BAYESIAN RELATION PROPAGATION FOR INTERPRETABLE MULTIMODAL MEDICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal medical models benefit from flexible interactions among heterogeneous inputs, but predictive objectives alone may lead to unstable or incidental feature dependencies. On the other hand, imposing predefined relational structures can improve structural transparency while suppressing useful predictive interactions. We propose a method of **B**ayesian **R**elation **P**ropagation (BRP) that preserves the original predictive pathway while adaptively incorporating relational information through a separate relation-propagation branch. BRP first identifies stable feature relations that persist under repeated subsampling and uses them to determine a relational direction. A disease prior provides feature relevance retrieved using a calibrated predictive posterior, while predictive uncertainty and Bayesian network reliability jointly control the propagation strength. The relational direction and propagation strength are then combined through residual relation propagation to adaptively update the predictive representations. We further characterize sufficient conditions under which relation propagation can simultaneously reduce local relational discrepancy and prediction loss. On MILK10k and CheXchoNET, BRP achieves a Macro-F1 of 0.55 and an exact match ratio of 0.87, improving predictive performance over the strongest competing methods by 1.85% and 2.35%, respectively. Compared with a strong predictive baseline of SOTA, BRP further reduces normalized Structural Hamming distance by 51.9% and 56.5%, respectively, implying stronger agreement with expert-informed relational structures. Additional analyses further support the robustness and model-agnostic applicability of BRP. Our code will be released upon publication.

1 INTRODUCTION

Recent advances in multimodal learning have shifted medical AI toward integrating clinical information, including medical imaging, electronic health records, and metadata Liang et al. (2023); Dufumier et al. (2025). To exploit complementary evidence, models use adaptive fusion mechanisms to capture intra- and inter-modal dependencies. These methods range from joint-modality representation alignment Nguyen et al. (2026) to mixture-of-experts architectures modeling distinct cross-modal interactions Xin et al. (2025), achieving strong predictive performance (see Fig. 1(a)).

Despite these predictive gains, conventional predictive representation learning (PRL) methods operate as black boxes, making it difficult to understand which feature relationships drive their predictions. Optimizing only for predictive performance does not guarantee that the learned relationships are clinically meaningful. Post-hoc attribution methods can identify important features, but they do not explain how features interact or whether the predictions rely on spurious correlations Sridhar et al. (2026); Zhu et al. (2026); Sepehri et al. (2025). This limitation is especially important in medical AI, where trustworthy decisions require clear and clinically plausible reasoning.

To cope with these limitations, recent studies in causal representation learning (CRL) incorporate causal or relational knowledge into multimodal prediction. Interventional CRL methods modulate modality- or feature-level contributions through counterfactual intervention Wang et al. (2026); Liao et al. (2026) (Fig. 1(b)), whereas graphical CRL methods encode explicit relational structures Wang et al. (2025); Dominici et al. (2025a) (Fig. 1(c)). However, they either do not explicitly model feature-level relations or impose restrictive structural constraints, leaving the trade-off between predictive flexibility and relational consistency unresolved.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

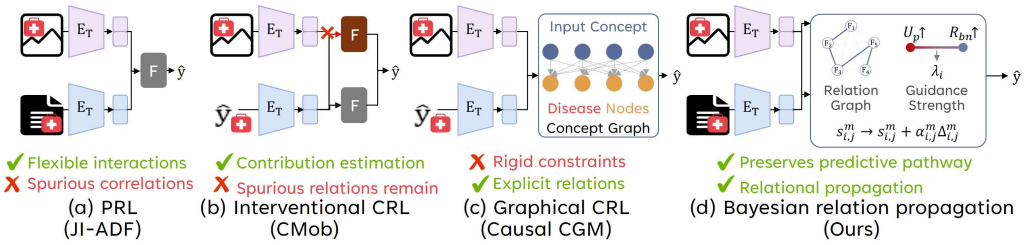


Figure 1: Comparison of multimodal learning methods. (a) PRL methods flexibly fuse multimodal features but may retain spurious correlations. (b) Interventional CRL methods estimate modality contributions, while spurious dependencies may remain. (c) Graphical CRL methods impose explicit relational graphs, improving structural transparency but limiting predictive flexibility. (d) BRP preserves the predictive pathway and adaptively controls relation propagation using predictive uncertainty and BN reliability.

Bayesian networks (BNs) provide a probabilistic framework for representing dependencies among variables and performing posterior inference Heckerman et al. (1995); Bielza & Larranaga (2014); Rohekar et al. (2018). As directed probabilistic graphical models, BNs represent explicit dependency structures and factorize the joint distribution accordingly, enabling posterior inference from observed evidence Kitson et al. (2023). The resulting BN posterior provides structure-informed probabilistic evidence complementary to the predictive posterior of the PRL model, supporting sample-specific control of relation propagation.

Building on these properties, we propose a new method of **B**ayesian **R**elation **P**ropagation (BRP) that adaptively propagates relational information into predictive representations while preserving the original predictive pathway. The relational direction is derived from stable feature dependencies that persist under repeated subsampling, reducing reliance on unstable and potentially spurious correlations. In parallel, a calibrated predictive posterior retrieves feature relevance from a disease prior. Predictive uncertainty and BN reliability determine the guidance strength, which is combined with feature relevance to adaptively control relation propagation. We will demonstrate the effectiveness with experiments on MILK10k Philipp et al. (2025) and CheXchoNET Elias & Bhave (2024), and analysis of normalized Structural Hamming Distance (nSHD) to show the stronger agreement with expert-informed relational structures. Further, applying BRP to multiple predictive architectures yields consistent performance improvements without architecture-specific modifications, demonstrating its model-agnostic applicability.

2 Related Work

We review multimodal medical learning from two perspectives relevant to our study: PRL and CRL.

Medical PRL. Multimodal medical AI has advanced through stronger representation backbones for extracting rich representations from heterogeneous clinical data Hatamizadeh & Kautz (2025); Shi (2024). Recent studies further employ flexible multimodal integration mechanisms to capture complementary information from medical images, clinical variables, and other modalities Huang et al. (2024); Yan et al. (2025). Nguyen et al. Nguyen et al. (2026) introduced adaptive fusion of joint and modality-specific representations to capture complementary information from multimodal medical inputs. Xin et al. Xin et al. (2025) modeled distinct cross-modal interactions through an architecture of mixture-of-experts that separately captures modality-specific and multimodal information. Wei et al. Wei et al. (2024) proposed sample-level modality valuation to estimate the contribution of each modality and improve multimodal cooperation through contribution-aware re-sampling. Yu et al. Yu et al. (2026) formulated multimodal learning through correlation maximization to strengthen task-relevant information shared among modalities. Du et al. Du et al. (2026) introduced dynamic modality selection at inference time, progressively incorporating modalities according to their incremental predictive contribution. Dufumier et al. Dufumier et al. (2025) proposed cross-modal representation alignment to learn more consistent and complementary representations from heterogeneous medical data. However, these methods do not explicitly regulate whether learned feature dependencies follow clinically meaningful relations.

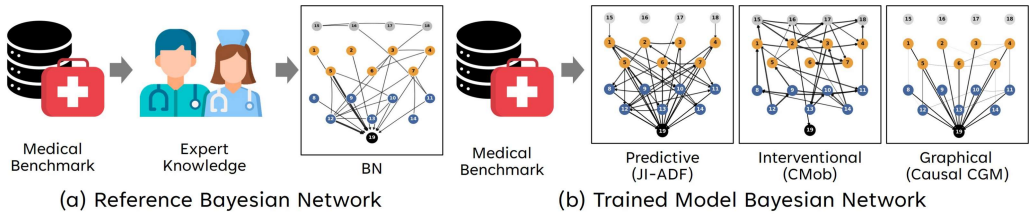


Figure 2: Structural comparison between the expert BN and model-derived BNs from existing PRL and CRL methods.

Medical CRL. To overcome the limitation of PRL, CRL methods incorporate causal or relational information through concept-based representations and intervention-aware learning Koh et al. (2020); Xu et al. (2024). For example, interventional and counterfactual methods incorporate causal effects into predictive learning to promote structured reasoning Lai et al. (2024); Dominici et al. (2025b). Hu et al. Hu et al. (2025) introduced editable concept representations, while Barbiero et al. Barbiero et al. (2024) modeled explicit relations among interpretable concepts. Graphical approaches further organize these concepts through directed or hierarchical structures Wang et al. (2025); Hill et al. (2026). Recent studies also explore latent causal factors and invariant multimodal representations Yao et al. (2024; 2025). In contrast, our approach preserves the original predictive pathway and incorporates relational structure through adaptive relation propagation guided by Bayesian reasoning, without directly imposing structural constraints on predictive representations.

3 Motivation

To examine the relationship between predictive performance and relational structure, we compare representative PRL and CRL methods on MILK10k. Predictive performance is evaluated using Macro-F1, while structural discrepancy is measured by nSHD between the expert-informed reference BN and model-derived BNs (Sup. G.3). As shown in Table 1, JI-ADF Nguyen et al. (2026) achieves strong predictive performance but exhibits greater structural discrepancy, whereas Causal CGM Dominici et al. (2025a) more closely follows the expert-informed relational structure at the cost of predictive performance, revealing a trade-off between the two objectives.

Table 1: Trade-off analysis.

Method	Macro-F1 ↑	nSHD ↓
JI-ADF	**0.540**	0.521
CMoB	0.532	0.483
Causal CGM	0.080	**0.194**

To examine whether all learned feature dependencies are necessary for strong predictive performance, we conduct a controlled relation-mask experiment on two PRL methods, JI-ADF and I^{2}MoE Xin et al. (2025). Using an LLM with domain knowledge, we identify supported feature–feature dependencies and disease–feature relations and encode them into a binary relation mask $\mathbf{M}$. The mask is incorporated into attention as $\mathrm{Softmax}((\mathbf{QK}^\top/\sqrt{d}) + \mathcal{B}(\mathbf{M}))\mathbf{V}$, where $\mathcal{B}(\cdot)$ assigns 0 to supported relations and $-\infty$ to unsupported relations.

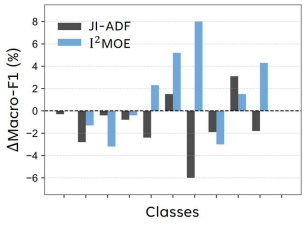


Figure 3: Performance change after applying the relation mask.

As shown in Fig. 3, despite substantially restricting feature interactions, applying $\mathbf{M}$ preserves overall predictive performance and even improves class-wise F1 for several classes. This result suggests that strong predictive performance does not necessarily require all dependencies acquired from the training data. However, hard masking completely blocks interactions outside the predefined relation set, requiring the predefined relations to be sufficiently complete across different samples and diseases and potentially suppressing predictive interactions that remain useful for individual cases. This motivates an adaptive use of relational structure: rather than imposing it as a hard constraint on the predictive pathway, we preserve the original predictive representations and incorporate stable feature relations through a separate relation-propagation pathway, while adaptively controlling their influence using Bayesian guidance. Details of the motivation experiment are provided in Sup. C.2.

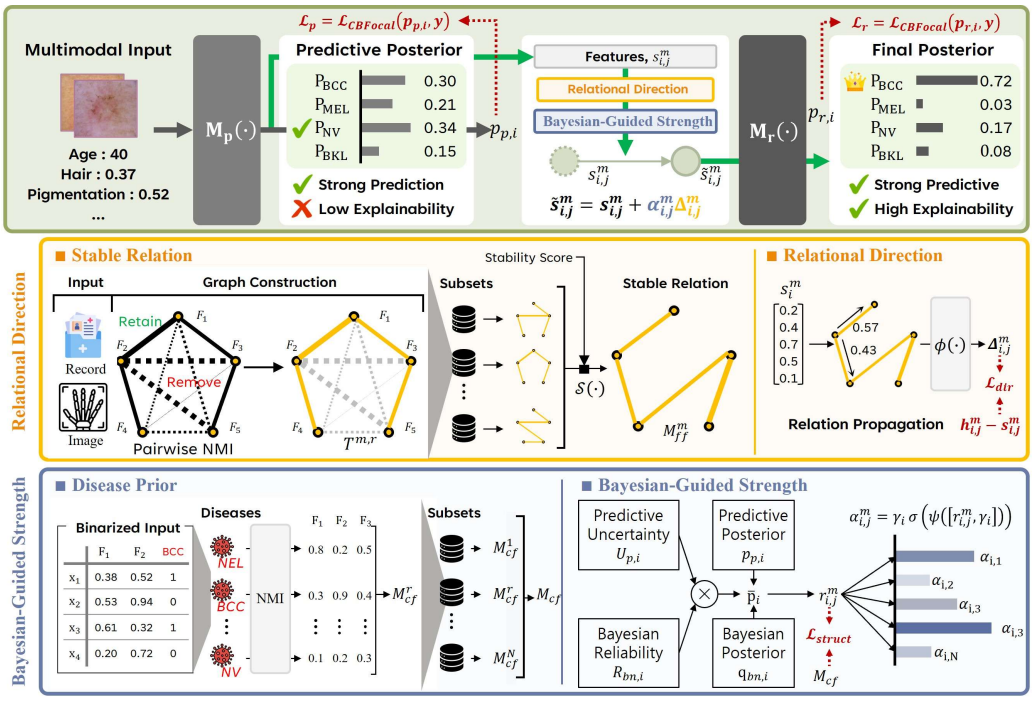


Figure 4: Overview of BRP. Repeated subsampling identifies stable relations $\mathbf{M}^m_{ff}$, which are used to form the relational direction $\Delta^m_{i,j}$. A calibrated predictive posterior retrieves feature relevance from the disease prior $\mathbf{M}_{cf}$. Predictive uncertainty and BN reliability determine the sample-level guidance strength γ_i, which is combined with feature relevance to obtain the Bayesian-guided propagation strength $\alpha^m_{i,j}$. The relational direction and propagation strength are then combined through residual relation propagation while preserving the original predictive pathway.

4 Method

Problem Setup. We use a relational direction to determine how each feature representation is updated based on stable dependencies identified through repeated subsampling, and a Bayesian-guided propagation strength to control the update magnitude using feature relevance retrieved from the disease prior, predictive uncertainty, and BN reliability. Let $\mathbf{x}_i = \{\mathbf{x}^m_i\}_{m=1}^M$ denote a multimodal input with label $y_i \in \{1, \dots, C\}$, where the predictive branch $\mathcal{M}_p$ produces the posterior $\mathbf{p}_{p,i}$ and feature representations $\mathbf{s}^m_i = [\mathbf{s}^m_{i,1}, \dots, \mathbf{s}^m_{i,K_m}]$ for each modality. To reduce reliance on unstable and potentially spurious dependencies while preserving predictive performance, BRP applies residual relation propagation to these representations.

4.1 Relational Direction

To derive a relational direction, we first identify feature dependencies that persist under repeated subsampling and use these stable relations to determine how each feature representation is propagated.

Stable Relation Estimation. A dependency estimated from a single training set may reflect either a reproducible dependency or an incidental correlation induced by sample composition. To reduce reliance on such correlations, we repeatedly estimate the relation structure from R subsampled training sets. For each subsampling iteration $r \in \{1, \dots, R\}$ and modality m, pairwise dependencies $\Pi^{m,(r)}_{jk}$ are computed using normalized mutual information (NMI). These dependencies are used as edge weights to construct a maximum-weight spanning tree $T^{m,(r)}$ with exactly $K_m - 1$ relations within each modality, following the Chow-Liu algorithm Bhattacharyya et al. (2021). Across the R trees, each candidate relation is scored according to how consistently it is selected and its dependency strength. The selection operator $\mathcal{S}(\cdot)$ then retains the $K_m - 1$ most stable relations, yielding the stable relation matrix $\mathbf{M}^m_{ff} \in \mathbb{R}^{K_m \times K_m}$.

Using this stable relation matrix, for sample $\mathbf{x}_i$, the relational context of feature j is formed as $\mathbf{h}_{i,j}^m = \sum_{k=1}^{K_m} M_{ff,jk}^m \mathbf{s}_{i,k}^m$. The original feature representation and its relational context are then combined to obtain the relational direction $\Delta_{i,j}^m = \phi^m([\mathbf{s}_{i,j}^m, \mathbf{h}_{i,j}^m])$, where $\phi^m(\cdot)$ maps the current feature representation and its relational context to a propagation direction.

4.2 Disease Prior and Bayesian Guidance

While $\mathbf{M}_{ff}^m$ determines the relational information used to form the direction, the update strength should depend on how relevant each feature is to the disease hypothesis of the current sample. We therefore adapt the strength for each sample according to disease relevance and posterior reliability. To estimate disease relevance, we construct a disease prior that characterizes feature relevance for each disease, and retrieve the corresponding relevance using posterior evidence. Since the ground-truth label is unavailable at inference time, the predictive posterior serves as a soft estimate of the disease hypothesis. However, this posterior may be uncertain or assign high probability to an incorrect disease. To complement this predictive evidence, we use the BN posterior, which provides probabilistic evidence under an explicit dependency structure.

Disease Prior. For each subsampling iteration r, we compute the relevance between feature F_j and disease c using NMI with the corresponding one-vs-rest disease indicator. These relevance scores are normalized within each disease to form a prior $\mathbf{M}_{cf}^{(r)} \in \mathbb{R}^{C \times K}$. The final disease prior is obtained by averaging the R priors as $\mathbf{M}_{cf} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{M}_{cf}^{(r)}$, where $K = \sum_{m=1}^{M} K_m$. Each row of $\mathbf{M}_{cf}$ therefore represents the relative relevance of the K features to a particular disease.

Bayesian-Guided Propagation Strength. Let $\mathbf{p}_{p,i}$ denote the posterior from the predictive branch and $\mathbf{q}_{bn,i}$ the posterior by the BN. Predictive uncertainty is measured by normalized entropy as $U_{p,i} = -\sum_{c=1}^{C} p_{p,i}^c \log p_{p,i}^c / \log C$, where $U_{bn,i} \in [0,1]$ denotes BN uncertainty and BN reliability is defined as $R_{bn,i} = 1 - U_{bn,i}$. We define the sample-level guidance strength as $\gamma_i = U_{p,i} R_{bn,i}$, so that guidance becomes stronger when the predictive branch is uncertain and the BN is reliable.

Using this strength, we calibrate the predictive posterior with BN evidence as $\bar{\mathbf{p}}_i = \mathrm{Softmax}(\log(\mathbf{p}_{p,i} + \epsilon) + \gamma_i \, \eta([\mathbf{p}_{p,i}, \mathbf{q}_{bn,i}, \gamma_i]))$, where $\eta(\cdot)$ is a learnable calibration function. The calibrated posterior retrieves feature relevance from the disease prior as $\mathbf{r}_i = \bar{\mathbf{p}}_i^\top \mathbf{M}_{cf}$. Here, $r_{i,j}^m$, the element of $\mathbf{r}_i$ corresponding to feature j of modality m, represents its relevance to the disease hypothesis of sample i. The Bayesian-guided propagation strength is then defined as $\alpha_{i,j}^m = \gamma_i \, \sigma(\psi([r_{i,j}^m, \gamma_i]))$, where $\psi(\cdot)$ is a learnable feature-wise gate. Accordingly, disease-relevant features can receive different propagation strengths even within the same sample.

Importantly, $\mathbf{M}_{ff}^m$ and $\mathbf{M}_{cf}$ play complementary roles: $\mathbf{M}_{ff}^m$ determines the stable feature relations used to form the relational direction $\Delta_{i,j}^m$, whereas $\mathbf{M}_{cf}$ provides posterior-weighted feature relevance from the disease prior that contributes to the propagation strength $\alpha_{i,j}^m$. Combining these two components, residual relation propagation updates the representation as $\widetilde{\mathbf{s}}_{i,j}^m = \mathbf{s}_{i,j}^m + \alpha_{i,j}^m \Delta_{i,j}^m$, and the resulting representations are fused to produce the relation-propagated posterior $\mathbf{p}_{r,i}$. By propagating information through stable relations identified under repeated subsampling, this formulation provides a more structurally interpretable prediction process, revealing which features are relevant to the current disease hypothesis and how stable relations contribute to the propagated representation.

4.3 Training Objective

Our training objective consists of task losses for classification and relational losses for guiding relation propagation. For the **task losses**, we train both the predictive and relation-propagated posteriors using class-balanced focal loss to account for class imbalance in the training data. We define $\mathcal{L}_p = \mathcal{L}_{\mathrm{CBFocal}}(\mathbf{p}_{p,i}, y_i)$ and $\mathcal{L}_r = \mathcal{L}_{\mathrm{CBFocal}}(\mathbf{p}_{r,i}, y_i)$. For the **relational losses**, we use two auxiliary objectives to guide relation propagation. First, $\mathcal{L}_{\mathrm{dir}}$ encourages the relational direction $\Delta_{i,j}^m$ to align with the relational residual $\mathbf{h}_{i,j}^m - \mathbf{s}_{i,j}^m$. This is further analyzed in § 4.4. Second, the prior-consistency loss $\mathcal{L}_{\mathrm{struct}}$ encourages the feature relevance $\mathbf{r}_i$ to remain close to the prior of the ground-truth disease, $\mathbf{M}_{cf}[y_i]$. The overall objective is $\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_r \mathcal{L}_r + \lambda_{\mathrm{struct}} \mathcal{L}_{\mathrm{struct}} + \lambda_{\mathrm{dir}} \mathcal{L}_{\mathrm{dir}}$. For MILK10k, we set $\lambda_p = \lambda_r = 1.0$, $\lambda_{\mathrm{struct}} = 0.3$, and $\lambda_{\mathrm{dir}} = 0.2$, and for CheXchoNET, we set $\lambda_p = \lambda_r = 1.0$, $\lambda_{\mathrm{struct}} = 0.1$, and $\lambda_{\mathrm{dir}} = 0.1$.

4.4 Theoretical Analysis

We analyze when the proposed relation propagation can reduce both local relational discrepancy and prediction loss. For a local propagation step, we treat the relation-informed context $\mathbf{h}_{i,j}^m$ computed from the current representations as fixed and define the relational residual as $\mathbf{e}_{i,j}^m = \mathbf{h}_{i,j}^m - \mathbf{s}_{i,j}^m$, with local relational discrepancy $\|\mathbf{e}_{i,j}^m\|_2$. Under $\widetilde{\mathbf{s}}_{i,j}^m = \mathbf{s}_{i,j}^m + \alpha_{i,j}^m \Delta_{i,j}^m$, the residual becomes $\widetilde{\mathbf{e}}_{i,j}^m = \mathbf{e}_{i,j}^m - \alpha_{i,j}^m \Delta_{i,j}^m$. Thus, successful propagation depends jointly on the alignment of $\Delta_{i,j}^m$ with $\mathbf{e}_{i,j}^m$ and the propagation strength $\alpha_{i,j}^m$.

Proposition 1. *Assume that the gradient of the relation-propagated prediction loss $\mathcal{L}_r$ is L-Lipschitz with respect to $\mathbf{s}_{i,j}^m$, and let $\mathbf{g}_{i,j}^m = \nabla_{\mathbf{s}_{i,j}^m} \mathcal{L}_r(\mathbf{s})$. If $\langle \mathbf{e}_{i,j}^m, \Delta_{i,j}^m \rangle > 0$ and $\langle \mathbf{g}_{i,j}^m, \Delta_{i,j}^m \rangle < 0$, then both local relational discrepancy and prediction loss decrease whenever $0 < \alpha_{i,j}^m < \min\{1, \alpha_{\text{rel}}^{\max}, \alpha_{\text{pred}}^{\max}\}$, where $\alpha_{\text{rel}}^{\max} = 2\langle \mathbf{e}_{i,j}^m, \Delta_{i,j}^m \rangle / \|\Delta_{i,j}^m\|_2^2$ and $\alpha_{\text{pred}}^{\max} = -2\langle \mathbf{g}_{i,j}^m, \Delta_{i,j}^m \rangle / (L\|\Delta_{i,j}^m\|_2^2)$.*

The first condition is encouraged by $\mathcal{L}_{\text{dir}}$ in § 4.3, which aligns $\Delta_{i,j}^m$ with the relational residual $\mathbf{h}_{i,j}^m - \mathbf{s}_{i,j}^m$. As shown in Fig. 5(a), when the propagation direction is positively aligned with the relational residual ($\theta < 90°$), a smaller angle leads to a larger reduction in relational discrepancy for a fixed strength. Fig. 5(b) illustrates the effect of propagation strength $\alpha_{i,j}^m$, which is controlled by the Bayesian-guided feature-wise gate; even with a favorable direction, the strength must remain within a direction-dependent range, since an excessive update can increase relational discrepancy.

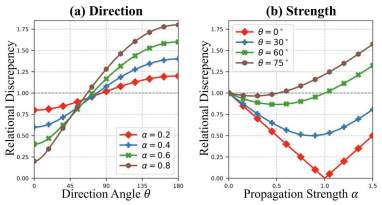


Figure 5: Effect of direction and strength on relational discrepancy.

The predictive condition is promoted through optimization of the relation-propagated task loss $\mathcal{L}_r$, encouraging $\Delta_{i,j}^m$ toward a direction that decreases prediction loss. Together, these conditions define a region in which relation propagation can improve relational consistency without increasing prediction loss. Detailed proofs are provided in Sup. D.

5 Experiments

We evaluate BRP on two multimodal prediction benchmarks from distinct medical domains: MILK10k, an 11-class skin lesion classification dataset, and CheXchoNET, a multi-label chest X-ray dataset with four joint disease states defined by severe left ventricular hypertrophy (SLVH) and dilated left ventricle (DLV). The evaluation focuses on two complementary aspects:

- **Predictive performance:** For MILK10k, we evaluate both class-wise F1 and Macro-F1 due to substantial class imbalance. For CheXchoNET, which involves multi-label classification over two disease targets, we report class-wise F1 and exact match ratio (EMR) Eissa et al. (2024).
- **Structural agreement:** We perform BN-based structural analysis using pyAgrum aGrUM/pyAgrum Developers. We construct an expert-informed reference BN from expert-annotated data and a model-derived BN from each model's predictions. Structural discrepancy is quantified using nSHD, where a lower value indicates stronger agreement with the expert-informed reference structure. For single-modality PRL, nSHD is not evaluated because they use only image inputs.

We compare against 19 baselines, including 9 single-modality PRL methods (MVFA-AD Huang et al. (2024), MambaVision Hatamizadeh & Kautz (2025), TransNeXt Shi (2024), PanDerm Yan et al. (2025), TMCEK Liang et al. (2025), ABNN Franchi et al. (2024), IVON Shen et al. (2024), and HENN Li et al. (2024),MLR-GCN Rawlekar et al. (2024)), 6 multimodal PRL methods (JI-ADF, I^{2}MoE, SMV Wei et al. (2024), TCMax Yu et al. (2026), DyMo Du et al. (2026), and LATA Bozorgtabar et al. (2026)), and four CRL methods (DARC Liao et al. (2026), NCG Wang et al. (2025), CMoB Wang et al. (2026), and Causal CGM Dominici et al. (2025a)). Dataset, baseline, implementation, and evaluation details are provided in Sup. E–G.

5.1 Comparison Results

Quantitative Analysis. Table 2 compares predictive performance and relational consistency on MILK10k. Among PRL methods, multimodal PRL generally outperform single-modality PRL, demonstrating the benefit of integrating complementary multimodal information. Compared with

Table 2: Classification performance on MILK10k. Values in parentheses indicate the class distribution (%). Methods are grouped into single-modality PRL, multimodal PRL, relational methods, and Ours. Light blue and light red denote the best and second-best results. All results are reported as the mean ± standard deviation over five folds.

Method	BCC (48.1%)	SCCKA (9.0%)	MEL (8.6%)	AKIEC (5.8%)	MAL-OTH (0.2%)	NV (14.2%)	BKL (10.4%)	DF (1.0%)	INF (1.0%)	VASC (0.9%)	BEN-OTH (0.8%)	Macro-F1 ↑	nSHD ↓
Single-Modality PRL													
MVFA-AD	0.88±0.13	0.66±0.13	0.58±0.15	0.46±0.08	0.00±0.00	0.78±0.17	0.45±0.05	0.33±0.12	0.27±0.06	0.78±0.04	0.36±0.11	0.50±0.10	–
MambaVision	0.86±0.01	0.60±0.06	0.50±0.08	0.40±0.04	0.00±0.00	0.71±0.05	0.37±0.08	0.07±0.10	0.07±0.10	0.44±0.32	0.04±0.10	0.37±0.09	–
TransNeXt	0.86±0.02	0.62±0.06	0.48±0.14	0.43±0.05	0.00±0.00	0.72±0.04	0.36±0.10	0.14±0.19	0.07±0.10	0.50±0.34	0.04±0.10	0.38±0.11	–
PanDerm	0.87±0.02	0.64±0.06	0.67±0.07	0.53±0.08	0.00±0.00	0.75±0.02	0.51±0.05	0.63±0.11	0.00±0.00	0.89±0.08	0.43±0.12	0.53±0.06	–
TMCEK	0.81±0.01	0.58±0.02	0.48±0.06	0.18±0.04	0.00±0.00	0.75±0.01	0.23±0.08	0.04±0.08	0.04±0.08	0.37±0.25	0.22±0.15	0.34±0.07	–
ABNN	0.83±0.01	0.56±0.02	0.47±0.09	0.27±0.09	0.00±0.00	0.72±0.02	0.36±0.05	0.00±0.00	0.00±0.00	0.37±0.21	0.00±0.00	0.33±0.05	–
IVON	0.79±0.02	0.41±0.11	0.30±0.07	0.02±0.03	0.00±0.00	0.64±0.03	0.10±0.03	0.00±0.00	0.00±0.00	0.10±0.14	0.00±0.00	0.21±0.04	–
HENN	0.84±0.01	0.56±0.03	0.48±0.04	0.30±0.09	0.00±0.00	0.70±0.03	0.32±0.03	0.18±0.09	0.14±0.10	0.67±0.14	0.08±0.11	0.39±0.06	–
MLR-GCN	0.74±0.02	0.16±0.03	0.21±0.02	0.00±0.02	0.00±0.00	0.59±0.05	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.16±0.01	–
Multimodal PRL													
JI-ADF	0.89±0.01	0.69±0.03	0.61±0.03	0.53±0.02	0.00±0.00	0.78±0.02	0.52±0.06	0.65±0.06	0.27±0.05	0.72±0.14	0.26±0.19	0.54±0.06	0.52±0.14
I^2MoE	0.83±0.01	0.61±0.06	0.52±0.06	0.28±0.08	0.00±0.00	0.75±0.01	0.36±0.07	0.26±0.26	0.08±0.12	0.46±0.26	0.14±0.13	0.39±0.10	0.58±0.04
SMV	0.82±0.01	0.55±0.04	0.51±0.09	0.22±0.08	0.00±0.00	0.74±0.02	0.29±0.07	0.23±0.16	0.23±0.11	0.53±0.14	0.21±0.16	0.39±0.08	0.73±0.17
TCMax	0.78±0.01	0.47±0.03	0.42±0.02	0.21±0.06	0.00±0.00	0.60±0.05	0.19±0.08	0.05±0.07	0.12±0.12	0.25±0.12	0.04±0.06	0.28±0.06	0.54±0.18
DyMo	0.81±0.01	0.51±0.08	0.42±0.09	0.13±0.14	0.00±0.00	0.70±0.02	0.18±0.10	0.00±0.00	0.00±0.00	0.13±0.08	0.00±0.00	0.26±0.06	0.75±0.07
LATA	0.03±0.01	0.01±0.01	0.36±0.02	0.01±0.01	0.00±0.00	0.12±0.02	0.16±0.02	0.04±0.04	0.02±0.02	0.01±0.02	0.02±0.04	0.07±0.02	0.73±0.12
Relational Methods													
DARC	0.88±0.00	0.64±0.00	0.67±0.00	0.55±0.00	0.00±0.00	0.73±0.00	0.49±0.00	0.67±0.00	0.14±0.00	0.62±0.00	0.31±0.00	0.52±0.00	0.73±0.16
NCG	0.82±0.01	0.53±0.05	0.39±0.08	0.17±0.12	0.00±0.00	0.72±0.02	0.31±0.06	0.00±0.00	0.00±0.00	0.04±0.09	0.00±0.00	0.27±0.04	0.38±0.21
CMoB	0.86±0.02	0.65±0.03	0.61±0.04	0.47±0.03	0.22±0.05	0.76±0.03	0.47±0.04	0.59±0.03	0.31±0.04	0.64±0.03	0.29±0.05	0.53±0.04	0.48±0.28
Causal CGM	0.60±0.35	0.04±0.06	0.05±0.11	0.00±0.00	0.00±0.00	0.14±0.21	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.08±0.07	0.19±0.07
Ours	0.89±0.01	0.69±0.04	0.63±0.05	0.50±0.03	0.13±0.30	0.78±0.03	0.51±0.04	0.62±0.17	0.25±0.11	0.69±0.16	0.28±0.08	0.55±0.09	0.25±0.21

Table 3: Classification performance and computational cost on CheXchoNET. EMR evaluates exact agreement for the two disease labels. Light blue and light red denote the best and second-best.

Method	GFLOPs	Params (M)	Inference Time (s)	SLVH (8.65%)	DLV (6.07%)	Avg. F1	Neither (86.23%)	SLVH Only (7.71%)	DLV Only (5.13%)	Both (0.94%)	EMR ↑	nSHD ↓
Single-Modality PRL												
MVFA-AD [CVPR'24]	475.77	440.23	0.04	0.56±0.28	0.59±0.37	0.58	0.86±0.13	0.17±0.11	0.14±0.14	0.07±0.08	0.74±0.12	–
MambaVision [CVPR'25]	8.92	31.15	0.01	0.02±0.03	0.16±0.04	0.09	0.91±0.03	0.01±0.02	0.15±0.04	0.02±0.02	0.84±0.03	–
TransNeXt [CVPR'24]	5.31	12.43	0.03	0.12±0.00	0.18±0.07	0.15	0.90±0.11	0.00±0.00	0.17±0.07	0.00±0.00	0.82±0.07	–
PanDerm [Nature'25]	35.14	85.91	0.02	0.24±0.01	0.20±0.02	0.22	0.84±0.03	0.08±0.06	0.01±0.00	0.05±0.01	0.69±0.03	–
TMCEK [ICML'25]	4.47	27.82	0.01	0.48±0.00	0.49±0.00	0.49	0.89±0.11	0.00±0.00	0.00±0.00	0.00±0.00	0.83±0.03	–
ABNN [CVPR'24]	349.16	33.53	0.44	0.06±0.05	0.19±0.02	0.13	0.88±0.04	0.05±0.04	0.17±0.04	0.03±0.06	0.81±0.10	–
IVON [ICML'24]	117.14	11.18	0.20	0.09±0.00	0.10±0.06	0.10	0.87±0.18	0.00±0.00	0.10±0.06	0.00±0.00	0.80±0.06	–
HENN [ICLR'24]	3.63	11.18	0.01	0.02±0.02	0.17±0.03	0.10	0.91±0.05	0.02±0.02	0.16±0.03	0.00±0.00	0.82±0.03	–
MLR-GCN [ICPR'24]	96.09	38.41	0.04	0.51±0.07	0.57±0.01	0.54	0.88±0.01	0.18±0.02	0.21±0.02	0.04±0.02	0.78±0.02	–
Multimodal PRL												
JI-ADF [arXiv'26]	16.88	47.03	0.11	0.52±0.12	0.58±0.01	0.55	0.90±0.01	0.20±0.02	0.21±0.03	0.11±0.05	0.80±0.03	0.62±0.11
I^2MoE [ICML'25]	9.46	61.37	0.31	0.50±0.01	0.57±0.01	0.54	0.92±0.00	0.04±0.02	0.17±0.02	0.00±0.00	0.85±0.01	0.64±0.05
SMV [CVPR'24]	7.25	11.51	0.01	0.21±0.12	0.25±0.18	0.23	0.90±0.11	0.15±0.09	0.17±0.02	0.15±0.04	0.81±0.07	0.61±0.13
TCMax [ICLR'26]	3.63	11.18	0.01	0.11±0.02	0.17±0.06	0.14	0.93±0.00	0.01±0.03	0.07±0.06	0.00±0.00	0.79±0.02	0.54±0.12
DyMo [ICLR'26]	8.47	38.40	0.02	0.29±0.00	0.22±0.03	0.26	0.89±0.05	0.00±0.00	0.02±0.03	0.00±0.00	0.81±0.02	0.58±0.12
LATA [CVPR'26]	16.87	87.67	0.01	0.51±0.12	0.58±0.01	0.55	0.88±0.01	0.10±0.02	0.16±0.03	0.06±0.02	0.76±0.02	0.51±0.12
Relational Methods												
DARC [CVPR'26]	127.61	29.76	0.06	0.52±0.01	0.62±0.02	0.57	0.90±0.01	0.18±0.02	0.19±0.03	0.08±0.02	0.81±0.02	0.58±0.15
NCG [ICLR'25]	3.63	11.18	0.01	0.48±0.01	0.16±0.03	0.32	0.91±0.04	0.01±0.01	0.15±0.03	0.00±0.01	0.84±0.02	0.57±0.04
CMoB [NeurIPS'25]	102.15	52.84	0.02	0.23±0.02	0.27±0.04	0.25	0.91±0.00	0.20±0.03	0.23±0.02	0.06±0.03	0.84±0.02	0.42±0.15
Causal CGM [ICLR'25]	51.58	209.47	0.01	0.13±0.04	0.20±0.02	0.17	0.93±0.00	0.11±0.04	0.16±0.03	0.02±0.03	0.77±0.03	0.37±0.18
Ours	26.18	28.53	0.06	0.53±0.00	0.59±0.03	0.56	0.92±0.00	0.22±0.00	0.27±0.00	0.06±0.00	0.87±0.01	0.27±0.15

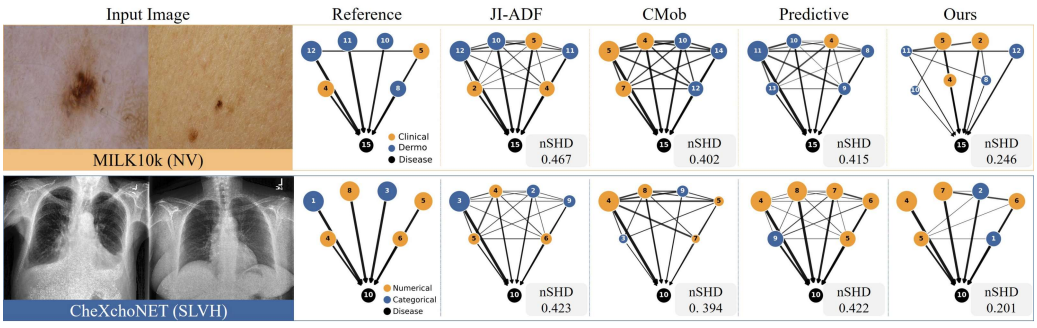


Figure 6: Comparison of expert-informed reference and model-derived BN structures on MILK10k and CheXchoNET.

PRL methods, several CRL methods achieve lower nSHD at the cost of predictive performance, illustrating a trade-off between predictive performance and relational consistency. In contrast, BRP achieves the highest Macro-F1 while attaining the second-lowest nSHD on MILK10k. Table 3 shows a similar pattern on CheXchoNET, where BRP maintains competitive predictive performance while achieving the lowest nSHD. These results indicate that BRP alleviates this trade-off, achieving strong predictive performance with dependency structures that are more consistent with the expert-informed reference structure.

Qualitative Analysis. We further compare relational structures by visualizing BNs derived from the trained models. Fig. 6 compares the expert-informed reference BN, model-derived BNs from representative high-performing methods, and the BN obtained from the predictive branch before

Table 4: Effect of the disease prior and Bayesian guidance. Light gray indicates the best result.

Case	$\mathbf{M}_{cf}$	γ	Feature-wise Gate	MILK10k Macro-F1	CheXchoNET EMR
1	–	–	–	0.507±0.023	0.857±0.003
2	✓	–	–	0.521±0.017	0.868±0.001
3	✓	✓	–	0.524±0.024	0.872±0.005
4	✓	✓	✓	0.549±0.001	0.877±0.015

Table 5: Loss weights analysis. Light gray indicates the best result.

λ_{struct}	λ_{dir}	MILK10k Macro-F1	CheXchoNET EMR
0.0	0.0	0.526±0.018	0.852±0.018
0.1	0.1	0.532±0.031	0.868±0.016
0.3	0.1	0.531±0.023	0.866±0.104
0.3	0.2	0.535±0.018	0.865±0.018
0.5	0.2	0.534±0.021	0.845±0.103

relation propagation on MILK10k and CheXchoNET. For visualization, only the Top-6 nodes are shown. To examine the effect of BRP, we further compare the predictive branch with the relation-propagated branch by constructing BNs from the original representations $\mathbf{s}_{i,j}^{m}$ and the propagated representations $\widetilde{\mathbf{s}}_{i,j}^{m}$.

The BNs derived from existing PRL methods contain numerous and highly interconnected relations, making it difficult to interpret the feature dependencies associated with their predictions. Although CMoB is a CRL method, its intervention-based design does not explicitly constrain feature-feature dependencies, and its derived BN therefore remains highly interconnected. Before relation propagation, the predictive BN similarly exhibits dense feature dependencies and highly connected relations. In contrast, the BN derived after relation propagation exhibits a sparser dependency structure with stronger agreement with the expert-informed reference structure, making the resulting structural consistency more explicit. We provide visualizations of the full BNs for all models in Sup. H.4.

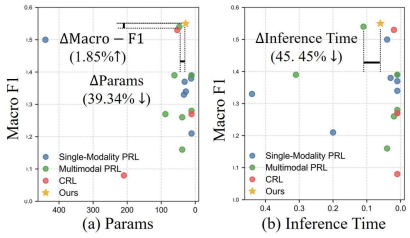


Figure 7: Predictive performance versus computational cost.

Computational Efficiency. Fig. 7 compares the predictive performance of BRP with model size and inference time. Compared with JI-ADF, BRP improves Macro-F1 by 1.85% while reducing the number of parameters and inference time by 39.34% and 45.45%, respectively. These results show that the additional relation propagation pathway remains computationally efficient while improving predictive performance.

5.2 Ablation Study

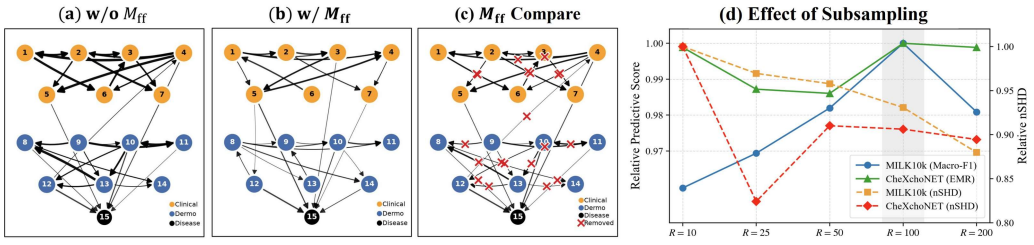


Figure 8: (a–c) Changes in the relation structure by $\mathbf{M}_{ff}$, where red crosses indicate removed relations. (d) Effect of the number of subsampling iterations.

Effect of Stable Relation Estimation. To analyze stable relation estimation, we compare settings with and without $\mathbf{M}_{ff}$ while excluding the disease prior $\mathbf{M}_{cf}$ and Bayesian guidance. Fig. 8(a–c) shows that applying $\mathbf{M}_{ff}$ removes incidental and unstable relations, reducing nSHD by 10.2%, with only a slight change in Macro-F1. Fig. 8(d) further evaluates the effect of repeated subsampling by varying R on both datasets. As R increases, relation estimation becomes more stable, while predictive performance improves up to $R = 100$. We therefore use $R = 100$.

Effect of Prior and Guidance. To evaluate the effect of the disease prior $\mathbf{M}_{cf}$, the sample-level guidance strength γ_i, and the feature-wise gate, we conduct the ablation study without $\mathbf{M}_{ff}$. As shown in Table 4, progressively introducing these components improves predictive performance, with the full configuration achieving a Macro-F1 of 0.549 on MILK10k and an EMR of 0.877 on CheXchoNET. In particular, the gain from the feature-wise gate demonstrates the benefit of adaptively controlling propagation strength according to feature relevance and guidance. These results are consistent with the theoretical analysis, highlighting the importance of controlling the propagation strength.

Effect of Loss Weights. We analyze the effect of the relational losses, fixing $\lambda_p = \lambda_r = 1.0$ and varying λ_{struct} and λ_{dir}. As shown in Table 5, incorporating relational losses improves performance over the setting without them, with the best results obtained at $\lambda_{\text{struct}} = 0.3$ and $\lambda_{\text{dir}} = 0.2$ on MILK10k and at $\lambda_{\text{struct}} = 0.1$ and $\lambda_{\text{dir}} = 0.1$ on CheXchoNET. These results demonstrate the complementary roles of prior consistency and directional alignment in guiding relation propagation.

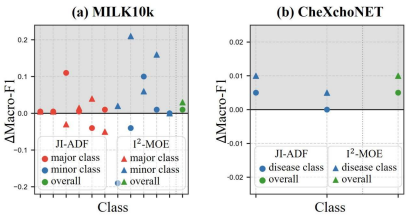


Figure 9: Model-agnostic analysis.

5.3 Additional Experiments

Model-Agnostic Generality. To evaluate generality, we apply BRP using JI-ADF and I^{2}MoE as the predictive models. As shown in Fig. 9, both models show improved predictive performance, particularly for minority classes, indicating that relation propagation is not tied to a specific architecture.

Case Study. Fig. 10 analyzes four prediction transitions between the predictive branch and the relation-propagated branch: C→C (75.23%), W→C (5.12%), C→W (1.37%), and W→W (18.28%). Here, C and W denote correct and wrong predictions, respectively, and the values in parentheses indicate the proportion of samples in each transition. We define W→C as the success case and C→W as the failure case. In the success case, high predictive uncertainty and BN reliability result in strong sample-level guidance γ_i and a relatively large propagation strength α through the feature-wise gate. Together with a well-aligned relational direction, this is associated with the largest increase in true-class probability and correction of the initial prediction. In contrast, in the failure case, the true-class probability decreases despite a positive direction cosine, indicating that relational alignment alone does not guarantee a direction favorable to prediction. Overall, the case study suggests that successful relation propagation depends on both appropriately controlled propagation strength and a relational direction that supports predictive improvement. These empirical observations are consistent with the sufficient conditions characterized in Proposition 1. Additional results are provided in Sup. H–I.

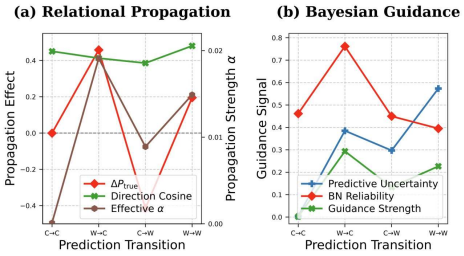


Figure 10: Relation Propagation Effect.

6 Limitations and Conclusion

We have proposed a Bayesian-guided relation propagation method that preserves the original predictive pathway while incorporating stable relational structure through adaptive propagation. By combining stable relation estimation, feature relevance retrieved from the disease prior, and Bayesian guidance, BRP improves predictive performance while increasing agreement with expert-informed relational structures, providing a more structurally interpretable view of the feature dependencies represented by the model. Theoretical and empirical analyses further indicate that effective relation propagation depends jointly on an appropriate relational direction and propagation strength.

Although BRP improves both predictive performance and structural interpretability, several directions remain for future research. The disease prior $\mathbf{M}_{cf}$ is estimated from training data and may therefore reflect dataset-specific characteristics. Consequently, the retrieved feature relevance may be influenced by associations specific to the observed data rather than generalizable disease-feature relationships. Future work can address this limitation by incorporating more interpretable and clinically grounded relation estimation. Combining such relational knowledge with BRP can further improve reliability and interpretability.

AI USE STATEMENT

Generative AI tools were used to assist with literature organization, reference formatting, language editing, and the preparation of scientific visualizations. All AI-assisted content was reviewed and validated by the authors. In addition, GPT-5.6 Sol was used as part of the experimental methodology for two specific purposes. First, it was used in the motivation experiment to identify supported feature dependencies and disease–feature relations under predefined domain knowledge and prompting constraints. Second, it was used to generate expert-informed reference annotations based on fixed ground-truth disease labels, observed multimodal evidence, and expert-provided domain knowledge for constructing the reference BN used in the structural consistency evaluation. The prompting procedures, generation constraints, generated examples, and corresponding validation procedures are described in the supplementary material.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the model architecture, relation construction, training procedure, hyperparameters, dataset splits, and evaluation protocols in the supplementary material.

REFERENCES

aGrUM/pyAgrum Developers. pyAgrum: A python library for probabilistic graphical models. `https://pyagrum.gitlab.io/`. Accessed: 2026-09-21.

Pietro Barbiero, Francesco Giannini, Gabriele Ciravegna, Michelangelo Diligenti, and Giuseppe Marra. Relational concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:77663–77685, 2024.

Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and NV Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*, pp. 147–160, 2021.

Concha Bielza and Pedro Larranaga. Discrete bayesian network classifiers: A survey. *ACM Computing Surveys (CSUR)*, 47(1):1–43, 2014.

Behzad Bozorgtabar, Dwarikanath Mahapatra, Sudipta Roy, Muzammal Naseer, Imran Razzak, and Zongyuan Ge. Lata: Laplacian-assisted transductive adaptation for conformal uncertainty in medical vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36311–36320, 2026.

Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Causal concept graph models: Beyond causal opacity in deep learning. In *International Conference on Learning Representations*, volume 2025, pp. 55102–55131, 2025a.

Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Counterfactual concept bottleneck models. In *International Conference on Learning Representations*, volume 2025, pp. 5184–5207, 2025b.

Siyi Du, Xinzhe Luo, Declan O'regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, volume 2026, pp. 116887–116913, 2026.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Tarek Eissa, Cristina Leonardo, Kosmas V Kepesidis, Frank Fleischmann, Birgit Linkohr, Daniel Meyer, Viola Zoka, Marinus Huber, Liudmila Voronina, Lothar Richter, et al. Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. *Cell Reports Medicine*, 5(7), 2024.

Pierre Elias and Shreyas Bhave. Chexchonet: A chest radiograph dataset with gold standard echocardiography labels. *PhysioNet*, 2024.

Gianni Franchi, Olivier Laurent, Maxence Leguéry, Andrei Bursuc, Andrea Pilzer, and Angela Yao. Make me a bnn: A simple strategy for estimating bayesian uncertainty from pre-trained models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12194–12204. IEEE, 2024.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *2025 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 25261–25270. IEEE, 2025.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Hierarchical concept-based interpretable models. In *International Conference on Learning Representations*, volume 2026, pp. 11581–11609, 2026.

Lijie Hu, Chenyang Ren, Zhengyu Hu, Hongbin Lin, Cheng-Long Wang, Zhen Tan, Weimin Lyu, Jingfeng Zhang, Hui Xiong, and Di Wang. Editable concept bottleneck models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24678–24726. PMLR, 2025.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11375–11385. IEEE, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, pp. 5574–5584, 2017.

Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Tarald O. Kvålseth. Entropy and correlation: Some comments. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(3):517–519, 1987. doi: 10.1109/TSMC.1987.4309069.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Equille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *International Conference on Learning Representations*, volume 2024, pp. 23756–23779, 2024.

Changbin Li, Kangshuo Li, Yuzhe Ou, Lance Kaplan, Audun Jøsang, Jin-Hee Cho, Dong Hyun Jeong, and Feng Chen. Hyper evidential deep learning to quantify composite classification uncertainty. In *International Conference on Learning Representations*, volume 2024, pp. 8826–8860, 2024.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

Xinyan Liang, Shijie Wang, Yuhua Qian, Qian Guo, Liang Du, Bingbing Jiang, Tingjin Luo, and Feijiang Li. Trusted multi-view classification with expert knowledge constraints. In *Forty-second International Conference on Machine Learning*, 2025.

Zhifang Liao, Junhao Li, HaoKang Ding, and Yucheng Song. Darc: Dual adjustment reasoning with counterfactuals for trustworthy chest x-ray classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28234–28243, 2026.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Phan Nguyen, Dat Cao, Hien Kha, Hien Chu, Minh Le, Trang Pham, and Nguyen Quoc Khanh Le. Ji-adf: Joint-individual learning with adaptive decision fusion for multimodal skin lesion classification. *arXiv preprint arXiv:2604.27343*, 2026.

Juri Opitz. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836, 2024. doi: 10.1162/tacl_a_00675.

Tschandl Philipp, Akay Nisa Bengü, Rosendahl Cliff, Rotemberg Veronica, Todorovska Verche, Weber Jochen, Wolber Anna Katharina, Müller Christoph, Kurtansky Nicholas, Halpern Allan, et al. Milk10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators. *Journal of Investigative Dermatology*, 2025.

Samyak Rawlekar, Shubhang Bhatnagar, Vishnuvardhan Pogunulu Srinivasulu, and Narendra Ahuja. Improving multi-label recognition using class co-occurrence probabilities. In *International Conference on Pattern Recognition*, pp. 424–439. Springer, 2024.

Raanan Y Rohekar, Yaniv Gurwicz, Shami Nisimov, Guy Koren, and Gal Novik. Bayesian structure learning by recursive bootstrap. *Advances in Neural Information Processing Systems*, 31, 2018.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 72145–72169, 2025.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Yuesong Shen, Nico Daheim, Bai Cong, Peter Nickl, Gian Maria Marconi, Bazan Clement Emile Marcel Raoul, Rio Yokota, Iryna Gurevych, Daniel Cremers, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational learning is effective for large deep networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44665–44686. PMLR, 2024.

Dai Shi. Transnext: Robust foveal visual perception for vision transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17773–17783. IEEE, 2024.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Shailesh Sridhar, Anton Xue, and Eric Wong. Missingness bias calibration in feature attribution explanations. In *International Conference on Learning Representations*, volume 2026, pp. 53299–53316, 2026.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1073–1080, 2009. doi: 10.1145/1553374.1553511.

Jiawei Wang, Shaofei Lu, Da Cao, Dongyu Wang, Yuquan Le, Zhe Quan, and Tat-Seng Chua. Neural causal graph for interpretable and intervenable classification. In *International Conference on Learning Representations*, volume 2025, pp. 97398–97418, 2025.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. Cmob: Modality valuation via causal effect for balanced multimodal learning. *Advances in Neural Information Processing Systems*, 38:82585–82607, 2026.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27328–27337. IEEE, 2024.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *Proceedings of machine learning research*, 267:68870, 2025.

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *International Conference on Learning Representations*, volume 2024, pp. 6944–6966, 2024.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, volume 2024, pp. 34817–34848, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, volume 2025, pp. 53847–53890, 2025.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *International Conference on Learning Representations*, 2026.

Zhiyu Zhu, Zhibo Jin, Jiayu Zhang, Bartlomiej Sobieski, Przemyslaw Biecek, Fang Chen, and Jianlong Zhou. Faithfulness under the distribution: A new look at attribution evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 38289–38312, 2026.