Casual Discovery and Explainable Diagnosis System for Medical Data

Ms.S.visali[1] and Mr.J.Lin Eby Chandra [2]
[1]PG Scholar, Department of Computer Science and Engineering, Jaya Engineering College.
[2] Assistant Professor, Department of Computer Science and Engineering, Jaya Engineering College

Abstract*— Most AI-based diagnostic tools in healthcare are built on statistical correlations, which can produce misleading results when hidden variables influence the data. This work proposes a medical diagnosis framework that combines causal discovery with explainable AI to address that gap. Causal relationships among patient risk factors and outcomes are identified using the PC algorithm, LiNGAM, and DoWhy, and represented as Directed Acyclic Graphs to make the reasoning transparent. The system integrates E*xplainable AI (XAI), which can help users understand and interpret such autonomous predictions, helping to restore the users' trust as well as making the decision-making process of such systems transparent. The addition of the XAI layer on top of the Machine Learning models in an autonomous system can also work as a decision support system for medical practitioners to aid the diagnosis process. In this research paper, we have analyzed the two most popular model explainers, Local Interpretable Model- agnostic Explanations (LIME) and SHAPley Additive explanations (SHAP), for their applicability in autonomous disease prediction.

Keywords: ***PC algorithm, LiNGAM, and DoWhy, analyze clinical datasets and construct Directed Acyclic Graphs (DAGs), Explainable AI (XAI) methods like SHAP and LIME.***

I. INTRODUCTION

In the realm of healthcare, AI systems have shown considerable promise in processing clinical data and supporting diagnostic decisions. However, a long-standing concern among medical professionals is that these models rarely explain the reasoning behind their conclusions. When a model relies purely on statistical patterns in training data, it may inadvertently learn from confounding variables rather than medically meaningful signals. This can result in predictions that perform well in controlled experiments but break down in real hospital environments, where patient demographics, data collection methods, and disease prevalence differ significantly from the training set.

To address this, the proposed system integrates causal inference algorithms—the PC algorithm, LiNGAM, and the DoWhy framework—to map how individual risk factors relate to health outcomes. These relationships are visualized as Directed Acyclic Graphs, giving clinicians a structured view of the diagnostic logic. SHAP and LIME are applied for model-level and patient-level explanations respectively, while counterfactual reasoning allows practitioners to test hypothetical scenarios. The focus is not only on improving prediction accuracy, but on building a system that clinicians can genuinely trust and confidently apply in practice.

II. EXISTING SYSTEM

Current AI-based medical diagnosis systems primarily rely on correlation-based approaches rather than on true cause-and-effect relationships. This creates limitations in reliability, interpretability, and clinical usability.

2.1 Black-Box Machine Learning Models

Models such as Deep Learning and XGBoost provide high prediction accuracy. However, they behave as black boxes, meaning they do not explain how decisions are made. In healthcare, this lack of transparency makes it difficult for doctors to trust and use these systems for critical diagnosis.

2.2 Statistical Correlation Models

Traditional models like regression are more interpretable but only identify statistical correlations. They often fail to distinguish between true causation and spurious relationships caused by hidden variables (confounders). This can lead to misleading conclusions in medical analysis.

2.3 Rule-Based Expert Systems

Rule-based systems follow predefined IF–THEN rules and are fully explainable. However, they require manual knowledge input, are rigid, and cannot adapt to complex or novel medical patterns in data.

2.4 Limitations of the Existing System

- Lack of Trust: Clinicians hesitate to rely on systems without clear explanations.
- Regulatory Issues: Modern healthcare regulations require transparency and explainability.
- Poor Decision Making: Correlation-based models may produce incorrect or unsafe recommendations.

III. PROPOSED SYSTEM

The proposed system is a Causal Discovery and Explainable Diagnosis Framework designed specifically for clinical data environments. Rather than treating the prediction pipeline as a single black-box process, it is organized into distinct stages, each handling a specific aspect of the diagnostic workflow. Raw clinical data first passes through a preprocessing stage where missing values are handled, features are normalized, and redundant attributes are removed. This step is critical because causal discovery algorithms are particularly sensitive to noise and inconsistent data formatting. Once the data is clean, causal structure learning is performed using the PC algorithm, LiNGAM, and DoWhy. Each algorithm approaches the problem differently—PC uses conditional independence tests, LiNGAM assumes linear non-Gaussian noise, and DoWhy provides a formal framework for estimating causal effects—and together they produce a more reliable picture of causal structure than any single method alone. The identified relationships are visualized as a Directed Acyclic Graph (DAG), making it easier for clinicians to see which factors directly influence a given outcome. Causal features extracted from this graph are fed into a machine learning classifier for the actual diagnosis. To bridge the gap between model output and clinical reasoning, SHAP values capture global feature importance while LIME provides instance-level explanations for individual patient predictions. A counterfactual module allows users to ask questions such as “how would the diagnosis change if this patient’s blood pressure were lower?”, which can directly support treatment planning. All outputs are presented through a user interface that shows the diagnosis, the supporting causal graph, and the explanation results side by side.

ADVANTAGES

Understanding true cause-and-effect (not just correlation) using causal discovery, which reduces the risk of the model learning spurious patterns from the data.

Predictions are explained at both the population level and per-patient level using SHAP and LIME, making the system’s reasoning accessible even to clinicians without a machine learning background.

Counterfactual analysis allows exploration of treatment alternatives by simulating “what-if” scenarios, giving clinicians a practical tool for evaluating interventions before applying them.

Clinicians are more likely to adopt AI tools when they can audit the reasoning behind recommendations, and this system’s transparent design directly supports that requirement.

Because causal relationships tend to be more stable across different populations than statistical correlations, the system may generalize better when applied to clinical settings beyond those seen during training.

DISADVANTAGES

Causal discovery algorithms rely on assumptions about the data’s underlying structure (e.g., no hidden confounders, linear relationships in LiNGAM) that may not hold in real clinical datasets, potentially leading to incorrect causal graphs.

The system’s performance depends heavily on input data quality. Incomplete records, measurement errors, or selection bias in the training dataset can distort the causal structures the algorithms learn.

Counterfactual predictions are only as reliable as the causal model they are based on. If the DAG incorrectly represents the relationships between variables, the “what-if” scenarios produced may be misleading rather than helpful.

SHAP and LIME explain what the model has learned from data, not necessarily what is happening biologically. These explanations can guide interpretation but should not be treated as established medical causation.

A model trained on one hospital's patient population may not transfer well to institutions with different demographics or data collection practices, limiting how broadly it can be deployed.

Deploying the system in a real hospital setting requires careful workflow integration, staff training, and regulatory approval—factors that add significant time and cost beyond the technical development itself.

Building and maintaining such a system requires expertise spanning both causal inference methodology and clinical domain knowledge—a combination that is uncommon and difficult to find within a single team.

IV.SYSTEM ARCHITECTURE

The system architecture of the system is shown in Figure 1 below. When a user or medical practitioner has received the test results and preliminary diagnosis, the user may choose to further interpret these results. Once the user logs in, the user needs to select a disease for which the user wants to interpret the predictions.

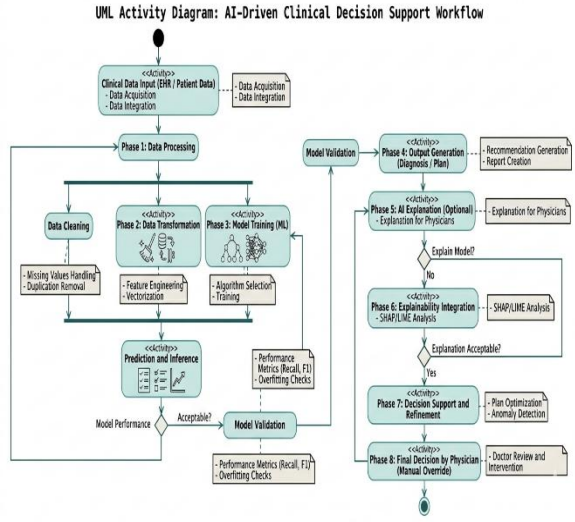

Figure 1

IV.OUTPUT (SAMPLE)

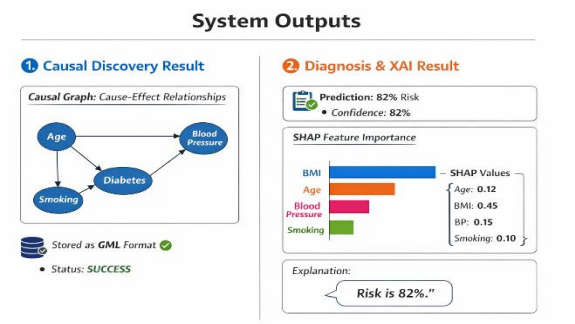

Figure 2

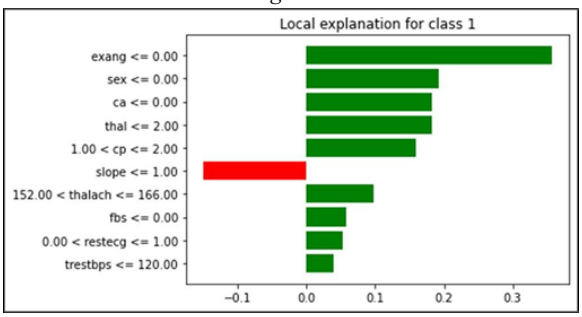

Figure 3 A detailed explanation obtained using Heart Disease Prediction.

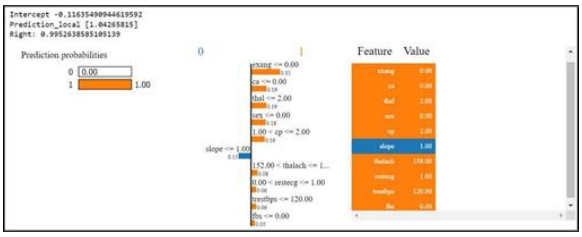

Figure 4 A detailed explanation obtained using Heart Disease Prediction.

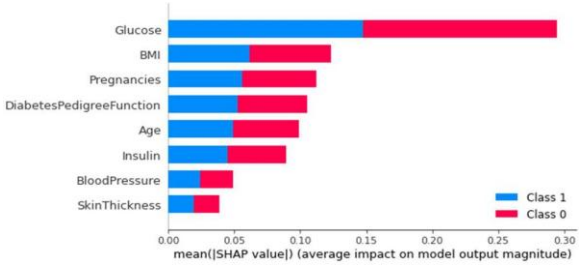

Figure 5: Visualization using SHAP for the dataset

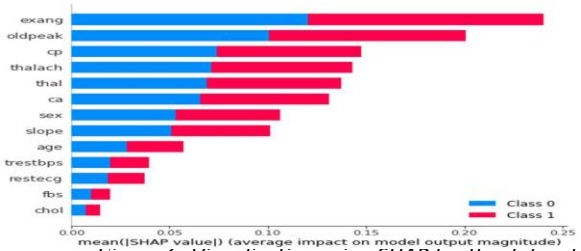

Figure 6: Visualization using SHAP for the dataset

REFERENCES

[1] Rzepiński Tomasz M. The structure of diagnosis in medicine: Introduction to interrogative characteristics. Theoretical Medicine and Bioethics. 2007;28:63–81. doi: 10.1007/s11017-006-9024-7. [DOI] [PubMed] [Google Scholar]

[2] Stanley Donald E, Campos Daniel G. The logic of medical diagnosis. Perspectives in Biology and

Medicine. 2013;56:300–315. doi: 10.1353/pbm.2013.0019. [DOI] [PubMed] [Google Scholar]

[3] Reggia James A, Nau Dana S, Peng Yun, Perricone Barry. A theoretical foundation for abductive expert systems. In: Gupta MM, Kandel A, Bandler W, Kiszka JB, editors. Approximate reasoning in expert systems. New York: Elsevier; 1985. pp. 459–472. [Google Scholar]

[4] Li C, Wang S, Xia Y, Shi F, Tang L, Yang Q, et al. et al. Risk factors and predictive models in the progression from MCI to alzheimer's disease. Neuroscience. 2025;565:312-319. [CrossRef] [Medline]

[5] Ribeiro-Dantas MDC, Li H, Cabeli V, Dupuis L, Simon F, Hettal L, et al. et al. Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients. iScience. 2024;27(5):109736. [FREE Full text] [CrossRef] [Medline]

[6] Hernán MA. The c-word: scientific euphemisms do not improve causal inference from observational data. Am J Public Health. 2018;108(5):616-619. [CrossRef] [Medline]

[7] Lundberg, S. M., Lee, S. I.: A unified approach to interpreting model predictions. In: Proceedigs of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017). https://doi.org/10.5555/3295222.3295230

[8] Ribeiro, M.T., Singh, S., 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, pp. 1135—1144 (2016). https://doi.org/10.1145/2939672.2939778

[9] Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[10] Breuer, N.O., Sauter, A., Mohammadi, M., Acar, E.: CAGE: causality-aware Shap-ley value for global explanations. In: Proceedings of Explainable Artificial Intel-ligence, xAI 2024, Communications in Computer and Information Science, vol.2155. Springer, Cham5. Springer, Cham (2024). https://doi.org/10.1007/978-3- 031-63800-8_8

[11] Spirtes, P., Meek, C., Richardson, T.: Causal inference in the presence of latent variables and selection bias. In: Proceedings of the Eleventh Conference on Uncer-tainty in Artificial Intelligence, pp. 499–506 (1995)

[12] Tyrovolas, M., Kallimanis, N., Stylios, C.: Advancing explainable AI with causal analysis in large-scale fuzzy cognitive maps. ArXiv, abs/2405.09190 (2024).https://doi.org/10.48550/arXiv.2405.09190

[13] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems 30 (2017).

[14] Explainable AI Frameworks Driving A New Paradigm for Transparency In AI. (Online) Available at: https://analyticsindiamag.com/8-explainable-aiframeworks-driving-a-new-paradigm-for-transparencyin-ai/ Accessed on 20th June 2022.

[15] Pima Indians Diabetes Database.Available at:= https://www.kaggle.com/uciml/pima-indians-diabetes-database Accessed on 5th July 2022.