel elements in its design. It can evaluate a given explanation from any XAI algorithm that provides feature-based relevance scores. The evaluation is in terms of clinical *concepts* at a granularity that may be coarser than the features used by the XAI algorithm. E.g., in the case of disease prediction from MRI images, the decision-making process of clinicians, the users of the predictive algorithm and its explanations, may be at the level of brain regions (not specific points in the brain given by the voxels). So, the brain regions may be used as concepts to evaluate the explanations. This facilitates evaluation using the same concepts that are used by clinicians while making decisions. As suggested by theories for explainability, we use multiple predictive tasks (*facts* and *foils*) to evaluate the contrastive nature of explanations, which are associated with causality. We develop a new scoring method that uses biomedical knowledge graphs to score an explanation with respect to its predictive value for the considered facts and foils.

As a case study, we use our metric to evaluate explanations generated by two popular and representative XAI algorithms (LIME and LRP) in the task of predicting the onset of Alzheimer's disease using genetic data. Our preliminary findings are promising and illustrate the versatility of our metric with respect to the XAI algorithms it can evaluate and its utility in clinical contexts. From a design science perspective, Hevner et al. (2004) highlight the importance of evaluation of a design artifact. The development of standardized evaluation methodologies allow for analytical evaluation and benchmarking of the design artifacts (in this case the XAI algorithms) to allow for useful feedback during the iterative design procedure. Thus, evaluation approaches, such as ours, for comparing explanations have the potential to play a significant role in the practical and theoretical development of both XAI algorithms and CDSS.

Background

Explainable AI (XAI) Techniques and their Evaluation

Explanations from an XAI algorithm may be characterised by three different criteria (Molnar 2022): (i) Explanation may be for prediction on a particular instance (*local*) or for the entire model (*global*); (ii) Explanation may be obtained by restricting complexity of the machine learning model (*intrinsic*) or may be generated using a technique after the model is trained (*post-hoc*); (iii) The XAI technique may be limited to a specific model class (*model-specific*) or can be applied to any model (*model-agnostic*).

Many XAI techniques have been developed for each of the above categories, with implementations in software libraries such as AIX360 (Arya et al. 2019) and innvestigate (Alber et al. 2019). In our case study presented later to illustrate the use our metric, we evaluate two popular representative techniques: Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016) and Layer-Wise Relevance Propogation (LRP) (Bach et al. 2015). LIME is a local, post-hoc, model-agnostic technique that estimates the

importance of input features for a prediction using a surrogate model. LRP is a local, post-hoc, model-specific method designed to obtain feature relevance values for predictions from neural networks. Both LIME and LRP provide relevance scores that indicate how relevant each feature is in predicting the output for a specific subject/observation.

Chen et al. (2022) categorize the evaluation of explanations based on two aspects, viz., *faithfulness* of the explanation to the prediction and *usefulness* of the explanation to the user. Faithfulness is the ability of the explanation to correctly represent the reason for prediction, such as explanations addressing how prediction is affected when features are missing. Usefulness relates to the applicability of explanations in addressing practical use-cases and in aiding decision-making. Many strategies for evaluating faithfulness have been proposed, such as those based on similarity and stability; a review can be found in Zhou et al. (2021). These metrics are useful "sanity checks", but may not be indicative of usability in real-world applications (Chen et al. 2022). Thus the importance of the usefulness criterion for explanations has been emphasized.

Usefulness of explanations plays an important role in the design of user-centric systems. Bauer et al. (2021) discuss the importance of explainability in AI applications and outline potential research avenues for Information Systems researchers, particularly for user-centric model interpretation. Asatiani et al. (2020) illustrate the need and challenges of explaining the black-box behavior of AI systems through a case study that highlights legal and ethical implications. They suggest human scrutiny of outputs from AI systems to provide meaningful explanations. Förster et al. (2020a) conduct a human-based study to evaluate explanations from XAI methods and derive characteristics of explanations, such as relevance, coherence and length, that users appreciate. Förster et al. (2020b) design and evaluate a process to calibrate and control the quality of user-centric XAI systems, which relies on a scalable, quantitative metric to evaluate explanations based on its characteristics. Many studies evaluate the usability of explanations by conducting user-studies (John-Mathews 2021). However, such evaluations do not scale well due to their dependence on human observers for evaluation. Lage et al. (2018) evaluate the usability of explanations in terms of their cognitive load. These metrics are specific to text or rule-based explanations, which may not necessarily translate to other modalities. Further, all these criteria generally use the input features (as used by the XAI method), which may not be the most appropriate granularity for evaluating an explanation. To summarize, extant approaches lack strategies to evaluate usability of explanations in an objective and scalable manner.

Theories for Explainability

Explanation, in general, is a justification or reason for a belief or action or decision; and has been studied outside the context of machine learning. Miller (2019) consolidates the frameworks of explanations from social sciences based on a deep understanding of how people define, generate, select, evaluate and present explanations, and lists four requirements of a good explanation:

- Explanations are **contrastive**: Explanations do not seek the answer of 'Why P?' but rather 'Why P rather than Q?', though Q may often be implicit in the context. Following Lipton (1990), we refer to P as the *fact* and Q as the *foil*.
- Explanations are **selective**: People rarely expect a complete explanation consisting of all causes, because the cognitive load of a complete explanation is too large (Keil 2006).
- **Cause** rather than probability matters: The most likely explanation is not necessarily the best because causal reasoning is used to identify if an explanation is satisfactory, not the actual nature of causation (Hoffman and Klein 2017).
- Explanations are **social**: There are two processes in an explanation, the cognitive process where abductive inference is used to deduce the explanation for a given case and the social process where the explanation leads to a transfer of knowledge between the explainer and the explainee.

Many theory-based explanation generation techniques, especially for contrastive and counterfactual explanations, have been designed, as surveyed by Stepin et al. (2021). These explanation generation techniques are based on theories of what a good explanation ought to be and target a general audience. They do not discuss criteria for evaluating explanations. The lack of literature on theory-based evaluation of explanations leads to a gap in evaluating the usability of explanations in clinical settings.

Explainability in Clinical Contexts

The clinical setting poses unique challenges because the end users of predictive methods and explanations (clinicians and nurses), through Clinical Decision Support Systems (CDSS) have relatively high domain knowledge, often have to make time-critical decisions and the decisions have high stakes, i.e. incorrect decisions can have adverse impact on both patient care and operational efficiency of the healthcare provider. Shortliffe and Sepúlveda (2018) assert that CDSSs should fit into existing workflows, enable time-sensitive usage and provide domain-aware information in terms of concepts relevant to clinicians. A good explanation can aid clinical decision-making by providing relevant and domain-aware information. To provide such an explanation in CDSS, designers have to select a suitable XAI algorithm from the many choices available.

As discussed earlier, most extant XAI literature focus on explanations being faithful to the prediction model. The usefulness criterion is important in the clinical context because of the contextual requirements and the profile of the end users. Chen et al. (2022) suggest developing simulation-based evaluation for evaluating the usefulness of explanations. However, as highlighted by Jussupow et al. (2021), there is inadequate understanding of how clinicians use the generated explanations in their decision-making and so, designing such simulations may be difficult. Previous reviews on XAI in CDSS, e.g., (Lai et al. 2020), highlight that studies evaluating usefulness did so through experiments, interviews or surveys.

To summarize, there is a lack of evaluation strategies for evaluating the usefulness of explanations, which requires a theory-based approach to suitably account for human interaction with the generated explanations. In CDSS, the end-users, clinicians, have a specific profile with characteristics and requirements that are different from many general users. They require explanations in terms of domain-aware concepts at the level of granularity that aligns with their decision-making. The number of concepts used should be appropriate to prevent cognitive burden during time-critical decision making. Table 1 compares the most recent and relevant studies on evaluation of explanations along these dimensions.

Evaluation (reference)	Faithfulness or Usefulness	Theory-based	Domain-Aware	Explanation Granularity
ElShawi et al. (2021)	Faithfulness	N	N	inputs
Lage et al. (2018)	Usefulness	Y	N	words, cognitive chunks
Our Approach	Usefulness	Y	Y	domain-based concepts

Table 1. Comparison of our work with recent literature

Our Evaluation Method

Our aim is to develop a method to evaluate explanations from an XAI algorithm used in a CDSS. Adopting a design science approach (Hevner et al. 2004), we first map the four requirements of a "good explanation" from Miller (2019) to design choices for our evaluation method, listed below and summarized in Figure 1.

- To evaluate how contrastive the explanation is, we propose to identify the *foils* for the given prediction task, where the *fact* corresponds to the prediction task itself. The requirement is that the relevant features from the explanation must be more predictive for the fact than for the foil.
- To evaluate selectivity in the concepts, we penalize explanations involving too many concepts.
- To support causal explanations, we restrict our explanations to contrastive explanations, which have been associated with causality in previous studies (Hoffman and Klein 2017).
- The social aspect of the explanation requires careful consideration for the clinical context. To be relevant to clinicians, the explanation must use "concepts" at the right granularity in the explanation. So, we score explanations using domain-derived concepts, which in turn are obtained from biomedical Knowledge Graphs (KG). Knowledge graphs are large, heterogeneous graphs with multiple node types representing clinical concepts (e.g., diseases, drugs) from standardized vocabularies and multiple edge types (e.g., 'treats', 'predisposes') representing associations between pairs of clinical concepts. These KGs are being actively developed by a combination of manual curation and automated methods.

Using these theory-derived design choices, we now develop a scoring method to evaluate explanations. We consider a CDSS that takes as input x (e.g., data or features corresponding to a patient) and has two outputs: the prediction y (from an AI model) and the explanation of the prediction given by r (from an XAI technique).

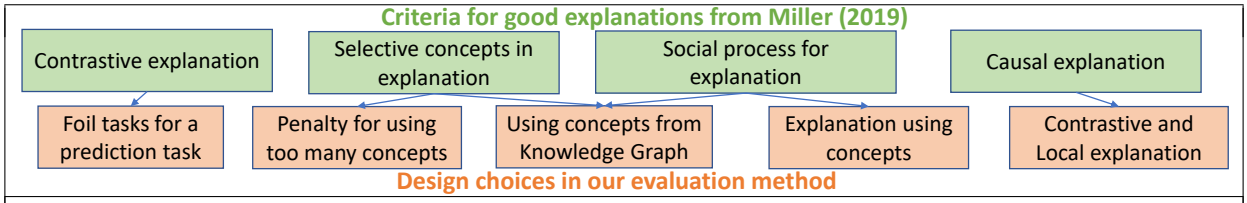

Figure 1. Mapping the Requirements to Design Decisions

We assume feature-based relevance scores as explanation, i.e., r is a k-dimensional vector where the i^{th} dimension, r_i gives the relevance score for the i^{th} feature. As noted previously, the granularity of the features may not necessarily be the apt granularity for an explanation. In such cases, simple post-processing steps can be used to pool these relevance scores to the required "concepts". For example, if we consider a brain MRI, then the features may be the brain voxels for which the XAI algorithm generates explanations, but these can be pooled to different brain regions using an atlas. We represent the set of m domain-specific *concepts* by $c = \{c_1, \cdots, c_m\}$, where $m \leq k$ (depending on the concepts chosen), and the explanation, at concept granularity, $e = [e_1, \cdots, e_m]$ by an m–dimensional vector, where the i^{th} real-valued explanation score, $e_i \in [0, 1]$, gives the (aggregated) relevance score, of concept c_i for the prediction task.

We consider the main prediction task as $fact$ and tasks contrastive to the main task as $foil$ tasks. The explanation needs to be scored such that the relevant features are more predictive of the $fact$ than the $foil$ tasks. For example, if the task is to predict onset of sepsis for an ICU patient, the foil may be to predict if the patient will develop anaphylaxis (a common misdiagnosis). Predictions leading to irrelevant explanations, e.g., reasons for admission to ICU, that may not be useful to clinicians could also be used as foils. The choice of foil tasks is a design decision that depends on the application and clinical context.

Our method evaluates the following in an input explanation from an XAI algorithm:

- We check if the scores e_i in the explanation are only for the relevant concepts c_i by checking if $\sum_{i=1}^{m} e_i \leq 1$. If the explanation does not satisfy this condition, then it may violate the selectivity criterion and may have high cognitive burden.
- A limit on the cognitive burden may be numerically set by setting the maximum number of concepts $max_{\#c}$ for which relevance scores are allowable beyond a pre-fixed limit ϵ. These values have to be determined based on the application and users. Let $n_c(e)$ be the number of concepts in a given explanation e for which relevance scores are greater than ϵ. Then, we check if $n_c(e) \leq max_{\#c}$.
- We check how relevant the explanation is to the $fact$ in comparison to the various $foil$ tasks selected by the user. Consider any one $foil$. We define a function f that uses the explanation to determine its *predictive* value for a task. We call it Predictive Value of Explanations (PVE) function. Note that the relevance scores are used here instead of the input feature values. Using f, we can compare the predictive value of e for the $fact$ as well as for the $foil$. We define a *contrastive score* for the explanation $CS = f(fact, e) - f(foil, e)$. There are multiple ways to design the function f and we describe one based on KGs below.

We first identify nodes, referred to as task-specific nodes in the KG for concepts corresponding to $fact$ and $foil$ tasks. They are respectively called fact node and foil nodes. For instance, in the sepsis prediction task, in the KG, we would choose disease nodes for sepsis ($fact$) and anaphylaxis (a $foil$). Let t denote a task-specific node for either the fact or the foils. Similarly, we identify the nodes in the KG for all the concepts $\{c_1, \ldots, c_m\}$. To determine the strength of association between the concepts and a task, we use a link prediction algorithm to obtain the prediction probability $p_i(c_i, t)$ of a link between concept c_i and task t. The predictive value of the explanation with respect to a task may be computed using these concept-specific link predictions. Thus, we define the PVE function as $f(t, e) = \sum_{i=1}^{m} e_i p_i(c_i, t)$.

There are various techniques to obtain the link prediction probability $p_i(c_i, t)$ from a KG. Graph embedding techniques have produced state-of-the-art results in KG link prediction (Wang et al. 2021) and we use one such graph embedding method, TransE (Bordes et al. 2013). A graph embedding learns vectorial reprsentations of nodes and edges, which capture global structural and semantic relations in the KG. Each edge in a KG may be represented by a triplet (h, l, t) where h, t are head and tail nodes (e.g., age and sepsis) and l is

a relation (e.g., risk factor). The key idea of TransE is that the relationship induced by the edges correspond to translations in vector space: for an edge (h, l, t) present in the KG, we want $v_h + v_l \sim v_t$ and for an absent edge (h, l, t), $v_h + v_l$ should be far away from v_t in the representation vector space. The norm of $v_h + v_l - v_t$, indicating the distance between $v_h + v_l$ and v_t, can be used to obtain the link prediction probability.

Case Study: Explanations for Alzheimer's Disease Prediction

We demonstrate our evaluation method to compare explanations from XAI methods LIME and LRP on the task of predicting the onset of Alzheimer's Disease. We use the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (ADNI 2022) that consists of patients in 3 groups: Cognitively Normal (CN), Mild Cognitive Impairement (MCI) and Alzheimer's Disease (AD). The binary classification task is to distinguish between CN and MCI groups, which may be used in CDSS to predict the onset of Alzheimer's disease. We use genetic data as features, that are available as Single Neucleotide Polymorphisms (SNPs), which indicate genetic variation among patients and are obtained from blood samples. We follow the preprocessing steps outlined in Qiu et al. (2019). We consider a random 80:20 split of train and test data. Multiple classifiers are trained on the train set. The performance of the classifiers, in terms of Area under the ROC Curve (AUROC), is shown in Table 2. The neural network has the best performance; similar results have been reported in previous studies, e.g., Jo et al. (2022) and Qiu et al. (2019).

Classifier	Logistic Regression	Support Vector Machine	Decision Tree	Adaboost	Neural Network
AUROC	0.5139	0.5162	0.5019	0.5690	**0.6629**

Table 2. Classifier Performance on Predicting Onset of Alzheimer's Disease

We obtain explanations for the model from LIME and LRP using their implementations from AIX360 and innvestigate libraries with their default settings. We normalize the outputs using softmax to get relevance scores between 0 and 1, which sum to 1. Figure 2 (left) shows the relevance scores obtained for the input features (SNPs) for a particular patient. Since our first check ($\sum_i e_i \leq 1$) is satisfied for both LIME and LRP, and we do not have relevant values for $\epsilon, max_{\#c}$, we restrict our discussion to the contrastive score.

Concept Identification

SNPs represent single locations in the DNA and may or may not be part of genes. Genes represent functionally important segments of the DNA and their roles in diseases are actively studied. Thus, we choose genes as the "concepts" with which explanations are presented. Not all SNPs in the data may be mapped to a particular gene, and multiple SNPs may be mapped to the same gene. We use the NCBI Entrez DB API available through BioPython (Cock et al. 2009) to obtain the mapping from SNPs to genes. We generate gene-level relevance scores in the explanation by adding the relevance scores for the SNPs which belong to same gene. We do not consider SNPs that cannot be mapped. Since the chosen concept level for the evaluation is genes, information not pertaining to any gene is not considered important to understand the prediction. Figure 2 (middle) shows the computed relevance scores for the genes that are used to evaluate the explanations.

Fact and Foil Identification in a Knowledge Graph

We use the "Disease resembles Disease" relation in the Hetionet KG (Himmelstein et al. 2017) to find diseases that are commonly confused with Alzheimer's Disease during diagnosis. Out of the 3 diseases found, we choose Parkinson's disease as a foil as it is also reported in Klatka et al. (1996) as a common misdiagnosis. We use pretrained TransE embeddings for each concept (gene) and each task (fact and foil) from the Drug Repurposing Knowledge Graph (DRKG) (Ioannidis et al. 2020), which combines multiple biomedical KGs. We use these embeddings to obtain the link prediction probabilities, which are then used to compute the contrastive score with respect to the foil.

Evaluating the Explanations

We see in Figure 2 that the explanations from the two techniques are quite different. We get 9 concepts from LIME and 144 from LRP. Despite these differences, our method can be used to evaluate them. The

contrastive score for LIME is less than that of LRP indicating that based on our evaluation the explanation generated by LRP is preferred over LIME for this patient. The contrastive score for LIME is negative indicating that the explanation concepts in LIME are closer to Parkinson's than to Alzheimer's Disease. The zero score for LRP indicates that its explanation may indicate either of the two diseases.

We find literature-based evidence that supports the conclusions drawn using our contrastive score. Mutations in the top scoring genes in LIME (MDGA2, MCTP2, HTR2C) have been associated with Parkinson's Disease (Cacabelos 2017; Hendrickx et al. 2021; Latourelle et al. 2009), but we could not find any literature associating them to Alzheimer's Disease. Genes BRI3BP and CELSR1 are associated with both Parkinson's Disease and Alzheimer's Disease (Chin et al. 2008; Hu et al. 2017; Jia et al. 2014; Patel et al. 2019). Thus, the literature suggests that the genes considered to be relevant by LIME are likely more relevant to our foil task, while the genes highlighted by LRP are equally relevant to both fact and foil tasks.

SNP	Relevance (LIME)
kgp11729816_C=2	0. 10000035
rs10086985_A=2	0. 10000033
kgp11313442_T=0	0. 10000029
rs13260210_A=0	0. 09999989
rs6541014_T=0	0. 09999988
kgp22824191_G=0	0. 09999987
rs4966020_G=0	0. 09999987
kgp12268363_T=0	0. 09999986
rs7201067_C=0	0. 09999984
rs1021687_T=0	0. 09999982

Gene	Relevance (LIME)	e
MDGA2	0. 10000035	e1
CSMD1	0. 10000033	e2
KMO	0. 10000029	e3
TNFRSF8	0. 09999988	e4
HTR2C	0. 09999987	e5
IGF1R	0. 09999987	e6
MCTP2	0. 09999986	e7
CDH13	0. 09999984	e8
GRM7	0. 09999982	e9

Our Evaluation of LIME

PVE(fact,e)	752.58
PVE(foil,e)	755.51
ContrastiveScore (fact,foil)	-2.933

SNP	Relevance (LRP)
kgp5828144_T	9.843215e-13
rs7133268_G	9.730556e-10
kgp10421358_A	9.707379e-11
rs17522973_A	9.551628e-12
rs2304195_C	9.410357e-13
kgp4258913_G	9.4041074e-14
kgp758913_A	9.376387e-13
rs2488100_T	9.3113985e-15
rs3006069_C	9.297414e-15
kgp3533779_T	9.273616e-14

Gene	Relevance (LRP)	e
NSUN2	9.843215e-13	e1
BRI3BP	9.730556e-10	e2
CELSR1	9.707379e-11	e3
PPP4R2	9.4041074e-14	e4
ZNF423	9.376387e-13	e5
PHACTR2	9.3113985e-15	e6
ANKH	9.297414e-15	e7
DISP1	9.273616e-14	e8
C6orf132	9.23371e-14	e9
ZNF354A	8.956648e-19	e10

Our Evaluation of LRP

PVE(fact,e)	0.03837
PVE(foil,e)	0.03863
ContrastiveScore (fact,foil)	-0.00026

Figure 2. Evaluation of Explanations from LIME and LRP using our Contrastive Score

Initial Contributions and Future Work

Our study makes several contributions to the Healthcare IS literature. To our knowledge, we are the first to develop a principled scoring metric to evaluate explanations that is both grounded in theory of explanations from social science and caters to specific requirements of clinical contexts. With increasing number of Explainable AI techniques, it becomes important for designers of CDSS to objectively evaluate explanations for use in clinical contexts and our method presents initial steps towards fulfilling this unmet need. Thus, our method can play a key role in user-centric design within Healthcare IS research, such as in the process proposed for XAI systems by Förster et al. (2020b). Our method is generic and can be used to compare explanations from any XAI method that provides feature-based relevance scores such as LIME, LRP, SHAP, Saliency map and others. Our method presents a novel use of knowledge graphs for choosing clinical concepts at the right granularity for clinicians and for evaluating the value of the chosen explanation concepts for the prediction task.

This work can be extended in many ways. Empirical evaluation of the metric needs to be performed in collaboration with clinicians for multiple clinical tasks. In particular, the impact of the use and presentation of our metric on CDSS usage patterns of clinicians remains to be evaluated. Our concept identification and scoring method may be developed further by investigating intrinsic design choices for knowledge graph, graph embeddings and link prediction algorithm used. Clinically meaningful ways to choose internal pa-

rameters ($\epsilon, max_{\#c}$) and foils need to be designed. Other approaches for contrastive scoring without the use of knowledge graphs and extensions to settings beyond those with relevance scores may also be explored.

References

ADNI, 2022. *Alzheimer's Disease Neuroimaging Initiative dataset*. URL: http://www.loni.ucla.edu/ADNI.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J., 2019. "iNNvestigate Neural Networks!" *Journal of Machine Learning Research* (20:93), pp. 1–8.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilovic, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J. T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., and Zhang, Y., 2019. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques". *CoRR* (abs/1909.03012).

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., and Salovaara, A., 2020. "Challenges of Explaining the Behavior of Black-Box AI Systems". *MIS Quarterly Executive* (19:4), pp. 259–278.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W., 2015. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". *PLOS ONE* (10:7).

Bauer, K., Hinz, O., Aalst, W. van der, and Weinhardt, C., 2021. "Expl (AI) n it to me–explainable AI and information systems research". *Business & Information Systems Engineering* (63:2), pp. 79–82.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O., 2013. "Translating embeddings for modeling multi-relational data". *Advances in Neural Information Processing Systems* (26).

Cacabelos, R., 2017. "Parkinson's disease: from pathogenesis to pharmacogenomics". *International Journal of Molecular Sciences* (18:3).

Chen, V., Li, J., Kim, J. S., Plumb, G., and Talwalkar, A., 2022. "Interpretable Machine Learning: Moving from Mythos to Diagnostics". *Queue* (19:6).

Chin, M. H., Qian, W.-J., Wang, H., Petyuk, V. A., Bloom, J. S., Sforza, D. M., Lacan, G., Liu, D., Khan, A. H., Cantor, R. M., et al., 2008. "Mitochondrial dysfunction, oxidative stress, and apoptosis revealed by proteomic and transcriptomic analyses of the striata in two mouse models of Parkinson's disease". *Journal of Proteome Research* (7:2), pp. 666–677.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al., 2009. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". *Bioinformatics* (25:11), pp. 1422–1423.

ElShawi, R., Sherif, Y., Al-Mallah, M., and Sakr, S., 2021. "Interpretability in Healthcare: A Comparative Study of Local Machine Learning Interpretability Techniques". *Computational Intelligence* (37:4), p. 1633.

Förster, M., Klier, M., Kluge, K., and Sigler, I., 2020a. "Evaluating Explainable Artifical Intelligence – What Users Really Appreciate". In: *ECIS 2020 Proceedings*, p. 19.

Förster, M., Klier, M., Kluge, K., and Sigler, I., 2020b. "Fostering Human Agency: A Process for the Design of User-Centric XAI Systems". In: *ICIS 2020 Proceedings*. P. 18.

Greenspan, H., Van Ginneken, B., and Summers, R. M., 2016. "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique". *IEEE Transactions on Medical Imaging* (35:5), pp. 1153–1159.

Gregor, S. and Benbasat, I., 1999. "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice". *MIS Quarterly* (23:4), pp. 497–530.

Hakkoum, H., Abnane, I., and Idri, A., 2022. "Interpretability in the Medical Field: A Systematic Mapping and Review Study". *Applied Soft Computing* (117), p. 108391.

Hendrickx, D. M., Garcia, P., Ashrafi, A., Sciortino, A., Schmit, K. J., Kollmus, H., Nicot, N., Kaoma, T., Vallar, L., Buttini, M., et al., 2021. "A new synuclein-transgenic mouse model for early Parkinson's reveals molecular features of preclinical disease". *Molecular Neurobiology* (58:2), pp. 576–602.

Hevner, A. R., March, S. T., Park, J., and Ram, S., 2004. "Design Science in Information Systems Research". *MIS Quarterly* (28:1), pp. 75–105.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E., 2017. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing". *eLife* (6), e26726.

Hoffman, R. R. and Klein, G., 2017. "Explaining Explanation, Part 1: Theoretical Foundations". *IEEE Intelligent Systems* (32:3), pp. 68–73.

Hu, Y.-S., Xin, J., Hu, Y., Zhang, L., and Wang, J., 2017. "Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach". *Alzheimer's Research & Therapy* (9:1).

Ioannidis, V. N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., and Karypis, G., 2020. *DRKG - Drug Repurposing Knowledge Graph for Covid-19*. https://github.com/gnn4dr/DRKG/.

Jia, Z., Guo, Y., Tang, Y., Xu, Q., Li, B., and Wu, Q., 2014. "Regulation of the protocadherin Celsr3 gene and its role in globus pallidus development and connectivity". *Molecular and Cellular Biology* (34:20).

Jin, D., Sergeeva, E., Weng, W.-H., Chauhan, G., and Szolovits, P., 2022. "Explainable Deep Learning in Healthcare: A Methodological Survey from an Attribution View". *WIREs Mechanisms of Disease* (14:3), e1548.

Jo, T., Nho, K., Bice, P., Saykin, A. J., and Initiative, A. D. N., 2022. "Deep learning-based identification of genetic variants: application to Alzheimer's disease classification". *Briefings in Bioinformatics* (23:2).

John-Mathews, J.-M., 2021. "A Critical Empirical Study of Black-box Explanations in AI". In: *ICIS 2021 Proceedings*.

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J., 2021. "Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence". *Information Systems Research* (32:3), pp. 713–735.

Keil, F. C., 2006. "Explanation and Understanding". *Annual Review of Psychology* (57:1), pp. 227–254.

Kim, T. W. and Routledge, B. R., 2018. "Informational privacy, a right to explanation, and interpretable AI". In: *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*. IEEE, pp. 64–74.

Klatka, L. A., Schiffer, R. B., Powers, J. M., and Kazee, A. M., 1996. "Incorrect diagnosis of Alzheimer's disease: a clinicopathologic study". *Archives of Neurology* (53:1), pp. 35–42.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez, F., 2018. "An Evaluation of the Human-Interpretability of Explanation". In: *Proceedings of Conference on Neural Information Processing Systems (NeurIPS) 2018*, p. 7.

Lai, Y., Sukhwal, P. C., and Kankanhalli, A., 2020. "Explanations in Clinical DSS: A Review and Research Agenda". In: *PACIS 2020 Proceedings*, p. 9.

Latourelle, J. C., Pankratz, N., Dumitriu, A., Wilk, J. B., Goldwurm, S., Pezzoli, G., Mariani, C. B., DeStefano, A. L., Halter, C., Gusella, J. F., et al., 2009. "Genomewide association study for onset age in Parkinson disease". *BMC Medical Genetics* (10:1), pp. 1–14.

Lipton, P., 1990. "Contrastive Explanation". *Royal Institute of Philosophy Supplement* (27), pp. 247–266.

Miller, T., 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences". *Artificial Intelligence* (267), pp. 1–38.

Molnar, C., 2022. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.

Patel, D., Mez, J., Vardarajan, B. N., Staley, L., Chung, J., Zhang, X., Farrell, J. J., Rynkiewicz, M. J., Cannon-Albright, L. A., Teerlink, C. C., et al., 2019. "Association of rare coding mutations with Alzheimer disease and other dementias among adults of European ancestry". *JAMA Network Open* (2:3).

Qiu, L., Rajan, V., and Tan, B. C., 2019. "Battling Alzheimer's Disease through Early Detection: A Deep Multimodal Learning Approach." In: *ICIS 2019 Proceedings*.

Ribeiro, M. T., Singh, S., and Guestrin, C., 2016. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Shortliffe, E. H. and Sepúlveda, M. J., 2018. "Clinical Decision Support in the Era of Artificial Intelligence". *JAMA* (320:21), pp. 2199–2200.

Stepin, I., Alonso, J. M., Catala, A., and Pereira-Fariña, M., 2021. "A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence". *IEEE Access* (9), p. 11974.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I., 2020. "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success". *npj Digital Medicine* (3), p. 17. pmid: 32047862.

Wang, M., Qiu, L., and Wang, X., 2021. "A survey on knowledge graph embeddings for link prediction". *Symmetry* (13:3), p. 485.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A., 2021. "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics". *Electronics* (10:5) (5), p. 593.