A Comparison of Causal Discovery and Explainable AI (XAI) for Image Datasets.

Atul Rawal[a], Adrienne Raglin[b], Qianlong Wang[a], and Ziying Tang[a]

[a,]Department of Computer & Information Sciences, Towson University, Towson, MD, USA.
[b]DEVCOM Army Research Laboratory, Adelphi, MD, USA

ABSTRACT

The recent push for fair, trustworthy, and responsible Artificial Intelligence (AI) and Machine Learning (ML) systems have pushed for more explainable systems that are capable of explaining their predictions/decisions and inner workings. This led to the field of Explainable AI (XAI) going through an exponential growth in the past few years. XAI has been crucial in making AI/ML systems more comprehensible. However, XAI is limited to the model that it is being applied to, for both post-hoc or transparent models. Even though XAI can explain the decisions being made by the ML systems, these decisions are based on correlation and not causation. For applications such as tumor classification in the medical field, this can have serious consequences as people's lives are affected. A potential solution for this challenge is the application of causal learning, which goes beyond the limitations of correlation for ML systems. Causal learning can generate analysis based on cause-and-effect relations within the data. This study compares the results of explanations given by post-hoc XAI systems to the causal features derived from causal graphs via causal discover for image datasets. We investigate how well XAI explanations/interpretations are able to identify the pertinent features within images. Causal graphs are generated for image datasets to extract the causal features that have a direct cause- and-effect relation with the label. These features are then compared to the features highlighted by XAI via feature relevance. The addition of causal learning for image datasets can aide in achieving fairness, bias detection, & mitigation to provide a robust and trustworthy system. We highlight the limitation of XAI tools such as LIME to make predictions based on physical features from images, whereas causal discovery can go beyond the simple pixel based perturbations to identify causal relations from image attributes.

Keywords: Causality, Causal Learning, Machine Learning, Artificial Intelligence,

1. INTRODUCTION

An increased utilization of artificial intelligence (AI) and machine learning (ML) systems within the last decade has given rise to concerns regarding their safety and trustworthiness. These systems which have been used in a plethora of applications, were found to be biased and unfair against different groups of populations. Examples include Amazon's hiring AI system which was shown to be biased against women, and META's advertising AI system which was biased against people of color and other minorities. These concerns have caused federal agencies, commercial companies and academic institutions to research and implement new methodologies and techniques that can ensure the safe use of these systems. Specific fields of research and development such as Responsible AI (RAI), bias identification/mitigation and fairness for AI/ML systems have seen tremendous growth. Towards this, causal learning and explainable AI (XAI) have been highlighted in the literature as potential solutions for ensuring robustness and trustworthiness of artificial reasoning systems.[1,2]

Even with the recent advancements in the fields of causal learning and XAI, challenges related to the limitations of current methodologies and modalities hinder the further advancement. While causal learning for experimental data via randomized control trials has been well-studied and perfected, generating causal relations between variables in observational datasets remains a challenge. The lack of ground-truth for most observational datasets presents a major hurdle for artificial reasoning systems. Due to these challenges most of the existing methodologies and techniques for both causal discovery and causal inference are limited to tabular data only. Even with a plethora of studies describing the use of causal discovery for different applications, there is a lack of studies highlighting the use of causal discovery with image datasets. A few studies within literature have highlighted the use of causal discovery for images.[3–5] However, extensive work is still needed to advance this research

forward. In this study we propose using causal learning and XAI to generate robust explanations by comparing and validating the causally relevant features with the results from the correlation-based model explanations.

Here, we propose the use of causal learning and XAI for image datasets using existing tools and methodologies. Since there are no available tools for extracting causal relations from observational image datasets, features/attributes can be extracted from the images and exported into a tabular format which allows for the application of existing causal learning tools. XAI can be readily applied to image classifiers based on the specific needs and applications. Even for more advanced, black-box deep learning image models, XAI tools such as DeepLIFT and SHAP are applicable to generate explanations for highlighting the features that have an impact on the image classification. The paper is organized as follows. Section 2 provides the background into causal learning via causal discovery, Section 3 provides an overview of XAI, Section 4 provides the proof-of-concept example, while Section 5 provides the results and discussion. Section 6 includes concluding remarks.

2. OVERVIEW OF CAUSAL LEARNING

Causality can be defined as the relationship between a cause and an effect.[6] Causal learning refers to the utilization of causality for AI/ML applications, where artificial reasoning models are capable of generating reasoning due to causation and not simple correlation. Here the causal learning portion of the reasoning involves highlighting the causal relations between the variables in the dataset. It can highlight the change in the model's prediction caused by a change in a feature due to modifications to another feature. Here the feature that is being modified is called the *treatment*, and the feature that is being investigated is called the *outcome*. Other features within the dataset that can cause a change in both the treatment and the outcome are called *confounders*, while background/noise variables are referred to as the *covariates*.

Causal relations between features in a dataset can be classified into three distinct categories referred to as the causal hierarchy: *association, intervention,* and *counterfactuals*.[7–9] The first level is called association and refers to the simple statistical relations between the variables. The second level, called intervention uses the causal structure of the variables to highlight the effect of changes to the treatment on the outcome. The last level of the causal hierarchy is called counterfactuals and it encompasses both association and interventions to derive the causal relations from the two former levels and make predictions based on unknown outcomes.

Causal learning for AI/ML systems can be done via two form of inquiry *causal discovery* and *causal inference*.[10,11] The first one, causal discovery is utilized to discover and highlight the causal relations that exist between the features in a dataset. The second one, causal inference can be used to investigate the extent to which a treatment can be modified to cause a change in the outcome. To investigate causal discover and inference two causal frameworks are available: *structural causal models (SCMs)* and *potential outcome framework*. Structural causal models utilize structural equations and causal graphs to derive a theory for causality.[6,7,12,13] Here the causal graphs are used to highlight the causal relations between the features via a directed graph which includes all the features in the dataset as individual nodes.[6] The potential outcome framework defined the potential outcome for an event as the outcome of the instance if the specific treatment was applied.[14] The estimation of the treatment effect on the outcome can be differentiated into three separate categories based on the population: *individual treatment effect* (ITE), *average treatment effect* (ATE), and *conditional average treatment effect* (CATE).[6,7,12,13]

3. OVERVIEW OF EXPLAINABLE AI (XAI)

With the tremendous advancements in the past decade for AI/ML systems, there has been an increased call to make them more interpretable and explainable. Even though the complexity of the models has increased, their explainability still remains a major challenge in achieving trustworthy AI. Especially for deep learning models which are inherently black-box, there is a need for explainable predictions. To address this challenge, the The U.S Defense Advanced Research Projects Agency (DARPA) started the Explainable AI (XAI) program in 2017. The program defined XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weakness, and convey an understanding of how they will behave in the future".[15] Here we provide a brief overview of XAI, for a more in-depth review of XAI, readers are encouraged to read the survey paper by authors on XAI.[1] Explainability has been described as a vital component for achieving accountable

and responsible AI where explainable systems can aid in identifying and mitigating bias from real-world data.[1] Madalina Busuioc listed explainability as one of the criteria for accountable AI:[1,16]

Multiple studies/tools/methodologies have been published to aid in achieving explainability for AI/ML models via transparency or post-hoc explanations. Algorithmic transparency makes the models inherently explainable, however for models that are not inherently transparent, post-hoc explanations can be utilized via tools such as LIME, DeepLIFT, SHAP, and DeepSHAP can be used.[17–20]

SHapley Additive exPlanations (SHAP) presented by Lundberg, et al., is one of the most commonly used post-hoc XAI framework for interpreting predictions. It generates feature importance scores for the model predictions based on the Shapley values by calculating a relevance score of the input features. These scores can be compared for individual features to investigate the impact each feature makes on the model predictions. Additionally, it also generates additive feature importance values to provide further insights into the relevance of the features.[17] DeepSHAP, presented by Chen et al., can be utilized for explaining complex deep learning models by layer wise propagation of the shapley values.[21]

Local Interpretable Mode-Agnostic Explanations (LIME) presented by Riberio et al., is another commonly used tool/library for generating robust explanations for classification models. It is a novel explanation method for generating explanations of classifiers by learning a local interpretable model around the classification.[18,19]

4. PROOF OF CONCEPT EXAMPLE

4.1 Data Source

The dataset for this study, the animals with attributes dataset was derived from The Institute of Science & Technology Austria's open-source data repository Works open-source data repository (https://cvml.ista.ac.at/AwA/). It consists of 37322 images of 50 different classes of animals with pre-extracted feature representations for each class. All the animal classes are characterized into 85 different attributes. Attributes that are shared across the different classes allow for transferring information between the classes. For the current study the dataset was divided into only two classes of Tiger and Zebra with 500 images for each class. The number of features/attributes was narrowed down from 85 to 17 as shown in (Fig 2).

4.1.1 Methods

As stated previously there is still a lack of existing tools and methodologies for extracting causal relations directly from images to perform causal discovery. To address this challenge we propose a workflow/framework of existing methodologies to generate causal relations via causal graphs from observational image datasets and comparing them to XAI based interpretations to compare if the attributes/features in both causal-based and model-based are the same (Fig 1). Mainly we propose to use existing tools/techniques in a workflow to generate causal graphs from observational image datasets. This consists of the following components:

- Attribute extraction from images.
- Conversion of extracted features to tabular data.
- Causal discovery on the tabular data.
- Comparison of causal features to the features from XAI.

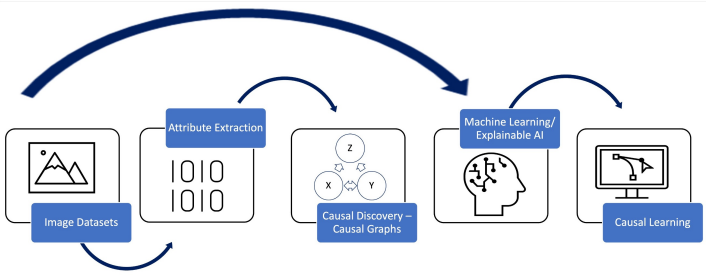

Figure 1: Proposed workflow for investigating causal relations in image datasets

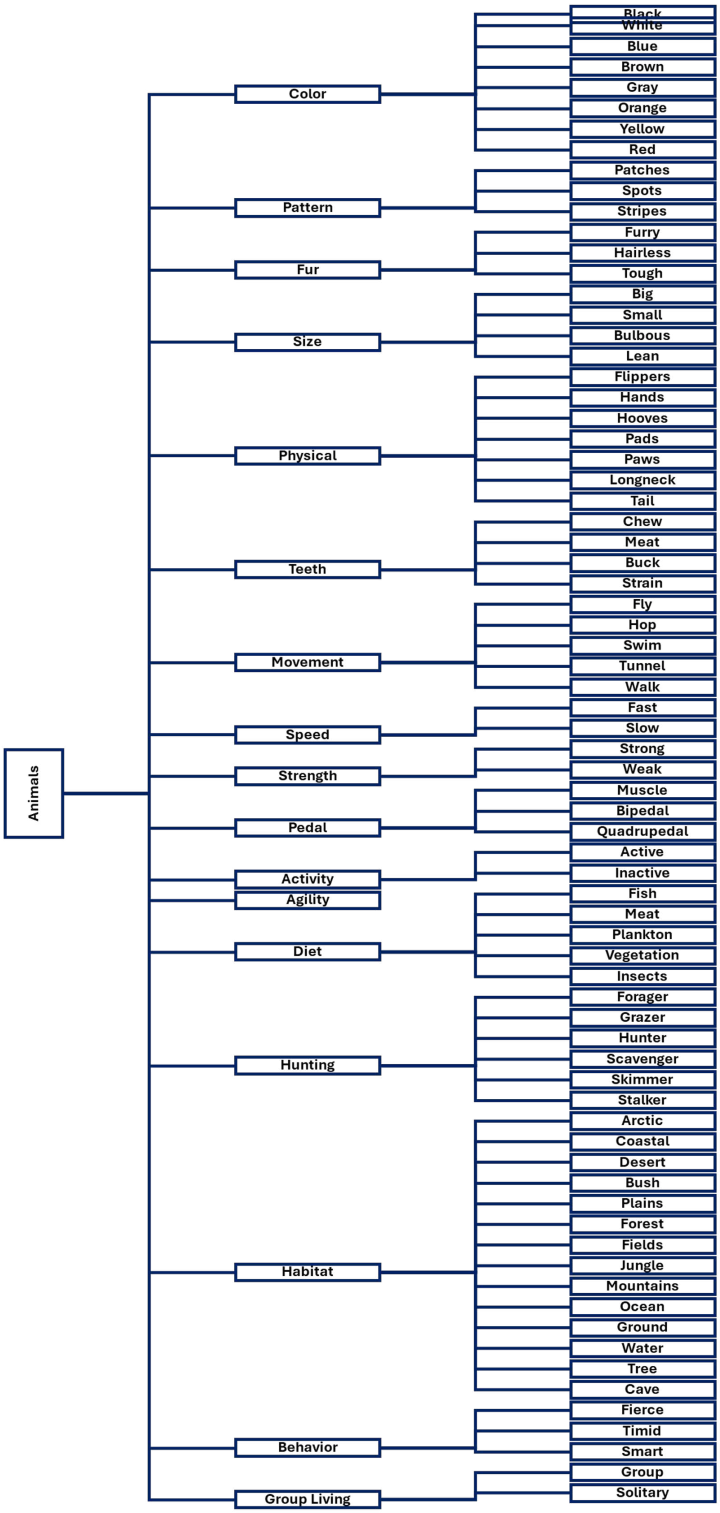

Figure 2: Extracted attributes for the images.

The first step of the proposed framework is attribute/annotation extraction from images. Attributes or annotations for images were extracted and exported in tabular format for investigating causal relations within the dataset. Attributes for each image/class are used to generate the causal images via existing causal discovery

tools. A skeleton graph from the raw data was generated to perform pairwise independence test. Directed causal graphs with causal relations between the different variables are generated using the Peter-Clark (PC) algorithm. Once the causal graphs have been generated, they can be compared to the features highlighted via XAI for the machine learning explanations.

Once the causal graphs are created, the image dataset is applied to a deep learning image classifier. For this study we utilized two TensorFlow-Keras sequential models for binary classification with different parameters and number of layers to choose the optimal model to apply explainability. The first model had 3 layers; keras, dropout and dense. with 2260546 parameters. The second model had 10 layers; rescaling, conv2d_1, max pooling_1, conv2d_2, max pooling_2, conv2d_3, max pooling_3, flatten, dense_1, and dense_2 with 6446498 parameters. Both models were trained for 50 epochs. The finalized dataset was then randomly split for training and testing at a 80%-20% split with 80% of the dataset used for training the model and 20% used for validation. The top performing Keras model was then used to apply XAI for further evaluation via LIME. Other libraries such as Lime, Matplotlib, sklearn, and PIL, were installed directly in the notebook using pip. Once the training and validation were completed the models were evaluated using performance metrics of accuracy, precision, recall and F1-score.

- Accuracy - Accuracy describes the number of correct predictions over all predictions.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

- Precision – Precision measures how many of the positive predictions made are correct.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

- Recall - Recall measures how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

- F1 Score - F1-Score combines both precision and recall and is the harmonic mean of the two. It is between 0 and 1, where 0 is the worst score and 1 indicates that the model predicts each observation correctly.

$$F1 - score = 2 * \frac{Precision \times Recall}{Precision + Recall}.$$

The top performing Keras model was chosen for further explainablity analysis using LIME. This allowed for the investigation of the features relevant to identify the image as either a Tiger or a Zebra. As mentioned in the earlier section LIME is an open-source post-hoc explanation method used for explaining ML and DL models, where explanations can be provided by perturbing the input images and highlighting the parts of the image that contribute most towards a specific classification. For causal discovery, the causal graphs were generated from tabular with the finalized features using the Causal Discovery Toolbox (CDT). The directed causal graphs with causal relations between the different variables can be compared to the correlation-based ML classifiers via LIME.

5. RESULTS & DISCUSSION

The performance metrics for both the classification models are - highlighted in Table 1. Even though neither of the classifiers were able to achieve a perfect score of 1 for the performance metrics, the Seq-10 Layer model with a 0.97 F1-score was deemed good for this proof-of-concept study.

Model	Accuracy	Precision	Recall	F1-Score
Seq - 3 Layers	0.56	0.52	0.52	0.52
Seq - 10 Layers	0.99	0.98	0.97	0.97

Table 1: Performance metrics for both Keras models

Upon completion of the model training, testing and evaluation, the best performing model (Seq-10 layers) was chosen for further analysis using XAI. Perturbation heat-maps were generated using LIME for the deep learning model. Causal graphs were generated via causal discovery to highlight the causally relevant features in the images that had a causal relation with the classification. Figure 3 shows the causal graph with the features that have a direct causal relation with the classification.

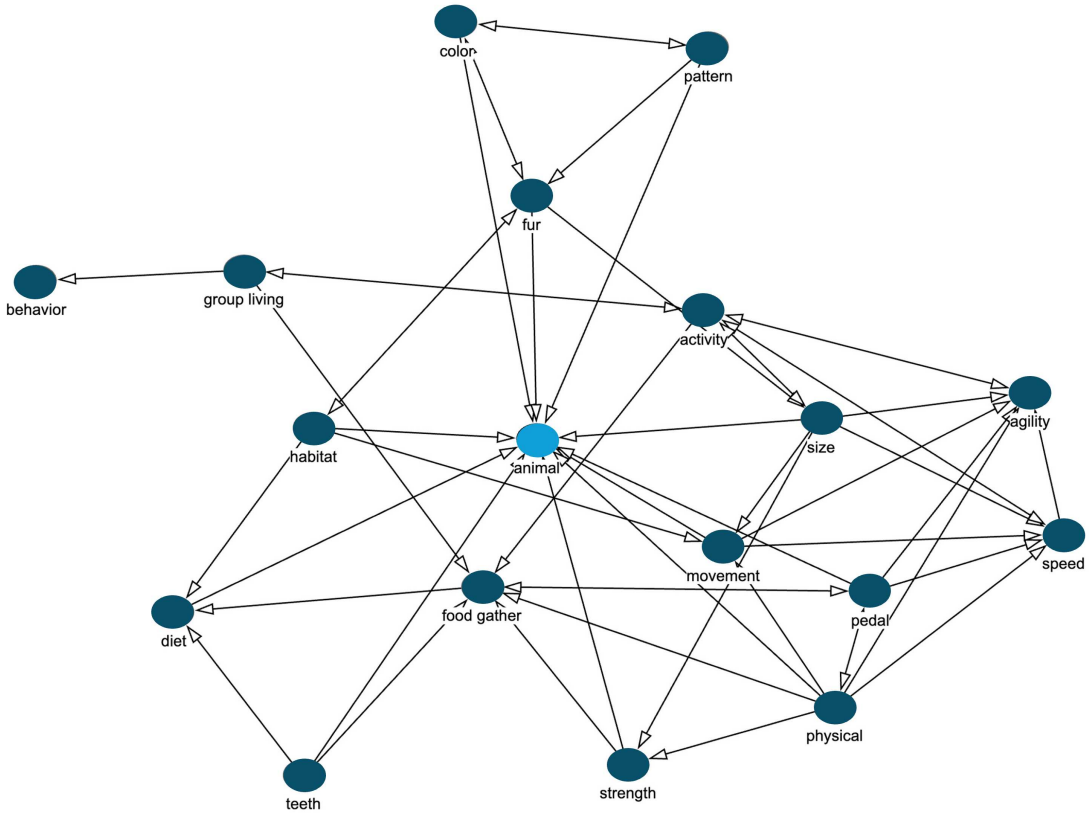

Figure 3: Causal graph for the animal classification.

The generated causal graph yielded eleven features with a direct causal relation to the animal classification; a) pattern, b) color, c) fur, d) movement, e) size, f) strength, g) teeth, h) diet, i) habitat, j) pedals k) physical. These eleven features can be ascribed to having a direct impact on the classification of an animal. While the other features still play a role in the classification, these features play a direct role. The graph highlighted the indirect causal relation to the classification for food gather and activity as food gather impacts the diet, which impacts the classification. Similarly activity impacts the size which impacts the classification. Activity had a

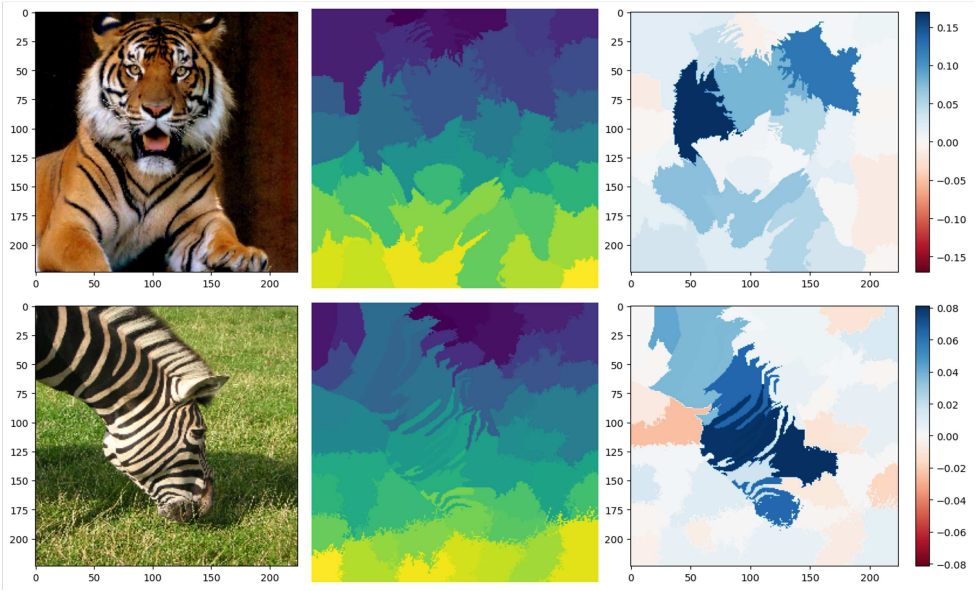

Figure 4: Explanation heat map for an example image of Tiger & Zebra using LIME.

bi-directional relation with agility and group living. From the graph we can deduce that eleven features have a direct impact on the classification while agility, activity, behavior, speed, food gather, and group living had a lower impact. The perturbation-based feature highlights from LIME for the Seq-10 layer model highlighted the head, pattern, color, and fur as the most impactful features for the classification of both the Tiger and Zebra example as shown in Figure 4. The perturbation based heatmap explanations from LIME also highlight some of the limitations of the correlation based AI/ML models for image classifications these models do not provide information on any non-physical attributes for images. For example, while the LIME explanation for the zebra image highlights the color, pattern, fur and head of the zebra in the heatmap other attributes such as the diet, speed, strength, and size cannot be highlighted since they are not physically presented in the images. For the Tiger image, even though the tiger's paws were in the picture along with the teeth, these were not highlighted in the heatmap to be as impactful as the head of the tiger. A potential solution towards this challenge is the presence of a human-in-the-loop when interpreting and analyzing the explanations from XAI based tools. With a human-in-the-loop, a more robust and comprehensive analysis can be achieved for image datasets by comparing/contrasting the causal and correlation based features.

Even though the current study is intended to serve as a simple-proof-of concept, it highlights the functionality and effectiveness of existing tools to highlight causal relations in image datasets for artificial reasoning systems to achieve "human-like" intelligence. The utilization of causal learning can provide an additional layer of robustness for the explanations generated for the image classifiers via XAI tools such as LIME and SHAP. This can aid in ensuring the trustworthiness and fairness of the AI/ML systems by helping them go beyond simple correlation and achieve causal reasoning.

In this paper we have provided a framework to extract causal relations from observational image datasets. The addition of causal learning for artificial reasoning provides a more robust AI/ML model where trustworthiness is achieved. While traditional AI/ML models are progressing towards achieving robustness and trustworthiness, sometimes correlation based explanations are insufficient. Therefore, the extra layer of robustness from the causally relevant features can be a simple yet effective method to ensure trustworthiness. We presented a simple and efficient yet novel method to ensure the use of observational data for trustworthy AI/ML systems. Due to the lack of large observational datasets with available ground-truth, causal graphs generated via causal discover can play a vital role in generating robust and trustworthy explanations.

6. CONCLUSION

There has been a recent rise in interest in fair, trustworthy and responsible AI/ML systems due to new laws/policies such as the Executive Order 13960 for Safe and Trustworthy AI. This in-turn has caused for calls to have AI/ML models be explainable and interpretable. While XAI has been crucial in making AI/ML systems be more interpretable, explainable and transparent, there are still challenges that need to be addressed. These include the lack of XAI systems to generate explanations based on causal-relations, which are crucial for achieving human-like reasoning. To address this issue, causality can be utilized to provide an added layer of robustness to ensure AI/ML models make predictions based on causal reasoning. The addition of causal learning for image datasets can aid in achieving fairness, bias detection, and mitigation to provide a robust and trustworthy system. This paper provides a proof-of-concept study to highlight the use of causal learning in addition to XAI for image classification models. We highlight the limitation of XAI tools such as LIME in making predictions based on physical features from images, whereas causal discovery can go beyond the simple pixel based perturbations to identify causal relations from image attributes.

ACKNOWLEDGMENTS

This research was conducted in collaboration with the Army Research Laboratory under CRADA Number 23-025 with the U.S. Army Research Laboratory. However, any opinion, finding, and conclusions or recommendations expressed in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the CRADA agency.

REFERENCES

[1] Rawal, A., Mccoy, J., Rawat, D. B., Sadler, B., and Amant, R., "Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives," *IEEE Transactions on Artificial Intelligence* **1**(01), 1–1 (2021).

[2] Rawal, A., Raglin, A., Rawat, D. B., and Sadler, B. M., "Causality and Machine Learning Review," tech. rep., DEVCOM Army Research Laboratory (07 2022).

[3] Castro, D. C., Walker, I., and Glocker, B., "Causality matters in medical imaging," *Nature Communications* **11**(1), 3673 (2020).

[4] Chalupka, K., Perona, P., and Eberhardt, F., "Visual causal feature learning," *arXiv preprint arXiv:1412.2309* (2014).

[5] Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., and Bottou, L., "Discovering causal signals in images," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 6979–6987 (2017).

[6] Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H., "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)* **53**(4), 1–37 (2020).

[7] Pearl, J., "Causal inference in statistics: An overview," *Statistics surveys* **3**, 96–146 (2009).

[8] Pearl, J., "Theoretical impediments to machine learning with seven sparks from the causal revolution," *arXiv preprint arXiv:1801.04016* (2018).

[9] Pearl, J., "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM* **62**(3), 54–60 (2019).

[10] Gelman, A., "Causality and statistical learning," (2011).

[11] Peters, J., Janzing, D., and Schölkopf, B., [*Elements of causal inference: foundations and learning algorithms*] (2017).

[12] Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A., "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)* **15**(5), 1–46 (2021).

[13] Pearl, J., [*Causality*], Cambridge university press (2009).

[14] Imbens, G. W. and Rubin, D. B., [*Causal inference in statistics, social, and biomedical sciences*], Cambridge University Press (2015).

[15] Gunning, D. and Aha, D., "Darpa's explainable artificial intelligence (xai) program," *AI magazine* **40**(2), 44–58 (2019).

[16] Busuioc, M., "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review* **81**(5), 825–836 (2021).

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

[18] Ribeiro, M. T., Singh, S., and Guestrin, C., "" 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*], 1135–1144 (2016).

[19] Ribeiro, M. T., Singh, S., and Guestrin, C., "Nothing else matters: model-agnostic explanations by identifying prediction invariance," *arXiv preprint arXiv:1611.05817* (2016).

[20] Shrikumar, A., Greenside, P., and Kundaje, A., "Learning important features through propagating activation differences," in [*Proceedings of the 34th International Conference on Machine Learning*], Precup, D. and Teh, Y. W., eds., *Proceedings of Machine Learning Research* **70**, 3145–3153, PMLR (06–11 Aug 2017).

[21] Chen, H., Lundberg, S., and Lee, S.-I., "Explaining models by propagating shapley values of local components," *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* , 261–270 (2021).