REGRESSION VIA CAUSALLY INFORMED NEURAL NETWORKS

Shahenda Youssef[1], **Frank Doehner**[1], **and Jürgen Beyerer**[1,2]
[1]Karlsruhe Institute of Technology KIT, Karlsruhe, Germany
[2]Fraunhofer IOSB, Fraunhoferstraße 1, Karlsruhe

ABSTRACT

Neural Networks have been successful in solving complex problems across various fields. However, they require significant data to learn effectively, and their decision-making process is often not transparent. To overcome these limitations, causal prior knowledge can be incorporated into neural network models. This knowledge improves the learning process and enhances the robustness and generalizability of the models. We propose a novel framework RCINN that involves calculating the inverse probability of treatment weights given a causal graph model alongside the training dataset. These weights are then concatenated as additional features in the neural network model. Then incorporating the estimated conditional average treatment effect as a regularization term to the model loss function, the potential influence of confounding variables can be mitigated, leading to bias minimization and improving the neural network model. Experiments conducted on synthetic and benchmark datasets using the framework show promising results.

Keywords Neural Networks · Causal graphs · Prior knowledge · Causal inference · Propensity score weighting · Regression

1 Introduction

Despite the promising results of Neural Networks (NNs) across various fields, they still face unresolved challenges [17]. A key issue is their limited performance when there is a lack of training data, which affects their ability to generalize [27]. Additionally, the black-box nature of NNs prevents a precise explanation of their mechanism [27]. While they excel at uncovering concealed features and their co-occurrences within the input data, they struggle to uncover and clarify any underlying causal relationships between those features. To address these challenges, it's critical to incorporate causality into machine learning frameworks [26, 19].

Incorporating NN models with causality enhance their robustness and ability to generalize [27]. Furthermore, such frameworks can accurately model shifts in data distributions by concentrating on causal relationships, which are based on a sequence of cause and

effect rather than correlation, which simply observes patterns without implying direction. Therefore, causality provides NN with the capability to properly reason beyond its training data [26].

The adoption of deep learning in manufacturing systems is still in its early stages [15]. This could be attributed not only to its exclusive reliance on data-driven techniques but also to the lack of research conducted on incorporating domain experts' causal knowledge into deep learning models. Recent research is focused on improving model performance and explainability by integrating prior knowledge into the learning process [6, 1].

The methodologies to incorporate prior knowledge within NNs are varied and innovative [31, 5]. This involves embedding prior knowledge into NNs by transforming the input data [12, 9], imposing informed constraints on the loss function to direct the optimization process towards solutions that respect established relationships and theoretical frameworks [20, 8], and the integration into the architecture of NNs itself or constraining model parameters [3, 12]. The ongoing studies aim to combine different methods to ensure that the incorporated prior knowledge effectively guides the learning process, while still allowing the NNs to uncover new insights based on the data. However, the development of models that incorporate causal prior knowledge continues to be a challenge [27].

The subsequent sections of the paper are structured in the following manner: Section 2 provides an overview of causal inference, and section 3 addresses related work. In section 4 we present the proposed framework, followed by the experimental results in section 5. We conclude with section 6 outlining some major unsolved challenges and provide insight into potential directions for future research.

2 Causal Inference

Causal inference determines cause-and-effect relationships between variables based on observational data. It aims to estimate the causal effects of a specific treatment or intervention on an outcome of interest while considering potential factors that may introduce bias or influence the relationship [24].

2.1 Causal Graphs

Causal graphs, that do not contain any cycles between variables, can be represented as Directed Acyclic Graphs (DAGs). Such a DAG can be defined as $G = (V, E)$, where the causal graph G depicts the causal relationships between variables, nodes, V, with directed edges E between the nodes depicting the direction of cause and effect [32].

When estimating causal effects, it is important to consider two types of variables: Confounders and instrumental variables (IVs) [22]. As shown in figure 1, a confounder $C \in V$ is a variable that influences both the treatment $X \in V$ and the outcome $Y \in V$, which implies that any observed correlation between the treatment and the outcome might be due to the confounder's spurious correlations rather than a causal relationship. On the other hand, the instrumental variable $Z \in V$ is a variable that only affects the treatment and is not directly linked to the outcome.

2.2 Conditional Average Treatment Effect

Conditional Average Treatment Effect (CATE) measures the causal effect of the treatment on the outcome, conditioned on confounding variables or covariates [18]. The backdoor

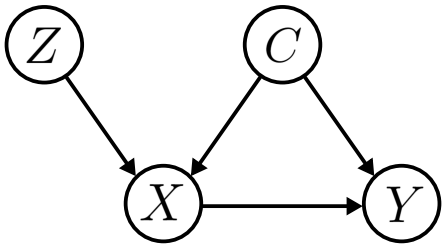

Figure 1: Causal Graph, $X \to Y$ indicates that the treatment node X is the cause of the outcome node Y. Node C is a confounding variable with an effect on X and Y, and Node Z is an instrumental variable with effect on X.

criterion introduced by Pearl [24] controls for confounding variables to obtain a more precise estimation of the causal effect. It identifies and blocks potential paths that might introduce bias between the treatment and outcome variables. The CATE is then defined by,

$$\tau_{\text{CATE}} = \mathbb{E}[Y \mid \text{do}(X = x)] = \mathbb{E}_C \mathbb{E}[Y \mid X = x, C = c], \tag{1}$$

where $\mathbb{E}[Y|\,\text{do}(X = x)]$ represents the expectation of the outcome Y under the intervention $\text{do}(X = x)$. The do-operator allows for the identification and estimation of causal effects from observational data under certain conditions. C represents the set of confounders of variables X and Y.

The instrumental variable method [10] estimates the effect of the instrument on the treatment and the outcome to estimate the effect of the treatment on the outcome. The CATE is then defined by,

$$\tau_{\text{CATE}} = \mathbb{E}[Y \mid \text{do}(X = x)] = \frac{\mathbb{E}[Y \mid Z = z]}{\mathbb{E}[X \mid Z = z]}\,, \tag{2}$$

where Z represents the set of IVs. It helps in identifying causal effects where the backdoor method fails [18].

2.3 Inverse Probability Weighting

Inverse Probability Weighting (IPW) [22] is a statistical technique that can reduce bias and provide more accurate estimates of causal effects. It involves assigning weights for each observation based on their estimated propensity scores $e(C)$, which represent the probability of receiving a particular treatment given a set of observed confounders, and it is represented as

$$e(C) = P(X = x \mid C = c). \tag{3}$$

IPW is a method for controlling the confounders by constructing pseudo-populations within the data and weighing them based on the inverse of their propensity scores α_{IPW} in order to decounfounded the data

$$\alpha_{\text{IPW}} = \frac{1}{P(X \mid C)}\,. \tag{4}$$

Pseudo-populations are created by upweighting the underrepresented and downweighted the uprepresented groups in the dataset [18]. This weighting aims to balance the distribution of covariates between the treatment and untreated groups, making them more comparable, and thereby reducing potential confounding effects.

3 Related Work

Deng et al. [7] present a deep learning framework for forecasting societal events that leverage causal inference. Their approach utilizes Individual Treatment Effects (ITE) to estimate the impact of different treatments or events on societal outcomes within spatiotemporal environments. The model integrates causal information into event predictions, by predicting potential outcomes for different treatment scenarios.

The work by Richens et al. [25] discusses the use of Structural Causal Models (SCM) as a means to encode the relationships between diseases, symptoms, and risk factors, enabling more accurate diagnostic reasoning. The authors develop algorithms that prioritize causal inference and highlight the importance of addressing confounding issues and the necessity of causal knowledge in the diagnostic process.

Kyono et al. [13] show how causal graphs can be used as prior knowledge to improve model selection and enhance the reliability of NN performances. They propose incorporating this knowledge into a Structural Causal Model to calculate a score that evaluates how well a model's predictions align with the SCM and input variables.

Teshima et al. [30] introduce a model-independent method for data augmentation that leverages the conditional independencies relations in the data distribution, encoded in causal graphs, to enhance supervised learning.

In a recent study Terziyan et al. [29] proposed a novel framework for enhancing Convolutional Neural Networks (CNNs) by incorporating causality-awareness into their architecture. An additional layer of neurons is introduced to the architecture that is specifically designed to estimate asymmetric causality in images by leveraging convolutional layers to extract features from images and then using these features to estimate conditional probabilities, effectively improving image classification and generation.

The expanding number of research emphasizes the importance of integrating causal regularization strategies into the framework of predictive modeling to address the issue of confounding variables in causal inference. By regularizing the model to consider causal relationships, it can provide more reliable and interpretable results [11, 14].

4 Method

We propose the novel framework RCINN that combines the strengths of causal inference and neural networks in regression tasks. The initial step involves encoding domain knowledge into a causal graph. If a causal graph is unavailable, causal discovery methods can be utilized to learn it from the data [16]. However, in this work, we are assuming that the causal graph is known.

Figure 2 shows the structure of our proposed, causally informed neural network framework RCINN for integrating causal prior knowledge into a neural network. Besides the usual training data $(\mathcal{D}, Y)$, where $\mathcal{D}$ are input features and Y are the true labels, the framework leverages additional prior knowledge from an independent source given by the DAG G.

In order to identify the CATE using either the backdoor criterion or the IV method discussed in 2.2, by using the Dowhy library for causal inference from Microsoft [28] to conduct an analysis of the given causal graph G regarding confounders and IVs.

Each observation in the training dataset is assigned a weight based on its corresponding inverse probability weight α_{IPW} as shown in equation 4. This weighting approach

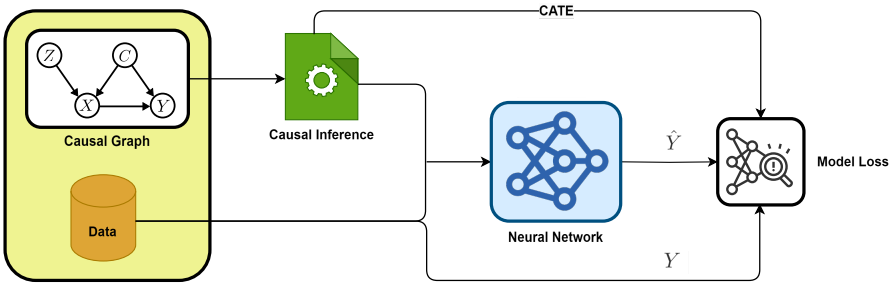

Figure 2: The proposed causally informed neural network framework RCINN.

prioritizes observations that are less likely, based on covariates, to receive the treatment they actually received. Adding these weighted feature values as an additional feature to the training dataset generates a new input features $\mathcal{D}_{\text{new}}$ that implies increasing the initial weights of the features with causal relationships when inputting them into the neural network model.

To estimate the causal effect τ_{CATE}, if the graph contains confounders, the CATE is estimated using the backdoor equation 1. On the other hand, for IVs, equation 2 is used. We scale the outcome by IPW to get an unbiased non-parametric CATE estimator

$$\tau_{\text{CATE}} = \mathbb{E}[\alpha_{\text{IPW}} \cdot \tau_{\text{CATE}}]. \tag{5}$$

To make sure that we have controlled for variables that contribute to bias. By using one of the refutation tests from Dowhy which is used to validate the causal estimates. It adds randomly generated covariates to the data, then reruns the estimator to return a tested causal estimate β and check if the causal estimate changes or not. The robust causal estimate should not change much with a small effect of the unobserved common cause. The neural network model is then trained with the new input features, by learning a function from the data $(\mathcal{D}_{\text{new}}, Y)$. The loss function of the regression neural network learning model is then computed as

$$\mathcal{L}_{\text{Pred}} = \mathcal{L}_{(Y,\hat{Y})} + \lambda \cdot W^2, \tag{6}$$

where $\mathcal{L}_{(Y,\hat{Y})}$ is the label-based loss that can be represented by the mean squared error, $Y, \hat{Y}$ are the actual labels and predicted values respectively. $\lambda \cdot W^2$ is the L2 regularization function used to control model complexity. It adds the squared values of the model weights W to the loss function and thereby encourages smaller weights and helps to prevent overfitting.

On top of the predictive loss, we added a regularization term causal loss $\mathcal{L}_{\text{Causal}}$ that penalizes the deviations from the causal estimate during neural network training. This is done by squaring the difference between the predicted outcomes and the estimated causal effect. The causal loss encourages the neural network to make predictions that are consistent with the causal relationships and robust to unobserved confounding, and it is defined as

$$\mathcal{L}_{\text{Causal}} = (\hat{Y} - \tau_{\text{CATE}})^2 + (\tau_{\text{CATE}} - \beta)^2, \tag{7}$$

By incorporating the causal loss term $\mathcal{L}_{\text{Causal}}$, we can mitigate the potential influence of confounding variables that may affect both the treatment and the outcome. We train the NN model by minimizing the following loss function

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Pred}} + \mu \cdot \mathcal{L}_{\text{Causal}}, \tag{8}$$

by multiplying $\mathcal{L}_{\text{Causal}}$ with the regularization hyperparameter μ, we control the strength of the regularization term. Higher values of μ increase the regularization effect, encouraging the model to respect the causal graph more strongly.

5 Experiments

We evaluate our method on both linear and non-linear regression datasets. The datasets include Linear, a linear synthetic dataset with three confounders and two IVs. Nonlinear, a non-linear synthetic dataset with three confounders. A benchmark Infant Health and Development Program (IHDP) dataset [28] with 25 confounders.

For a fair comparison, we adopt the same training setting and the same NN architecture for all models. For each task, we report the model loss with a standard deviation calculated over 5 different testing data. The data is divided into training, validation, and testing sets with proportions of 60%, 10%, and 30% respectively. The neural network architecture consists of two layers, with the first layer having 128 neurons and the second layer having 64 neurons. The output layer is the final layer of the neural network. During training, we adopt the Adam optimizer [2]. The training batch size is 32. We set λ as 0.1 and μ as 0.01 for all datasets.

Table 1: Performance of Regression tasks on three different datasets: Linear, Nonlinear, and IHDP. RF represents a Random Forest algorithm, Baseline_NN model is the neural network without causality and RCINN utilizes causal prior knowledge.

Datasets	RF	Baseline_NN	RCINN
Linear	0.227 ± 0.006	0.425 ± 0.006	$\mathbf{0.152 \pm 0.002}$
Nonlinear	34.853 ± 0.759	25.329 ± 0.593	$\mathbf{20.913 \pm 0.145}$
IHDP	1.95 ± 0.43	2.52 ± 0.11	$\mathbf{1.7 \pm 0.14}$

The results are reported in Table 1. We observe that our method demonstrates promising performance on both the synthetic datasets as well as on the benchmark dataset.

Overall, this approach allows us to leverage the strengths of both causal inference and neural networks, leading to a reduction in bias and an improvement in the performance of the neural network model.

6 Future Prospects

Much potential lies in exploring new combinations of approaches, that have not yet been investigated. One such example is merging causal prior knowledge with the architecture of NNs using the attention mechanism [23]. This allows the model to iteratively process knowledge by selecting only relevant content in each step. The inclusion of a knowledge-based attention layer improves prediction and overall model performance.

Another promising framework involves integrating a causal graph model in the form of an embedding graph layer, which can then be used as input for Bayesian Neural Networks (BNNs) [21]. This integration aims to enhance the incorporation of causal prior knowledge by refining the prior distribution during model training. Such a probabilistic approach takes the uncertainties in the model's predictions into account, recognizing that understanding the confidence level of a prediction is equally important as the prediction itself.

Recent research primarily focuses on data-driven approaches that assume independent and identically distributed (IID) data. However, when working with spatiotemporal data that deviates from this IID assumption, it becomes challenging [4]. Future work will be directed towards incorporating causal models capable of handling highly correlated values over time.

Acknowledgments

This work is funded by Research Group DFG FOR 5339.

References

[1] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Rueden. Explainable machine learning with prior knowledge: an overview. *arXiv preprint arXiv:2105.10172*, 2021.

[2] S. Bock and M. Weiß. A proof of local convergence for the adam optimizer. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[3] A. Borghesi, F. Baldo, and M. Milano. Improving deep learning models via constraint-based domain knowledge: a brief survey. *arXiv preprint arXiv:2005.10691*, 2020.

[4] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[5] T. Dash, S. Chitlangia, A. Ahuja, and A. Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, 2022.

[6] A. Daw, A. Karpatne, W. D. Watkins, J. S. Read, and V. Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. In *Knowledge Guided Machine Learning*, pages 353–372. Chapman and Hall/CRC, 2022.

[7] S. Deng, H. Rangwala, and Y. Ning. Causal knowledge guided societal event forecasting. *arXiv preprint arXiv:2112.05695*, 2021.

[8] E. Gallup, T. Gallup, and K. Powell. Physics-guided neural networks with engineering domain knowledge for hybrid process modeling. *Computers & Chemical Engineering*, 170:108111, 2023.

[9] M. Gaur, K. Faldu, and A. Sheth. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing*, 25(1):51–59, 2021.

[10] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Boca Raton: Chapman Hall/CRC, 2020.

[11] L. Kania and E. Wit. Causal regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees. *arXiv preprint arXiv:2205.01593*, 2022.

[12] S. W. Kim, I. Kim, J. Lee, and S. Lee. Knowledge integration into deep learning in dynamical systems: an overview and taxonomy. *Journal of Mechanical Science and Technology*, 35:1331–1342, 2021.

[13] T. Kyono and M. van der Schaar. Improving model robustness using causal knowledge. *arXiv preprint arXiv:1911.12441*, 2019.

[14] T. Kyono, Y. Zhang, and M. van der Schaar. Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

[15] R. Malhan and S. K. Gupta. The role of deep learning in manufacturing applications: Challenges and opportunities. *Journal of Computing and Information Science in Engineering*, 23(6), 2023.

[16] D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

[17] G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[18] A. Molak. Causal inference and discovery in python. *Small*, 344:344, 2023.

[19] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[20] N. Muralidhar, M. R. Islam, M. Marwah, A. Karpatne, and N. Ramakrishnan. Incorporating prior domain knowledge into deep neural networks. In *2018 IEEE international conference on big data (big data)*, pages 36–45. IEEE, 2018.

[21] K. P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[22] B. Neal. Introduction to causal inference. *Course Lecture Notes*, 2020.

[23] Z. Niu, G. Zhong, and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[24] J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[25] J. G. Richens, C. M. Lee, and S. Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

[26] B. Schölkopf. Causality for machine learning. *Probabilistic and Causal Inference*, 2022.

[27] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[28] A. Sharma and E. Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[29] V. Terziyan and O. Vitko. Causality-aware convolutional neural networks for advanced image classification and generation. *Procedia Computer Science*, 217:495–506, 2023.

[30] T. Teshima and M. Sugiyama. Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation. In *Uncertainty in Artificial Intelligence*, pages 86–96. PMLR, 2021.

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

[32] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.