Bridging Expertise and Data: Multi-Label Disease Detection via Causal Learning and Decision Fusion

Extended Abstract

Xin Zhang
Northwestern Polytechnical University
Xi'an, China
zhangxin1@mail.nwpu.edu.cn

Minhui Zhang
Northwestern Polytechnical University
Xi'an, China
zmh17@mail.nwpu.edu.cn

Jiaqi Liu*
Northwestern Polytechnical University
Xi'an, China
jqliu@nwpu.edu.cn

Zhiwen Yu
Harbin Engineering University & Northwestern Polytechnical University
Harbin, China
zhiwenyu@nwpu.edu.cn

Bin Guo
Northwestern Polytechnical University
Xi'an, China
guob@nwpu.edu.cn

ABSTRACT

Recent multi-label disease detection methods exploit disease causality and disease–image feature interactions, but causal learning is often inaccurate and computationally costly. Meanwhile, human–AI collaboration in diagnosis can outperform either clinicians or models alone. We propose a framework that combines expert-guided causal learning with Bayesian human–AI decision fusion. First, we learn an expert causal matrix via a GCN from expert labels and authoritative medical knowledge, and use it to regularize inter-disease causal learning. Second, we convert per-label probabilities into joint label-set probabilities and fuse them with expert decisions using a Bayesian scheme. Experiments on three medical datasets show that our method outperforms state-of-the-art multi-label disease detection models by up to 13.18%.

KEYWORDS

Multi-Label Disease Detection; Human-AI Collaboration; Causal Learning; Decision Fusion

ACM Reference Format:
Xin Zhang, Minhui Zhang, Jiaqi Liu, Zhiwen Yu, and Bin Guo. 2026. Bridging Expertise and Data: Multi-Label Disease Detection via Causal Learning and Decision Fusion: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/FPSF4954

1 INTRODUCTION

In recent years, Artificial Intelligence has achieved strong performance in speech recognition, image processing, and natural language processing, enabling broad deployment in healthcare and other domains. In healthcare, AI has been used to diagnose many diseases, including lung diseases [13], heart diseases [11], Alzheimer's diseases [16], diabetes [4], and retinal diseases [3]. However, medical scenarios remain challenging due to domain shift, bias, and limited interpretability. Human experts can be more robust in ambiguous or atypical cases, motivating human–AI collaboration that can outperform either alone [2].

Most prior work targets single-disease diagnosis [5, 8, 19], while real patients often have multiple coexisting conditions, making diagnosis a multi-label classification problem. This task is difficult due to complex inter-disease relations (co-occurrence, mutual exclusivity, and causality) and overlapping symptoms. Existing approaches exploit label relations and disease–image feature interactions: graph-based methods model inter-label structure [1, 18], and transformer-based methods learn label-specific representations via label–feature interactions [9, 15]. Recent work further combines inter-disease causality with feature interactions [17], but data-driven causal discovery can be brittle under sparse data or many labels.

We propose a multi-label disease detection method that combines the expert-guided causal learning and Bayesian human-AI decision fusion, thereby bridging human expertise and data. First, we compute the co-occurrence matrix as well as the mutual information matrix from human experts' predictions. We collect relevant authoritative medical knowledge to obtain the causal matrix based on medical knowledge. Then, we use GCN to estimate human causality between diseases from the three matrices described above, and use this matrix to guide the AI in the causal learning and classification. Finally, we transfer the probabilities of AI's results to the probabilities of different disease combinations, construct the human confusion matrix, and fuse human and AI classification results.

2 METHOD

We build upon a causal multi-label classification pipeline and introduce two human-integrated modules: Expert-Guided Causal Matrix Construction (ECMC) and Bayesian Human–AI Decision Fusion (BHDF).

*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/FPSF4954

Table 1: Overall Performance and Ablation Study (%). Bolded text indicates the best performance among all models.

Model	FFA					Chest-9					Chest-4				
	MAF1 ↑	mAP ↑	mAUC ↑	SA ↑	HL ↓	MAF1 ↑	mAP ↑	mAUC ↑	SA ↑	HL ↓	MAF1 ↑	mAP ↑	mAUC ↑	SA ↑	HL ↓
Resnet50	81.83	75.41	67.10	30	19.88	34.49	30.04	63.93	2	40.35	44.99	39.18	64.87	21	31.55
GCN	79.37	86.35	74.01	30	22.15	39.49	37.79	70.38	32	23.13	40.18	40.13	71.24	59	15.36
dyGCN	83.98	88.03	75.83	35	17.59	39.38	43.63	69.01	21	18.94	44.17	40.23	72.56	62	11.85
Q2Lcausal	84.44	87.60	76.43	34	17.32	45.30	43.52	72.21	3	25.59	49.45	51.81	70.63	67	10.25
MLDcausal	**84.73**	87.88	76.61	33	17.36	45.13	43.69	70.51	30	17.03	49.23	52.30	71.51	67	9.72
Q2Lcausal+CH	84.52	87.36	77.31	36	17.05	48.24	46.10	**75.27**	32	16.85	**59.86**	**62.92**	**78.05**	69	9.03
Q2Lcausal+FH	-	-	-	76	8.79	-	-	-	71	6.10	-	-	-	83	4.61
Q2Lcausal+CH+FH	-	-	-	80	7.77	-	-	-	**73**	**5.46**	-	-	-	**90**	**2.74**
MLDcausal+CH	84.15	**87.92**	**77.74**	35	17.12	**48.45**	**51.80**	74.92	32	16.95	58.11	59.70	76.38	67	10.59
MLDcausal+FH	-	-	-	81	7.71	-	-	-	66	6.83	-	-	-	82	4.69
MLDcausal+CH+FH	-	-	-	**82**	**7.31**	-	-	-	71	5.56	-	-	-	84	4.23

2.1 Base Pipeline

Given an image X, a backbone extracts features F_0, and a transformer decoder uses learnable disease queries Q_0 to obtain label-specific features Q_1 via disease–image interactions [9, 15, 17]. We model inter-disease causality with an adjacency matrix $W \in \mathbb{R}^{K\times K}$ (zero diagonal) and propagate $Q = WQ_1$ for causality-aware prediction. The base loss combines multi-label classification loss (e.g., asymmetric loss) and DAG regularization for causal learning [14].

2.2 ECMC: Expert-Guided Causal Matrix Construction

Purely data-driven causal discovery can be unstable and slow when K is large. We therefore estimate an expert causal prior W^h from (i) expert annotations and (ii) authoritative medical knowledge, and use it to guide learning of W. Concretely, we compute three relation matrices: co-occurrence M_{cooc}, mutual information M_{mi}, and a directional authority matrix M_{auth} extracted from medical sources. We construct a multi-relation directed graph (ESR-Graph) over diseases, where edges are added from M_{auth} (all non-zero), M_{cooc} (above a threshold), and M_{mi} (above a threshold). A multi-relation GCN encodes each relation, fuses the embeddings, and decodes pairwise causal strengths to obtain W^h. We integrate this expert prior by initializing W with W^h and adding a regularizer:

$$\mathcal{L}_w = \tau \|W - W^h\|_F^2, \tag{1}$$

which prevents the learned causality from drifting away from expert knowledge.

2.3 BHDF: Bayesian Human-AI Decision Fusion

To improve final diagnosis accuracy, we fuse expert decisions with the model output at the *label-set* level. Let $S = 0, 1^K$ denote all label combinations, and $s \in S$. Given per-label probabilities $p_a(x)$, we construct a distribution over S (e.g., under conditional independence). For example, when $K = 3$, $|S| = 2^3 = 8$, corresponding to no disease, any single disease, any pair of diseases, and all three diseases. Let the expert output be $\tilde{s} \in S$. We estimate a joint confusion matrix of size $2^K \cdot 2^K$, where each column (true s) parameterizes $\tilde{s} \mid s \sim \text{Discrete}(\boldsymbol{\phi}_{*s})$. To avoid high-variance MLE with limited expert labels, we use Dirichlet smoothing [7] by placing $\boldsymbol{\phi}_{*s} \sim \text{Dirichlet}(\boldsymbol{\alpha}_s)$ with $(\alpha_s)_s = \gamma$ and $(\alpha_s)_{s'} = \beta$ $(s' \neq s)$. Given human and AI posteriors $p_h(x)$ and $p_a(x)$ over S, we fuse them as

$$p(y = s \mid p_h(x), p_a(x)) = \frac{p_h(x)_s \, p_a(x)_s}{\sum_{s' \in S} p_h(x)_{s'} \, p_a(x)_{s'}}, \tag{2}$$

3 EXPERIMENTS

3.1 Settings

Datasets. We evaluate on LID-FFA [17] and two Chest X-ray datasets (Chest-9 [12] and Chest-4 [10]). **Baselines.** We compare against ResNet50 [6], GCN [1], dyGCN [18], Q2L [9], MLDecoder [15], and the causal variants Q2Lcausal/MLDcausal [17]. **Our method.** We denote expert-guided causal learning as **CH** (applied to Q2Lcausal and MLDcausal) and Bayesian human–AI decision fusion as **FH** (applied to all models). We report standard multi-label metrics (MAF1, mAP, mAUC) and set-level metrics (SA, HL).

3.2 Results

Results on the FFA and Chest datasets (Table 1) show consistent improvements. Adding human causal relations (CH) strengthens Q2Lcausal and MLDcausal on most label-wise metrics, indicating better discriminative ability and ranking quality. After applying FH, we observe a further and more pronounced gain at the set level: SA increases while HL decreases across models. Overall, these results suggest that human knowledge helps the model make more reliable predictions, and human-AI fusion further improves diagnosis accuracy, which is critical for real clinical multi-disease settings.

4 CONCLUSION

We integrate expert-guided causal learning with Bayesian human–AI decision fusion for multi-label diagnosis. This improves both accuracy and training efficiency over strong baselines.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (No.62372381).

REFERENCES

[1] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5177–5186.

[2] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593, 7857 (2021), 33–36.

[3] Li Dong, Wanji He, Ruiheng Zhang, Zongyuan Ge, Ya Xing Wang, Jinqiong Zhou, Jie Xu, Lei Shao, Qian Wang, Yanni Yan, et al. 2022. Artificial intelligence for screening of multiple retinal and optic nerve diseases. *JAMA network open* 5, 5 (2022), e229960–e229960.

[4] Wanglong Gou, Chu-wen Ling, Yan He, Zengliang Jiang, Yuanqing Fu, Fengzhe Xu, Zelei Miao, Ting-yu Sun, Jie-sheng Lin, Hui-lian Zhu, et al. 2021. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* 44, 2 (2021), 358–366.

[5] Along He, Tao Li, Ning Li, Kai Wang, and Huazhu Fu. 2020. CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging* 40, 1 (2020), 143–153.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems* 34 (2021), 4421–4434.

[8] Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, Lequan Yu, and Lei Xing. 2020. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging* 39, 12 (2020), 4023–4033.

[9] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834* (2021).

[10] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294, 2 (2020), 421–431.

[11] S Manimurugan, Saad Almutairi, Majed Mohammed Aborokbah, C Narmatha, Subramaniam Ganesan, Naveen Chilamkurti, Riyadh A Alzaheb, and Hani Almoamari. 2022. Two-stage classification model for the prediction of heart disease using IoMT and artificial intelligence. *Sensors* 22, 2 (2022), 476.

[12] Zaid Nabulsi, Andrew Sellergren, Shahar Jamshy, Charles Lau, Edward Santos, Atilla P Kiraly, Wenxing Ye, Jie Yang, Rory Pilgrim, Sahar Kazemzadeh, et al. 2021. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. *Scientific reports* 11, 1 (2021), 15523.

[13] Vidhi Pitroda, Mostafa M. Fouda, and Zubair Md Fadlullah. 2021. An Explainable AI Model for Interpretable Lung Disease Classification. In *2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*. 98–103. https://doi.org/10.1109/IoTaIS53735.2021.9628573

[14] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 82–91.

[15] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. 2023. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 32–41.

[16] Vimbi Viswan, Noushath Shaffi, Mufti Mahmud, Karthikeyan Subramanian, and Faizal Hajamohideen. 2024. Explainable artificial intelligence in Alzheimer's disease classification: A systematic review. *Cognitive Computation* 16, 1 (2024), 1–44.

[17] Jianyang Xie, Xiuju Chen, Yitian Zhao, Yanda Meng, He Zhao, Anh Nguyen, Xiaoxin Li, and Yalin Zheng. 2024. Multi-disease Detection in Retinal Images Guided by Disease Causal Estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 743–753.

[18] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 649–665.

[19] Yi Zhou, Boyang Wang, Lei Huang, Shanshan Cui, and Ling Shao. 2020. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. *IEEE transactions on medical imaging* 40, 3 (2020), 818–828.