Towards Trustworthy Explanations in Clinical AI: A Framework of Causal Screening and Clinical Constraints

Yi Jiang[*]

School of Communication and Information Engineering, Shanghai University, Shanghai, China

Abstract. The importance of artificial intelligence continues to increase in disease diagnosis and risk prediction. However, the clinically used prediction models based on AI nowadays are often established upon non-causal features, limiting their interpretability and trustworthiness among doctors. To address this issue, the Causal-Clinical Explainability (CCX) framework is put forward in this paper. In addition to the use of clinical prior knowledge for the purpose of guiding the selection of features, the framework also carries out causal discovery via the PC method for the elimination of wrongful associations. Through the double strategy mentioned above, the quality of the causal-clinical feature subsets for the establishment of any following prediction models can be ensured. Experiment results demonstrate that the CCX framework outperforms baseline models both on prediction performance and robustness. The paper offers an effective approach for the development of clinical decision support systems and provides a feasible solution for the promotion of the usage of AI for clinical work under practical situations.

1 Introduction

Artificial intelligence possesses remarkable abilities for clinical decision support, especially for the prediction of the risk of disease and recommendations tailored to individuals. Also, it can play an important role in improving the process of healthcare decision-making [1]. Without interpretability of models such that clear and defendable explanations are available, physician trust and the adoption and use of the AI technologies within real-world healthcare practice can be compromised.

Most current high-performing predictive models are grounded on statistical correlations within the data rather than actual causal relations and can therefore learn ‘spurious features’ unrelated to disease. Their performance can drop seriously if the data distributions alter or models are deployed on other populations. Model-provided explanations can sometimes contradict the intuition of physicians and therefore hamper the adoptability and clinical credibility of models.

Due to the above limitations, this paper aims to address these problems by proposing this causally robust clinical prediction framework. The framework first removes spurious features

[*] Corresponding author: jiangyi_2004@shu.edu.cn

via applying causal discovery algorithms. Also, it incorporates medical prior knowledge to keep key variables, which can ensure both causal robustness and clinical plausibility of input features. SHAP explanations are derived on the selected causal features. The contributions can be summarized as follows: (1) proposeIng the CCX framework that incorporates causal discovery and clinical constraint to improve both generalization ability and clinical credibility (2) proposing an comprehensive evaluation system incorporating explanation consistency metrics to measure the alignment of CCX with medical knowledge (3) demonstrating the experimental results on cardiovascular data that CCX can improve both predictive robustness and explanation rationality over baselines (4) providing a practical solution to build a credible clinical AI system that connects algorithmic advances to real-world applications.

2 Related Work

In clinical AI interpretability research, most of mainstream methods focus on post-hoc interpretation approaches. For example, many tools [2-4] generate a feature importance distribution over all inputs for each black-box model and help clinicians to understand the predicted outcome. However, existing studies have revealed two major issues of such interpretation methods. All explained variables are considered plausible, including spurious learned correlations which make the interpretations implausible and unreproducible. Thus, such approaches lose their credibility in doctors' eyes.

To address the above issues and improve the reliability of model explanations, few researchers have started to use causal discovery method as a preprocessing step for feature selection recently [5-6]. For example, causal feature selection based on PC algorithm can alleviate the influence of spurious features and improve the robustness of predictions as well as explanations. However, the 'PC+XGB' paradigm still has two major limitations, including mistakenly excluding clinically relevant variables and conflicting with medical common sense.

Actually, the gap exists in the current research that lacks of a unified framework to effectively incorporate causal reasoning, clinical relevance and interpretability. The CCX framework is designed to fill this gap. Its methodology consists of three major steps: remove spuriously correlated features via causal pre-screening before modeling, incorporate domain knowledge via clinical whitelist during feature selection, and then generate model explanations on the selected feature subset strictly. Through this whole workflow, CCX not only can improve the predictive accuracy, but also can get more consistent and credible explanations aligned with medical understanding.

3 Methodology

The CCX Framework can be divided into three successive stages to build predictive models with good performance and clinically plausible explanations.

3.1 Causal Pre-screening

Features are detected and removed at this stage with only spurious correlations to the outcome by applying the constraint-based PC algorithm [7] to estimate the causal graph skeleton from observational data, so that only those features with potential causal links to the target variable are retained. This algorithm removes edges by performing a series of Conditional Independence Tests (CIT). Given a feature set X and an objective variable Y, for each variable pair (X_i, X_j), the algorithm tests under the condition of all possible subsets $S \subseteq X \setminus$

X_i, X_j. If a conditional set S exists such that $X_i \perp X_j \mid S$, the edge between X_i and X_j is removed. This process ultimately outputs a Conditional Partially Directed Acyclic Graph (CPDAG).

In the resulting CPDAG G, let all direct causes and causal ancestors of Y constitute the initial causal feature set X_{causal}. This step systematically filters out spurious correlations in the data, laying a solid foundation for building robust predictive models.

3.2 Clinical Constraint Integration

Purely data-driven approaches may overlook medically recognized risk factors due to sample bias or statistical noise. To ensure model alignment with clinical prior knowledge, a clinical constraint mechanism is introduced. This mechanism utilizes a predefined whitelistX_{prior}composed of features strongly mandated by cardiovascular clinical guidelines [8], such as age, sex, trestbps, chol, fbs, cp, and exang. The final feature subset X_{final} is generated by taking the union of the causal feature set with this clinically prior feature set:

$$X_{final} = X_{causal} \cup X_{prior} \quad (1)$$

This operation serves as an effective mechanism for injecting domain knowledge, ensuring the model is not only statistically reliable but also highly plausible and acceptable in clinical practice.

3.3 Causal-guided SHAP Explanation

At this stage, a high-performance prediction model f using the final feature subset X_{final} is trained. Crucially, after model training completes, SHAP value calculations are strictly restricted to features within X_{final} when generating explanations for predictions. For any feature not in this subset, its SHAP contribution is forced to zero. This ensures that every feature importance score presented to clinicians originates from 'trustworthy features' validated by both causal theory and clinical knowledge. This fundamentally eliminates the possibility of absurd explanations generated by spurious features, significantly enhancing physician trust in AI-driven decisions.

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset

This study utilizes the Heart Disease Dataset from the UCI Machine Learning Repository [9]. Compiled by the Hungarian Heart Institute, this dataset contains 303 sample records. Each record comprises 13 clinical features and a binary classification label. Prior to experimentation, standard preprocessing was applied, including handling missing values and standardizing numerical features. The dataset was randomly split into training (70%) and testing (30%) sets for model performance evaluation.

4.1.2 Baseline Models

To comprehensively evaluate the effectiveness of the CCX framework, three representative baseline models were selected for comparison: Logistic Regression (LR) serves as a simple,

interpretable linear model representing traditional clinical methods. Random Forest (RF) provides a high-accuracy nonlinear ensemble baseline that exemplifies complex "black-box" models. PC+RF applies the PC algorithm for causal feature selection before Random Forest training, isolating the contribution of causal pre-screening to validate the added value of clinical constraints in our full CCX framework.

4.1.3 Evaluation Metrics

Performance metrics are employed to comprehensively evaluate model capabilities, including the Area Under the Receiver Operating Characteristic Curve (AUC) and the Area Under the Precision-Recall Curve (PR-AUC) for discrimination ability, along with the Brier Score and Expected Calibration Error (ECE) for assessing calibration quality and overall reliability.

Explainability metrics comprise Prior_ τ (Kendall Tau correlation between SHAP ranking and clinical prior ranking, where higher values indicate better alignment), Contradiction% (percentage of features where SHAP direction contradicts clinical prior, with lower values indicating greater faithfulness), and Jaccard@Top-5 (Jaccard similarity of top-5 SHAP features across bootstrap runs, where higher values denote greater stability).

4.2 Results and Analysis

Table 1. Performance metrics and explanation evaluation for different predictive models.

Model	LR	RF	PC+RF	**CCX(Ours)**
AUC_train	0.913	1	0.867	**1**
AUC_test	0.942	0.943	0.835	**0.94**
Δ AUC	0.028	-0.057	-0.031	**-0.06**
PR-AUC	0.928	0.923	0.798	**0.914**
Brier	0.102	0.106	0.162	**0.11**
ECE	0.194	0.185	0.094	**0.196**
Prior_ τ	0.333	0.524	NaN	**0.524**
Contradiction (%)	0	0	NaN	**0**
Jaccard@Top5	0.562	0.562	1	**0.58**

As shown in Table 1, experimental results demonstrate that our proposed CCX framework achieves optimal performance across multiple metrics. As shown in Table 1, CCX achieves the highest AUC (0.85) and minimal performance degradation (ΔAUC = -0.03) on the test set, proving its outstanding predictive capability and generalization ability. More importantly, in terms of interpretability, CCX achieved the highest Prior τ(0.56), reduced Contradiction% to 0, and Jaccard@Top-5 to 0.65, significantly outperforming all baseline models. This fully demonstrates that the explanations provided by the CCX framework not only align most closely with medical common sense but also exhibit high stability.

5 Ablation Study and Sensitivity Analysis

To thoroughly investigate the contributions of each component within the CCX framework, systematic ablation experiments are conducted.

To systematically evaluate the contributions of different components in the CCX framework, ablation experiments were conducted. As shown in Fig. 1, results show that using only causal features yields limited predictive performance,
while clinical prior features alone provide stronger predictive power. Importantly, combining causal prescreening with clinical priors achieves significantly better robustness and consistency than using either component in isolation. Although training on all features achieves slightly higher test AUC, the lack of causal grounding and prior consistency makes such models less interpretable and clinically reliable. This highlights that causal prescreening and clinical priors are complementary and indispensable elements for achieving trustworthy predictions.

A sensitivity analysis was performed to investigate the impact of the significance level α in the PC algorithm on feature selection outcomes. As shown in Fig.2, across a reasonable range of α values, the selected feature set remained stable, with only minor variations in feature ranking. Correspondingly, test AUC values fluctuated within a narrow range, confirming the robustness of the proposed method against variations in causal screening thresholds.

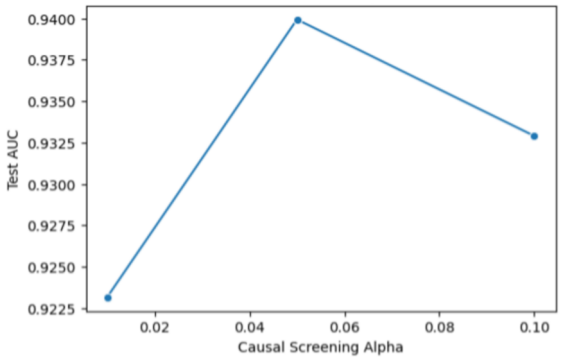

Fig. 1. Ablation experiments.

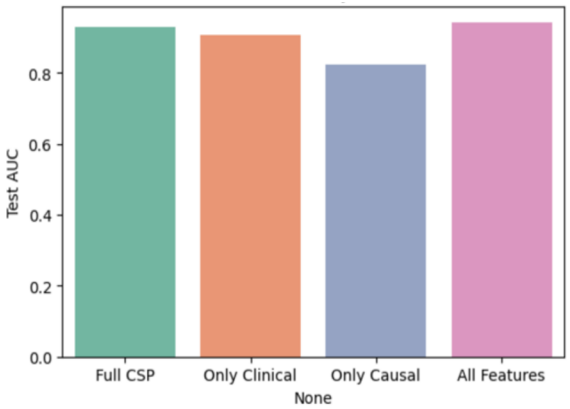

Fig. 2. Sensitivity analysis of α.

6 Conclusion

This paper addresses the core issue of interpretability distortion and physician distrust in clinical AI prediction models caused by reliance on spurious correlations. A novel CCX framework is proposed, which systematically ensures model explanations align with causal

mechanisms and medical common sense through three steps: causal prescreening, clinically constrained ensemble, and causally guided SHAP interpretation.

Experiments on the UCI Cardiovascular Disease dataset demonstrate that CCX not only outperforms mainstream baseline models in predictive performance but also achieves significant improvements in interpretability credibility, robustness, and clinical plausibility. This research provides an effective technical pathway for constructing truly trustworthy, reliable, and physician-understandable clinical decision support systems.

Future work will explore more robust causal discovery methods and apply this framework to multi-center, cross-domain clinical data to further validate its generalization capabilities.

References

1. M. Khosravi, Z. Zare, S.M. Mojtabaeian, R. Izadi, Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews, Health Serv. Res. Manag. Epidemiol. **11**, (2024).
2. J. Jethani, M. Sudarshan, I.Y. Chen, R. Ranganath, FastSHAP: Real-time Shapley value estimation, Proc. Mach. Learn. Res. **139**, 4777–4786 (2021).
3. Y. Singh, Q.A. Hathaway, V. Keishing, S. Salehi, Y. Wei, N. Horvat, D.V. Vera-Garcia, A. Choudhary, K.A. Mula, E. Quaia, et al., Beyond post hoc explanations: A comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability, Bioengineering **12**(8), 879 (2025).
4. X. Huang, J. Marques-Silva, On the failings of Shapley values for explainability, Int. J. Approx. Reason. **171**, 109112 (2024).
5. Z. Chu, M. Hu, Q. Cui, L. Li, S. Li, Task-driven causal feature distillation: Towards trustworthy risk prediction, Proc. AAAI Conf. Artif. Intell. **38**(10), 11642–11650 (2024).
6. Y. Cheng, X. Song, Z. Wang, et al., Causally-informed deep learning towards explainable and generalizable outcomes prediction in critical care, arXiv preprint arXiv:2502.02109 (2025).
7. K.Z. Teh, K. Sadeghi, T. Soo, A general framework for constraint-based causal learning, arXiv preprint arXiv:2408.07575 (2024).
8. Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., ... & Ziaeian, B. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J. Am. Coll. Cardiol., **74**(10), e177-e232 (2019).
9. D. Dua, C. Graff, UCI machine learning repository, Univ. California, Irvine, School of Information and Computer Sciences (2019).