The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 134-140
DOI 10.37547/tajet/Volume07Issue04-18

Combining causal analysis and machine learning to predict the effects of interventions

Ivan Kitov

Senior Data Scientist, Wolt Berlin, Germany.

OPEN ACCESS

SUBMITTED 18 February 2025
ACCEPTED 21 March 2025
PUBLISHED 28 April 2025
VOLUME Vol.07 Issue 04 2025

CITATION

Ivan Kitov. (2025). Combining causal analysis and machine learning to predict the effects of interventions. The American Journal of Engineering and Technology, 7(04), 134–140. https://doi.org/10.37547/tajet/Volume07Issue04-18

Abstract: This paper examines the integration of causal analysis (causality) and machine learning methods to accurately predict the effects of interventions. The first part introduces the rationale for the importance of the causal approach when classical statistical models and purely associative ML methods face problems of hidden factors and incorrect extrapolation of results. The second part discusses the basic theoretical concepts of causal graphs, do-operator, intervening and counterfactual distributions, and the role of identifiability assumptions in the presence of unobserved confounders. Next, methods for integrating causality and machine learning - causal supervised learning (to deal with spurious correlations and increase robustness to distributional shifts), causal generative modeling (with a focus on generating counterfactual data), and other state-of-the-art approaches (causal model explanation, causal fairness, causal reinforcement learning) - are discussed in detail. It is shown how such methods can better account for the real-world structure of the data and produce more reliable predictions, especially in heterogeneous environments. The results can be applied to medicine, economics, social sciences, and other fields where it is important to accurately predict the effects of potential interventions.

Keywords: causal analysis, machine learning, causal graphs, hidden confounders, interventions, counterfactual reasoning, invariant risk minimization, causal data generation.

Introduction: Over the past decade, there has been a rapid rise in interest in combining causal inference (causality) with machine learning (ML) methods to build predictive models for interventional effects [6, 12].

Conventional statistical approaches, which rely predominantly on correlation, often encounter challenges when trying to isolate genuine causal mechanisms or reliably extrapolate results to new conditions [14]. This becomes particularly evident in tasks where the data distribution can shift significantly: in such cases, predictions grounded in purely associative patterns tend to perform poorly in new environments [1].

Classical statistical methods—those based on conditional probabilities and linear regression—commonly overlook intricate structural (causal) relationships among variables [12]. For example, in evaluating how a medical therapy affects different patient groups, hidden confounding may arise, where unmeasured factors distort the true effect of an intervention [6]. Similar constraints are found in purely associative (or "black-box") ML models, where algorithms attempt to "fit" dependencies between features and the outcome variable without distinguishing genuine causal factors from spurious correlations [2]. Consequently, such models are at risk when transferred to new, out-of-sample conditions [15].

Meanwhile, methods from causal inference offer tools for formally identifying and estimating causal effects, enabling a clearer interpretation of observed data and stronger predictions for potential interventions. As a result, there is growing demand to integrate ML approaches with causal models. The aim of this article is to demonstrate how such integration improves the accuracy and robustness of predictions, especially when data are heterogeneous or external conditions vary significantly. This is particularly relevant in medicine, socio-economic analysis, and applied research, where the task is not just to predict an outcome but to understand how that outcome might change under an intervention (e.g., modifying the dosage of a drug or adjusting social policy).

In the literature on causal modeling, one finds a wide range of methods—spanning from Bayesian networks and structural equation approaches [12, 14] to cutting-edge hybrid techniques that merge deep neural networks with elements of causality [10]. Some studies focus on theoretical foundations, highlighting the do-operator, counterfactual reasoning, and identifiability issues [6, 12]. Others present practical applications in medicine, economics, and the social sciences [2, 15].

A particularly comprehensive overview of current developments in this field highlights several key directions:

- Causal supervised learning: Pursuit of invariant features and stable models that better handle distributional shifts in the data [1].
- Causal generative modeling: Generating data in a way that accounts for causal mechanisms, which enables realistic counterfactual examples and strategies for mitigating "spurious" patterns.
- Causal explanations: Interpreting complex ML models by pinpointing causally significant factors.
- Causal reinforcement learning: Enhancing reinforcement learning through structural (causal) dependencies.
- Causal fairness: Defining fairness in causal terms to prevent discrimination in automated decisions.

According to these overviews, there remain various open challenges—for instance, detecting hidden confounders, limited avenues to test causal hypotheses using real-world data, and the difficulty of integrating powerful neural networks with explicit causal graphs [7]. There is also a lack of benchmark datasets for validating causal models in different applied fields [15].

The novel contribution of this paper lies in an attempt to systematize contemporary solutions, bring them in line with the existing body of work on causal inference, and illustrate their practical use. First, we clarify and extend known methods for effect estimation by leveraging the potential of structural causal models (SCMs). Second, we synthesize progress in the domain of generative counterfactual techniques, which offer a deeper understanding of how possible interventions can influence outcomes [10]. To our knowledge, this is the first work to comprehensively review and unify these diverse threads of causal and machine learning research with an eye towards intervention outcome prediction.

1. Theoretical Foundations of Causal Inference and Machine Learning

Causal inference (causality) is closely interwoven with both classical and modern machine learning (ML) methods, enriching them with tools for uncovering not just associative but genuinely causal relationships in data [7, 12]. Below is a concise overview of the key concepts needed to understand and develop predictive models for interventional effects.

We begin by examining the basics of Bayesian networks

and causal graphs. A Bayesian network is defined by a directed acyclic graph (DAG), where each node X_i is connected to its parent nodes $Pa(X_i)$, encoding a factorization of the joint distribution:

$$p(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} p\big(X_i \mid Pa(X_i)\big)$$

where $X_1, X_2, \dots, X_n$ are the model variables and $Pa(X_i)$ is the set of parents of Xi in the graph. However, such a graph often captures only the structure of conditional dependencies, without guaranteeing a causal interpretation [14]. To move toward causal graphs, one adopts the perspective of structural causal models (SCMs) [12], in which each variable X_i is generated not only by $Pa(X_i)$ but also by its own "noise" (exogenous) factor εi. This framework helps describe the "mechanisms" of data generation by assuming that each node—either deterministically or stochastically—follows an equation of the form

$$X_i = f_i(Pa(X_i), \varepsilon_i)$$

A focal concept in causal analysis is the so-called do-operator, which enables modeling of interventions. If $p(y \mid x)$ denotes the observational distribution of an outcome *Y* given $X = x$, then $p\big(y \mid do(x)\big)$ is interpreted as the result of "forcibly" fixing *X* to the value *x*, removing any links by which *X* depends on other nodes in the graph [12]. In practice, this allows us to address questions like "what happens if we intervene and alter *X*?" and thereby distinguish causal effects from mere correlations [6]. In ML, where the objective is to predict outcomes under "new scenarios" (i.e., changing one or more factors), failing to incorporate this operator can lead to confusion between genuine causal impacts and indirect associations [2].

Closely tied to the intervention concept is counterfactual reasoning. Counterfactuals answer questions like "what if we changed *X* to *x'*, even though in reality *X* took another value *x*?" [12]. Such counterfactuals involve holding the "noise" terms εi constant exactly as in the actual scenario but altering certain structural equations. In ML, this perspective is useful, for example, in generating counterfactual examples that help interpret decisions by complex models or test how robust an algorithm is to changes in context [7].

It is critically important to distinguish observational distributions, $p(y \mid x)$, from interventional ones, $p\big(y \mid do(x)\big)$. The former essentially captures patterns gleaned from data in the absence of external manipulation, whereas the latter models a scenario where *X* is forcibly set to a given value [12, 14]. If the model contains hidden factors, then correlation between *X* and *Y* may be misleading: the simple conditional distribution $p(y \mid x)$ can deviate substantially from the "true" causal effect $p\big(y \mid do(x)\big)$ [7]. In predictive tasks such as "what is the outcome if *X* changes?" these discrepancies may yield incorrect conclusions and problematic decisions, essentially because changing X can influence other variables in the system [2].

A fundamental challenge in causal inference is posed by hidden (unobserved) confounders. A confounder is a variable that affects both *X* and *Y* but remains unobserved by the researcher [6]. If such a factor is not accounted for, observational analyses or standard ML models can produce spurious associations and mislabel them as "causal." To correctly estimate effects, one typically assumes identifiability: either all critical variables are observable, or the researcher has access to extra information (e.g., instrumental variables or known graph structure) that helps "untangle" paths from the hidden node [11].

Table 1 below summarizes several core distinctions between observational, interventional, and counterfactual distributions in the context of evaluating interventional effects and outcome prediction.

Table 1. Comparing Distribution Types in Causal Analysis

Type	Definition	Example Question
Observational $p(y \mid x)$	The distribution derived from data without any external manipulation (all natural connections remain).	"If we observe that patient A took drug X, what is the probability of improvement Y?"
Interventional $p\big(y \mid$	The distribution modeling the outcome	"If we make patient A

Type	Definition	Example Question
$do(x)\big)$	of forcefully setting X = x, disabling any paths that would normally affect X.	take drug X, what is the probability of improvement Y?"
Counterfactual	A hypothetical distribution that assesses Y given a different scenario for X, while holding exogenous factors consistent with the actual history.	"Had patient A (who in reality took X) nottaken it, while keeping all else equal, what would have happened to Y?"

From a practical standpoint, knowing precisely which type of distribution we aim to capture—observational, interventional, or counterfactual—is critical when designing predictive models. Observationally, we might see that patients who chose to exercise had better health outcomes. But interventionally, if we *force* someone to exercise (controlling for other factors), what is the effect? A counterfactual question would be: for a specific patient who did not exercise and got ill, would they have stayed healthy *had* they exercised (keeping their other characteristics the same)? In traditional ML approaches devoid of causal insights, one typically learns an approximation for $p(y \mid x)$ or $f(x) \approx E[Y \mid X = x]$, which often suffices for predictive tasks under the same conditions as those in the training data [12]. Yet when one needs to evaluate the effect of an intervention, $\Delta = p\big(y \mid do(x = 1)\big) - p\big(y \mid do(x = 0)\big)$, or answer "how exactly should we alter *X* to achieve a desired outcome *Y*?," only a causal model provides the appropriate framework [6, 7].

Accordingly, to achieve highly accurate and robust predictions under heterogeneous data or genuine manipulations of variables, it becomes essential to embed causal machinery into ML—whether through constructing causal graphs, reflecting the parent–child dependency, or using the do-operator when selecting predictors [1]. Otherwise, modeling may remain trapped by spurious correlations, especially when hidden confounders hinder a correct interpretation of the dependencies.

2. Methods for Integrating Causality and ML for Effect Prediction

A crucial path toward improving the reliability of predictions under changing conditions is to account for the causal structure of data in both classical and modern machine learning (ML) algorithms [7, 12]. Below, we consider three primary directions for integrating causality and ML.

In the first area, causal supervised learning addresses the problem of eliminating spurious correlations that arise from distribution shifts. In conventional ML, when training regression, decision trees, or neural networks, we typically minimize some empirical risk $L(f(x), y)$ over a dataset, ignoring the possibility that certain features may be "spurious"—i.e., only conditionally relevant under a specific, "local" distribution [2]. Under a new distribution, such features either become unhelpful or even detrimental. To tackle this, researchers have proposed methods that exploit causal invariants. One of the best-known approaches is invariant risk minimization (IRM) [1]. It assumes there exists a subsystem of "invariant" characteristics (features) that drive the causal link to the target variable. To detect them, the method uses data from multiple "environments," each with a different distribution, and seeks to learn a representation $\Phi(x)$ for which the optimal linear (or similar) model is the same across all environments. Formally, this can be expressed as solving

$$min_{\Phi} \sum_{e \in E} R^e(\Phi)\ subject\ to\ w_e^* \in arg\ min_w R^e(w \circ \Phi) \forall e$$

where R^e is the risk in environment e, and w_e^* denotes the same (or practically identical) classifier. In doing so, IRM aims to exclude "env-specific" correlations and retain only robust patterns [1, 11]. In practice, this yields benefits under severe shifts in distribution, as the model learns to focus on genuine causal drivers [2].

The availability of multi-environment data is of particular importance. As noted in [7], when all samples are collected from a single setting, uncovering causally invariant features is extremely difficult: it is often "easier" for the algorithm to learn a superficial, yet purely correlational, dependence. However, having

multiple datasets where the external context varies (different hospitals, seasons, or regions, for example) introduces the chance to distinguish "global" (invariant) dependencies from "local" (spurious) ones in a statistically rigorous way. Such ideas pervade many modern methods in causal supervised learning, enabling more accurate, long-term effect prediction when real interventions are performed on variables affecting the outcome [6].

The second area, causal generative modeling, emphasizes the generation of "causally plausible" data. If classical GANs (generative adversarial networks) [5] or variational autoencoders (VAEs) [8] produce samples by reproducing the overall statistics of the original data, then the causal version requires explicitly separating "content"—causally significant factors—from "style", i.e. contextual or background-dependent attributes [7, 10]. This separation allows the model architecture to incorporate causal constraints, eliminating undesirable correlations and yielding more accurate simulations of interventions.

One of the most striking outcomes of causal generative models is counterfactual generation. Suppose we have a complex neural network trained to forecast a certain variable Y based on inputs X. We might ask, "What if X differed in some critical component while all other details remained fixed?" [12]. Generative methods make this scenario possible: we can "fix" the exogenous noise and modify only selected causal variables (content), leaving "style" unchanged. Comparing the original sample to a counterfactual one provides valuable information as to whether altering that factor truly affects Y, or whether it was merely an illusion [10]. Recent work has proposed frameworks like CausalGAN that integrate DAG constraints into GAN training, or variational autoencoders that infer latent causal factors. These approaches ensure that generated samples obey certain causal relationships from the data, rather than just any statistical relationships. This is extremely useful in scenarios where real interventions are expensive or unethical; one can simulate counterfactual populations (e.g., what if we treat vs. don't treat a patient) to predict outcomes and variability, thus aiding decision-making. In a stricter ML context, counterfactual samples let us check, for instance, "What is the minimal feature shift required to change an instance's category?"—improving model interpretability and facilitating debugging.

A third group of other contemporary approaches examines the role of causality in interpretation (causal explanations), reinforcement learning (RL), and fairness in algorithms. For interpretation (and explainability), the "**counterfactual explanation**" concept is often invoked, pinpointing which input component actually caused the model's decision [13]. For instance, in a loan approval model, a causal counterfactual explanation might say: *'Had the applicant's income been $5,000 higher (with all else the same), the loan would have been approved.'* This identifies a causal factor in the model's decision. **In causal RL** [3], this framework helps an agent more quickly discern changes in the environment's dynamics: the agent builds a structural model and can predict the effect of particular actions in a "new" state. In fairness, **causal fairness methods** focus on whether discrimination arises from direct causal pathways rather than any difference in a protected attribute: i.e., if we change a sensitive attribute (race, gender) in a person's data while keeping all other attributes the same, a fair algorithm would produce the same outcome [9]. Causal methods help formalize this by explicitly modeling how sensitive attributes influence other variables. Finally, **causal discovery** algorithms [11, 14] aim to learn the causal graph from observational data. Incorporating such learned structures into ML models can improve effect prediction and even allow simulation of interventions in scenarios where the structure was not known a priori. This also opens the door to **long-term forecasting** under interventions, because once we trust the discovered causal model, we can propagate changes over time or across systems..

Table 2 summarizes some of the methods introduced in this section and the core tasks they address.

Table 2. Overview of Methods Integrating Causality and ML

Method / Approach	Key Idea	Typical Tasks
Invariant Risk Minimization (IRM) [1]	Identifying features invariant across different "environments," removing spurious dependencies	Reliable classification/regression under large distribution shifts

Method / Approach	Key Idea	Typical Tasks
Causal Generative Modeling [5, 10]	Separating style/content in data generation, incorporating structural constraints	Generating realistic counterfactuals; synthetic data augmentation
Counterfactual Explanations [13]	Showing what must be changed in X to shift a classification to another category while keeping other aspects fixed	Interpreting black-boxes (neural nets, GBMs), algorithmic audits
Causal RL [3]	Agent constructs a causal model of the environment and interventions, improving adaptation to new settings	Adaptive control strategies; evaluating actions when environmental dynamics change
Causal Fairness [9]	Focuses on causal pathways of discrimination, distinguishing allowable vs. disallowed dependencies	Fair candidate selection, risk assessment with attention to bias

From the perspective of predicting interventional effects, this group of techniques expands the scope of conventional ML. Rather than being limited to correlational models, we gain instruments for analyzing causal mechanisms, accounting for shifts over time and space, and introducing additional constraints on which factors genuinely drive changes in the target variable [6, 12]. This is especially relevant in medicine (optimizing therapy choice), economics (policy evaluation), user behavior analysis, industrial control systems, and so on [7].

CONCLUSION

The directions surveyed in this article underscore that traditional machine learning—fundamentally relying on associative pattern discovery—often remains vulnerable to distribution shifts and fails to offer reliable answers to “what if” questions. By contrast, the application of causal inference substantially extends the capabilities of ML. First, it provides a formal basis for assessing genuine interventional effects; second, it enhances model interpretability by distinguishing true causal drivers from mere correlates. Structural causal models, which enable the use of the do-operator and the construction of counterfactual scenarios, are key to this approach. Methods such as invariant risk minimization demonstrate how spurious correlations can be removed with multi-environment data, while causality-oriented generative models equip us to produce counterfactual samples and evaluate algorithmic behavior under new conditions.

Hence, integrating machine learning with causal inference not only raises the accuracy of predictions but also makes them more robust to potential environmental changes, improves interpretability, and supports well-grounded recommendations in fields such as medicine, economics, and social planning. By systematizing current methods and linking them to established causal theory, this work provides researchers and practitioners with a clearer roadmap for incorporating causality into machine learning models for decision-making. Despite advancements in causal machine learning, unresolved challenges remain regarding the limited identifiability of hidden factors and the lack of universal benchmark datasets. Nevertheless, the ongoing development of integrative approaches to causal inference and ML continues to reveal new opportunities for more precise, transparent, and reliable modeling of complex systems. Future research should explore techniques for identifying or mitigating hidden confounders (perhaps via advanced instrumentation or causal discovery), and develop standardized benchmarks to evaluate causal ML models across domains. Such efforts would accelerate progress in this field. In conclusion, as machine learning systems are increasingly used for critical decision-making, embedding causal reasoning into these systems is not just a theoretical luxury but a practical necessity for robust, ethical, and reliable AI.

REFERENCES

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D.

(2019). Invariant risk minimization. arXiv preprint, arXiv:1907.02893.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. Proceedings of the national academy of sciences, 113(27), 7345–7352.

Bareinboim, E., Forney, A., & Pearl, J. (2015). Bandit problems with causal background knowledge. In proceedings of the thirty-first conference on uncertainty in artificial intelligence (pp. 42–51).

Frye, C., Feige, I., Rowat, C., & de Figueiredo, D. (2019). Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. arXiv preprint, arXiv:1910.06358.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In advances in neural information processing systems, 27.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: an introduction.Cambridge University Press.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal Machine learning: a survey and open problems. arXiv preprint, arXiv:2206.15475.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In international conference on learning representations (ICLR).

Kusner, M. J., Loftus, J., Russakovsky, O., & Silva, R. (2017). Counterfactual fairness. In Advances in neural information processing systems, 30.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In advances in neural information processing systems, 30.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5), 947–1012.

Pearl, J. (2009). Causality: models, reasoning, and inference (2nd ed.). Cambridge University Press.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard journal of law & technology, 31(2), 841–887.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. MIT press.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2018, July). Learning independent causal mechanisms. In international conference on machine learning (pp. 4036-4044). PMLR.