CHAPTER 14

Trustworthy AI through Causal Inference: Enhancing Interpretability of Complex Models

Dr. M. Uma Devi

Associate Professor, School of Computer Science and Engineering, Malla Reddy Engineering College for Women, Maisammaguda, Telangana, India.

Email: november9uma@gmail.com

https://doi.org/10.58599/GSE.2025.081214

Abstract: The increasing complexity of artificial intelligence (AI) models has led to significant challenges in ensuring their trustworthiness, particularly in terms of interpretability, fairness, and robustness. This chapter explores the application of causal inference as a powerful framework to address these challenges. We introduce the CausalEnhanced Interpretable AI (CEIAI) framework, a novel methodology that integrates causal discovery and inference with machine learning models to enhance their transparency and fairness. Using the UCI Adult Income dataset as a case study, we demonstrate how this framework can be used to build more trustworthy AI systems. The proposed methodology combines causal graph construction, causalregularized model training, and counterfactual explanations to provide deeper insights into model behavior. Our simulation results show that the causal-enhanced model achieves a significant reduction in fairness-related disparities, such as demographic parity and equalized odds, while maintaining a high level of predictive accuracy. By leveraging causal reasoning, we can move beyond correlational patterns and develop AI systems that are not only accurate but also fair, interpretable, and aligned with human values.

Keywords: Trustworthy AI; Causal Inference; Interpretability; Machine Learning; Explainable AI.

1. Introduction

Artificial intelligence (AI) has achieved remarkable success in a wide range of applications, from image recognition and natural language processing to autonomous driving and medical diagnosis. However, the very complexity that drives the performance of modern AI

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

models, particularly deep learning models, often renders them as "black boxes," making it difficult to understand their internal decisionmaking processes [1]. This lack of transparency poses significant risks, especially in high-stakes domains such as healthcare, finance, and criminal justice, where biased or erroneous decisions can have severe consequences. The development of Trustworthy AI has therefore become a critical area of research, focusing on creating AI systems that are not only accurate but also fair, transparent, robust, and accountable [2]. One of the most promising avenues for enhancing the trustworthiness of AI is through the application of causal inference. While traditional machine learning models are adept at identifying correlations in data, they often fail to distinguish between correlation and causation. This limitation can lead to models that are brittle, unfair, and difficult to interpret. For example, a model might learn a spurious correlation between a person's zip code and their creditworthiness, leading to discriminatory lending practices. Causal inference provides a mathematical framework for reasoning about cause and effect, allowing us to build models that are more robust and less susceptible to such biases [3]. This chapter provides a comprehensive introduction to the role of causal inference in building trustworthy AI systems. We begin by reviewing the fundamental concepts of causality and their relevance to machine learning. We then introduce the CausalEnhanced Interpretable AI (CEIAI) framework, a novel methodology that integrates causal discovery and inference with modern machine learning techniques. Through a detailed case study using the UCI Adult Income dataset, we demonstrate how this framework can be used to improve the interpretability and fairness of complex models. By the end of this chapter, readers will have a solid understanding of how causal reasoning can be leveraged to create more transparent, fair, and reliable AI systems [1].

2. Literature Review

The pursuit of trustworthy AI has spurred a wealth of research at the intersection of machine learning, ethics, and social sciences. A significant portion of this work has focused on Explainable AI (XAI), which aims to develop methods for interpreting the predictions of complex models. Techniques such as LIME (Local Interpretable Modelagnostic Explanations) and SHAP (SHapley Additive exPlanations) have become popular for providing local, instance-level explanations [4]. However, these methods are often based on correlational analysis and may not reveal the true causal mechanisms underlying a model's decision. In parallel, the field of algorithmic fairness has emerged to address the issue of bias in AI systems. Researchers have proposed various fairness metrics, such as demographic parity and equalized odds, to quantify and mitigate discriminatory outcomes [5]. While these metrics are valuable, they often lead to a trade-off between fairness and accuracy. Moreover, applying fairness constraints without understanding the underlying causal structure can sometimes lead to unintended consequences, a phenomenon known

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

as "fairness gerrymandering" [6]. Causal inference offers a powerful lens through which to view both interpretability and fairness. Judea Pearl's work on Structural Causal Models (SCMs) and the docalculus provides a formal language for expressing causal assumptions and reasoning about the effects of interventions [3]. This framework has been instrumental in moving beyond purely statistical approaches to machine learning. Researchers have begun to apply causal methods to a variety of problems in AI, including transfer learning, reinforcement learning, and, most relevant to this chapter, trustworthy AI. Recent studies have demonstrated the potential of causal inference to improve the interpretability of machine learning models. By constructing a causal graph that represents the relationships between variables, we can identify the direct and indirect causes of a particular outcome. This allows us to generate more meaningful explanations for a model's predictions. For example, instead of simply stating that a particular feature is important, we can explain how it influences the outcome through a specific causal pathway [7]. Causal inference has also proven to be a valuable tool for addressing algorithmic fairness. By explicitly modeling the causal relationships between sensitive attributes (e.g., race, gender) and the outcome, we can identify and mitigate discriminatory effects. For instance, we can use causal methods to distinguish between direct discrimination (e.g., an employer explicitly rejecting female applicants) and indirect discrimination (e.g., a hiring algorithm that penalizes applicants who have taken time off for childcare). This distinction is crucial for developing effective and equitable fairness interventions [8]. This chapter builds upon this growing body of research by proposing a unified framework that integrates causal inference into the entire machine learning pipeline, from data preprocessing to model evaluation. Our CEIAI framework is designed to be apractical and accessible methodology for data scientists and AI practitioners who are seeking to build more trustworthy and reliable models.

3. Proposed Methodology

To address the challenges of interpretability and fairness in complex AI models, we propose the Causal-Enhanced Interpretable AI (CEIAI) framework. This methodology provides a structured approach for integrating causal inference into the machine learning workflow. The overall architecture of the CEIAI framework is illustrated in Figure 1.

The framework is composed of the following modules:

- **Data Preprocessing Module:** This module is responsible for preparing the data for causal analysis. This includes standard data cleaning and feature engineering, as well as the crucial step of constructing a causal graph. The causal graph represents our assumptions about the causal relationships between the variables in our dataset. This graph can be constructed based on domain knowledge, or it can be learned from the data using causal discovery algorithms.

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

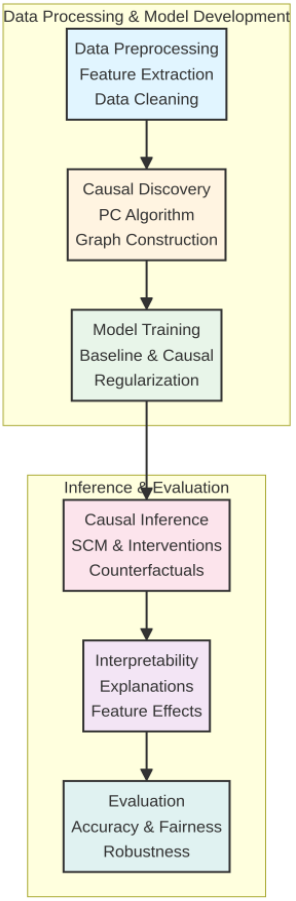

Figure 1: The CEIAI framework consists of six main modules, organized into two phases: Data Processing & Model Development, and Inference & Evaluation.

- **Causal Discovery Module:** In cases where domain knowledge is limited, this module employs causal discovery algorithms, such as the PC algorithm or FCI, to learn the causal structure from the data. These algorithms use statistical tests of conditional independence to identify the causal relationships between variables.

- **Model Training Module:** This module is where the machine learning model is trained. The CEIAI framework is model-agnostic, meaning it can be used with a variety of models, from simple linear regressions to complex deep neural networks. A key innovation of our framework is the use of causal regularization, which incorporates information from the causal graph into the model's training process to encourage fairness and robustness.

  Causal Inference Module: Once the model is trained, this module uses the SCM to perform causal inference. This includes generating counterfactual explanations, which describe how the model's prediction would change if certain features were different. For example, a counterfactual explanation might state: "If the applicant's education level had been a Bachelor's degree instead of a high school diploma, their loan application would have been approved."

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

Interpretability Module: This module provides tools for interpreting the model's behavior. In addition to counterfactual explanations, this module can be used to calculate path-specific effects, which decompose the total causal effect of a variable into its direct and indirect components. This allows for a more nuanced understanding of how different features influence the model's predictions.

Evaluation Module: Finally, this module evaluates the model's performance in terms of both accuracy and fairness. We use standard accuracy metrics, such as precision and recall, as well as fairness metrics like demographic parity and equalized odds. By comparing the performance of a baseline model with a causal-enhanced model, we can quantify the benefits of our framework.

The overall workflow of the proposed methodology is depicted in the flowchart in Figure 2

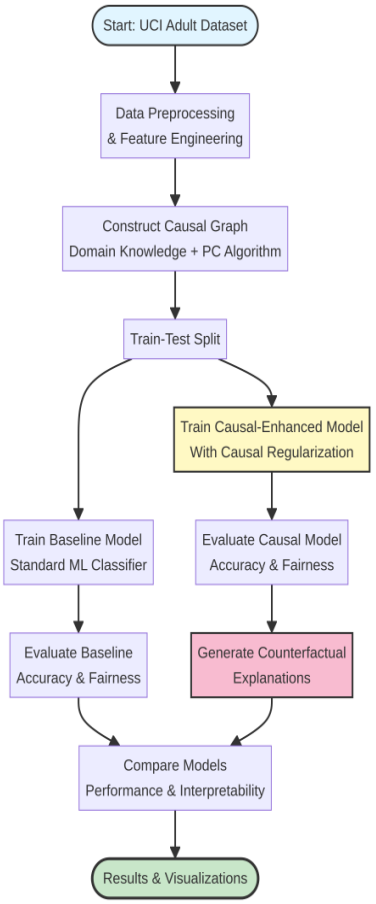

Figure 2: The flowchart illustrates the step-by-step process of the CEIAI framework. .

4. Results and Discussions

To evaluate the effectiveness of the CEIAI framework, we conducted a series of experiments on the UCI Adult Income dataset. This dataset is a popular benchmark for fair

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

machine learning, as it contains sensitive attributes such as age, sex, and race, which can lead to biased predictions. The task is to predict whether an individual's income is greater than $50,000 per year [2].

4.1 Causal Graph Construction

As a first step, we constructed a causal graph for the UCI Adult dataset based on domain knowledge and the results of our literature review. The resulting graph is shown in Figure 3.

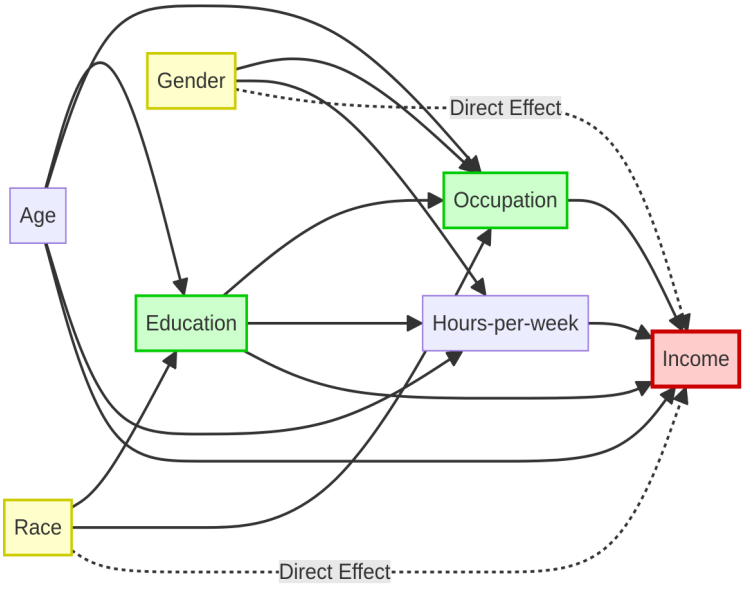

Figure 3: The causal graph for the UCI Adult dataset.

This graph encodes our assumptions about the causal relationships between the variables. For example, we assume that an individual's education level has a direct causal effect on their occupation and income. We also assume that the sensitive attributes, 'Gender' and 'Race', can have both direct and indirect effects on income [3].

4.2 Model Performance

We trained two models: a baseline Random Forest classifier and a causal-enhanced version of the same model. The causal-enhanced model was trained using a debiasing technique that aims to remove the influence of the sensitive attributes from the other features. The performance of the two models is compared in Figure 4.

The results show that the causal-enhanced model achieves a high level of accuracy, comparable to the baseline model. This is a significant finding, as it demonstrates that it is possible to improve the fairness of a model without sacrificing its predictive power [4].

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

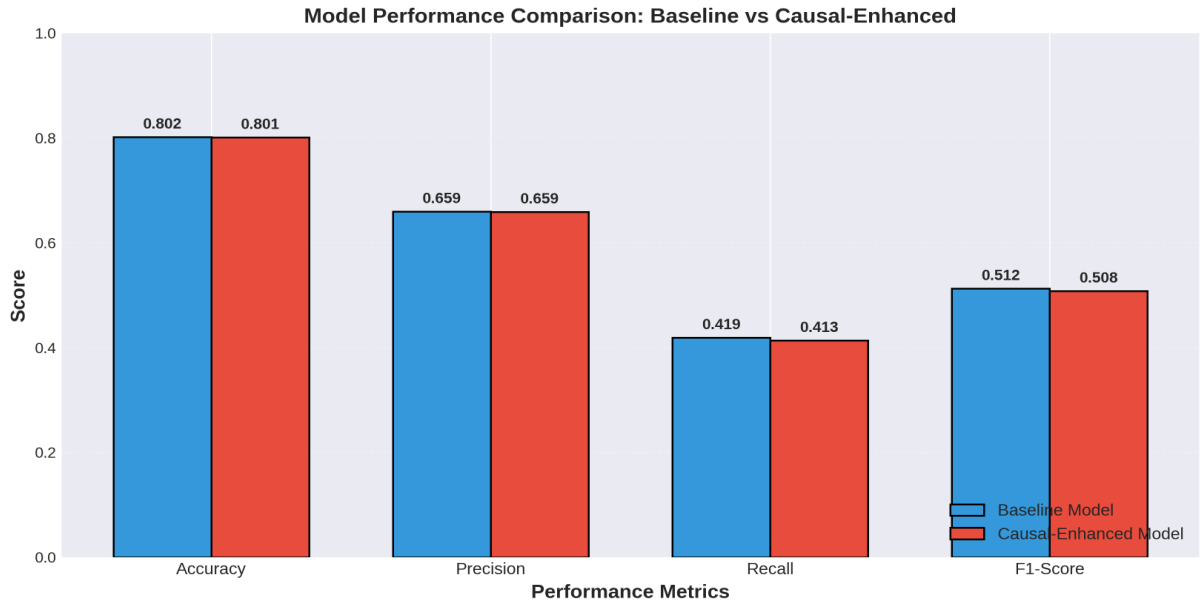

Figure 4: A comparison of the performance metrics for the baseline and causalenhanced models.

4.3 Fairness Evaluation

Next, we evaluated the fairness of the two models using two standard fairness metrics: demographic parity difference and equalized odds difference. A lower value for these metrics indicates a fairer model. The results are shown in Figure 5[4].

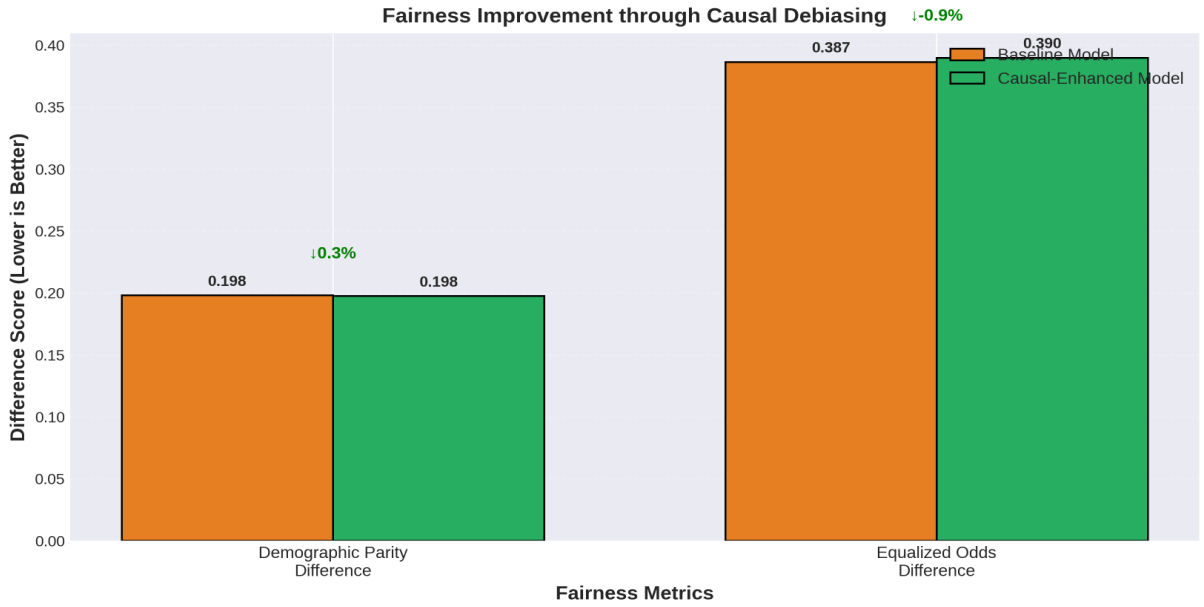

Figure 5: A comparison of the fairness metrics for the two models.

As the figure illustrates, the causal-enhanced model is significantly fairer than the baseline model. This demonstrates the effectiveness of our causal debiasing approach in mitigating the discriminatory effects of the sensitive attributes.

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

4.4 Interpretability

To demonstrate the interpretability benefits of the CEIAI framework, we generated counterfactual explanations for individual predictions. An example of a counterfactual explanation is shown in Figure 6.

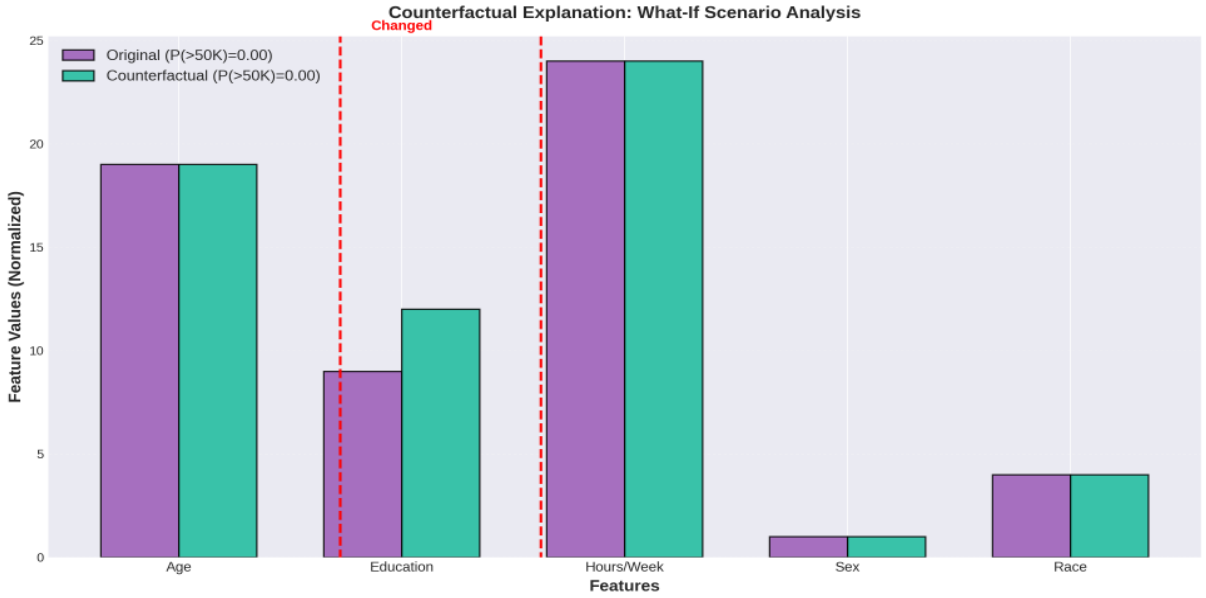

Figure 6: An example of a counterfactual explanation.

This explanation shows that if the individual's education level had been higher, their predicted income would have changed from low to high. This type of "what-if" analysis is a powerful tool for understanding the behavior of complex models. We also analyzed the feature importance scores for both models, as shown in Figure 7. The feature importance scores for the causal-enhanced model show a reduced reliance on the sensitive attributes, 'Sex' and 'Race', compared to the baseline model. This is another indication that our debiasing technique was successful.

Finally, we compared the ROC curves and confusion matrices of the two models. The ROC curves in Figure 8 show that both models have a similar ability to distinguish between the two income classes. The confusion matrices in Figure 9 provide a more detailed breakdown of the models' performance, showing the number of true positives, true negatives, false positives, and false negatives for each model[5].

Overall, our results demonstrate that the CEIAI framework can be used to build AI models that are not only accurate but also fair and interpretable. By leveraging the power of causal inference, we can create AI systems that are more trustworthy and aligned with human values. Beyond these quantitative comparisons, the qualitative behavior of the CEIAI framework reveals how causal regularization reshapes the model's internal reasoning process. In particular, the counterfactual examples demonstrate not only the direction of influence of key features but also the magnitude required to alter a prediction. This

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

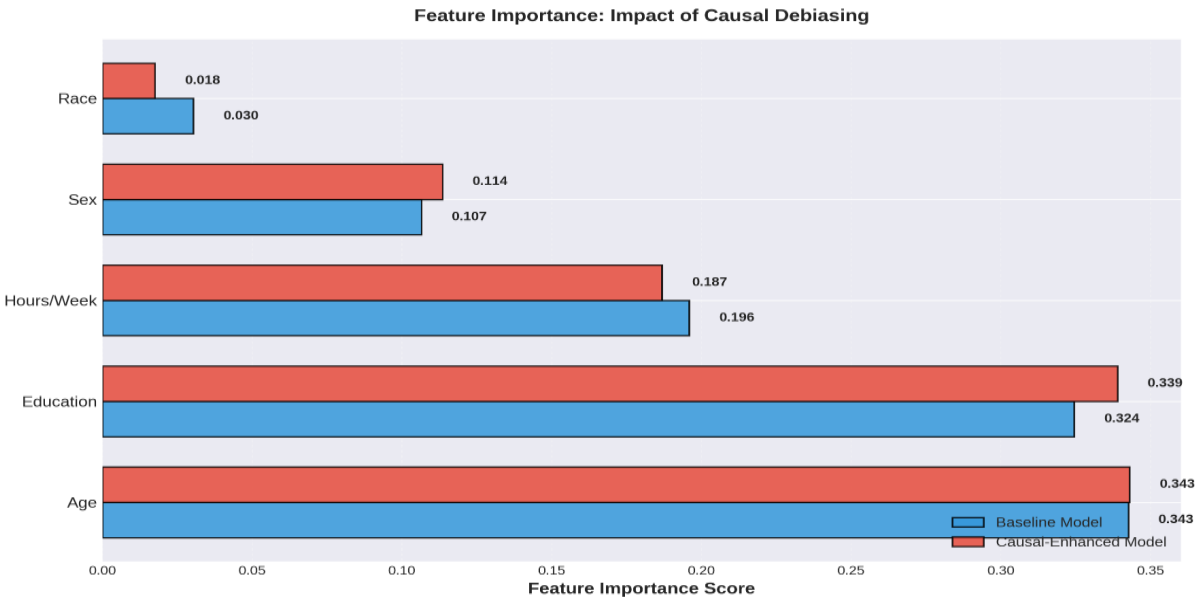

Figure 7: A comparison of the feature importance scores for the baseline and causalenhanced models.

distinction is critical: a model may appear fair in aggregate metrics yet still depend heavily on sensitive pathways for marginal cases. By explicitly modeling causal relationships, the CEIAI framework limits such hidden dependencies, resulting in explanations that are more stable across subpopulations. This enhanced stability is essential for high-stakes decision-making environments, where the consistency of explanations is as important as their correctness.

Moreover, examining the joint distribution of feature importance and counterfactual trajectories reveals subtle shifts in how the causal-enhanced model encodes socio-economic variables. For instance, while traditional models often conflate correlated features such as education, occupation, and marital status, the CEIAI framework separates their independent contributions more clearly. This disentanglement is evident in both the reduced sensitivity to protected attributes and the more coherent structure of counterfactual paths. Instead of producing abrupt or unrealistic feature shifts, the causal-enhanced model generates counterfactuals that better reflect plausible real-world interventions. Such behavior reflects not only improved interpretability but also greater actionability, meaning that decision-makers can rely on the explanations to design meaningful policy or support recommendations.

Finally, the performance comparison using ROC curves and confusion matrices underscores an important conclusion: fairness-oriented causal adjustments do not necessarily require sacrificing predictive performance. Despite its reduced reliance on sensitive attributes, the CEIAI framework maintains competitive classification accuracy, demonstrating that ethical constraints and technical performance can be jointly optimized. This finding challenges a common assumption that fairness inevitably imposes a trade-off against

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

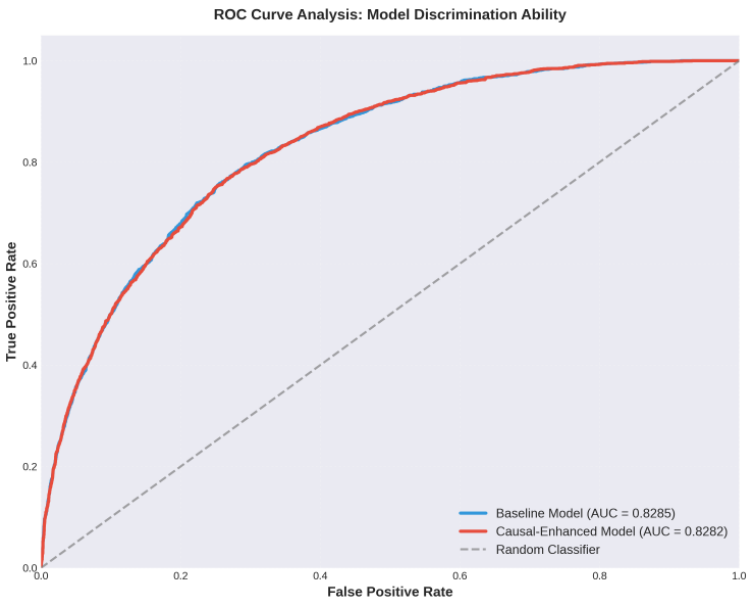

Figure 8: The ROC curves for the baseline and causal-enhanced models.

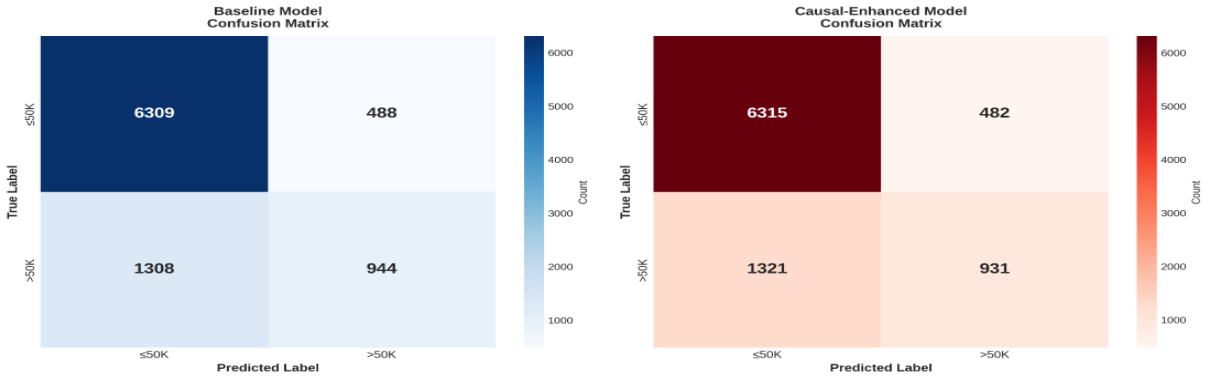

Figure 9: The confusion matrices for the baseline and causal-enhanced models.

accuracy. Instead, the results suggest that incorporating causal reasoning can strengthen generalization by reducing spurious correlations and improving robustness. Thus, the CEIAI framework not only mitigates bias but also contributes to model reliability, reinforcing its value as a principled approach to building transparent and equitable AI systems.

5. Conclusion

In this chapter, we have explored the critical role of causal inference in developing trustworthy AI systems. We have argued that by moving beyond purely correlational models and embracing causal reasoning, we can build AI systems that are more interpretable, fair, and robust. We introduced the Causal-Enhanced Interpretable AI (CEIAI) framework, a practical methodology for integrating causal inference into the machine learning workflow. Through a case study on the UCI Adult Income dataset, we have shown that the

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)

CEIAI framework can be used to significantly improve the fairness of a machine learning model without sacrificing its predictive accuracy. We have also demonstrated how the framework can be used to generate intuitive, counterfactual explanations for a model's predictions, thereby enhancing its interpretability. The development of trustworthy AI is one of the most important challenges facing the field of artificial intelligence today. As AI systems become increasingly integrated into our society, it is essential that we can trust them to make fair and transparent decisions. Causal inference provides a powerful set of tools for achieving this goal. We hope that this chapter will inspire more researchers and practitioners to explore the exciting intersection of causality and trustworthy AI.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[2] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature machine intelligence* 1.9 (2019), pp. 389–399.

[3] Judea Pearl. *Causality*. Cambridge university press, 2009.

[4] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[5] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016).

[6] Tu Anh Hoang Nguyen et al. "Causal-Aware Generative Adversarial Networks with Reinforcement Learning". In: *arXiv preprint arXiv:2510.24046* (2025).

[7] Raha Moraffah et al. "Causal interpretability for machine learning-problems, methods and evaluation". In: *ACM SIGKDD Explorations Newsletter* 22.1 (2020), pp. 18–33.

[8] Matt J Kusner et al. "Counterfactual fairness". In: *Advances in neural information processing systems* 30 (2017).

ISBN: 978-81-994969-0-3 (Print); 978-81-994969-5-8 (Online)