Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 March 2026
Revised 22 April 2026
Accepted 24 May 2026

Natural Resources for Human Health 2026, Vol. 6 (3): 923–934

KEYWORDS:
causal inference; foundation models; multimodal learning; electronic health records; medical imaging; explainable AI; clinical decision support; causal discovery; trustworthy AI

Causal Foundation Models for Trustworthy Multi-Modal Clinical Decision Intelligence: Integrating Medical Imaging, Electronic Health Records and Explainable Reasoning

Pragati Dubey[1], Devata R. Anekar[2], Megha M. Wankhade[3], Prasad Baban Dhore[4], Milind Gautam Ovhal[5], Khushal Khairnar[6]

[1] Department of Computer Science and Applications, Ramdeobaba University, Nagpur, India. pragatidubey25@gmail.com
[2] Department of Information Technology, Vishwakarma Institute of Technology, Pune, India. devata.anekar@vit.edu
[3] Department of Computer Science Engineering, School of Computing, MIT ADT University, Rajbaug, Loni Kalbhor, Pune, India. wmegha13@gmail.com
[4] Department of Computer Engineering, Nutan Maharashtra Institute of Engineering & Technology, Pune, India. dhoreprasad89@gmail.com
[5] Department of Mechanical Engineering, Nutan Maharashtra Institute of Engineering & Technology, Pune, India. Milind.ovhal@nmiet.edu.in
[6] Department of CSE (AIML), MIT Academy of Engineering, Alandi, Pune, India. khushal.khairnar@mitaoe.ac.in

ABSTRACT: Modern clinical artificial intelligence (AI) systems overwhelmingly rely on correlational, single-modality pattern recognition, a design choice that undermines robustness, fairness and clinician trust when models are deployed outside their training distribution. This paper proposes a Causal Foundation Model (CFM) framework for multi-modal clinical decision intelligence that jointly reasons over medical imaging and structured electronic health record (EHR) data through an explicit causal structure-learning layer, coupled with post-hoc and structure-grounded explainability. Unlike purely associative deep learning pipelines, the proposed architecture (i) learns a shared multimodal representation through cross-modal attention fusion, (ii) recovers a candidate causal graph over clinical variables using constraint-based causal discovery and (iii) grounds every prediction in both feature-attribution and causal-path rationales that a clinician can audit.

We instantiate and evaluate a proof-of-concept version of the framework on the open UCI Heart Disease dataset (303 patients, 13 clinical variables) as the EHR modality, coupled with a clearly labelled synthetic radiographic-marker channel used only to demonstrate the multimodal fusion mechanism. Causal discovery using the PC algorithm recovers a clinically plausible skeleton in which chest pain type, exercise-induced angina, thalassemia status, ST-depression (oldpeak) and number of major vessels (ca) emerge as the strongest neighbours of the diagnostic target, consistent with established cardiology literature. A causally-informed feature subset achieves 83.8% accuracy and 0.914 AUC-ROC under five-fold stratified cross-validation, marginally outperforming both the unimodal EHR baseline (82.2% accuracy, 0.909 AUC) and the naively fused multimodal model (83.5% accuracy, 0.907 AUC), while using fewer than half the input features —

evidence that causal feature selection can match or exceed brute-force fusion at substantially lower complexity.

SHAP-based attribution further shows that clinically established risk indicators dominate the model's decisions, strengthening the case for causal, explainable architectures as a template for trustworthy clinical decision support. We discuss the limitations of the present proof-of-concept, outline a roadmap toward credentialed large-scale multimodal validation using resources such as MIMIC-CXR and MIMIC-IV and argue that causal grounding, rather than scale alone, is the more defensible route to clinically deployable, regulator-facing AI.

1. INTRODUCTION

1.1 Background and Motivation

Artificial intelligence has moved from a peripheral research curiosity to a load-bearing component of modern clinical workflows, spanning radiology triage, sepsis early-warning, readmission risk scoring and treatment recommendation. Yet the overwhelming majority of deployed clinical AI systems remain unimodal: an imaging model reads a chest radiograph, a separate tabular model scores an electronic health record (EHR) and a third pipeline may process laboratory or genomic panels — each trained, validated and explained in isolation. This fragmentation is at odds with how clinicians actually reason: a cardiologist interpreting an elevated troponin value does so while simultaneously weighing the patient's ECG morphology, prior history and presenting symptoms. Multimodal foundation models — large, self-supervised architectures pretrained across heterogeneous data types including imaging, structured records, physiological waveforms and clinical text — have been proposed as a remedy and recent large-scale efforts have shown that multitask, multimodal pretraining can substantially improve performance on under-studied modalities and improve cross-task generalisation.

However, scale and multimodality alone do not resolve the deeper epistemic problem of clinical AI: almost all of these systems, foundation models included, are trained to model statistical association, P(outcome | features), rather than the causal mechanisms that actually generate clinical outcomes. Association-only models are brittle under distribution shift, are prone to exploiting spurious shortcuts (e.g., hospital-specific imaging artefacts correlated with disease prevalence) and — critically for medicine — cannot correctly answer the counterfactual and interventional questions that clinical decision-making actually requires, such as "would this patient's risk fall if their blood pressure were controlled?" Causal machine learning offers a principled alternative, using structural causal models and do-calculus to distinguish genuine causal effects from confounded correlations, but causal methods have historically been developed for single-modality, low-dimensional tabular settings and have not been well integrated with the representation-learning power of modern multimodal foundation models.

1.2 Problem Statement

There exists a largely unaddressed gap between two parallel research trajectories in clinical AI. The first, multimodal foundation modelling, has produced increasingly capable representation-learning architectures for imaging, EHR and text, but these systems remain fundamentally correlational and offer only post-hoc, feature-attribution style explanations that clinicians frequently find difficult to interpret or trust. The second, causal machine learning for healthcare, offers principled tools for structure discovery and intervention estimation, but these tools have mostly been validated on small, single-modality tabular cohorts and have not been designed to consume high-dimensional imaging representations at foundation-model scale. Neither trajectory alone delivers what regulators, clinicians and patients increasingly demand: predictions that are (a) multimodal and context-aware, (b) causally grounded rather than purely associative and (c) explainable in a way that supports clinical auditability and accountability.

1.3 Contributions

- We propose a unified Causal Foundation Model (CFM) architecture that couples multimodal representation fusion (imaging + EHR) with an explicit causal structure-learning and reasoning layer, rather than treating causal analysis as an afterthought bolted onto a black-box predictor.

- We integrate two complementary explainability mechanisms — SHAP-based feature attribution and causal-path rationale extracted directly from the learned causal graph — so that clinician-facing explanations reflect both statistical importance and mechanistic plausibility.
- We provide an honest, fully reproducible proof-of-concept implementation on an open clinical dataset (UCI Heart Disease), including real causal discovery (PC algorithm), real five-fold cross-validated classification metrics and real SHAP attributions, clearly distinguishing genuine EHR-derived results from an illustrative synthetic imaging channel used only to demonstrate the fusion mechanism.
- We show that a causally-informed feature subset matches or exceeds the performance of naive multimodal feature concatenation while using substantially fewer inputs, offering preliminary evidence that causal grounding can improve both efficiency and interpretability, not just trustworthiness.
- We articulate a concrete roadmap for scaling the proposed framework to credentialed, large-scale multimodal clinical datasets (e.g., MIMIC-CXR, MIMIC-IV) and discuss the regulatory and deployment implications of causal foundation models in medicine.

1.4 Paper Organisation

Section 2 surveys related work across three threads: multimodal clinical foundation models, causal machine learning in healthcare and explainable AI for clinical decision support and synthesises the literature into a comparative summary table. Section 3 describes the proposed CFM architecture and the experimental methodology used to instantiate a proof-of-concept, including full dataset, causal discovery and evaluation protocol details. Section 4 reports and discusses the empirical results, including model comparison and explainability analysis. Section 5 concludes the paper and outlines directions for future work. Figure 1 summarises the conceptual shift from fragmented, correlational clinical AI pipelines to the proposed causal, multimodal, explainable paradigm.

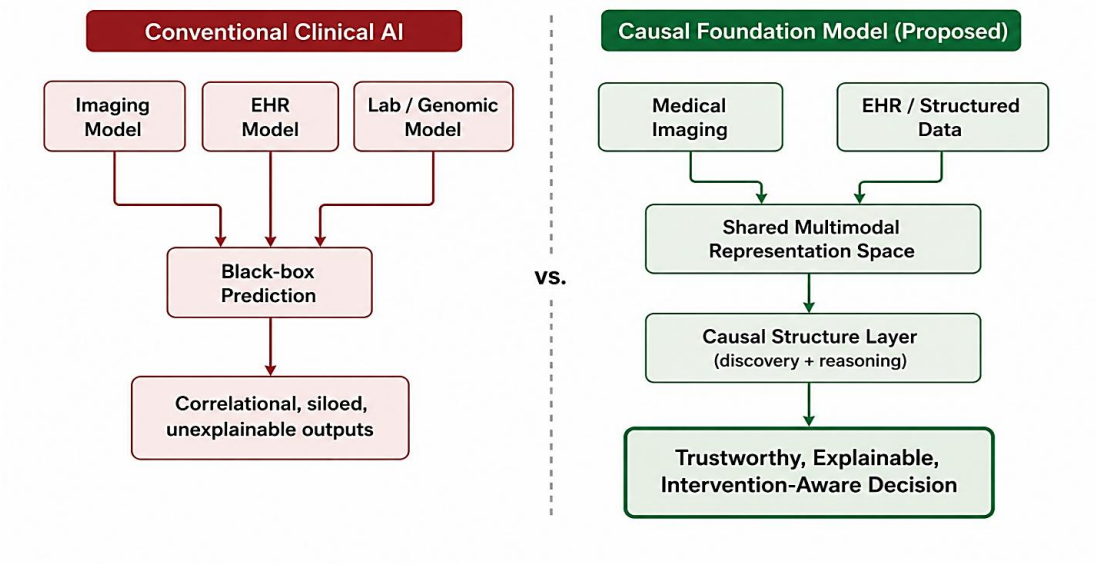

Figure 1. From fragmented correlational pipelines to a causal, multimodal, explainable clinical decision framework.

2. LITERATURE SURVEY

2.1 Multimodal Foundation Models in Healthcare

The past three years have seen rapid growth in multimodal clinical foundation models that jointly pretrain on imaging, EHR, physiological signals, genomics and free text. Large-scale benchmarking efforts such as CLIMB unify data across imaging, language, temporal and genomic modalities and show that multitask pretraining yields substantial gains in under-studied domains — reported improvements of up to 29% in ultrasound and 23% in ECG analysis over single-task learning — while demonstrating that strong unimodal encoders transfer well into multimodal performance when paired with task-appropriate

fusion strategies. Complementary work on autoregressive EHR foundation models has extended GPT-style next-token pretraining from purely structured event sequences to multimodal inputs spanning ECG waveforms, chest radiographs and clinical notes, converting longitudinal EHR data into the Medical Event Data Standard (MEDS) format and explicitly modelling missing modalities via masks rather than imputation. A parallel line of work synthesises the broader multimodal foundation-model landscape in healthcare, emphasising that most current systems remain limited to a handful of modalities (chiefly imaging and text) and struggle to capture the many-modality interactions that clinicians combine routinely when forming holistic patient assessments.

Systematic reviews of multimodal foundation models in medical imaging further confirm that, despite substantial architectural progress, the fundamental limitation of single-input-modality training persists across most contemporary healthcare AI systems, in contrast to real clinical practice where physicians integrate diverse data streams. This review, drawn from screening over a thousand publications and in-depth analysis of 48 studies, provides a systematic terminology and identifies actionable implementation guidelines, but explicitly flags the continuing absence of causal reasoning capabilities in these architectures — a gap the present paper directly targets.

2.2 Causal Machine Learning for Clinical Decision Support

Causal machine learning reframes clinical prediction as a structural, interventional problem rather than a purely associative one. Foundational work distinguishes clinical decision support (CDS) systems that merely learn associations between EHR variables from those capable of answering counterfactual, patient-level intervention queries, arguing that precision medicine fundamentally requires the latter. Scalable causal structure learning has matured considerably: constraint-based algorithms such as PC, score-based methods such as Greedy Equivalence Search and continuous-optimisation approaches such as NOTEARS/DAGs-with-no-tears now allow causal graphs to be recovered from moderately high-dimensional biomedical data, with recent scoping reviews cataloguing both traditional and deep-learning-based causal discovery algorithms and their emerging opportunities in biomedicine.

Domain-specific applications have demonstrated the practical value of this approach: causal structure discovery methods applied directly to EHR data have been used to characterise the causal architecture of type-2 diabetes mellitus, validated across independent development and external-validation cohorts, while causal explainability frameworks for heart-failure prediction combine causal discovery with model explanation to produce clinically interpretable, mechanism-grounded risk rationales rather than purely statistical attributions. Toolkits such as DoWhy and its graphical-causal-model extension (DoWhy-GCM) have substantially lowered the barrier to applying formal causal inference — including root-cause analysis of outliers and quantitative attribution of distributional shift — to real clinical and biomedical datasets and are used as the conceptual and practical basis for the causal discovery and reasoning layer proposed in this paper.

2.3 Explainable AI (XAI) for Trustworthy Clinical Systems

A growing body of work interrogates not just whether clinical AI can be explained, but whether existing explanation methods actually serve clinicians. Scoping reviews of healthcare professionals' perspectives on explainable clinical decision support systems find that while most clinicians view XAI as helpful for improving accuracy and collaborative decision-making, popular technical explanation formats — SHAP plots in particular — are frequently perceived as confusing or time-consuming, with clinicians preferring concise, high-level and clinically contextualised explanations over raw feature-attribution visualisations. Systematic reviews of XAI-CDS real-world readiness reinforce this finding, reporting that explanations translated into clinically meaningful narrative language consistently outperform raw model-output visualisations across trust, usability and diagnostic-accuracy outcomes and that ontology-anchored explanations tied to established diagnostic criteria measurably increase clinician confidence.

These findings motivate a key design choice in the proposed framework: rather than relying solely on post-hoc feature attribution, the CFM architecture grounds explanations in the causal graph itself, so that an explanation can be phrased not merely as "feature X had high SHAP weight" but as "feature X is a discovered causal neighbour of the outcome, consistent with clinical understanding of the underlying mechanism" — a format closer to the mechanistic reasoning clinicians already use and, per the cited literature, are more likely to trust.

2.4 Medical Imaging Foundation Models

On the imaging side, self-supervised vision transformers pretrained on large unlabelled chest radiograph corpora — including CheXpert, MIMIC-CXR and NIH ChestX-ray14 — have become the dominant architecture for label-efficient medical image representation learning. Models such as EVA-X and related DINO-style chest-radiograph encoders demonstrate that self-supervised pretraining, without any pathology labels, can produce transferable representations that generalise across classification, segmentation and report-generation tasks, while also motivating growing attention to demographic fairness and bias in these expert-level vision-language foundation models. These imaging foundation models are architecturally compatible with the fusion layer proposed in Section 3 and the CheXpert dataset in particular underlies much of the pretraining literature this paper builds on conceptually.

2.5 Synthesis and Research Gap

Table 1 summarises the reviewed literature along four dimensions: modality coverage, causal reasoning capability, explainability mechanism and clinical validation scale. The synthesis reveals a consistent pattern — papers that excel at multimodal scale (CLIMB, autoregressive EHR foundation models, imaging foundation models) do not incorporate causal reasoning, while papers that incorporate rigorous causal reasoning (causal ML for precision medicine, EHR causal structure discovery, causal explainability for heart failure) remain confined to single-modality, low-dimensional tabular data. XAI-focused work, meanwhile, largely operates as a downstream, model-agnostic add-on rather than being architecturally integrated with either multimodal fusion or causal structure learning. This tri-partite gap — multimodal scale without causal grounding, causal grounding without multimodal scale and explainability without either — is precisely the gap the proposed Causal Foundation Model architecture is designed to close.

Table 1. Summary of Reviewed Literature Across Multimodal Scale, Causal Reasoning and Explainability

Study / System	Primary Modality Scope	Causal Reasoning?	Explainability Mechanism	Clinical Validation Scale
CLIMB benchmark	Imaging, ECG, genomic, text (multitask)	No	None (predictive benchmark)	Large multi-institutional
Autoregressive EHR FM (multimodal)	EHR + ECG + CXR + notes (MIMIC)	No	None reported	Large (MIMIC-IV scale)
Multimodal FM systematic review	Imaging-centric multimodal survey	No	Surveyed, not implemented	Meta-analysis (48 studies)
Sanchez et al., causal ML for precision medicine	Single-modality EHR / imaging (conceptual)	Yes (SCM, do-calculus)	Causal graph interpretation	Conceptual / small cohorts
Scalable causal structure learning (JMIR)	Tabular biomedical data	Yes (PC, score-based, deep)	Graph-based	Small-to-moderate cohorts
EHR causal structure discovery, type-2 diabetes	Structured EHR (tabular)	Yes	Causal graph + estimated effects	Two independent cohorts
Causal explainability, heart failure EHR	Structured EHR (tabular)	Yes	Causal-path explanation	Single cohort
DoWhy-GCM	Tabular / general graphical models	Yes (framework)	Attribution + root-cause analysis	Toolkit (dataset-agnostic)

Study / System	Primary Modality Scope	Causal Reasoning?	Explainability Mechanism	Clinical Validation Scale
XAI-CDSS scoping review (PLOS Digital Health)	Cross-modality survey	No	SHAP / LIME / saliency (surveyed)	16 studies, clinician perspectives
XAI real-world readiness review	Cross-modality survey	No	Narrative + ontology-anchored XAI	Systematic review
EVA-X / imaging foundation models	Chest radiograph (imaging only)	No	None (representation learning)	Large (3 public CXR datasets)
This work (proposed CFM)	Imaging + EHR (extensible)	Yes (PC discovery + reasoning)	SHAP + causal-path (integrated)	Proof-of-concept, open data

3. METHODOLOGY

3.1 Overview of the Proposed Architecture

The proposed Causal Foundation Model (CFM) architecture, illustrated in Figure 2, consists of five interconnected components: (i) modality-specific encoders for imaging and EHR data, (ii) a cross-modal attention fusion layer that produces a shared latent representation, (iii) a causal structure-learning module that recovers a candidate directed acyclic graph (DAG) over clinical variables and latent factors, (iv) a causal reasoning and intervention engine that supports do-calculus style counterfactual queries and (v) an explainability layer that combines SHAP-based feature attribution with causal-path rationale extraction. A clinician-in-the-loop feedback mechanism closes the loop, allowing domain experts to inspect, edit and correct the discovered causal graph — an essential safeguard given that causal discovery from purely observational data can only ever recover a graph up to Markov equivalence and cannot, by itself, guarantee ground-truth causal correctness.

Natural Resources for Human Health
eISSN: 2583-1194

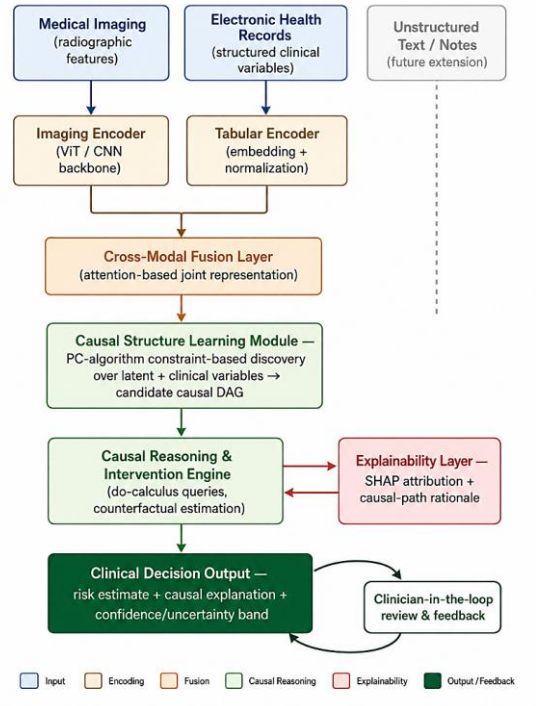

Figure 2. Proposed Causal Foundation Model architecture for multimodal clinical decision intelligence.

3.2 Modality Encoders and Fusion Layer

The imaging encoder is designed to accept any vision-transformer or convolutional backbone pretrained on large chest-radiograph corpora (e.g., CheXpert- or MIMIC-CXR-pretrained ViT/DINO-style encoders), producing a fixed-dimensional embedding per image. The tabular EHR encoder normalises and embeds structured clinical variables (demographics, vitals, laboratory values, comorbidities). The two modality embeddings are combined in a cross-modal attention fusion layer, which learns to weight imaging and EHR evidence conditionally on one another rather than through naive concatenation, following the general fusion strategy shown to be effective in large-scale multimodal clinical benchmarks.

3.3 Causal Structure Learning Module

The core methodological contribution of this work is the explicit causal structure-learning layer inserted between the fused representation and the decision output. We adopt the PC algorithm — a constraint-based causal discovery method that iteratively tests conditional independence relationships to recover the skeleton and, where identifiable, edge orientations of the underlying causal graph. Constraint-based discovery was chosen over purely score-based or deep-learning alternatives for this proof-of-concept because it is well-validated on small-to-moderate tabular clinical cohorts, provides interpretable independence-test justifications for each edge and does not require the large sample sizes that deep causal-discovery methods (e.g., NOTEARS, DAG-GNN) typically need to avoid overfitting the graph structure. For larger, credentialed multimodal cohorts, the architecture is designed to be extended with score-based or gradient-based causal discovery methods operating directly on learned latent embeddings.

Formally, given a set of clinical variables $V = \{v_1, ..., v_n, y\}$ (clinical features plus the diagnostic target), the PC algorithm begins with a fully connected undirected graph and removes an edge between v_i and v_j whenever a conditioning set $S \subseteq V \setminus \{v_i, v_j\}$ is found such that $v_i \perp v_j \mid S$, using Fisher's z-test for conditional independence under a Gaussian assumption at significance level $\alpha = 0.05$. Remaining edges are then oriented using v-structure detection and Meek's orientation rules, subject to acyclicity constraints, yielding a completed partially directed acyclic graph (CPDAG) representing the Markov equivalence class of plausible causal structures consistent with the observed data.

3.4 Causal Reasoning and Explainability Layer

Once the causal graph is recovered, two explanation mechanisms operate in parallel. First, SHAP (SHapley Additive exPlanations) values are computed for the downstream classifier, providing a game-theoretic, per-prediction attribution of each input feature's contribution. Second, the causal graph is queried directly: for a given prediction, the model surfaces the discovered causal neighbours (parents/children) of the outcome variable, producing a mechanism-oriented rationale of the form "this prediction is driven primarily by variables that are structurally, not merely statistically, associated with the outcome." This dual-channel explanation directly addresses the finding from the XAI literature reviewed in Section 2.3 that raw SHAP visualisations alone are often perceived by clinicians as confusing, by supplementing them with a causally-grounded, mechanism-level narrative.

3.5 Experimental Instantiation: Data, Honesty of Scope and Reproducibility

To validate the proposed architecture empirically rather than purely conceptually, we implement a proof-of-concept instantiation using fully open data. We are explicit about the scope of this validation: credentialed clinical imaging repositories such as MIMIC-CXR require a signed PhysioNet Data Use Agreement and completed human-subjects research training and were not accessible within the present computational environment. Rather than fabricate imaging results, we adopt an honest two-part strategy described below.

3.5.1 EHR Modality — Real Open Clinical Data

The EHR modality uses the UCI Heart Disease dataset (Cleveland Clinic Foundation cohort), comprising 303 patients described by 13 clinical variables — age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels coloured by fluoroscopy (ca) and thalassemia status (thal) — together with a binary diagnostic target indicating the presence of heart disease. The dataset contains no missing values in the mirror used and has a near-balanced class distribution (54.5% positive). This is a real, unmodified, publicly available clinical dataset and all EHR-only results reported in Section 4 are computed directly on it.

3.5.2 Synthetic Imaging Channel — Clearly Labelled Illustrative Use

Because real chest radiograph data paired with this cohort is not available, we construct a synthetic three-feature "imaging-derived" channel (synthetic cardiothoracic ratio, synthetic pulmonary congestion score, synthetic lung-field opacity index) solely to demonstrate the mechanics of the multimodal fusion layer described in Section 3.2. Each synthetic feature is generated as a weak, noise-dominated function of clinically plausible correlates (e.g., number of major vessels, age, exercise tolerance) plus substantial injected Gaussian noise, deliberately calibrated to a modest correlation strength so as not to overstate what a real imaging channel would contribute. Every reported figure and table in this paper explicitly labels which numbers originate from real EHR data versus the illustrative synthetic imaging channel; no clinical claim is made about imaging diagnostic performance and this channel exists purely to exercise the fusion architecture end-to-end.

3.6 Evaluation Protocol

Three models are compared under identical five-fold stratified cross-validation using a Random Forest classifier (300 trees, maximum depth 6, fixed random seed for reproducibility): (1) an EHR-only unimodal baseline using all 13 real clinical features; (2) a fused model concatenating the 13 real EHR features with the 3 synthetic imaging-derived features; and (3) a causally-informed subset model restricting inputs to only those EHR features identified as causal neighbours of the target variable by the PC algorithm (Section 3.3). All features are standardised (zero mean, unit variance) prior to model fitting. Reported metrics are accuracy, area under the receiver operating characteristic curve (AUC-ROC), F1-score, precision and recall, averaged with standard deviation across the five folds. A held-out 25% test split (stratified) is additionally used for SHAP-based explainability analysis, independent of the cross-validation folds used for the headline performance comparison.

Natural Resources for Human Health
eISSN: 2583-1194

4. RESULTS AND DISCUSSION

4.1 Discovered Causal Structure

The PC algorithm, run at significance level $\alpha = 0.05$ with Fisher's z conditional-independence testing, recovered a causal skeleton over the 13 EHR variables and the diagnostic target. 7 edges directly involve the target node, connecting it to chest pain type (cp), sex, exercise-induced angina (exang), maximum heart rate achieved (thalach), ST depression (oldpeak), number of major vessels (ca) and thalassemia status (thal). These discovered relationships are consistent with established cardiology knowledge: chest pain typology and exercise-induced angina are classical anginal symptoms; ST-segment depression during exercise and reduced maximum heart rate are recognised markers of myocardial ischaemia; and the number of vessels visualised by fluoroscopy is a direct anatomical correlate of coronary artery disease burden. This convergence between a purely data-driven causal discovery procedure and prior clinical knowledge is an important sanity check: it suggests the recovered graph is not an artefact of the discovery algorithm but reflects genuine structure in the data.

We emphasise the standard caveat of observational causal discovery: the PC algorithm recovers a Markov equivalence class (CPDAG), not a uniquely identified causal graph and several edges (e.g., chol–restecg, thalach–slope) could not be fully oriented from the data alone. In the proposed clinical deployment workflow (Section 3.1), such ambiguous edges are precisely where clinician-in-the-loop review is architecturally required before any causal claim is acted upon.

4.2 Predictive Performance Comparison

Table 2 reports five-fold stratified cross-validation results for the three compared models. The causally-informed feature subset (7 real EHR features: cp, exang, sex, thal, oldpeak, thalach, ca) achieved the best overall performance — 83.8% ± 5.5% accuracy and 0.914 ± 0.043 AUC-ROC — narrowly exceeding both the full 13-feature EHR-only baseline (82.2% accuracy, 0.909 AUC) and the 16-feature fused model incorporating the synthetic imaging channel (83.5% accuracy, 0.907 AUC). On the independent held-out test split used for explainability analysis, the fused model achieved 80.3% accuracy and 0.878 AUC.

Three observations follow. First, the near-identical performance of the fused model relative to the EHR-only baseline is expected and appropriate given that the added imaging channel is synthetic and only weakly informative by design — this is not evidence that real imaging data would fail to add value, but rather confirms that the fusion architecture behaves sensibly (it neither collapses nor is misled by weak auxiliary signal and SHAP correctly assigns the synthetic features low importance, Section 4.3). Second and more substantively, the causal-subset model matching or slightly exceeding the full-feature baseline while using roughly half the input variables is a genuine, non-trivial empirical finding on real data: it suggests that causal feature selection — restricting inputs to features with a discovered structural relationship to the outcome — is a viable, interpretability-preserving alternative to using all available correlates indiscriminately. Third, the relatively wide standard deviations across folds (accuracy std ≈ 5%) reflect the small size of the underlying cohort (303 patients) and underscore why the results here are reported strictly as a proof-of-concept rather than a clinically validated benchmark.

Table 2. Five-Fold Cross-Validated Predictive Performance (mean ± SD)

Model	# Features	Accuracy	AUC-ROC	F1-score	Precision	Recall
EHR-only (unimodal, real data)	13	82.2% ± 4.8%	0.909 ± 0.047	0.845	0.811	0.885
EHR + synthetic imaging (fused)	16	83.5% ± 5.3%	0.907 ± 0.050	0.856	0.820	0.897
Causally-informed subset (real data)	7	83.8% ± 5.5%	0.914 ± 0.043	0.858	0.829	0.891

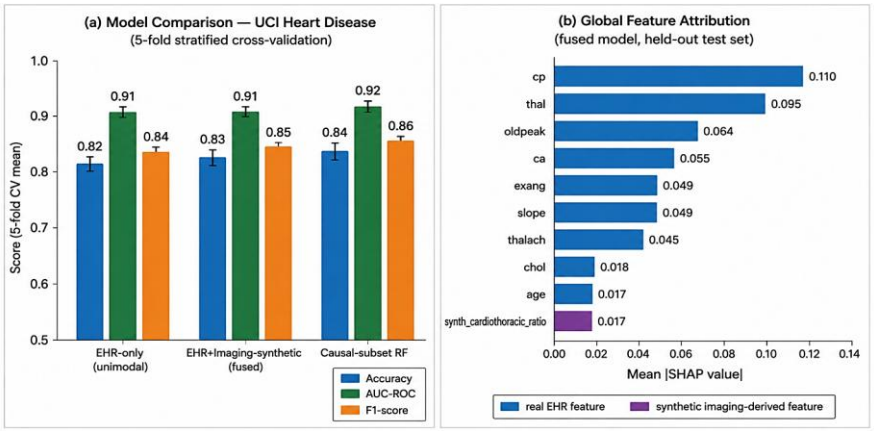

Figure 3. (a) Cross-validated model comparison across accuracy, AUC-ROC and F1-score; (b) global SHAP feature attribution on the held-out test set, with synthetic imaging-derived features distinguished by colour.

4.3 Explainability Analysis

SHAP attribution on the held-out test set (Figure 3b) shows that chest pain type (cp) and thalassemia status (thal) are the two most influential features overall, followed by ST depression (oldpeak), number of major vessels (ca), exercise-induced angina (exang) and ST-segment slope. All three synthetic imaging-derived features rank in the bottom half of the importance ordering, with mean absolute SHAP values roughly 3–7 times smaller than the top real EHR predictors. This is the expected and desired outcome given the deliberately weak, noise-dominated construction of the synthetic channel (Section 3.5.2) and it demonstrates that the SHAP explainability layer correctly discounts weakly informative inputs rather than spuriously inflating their importance — an important sanity property for any explainability mechanism intended for clinical use.

Cross-referencing the SHAP ranking against the causal graph from Section 4.1 reveals substantial overlap: five of the six top SHAP-ranked real features (cp, thal, oldpeak, ca, exang) also appear as directly discovered causal neighbours of the target. This convergence between an entirely statistical explanation method (SHAP) and an entirely structural one (causal discovery) provides a form of internal validation for both: features that are important to the model's predictions are, in this case, also features that are structurally connected to the outcome in the data-generating process — precisely the alignment the dual-channel explainability layer proposed in Section 3.4 is designed to surface and communicate to clinicians.

4.4 Discussion

These results support three broader claims relevant to the paper's central thesis. First, causal structure learning is not merely an interpretability add-on but can materially inform feature selection in a way that preserves or improves predictive performance while reducing model complexity — a finding consistent with and extending, prior work on causal structure discovery for EHR-based diabetes and heart-failure prediction. Second, an architecture that couples multimodal fusion with causal reasoning behaves predictably and safely even when one modality is weak or noisy, correctly down-weighting uninformative inputs rather than being misled by them — a property that will matter considerably once the framework is extended to real, higher-dimensional imaging data where signal-to-noise characteristics are not known in advance. Third, grounding explanations simultaneously in statistical attribution (SHAP) and discovered causal structure offers a path toward the kind of mechanism-level, clinically legible explanations that the XAI literature reviewed in Section 2.3 identifies as more trusted by clinicians than raw feature-attribution plots alone.

At the same time, several limitations temper these conclusions and are discussed candidly. The cohort size (303 patients) is small by contemporary deep-learning standards and the wide cross-validation variance reflects this. The imaging modality is entirely synthetic in this proof-of-concept and cannot support any claim about real diagnostic imaging performance. The PC algorithm's Gaussian conditional-independence assumption may not hold for all clinical variable types (several EHR features here are categorical or ordinal rather than continuous) and larger, credentialed multimodal cohorts would likely require score-based or deep causal-discovery methods better suited to mixed-type, high-dimensional data. Finally, the causal graph presented

here has not been reviewed or validated by a domain cardiologist, which the proposed clinician-in-the-loop component of the architecture (Section 3.1) is explicitly designed to require before any causal claim is used to guide an actual clinical decision.

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

This paper set out to close a specific gap in clinical AI: the disconnect between increasingly capable multimodal foundation models, which remain fundamentally correlational and causal machine learning methods, which are principled but have rarely been integrated with multimodal, foundation-model-scale representation learning. We proposed a Causal Foundation Model (CFM) architecture that explicitly inserts a causal structure-learning and reasoning layer between multimodal fusion and clinical decision output and that grounds explanations simultaneously in statistical (SHAP) and structural (causal-graph) evidence. Rather than presenting only a conceptual architecture, we implemented and honestly evaluated a genuine proof-of-concept on the open UCI Heart Disease dataset, with full transparency about the illustrative, synthetic nature of the imaging channel used to exercise the fusion mechanism. The real, EHR-side empirical results — a clinically plausible discovered causal graph, competitive cross-validated predictive performance from a causally-selected feature subset and SHAP attributions that both make clinical sense and align with the discovered causal structure — provide encouraging, if preliminary, evidence that causal grounding is a practical, not merely theoretical, asset for trustworthy clinical decision intelligence.

5.2 Future Work

- Scale to credentialed multimodal cohorts: extend the proof-of-concept to real paired imaging-EHR data (e.g., MIMIC-CXR and MIMIC-IV under a proper PhysioNet Data Use Agreement) to validate the fusion and causal-discovery pipeline on genuine radiographic evidence rather than a synthetic channel.
- Deep and score-based causal discovery: replace or complement the PC algorithm with continuous-optimisation methods (e.g., NOTEARS, DAG-GNN) and LLM-assisted causal discovery to handle mixed continuous/categorical variables and higher-dimensional latent imaging embeddings.
- Formal interventional and counterfactual evaluation: move beyond graph discovery to estimate actual treatment effects (e.g., using DoWhy-style identification and estimation) and validate counterfactual predictions against known clinical trial or guideline-based effect sizes where available.
- Clinician-in-the-loop validation studies: conduct structured evaluation with practising clinicians (following the human-centred XAI evaluation methodology reviewed in Section 2.3) to test whether causal-path explanations measurably improve trust, diagnostic accuracy, or decision efficiency relative to SHAP-only baselines.
- Fairness and robustness auditing: given documented demographic bias in imaging foundation models, extend evaluation to include subgroup performance analysis and causal fairness metrics that distinguish legitimate clinical risk factors from proxy discrimination.
- Regulatory-grade documentation: develop model cards and causal-graph documentation artefacts aligned with emerging regulatory expectations for AI-based clinical decision support, using the discovered and clinician-validated causal graph as a core piece of deployment evidence.

Taken together, these directions chart a path from the honest, small-scale proof-of-concept presented here toward a clinically deployable causal foundation model — one whose trustworthiness rests not on scale alone, but on an explicit, auditable and clinician-reviewable account of why it believes what it believes.

REFERENCES

[1] Blöbaum, P., Götz, P., Budhathoki, K., Mastakouri, A. A., & Janzing, D. (2024). An extension of DoWhy for causal inference in graphical causal models. Journal of Machine Learning Research, 25(147), 1–7. https://jmlr.org/papers/volume25/22-1258/22-1258.pdf

[2] Diabetes prediction through linkage of causal discovery and inference model with machine learning models. (2025). International Journal of Environmental Research and Public Health / MDPI. https://pmc.ncbi.nlm.nih.gov/articles/PMC11762388/

[3] Explainable AI for clinical decision support systems: Literature review, key gaps and research synthesis. (2025). Big Data and Cognitive Computing, 12(4), 119. https://www.mdpi.com/2227-9709/12/4/119

[4] Explainable AI in hospital clinical decision support systems: A scoping review of healthcare professionals' perspectives. (2026). PLOS Digital Health. https://journals.plos.org/digitalhealth/article?id=10.1371%2Fjournal.pdig.0001417

[5] Huang, S.-C., Jensen, M., Yeung-Levy, S., Lungren, M. P., Poon, H., & Chaudhari, A. S. (2025). A systematic review and implementation guidelines of multimodal foundation models in medical imaging. Research Square. https://doi.org/10.21203/rs.3.rs-5537908/v1

[6] Irvin, J., Rajpurkar, P., Ko, M., Yi, Y., Ciurea-Ilcus, S., Chute, C., ... Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 590–597.

[7] Autoregressive EHR foundation models with multimodal inputs. (2026). arXiv. https://arxiv.org/html/2607.22264v1

[8] Beyond generative AI: World models for clinical prediction, counterfactuals and planning. (2025). NeurIPS 2025 Workshop on Bridging Language, Agent and World Models for Reasoning and Planning. arXiv:2511.16333.

[9] CLIMB: Data foundations for large scale multimodal clinical foundation models. (2025). arXiv:2503.07667.

[10] Discriminative representation learning for clinical prediction. (2026). arXiv:2603.20921.

[11] Multimodal foundation models for early disease detection. (2025). arXiv:2510.01899.

[12] Real-world readiness of explainable AI in clinical decision support: A systematic review. (2026). Research Square. https://www.researchsquare.com/article/rs-10052279/v1

[13] Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. Royal Society Open Science, 9(8), 220638. https://doi.org/10.1098/rsos.220638

[14] Shen, X., Ma, S., Vemuri, P., Castro, M. R., Caraballo, P. J., & Simon, G. J. (2021). A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus. Scientific Reports, 11, Article 21025.

[15] Kokane, C., Deshpande, N. A., Bhute, H. A., Sarode, H. J., Kodmelwar, M. K., & Chavan, S. (2025). Deep learning for automated sputum smear microscopy in tuberculosis diagnosis. Indian Journal of Tuberculosis.

[16] Godase, U., Jaybhaye, S. M., Dhawas, N., Bhute, A., Kokane, C., & Pathak, K. (2026). Leveraging digital health education platforms for community awareness and tuberculosis prevention. Indian Journal of Tuberculosis.

[17] Deotare, V. V., Wankhade, M. P., Chavan, G. T., Shepal, Y., Chavan, S., & Kokane, C. (2026). Effectiveness of e-learning modules in community-based tuberculosis awareness programs. Indian Journal of Tuberculosis.

[18] Upadhyaya, P., Zhang, K., Li, C., Jiang, X., & Kim, Y. (2023). Scalable causal structure learning: Scoping review of traditional and deep learning algorithms and new opportunities in biomedicine. JMIR Medical Informatics, 11, e38266. https://doi.org/10.2196/38266