Medical Image Analysis 103 (2025) 103581

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

CausalMixNet: A mixed-attention framework for causal intervention in robust medical image diagnosis

Yajie Zhang [a], Yu-An Huang [b], Yao Hu [a], Rui Liu [a], Jibin Wu [a,c,d], Zhi-An Huang [e,*], Kay Chen Tan [a,d]

[a] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
[b] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[c] Department of Computing, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
[d] Research Center on Data Sciences and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
[e] Department of Computer Science, City University of Hong Kong (Dongguan), Dongguan, China

ARTICLE INFO

Keywords:
Unobserved confounders
Front-door adjustment
Medical image diagnosis
Patch mixing

ABSTRACT

Confounding factors inherent in medical images can significantly impact the causal exploration capabilities of deep learning models, resulting in compromised accuracy and diminished generalization performance. In this paper, we present an innovative methodology named CausalMixNet that employs query-mixed intra-attention and key&value-mixed inter-attention to probe causal relationships between input images and labels. For mitigating unobservable confounding factors, CausalMixNet integrates the non-local reasoning module (NLRM) and the key&value-mixed inter-attention (KVMIA) to conduct a front-door adjustment strategy. Furthermore, CausalMixNet incorporates a patch-masked ranking module (PMRM) and query-mixed intra-attention (QMIA) to enhance mediator learning, thereby facilitating causal intervention. The patch mixing mechanism applied to query/(key&value) features within QMIA and KVMIA specifically targets lesion-related feature enhancement and the inference of average causal effect inference. CausalMixNet consistently outperforms existing methods, achieving superior accuracy and F1-scores across in-domain and out-of-domain scenarios on multiple datasets, with an average improvement of 3% over the closest competitor. Demonstrating robustness against noise, gender bias, and attribute bias, CausalMixNet excels in handling unobservable confounders, maintaining stable performance even in challenging conditions.

1. Introduction

Computer-aided diagnosis systems are designed to automate the feature extraction and pattern recognition processes from medical images, streamlining the crucial analysis tasks such as lesion detection (Chen et al., 2020; Li et al., 2023), segmentation (Zhong et al., 2025; Jiang et al., 2025), and classification (Liu et al., 2024b,a; Huang et al., 2021). They enhance the efficiency of medical practitioners and also reduce subjectivity in disease diagnosis. Typically, deep learning-based computer-aided diagnosis systems have become pivotal in medical image analysis.

However, the pervasive presence of both observable and unobservable noise (Xu et al., 2020; Van der Velden et al., 2022) presents substantial obstacles for deep learning applications in disease diagnosis when they attribute diagnostic relevance to a confluence of observable noise and pathological indicators (Miao et al., 2023). The complex interplay between various tissues, organs, and imaging artifacts poses a significant challenge in precisely demarcating pathological regions (Chen et al., 2022b). Furthermore, performance degradation can occur due to domain shifts (Hu et al., 2023), where the well-learned patterns do not generalize well to new, unseen data. Additionally, these issues are further complicated by unobservable noise, which subtly undermines the learning process of pathological features. Unlike observable noise, unobservable noise cannot be directly detected or quantified, making it challenging for existing deep learning methods in directly countering its effects.

Currently, a myriad of methods exhibit promising potential in addressing observable noise. Wang et al. (2023) employed a multi-task learning paradigm to stimulate the model towards acquiring content-related features, prioritizing them over stylistic features. Chen et al. (2022a) identified invariant features in the feature matrix using gradient information, thereby alleviating the impact of noise. Mao et al. (2022) utilized counterfactual invariance/variance representations to enhance the model's robustness to noise. Nie et al. (2023) employed

* Corresponding author.
E-mail address: huang.za@cityu-dg.edu.cn (Z.-A. Huang).

https://doi.org/10.1016/j.media.2025.103581
Received 16 October 2024; Received in revised form 25 March 2025; Accepted 1 April 2025
Available online 8 May 2025

attention mechanism-based backdoor adjustments to learn the etiology in images. However, these methods generally overlook how to address unobservable noise.

Unobservable noise, as a confounding factor, significantly impacts the interpretation of causal relationships in deep learning models. Unobserved confounders usually refer to the unmeasured variables affecting both the input and the outcome variables (Bareinboim et al., 2015). In medical image classification, these confounders can include factors like gender, age (Pölsterl and Wachinger, 2021), and genetic background (Zardavas et al., 2015). While potentially obtainable through other sources, these factors cannot be directly discernible from the image data itself, thus acting as unobserved confounders in this work. Furthermore, variations in operational techniques can introduce biases in dataset annotations (Fabbrizzi et al., 2022), leading to misleading inference results. These biases may result in improved model performance in specific populations or conditions, thereby compromising generalizability and reliability in clinical applications. Such instances highlight the pervasive influence of unobservable confounding factors in disease diagnostic systems. Therefore, it is essential to develop algorithms that can effectively mitigate the impact of these unobservable confounders.

In this paper, we propose the CausalMixNet, a novel medical image diagnosis approach, integrated with query-mixed intra-attention and key&value-mixed inter-attention by eliminating the negative effects of unobserved confounders and exploring the true causality between input and output. CausalMixNet is underpinned by the concept of front-door adjustment (FDA) (Pearl, 1995) that introduces a mediator to bridge the gap between input and output, thereby controlling unobservable confounding factors. We first design a non-local reasoning module (NLRM) to extract the mediator in FDA to link the causal effect from image features to labels. Then, a patch-masked ranking module (PMRM) is proposed to identify the lesion patches and non-lesion patches by masking each sub-patch and calculating their individual importance score. To enhance the discernibility of the mediator, we propose the query-mixed intra-attention (QMIA) module to increase the diversity of query vectors. Finally, a key&value-mixed inter-attention (KVMIA) module is designed to conduct causal intervention on the refined mediator. Extensive experiments spanning four widely-used datasets for both in-domain and out-of-domain testing demonstrate that CausalMixNet achieves promising improvements in both scenarios. CausalMixNet also exhibits remarkable robustness in noise testing even in the presence of severe data noise, gender bias analysis, and attribute bias analysis. The main contributions of this work can be summarized as follows.

- The proposed deconfounding method CausalMixNet is designed to explore the causality for medical image diagnosis, especially when unobserved confounders exist in the system. This method aims to uncover the true causal relationships between image features and diagnosis labels.
- The seamless integration of four distinct modules NLRM, PMRM, QMIA, and KVMIA is demonstrated to effectively suppress the impact of unobservable confounding factors. The NLRM and KVMIA are the core components of the causal inference model following the strategy of FDA, while the PMRM and QMIA modules further enhance the causal inference process.
- Comprehensive experimental assessments demonstrate that CausalMixNet achieves accurate disease diagnosis not only within the domain it was trained on but also maintains stability and effectiveness when applied to out-of-domain data, gender bias analysis, attribute bias analysis, and in noisy testing conditions.

2. Related works

In this section, we systematically outline the applications of supervised deep learning in medical image diagnosis. Subsequently, we conduct a concise discussion of existing methods for image debiasing. Lastly, we provide a comprehensive summary of causal inference methods employed in the realm of medical imaging.

2.1. Supervised medical image diagnosis

Supervised medical image diagnosis has garnered significant attention in recent years, with numerous studies contributing to its advancement. Researchers have delved into leveraging annotated medical image datasets to train and fine-tune various deep learning architectures, empowering them to recognize patterns and features indicative of specific disease labels. Tajbakhsh et al. (2016) explored the effectiveness of fine-tuning pre-trained deep convolutional neural networks for various medical image analysis tasks. Previous work has also attempted to explore the adaptability of different deep model structures to medical images. Gao et al. (2020) investigated a temporal emphasis recurrent neural network to model temporal information for the lung cancer classification task. Wang et al. (2022) and He et al. (2021) introduced the attention mechanism guided by the radiologists' gaze tracks and category labels. Liu et al. (2020) modeled the class correlation by a graph residual network, where the class-dependency prior improves the effectiveness of the graph convolutional network. Liu et al. (2024c) introduced a spatial–temporal co-attention learning graph neural network for computer-aided diagnosis of mental disorders. He et al. (2022) captured long-range contextual information through the spatial pyramid transformer for medical image diagnosis. While these exemplary approaches have indeed yielded substantial advantages in the realm of medical image diagnosis, their efficacy in mitigating bias within medical images remains less than ideal. Bias in medical data increases the risk of model overfitting and undermines the accuracy of various deep learning architectures. Therefore, medical image debiasing has gained ongoing attention in recent years.

2.2. Image debiasing

Image bias generally refers to the systematic difference of a statistical estimation from its population value, stemming from subject selection, acquisition method, processing biases, etc. (Wachinger et al., 2019). Overfitting to data bias in deep neural networks can lead to decreased performance and reduced generalization ability (Ktena et al., 2024). In the early stages, many methods address biases in multi-source domain generalization. For instance, Arjovsky et al. (2019) aimed to develop a model that is unaffected by specific domains by maximizing the consistency of predictions across all source domains, thereby improving its generalization ability in new domains. Makar et al. (2022) enhanced model performance in multi-source domain generalization by incorporating auxiliary labels and conditional independence. However, the reliance on multi-source data in these methods limits their applicability in single-source scenarios. To address the bias issues in single-source domain generalization, a set of methods leverages techniques such as adversarial training (Gokhale et al., 2023), data augmentation (Che et al., 2023; Zhang et al., 2018; Zhou et al., 2020) and contrastive learning (Duboudin et al., 2021) to learn invariant features. Rame et al. (2022) established domain invariance in the gradient space of the loss function, enabling out-of-domain generalization. Additionally, Chen et al. (2023) and Nguyen et al. (2023) also aimed to mitigate the effects of data bias through causal inference techniques. However, these methods often overlook the influence of confounding factors in medical imaging, which can result in a failure to accurately capture causal relationships.

2.3. Causality for medical imaging

Causality (Pearl, 1995) in medical imaging represents a novel paradigm aimed at unraveling the intricate relationships between imaging features and underlying physiological phenomena. Based on the nature of confounding factors, current methods can be categorized into two types: methods addressing observable confounders and methods targeting unobservable confounders. To tackle the side-effect of observable confounding factors, backdoor adjustment (Qu et al., 2024) and

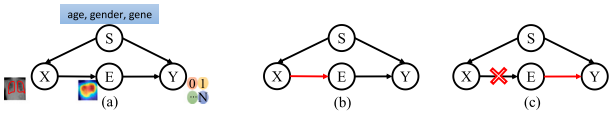

Fig. 1. (a) The structural causal model for medical image diagnosis: X is the input image, E denotes the mediator, Y is the predicted label, and S represents unobserved confounders. The front-door path is $X \rightarrow E \rightarrow Y$, and the backdoor path is $X \leftarrow S \rightarrow Y$. (b) The structural causal model of $X \rightarrow E$. (c) The structural causal model of $E \rightarrow Y$, where the backdoor path is $E \leftarrow X \leftarrow S \rightarrow Y$.

counterfactual reasoning (Singla et al., 2023) have gained significant attention in recent years. Chen et al. (2022b), Miao et al. (2023), and Qu et al. (2024) advocated for adopting backdoor-adjustment strategies to effectively decouple the misleading associations between organ features and background information. Thiagarajan et al. (2022) employed an interval calibration strategy based on uncertainty to effectively generate counterfactuals, ensuring desired alterations in model predictions. On the other hand, a set of methods specifically targets the unobservable confounding factors by leveraging auxiliary labels from biased sources. For example, Bissoto et al. (2020) addressed confounding in skin disease recognition by explicitly identifying and classifying image artifacts. While artifacts can be annotated, they are often not directly indicated by standard image category labels, thus acting as unobserved confounders unless specifically addressed. Similarly, Deng et al. (2023) aimed to decouple the bias features (e.g. race and gender) from image features through orthogonalization constraints. Even though race and gender might be obtainable from other sources, they are not directly derived from the image features themselves and can introduce confounding if not explicitly accounted for. However, the reliance of bias labels limits their applicability in scenarios with unknown or unlabeled bias sources. To address this limitation, some attempts have been made to identify and mitigate bias without relying on explicit bias labels. Luo et al. (2022) assumed that easily learned features are biased features and uses this assumption to perform bias pseudo-labeling, thereby debiasing for the model. Zhang et al. (2024) applied front-door adjustment to address unobserved confounders (such as gender) in the medical image classification task. However, it does not take into account the impact of unobserved confounders on the extraction of mediating factors. In contrast to these methods, this work adopts a causal inference approach by utilizing front-door adjustment, which introduces a mediator between the input and output. During the extraction of mediating factors, the QMIA module employs an intra-class sample exchange mechanism to mitigate the influence of unobserved confounding factors on the mediators. This mediator enables our method to perform causal interventions, effectively circumventing the need to directly address unobservable confounding factors.

3. Preliminaries of front-door adjustment

Front-door adjustment involves the identification and incorporation of an intermediate variable positioned on the causal pathway between a treatment and an outcome. The objective is to isolate the direct effect of the treatment by adjusting for the mediator variable. In this section, we introduce the fundamentals of front-door adjustment to learn the true causal effect from input images to predicted labels, as shown in Fig. 1.(a). Here, X refers to input images of a deep learning model encompassing both the lesion feature and additional irrelevant features introduced by unobserved confounders. Term S represents the unobserved confounders (age, gender, genetic background, etc.), and term Y corresponds to the ground-truth label of X. Deconfounding S necessitates causal intervention on X (a.k.a. $do(X)$) by calculating the average causal effects across various confounders. However, the unobserved confounders are not directly estimable, the implementation of causal intervention on X is impractical. To tackle this dilemma, we introduce the mediator E between X and Y, facilitating the knowledge transfer from X to Y. The causal effect of $P(Y|do(X))$ can be transferred to $\sum P(Y|do(E))P(E|do(X))$ by considering the joint effect of the causal relationships $X \rightarrow E$ and $E \rightarrow Y$. The causal relationship of $X \rightarrow E$ is shown in Fig. 1.(b). There is only one causal path between X and E, as the path $X \leftarrow S \rightarrow Y \leftarrow E$ is blocked by the collider $S \rightarrow Y \leftarrow E$. Therefore, the causal effect $P(E|do(X))$ can be directly expressed as $P(E|X)$.

The causal relationship of $E \rightarrow Y$ is shown in Fig. 1.(c). It can be observed that there are two paths connecting E and Y: the front-door path $E \rightarrow Y$ and the backdoor path $E \leftarrow X \leftarrow S \rightarrow Y$. To block the backdoor path, we should perform causal intervention on E by calculating the average causal effects across different X stratums or S stratums. Due to the unmeasurable nature of S, we can express $P(Y|do(E))$ as follows:

$$P(Y|do(E)) = \sum_{x} P(Y|E=e,x)P(x), \tag{1}$$

where x denotes one stratum of variable X.

Through a step-by-step causal reasoning approach (Pearl, 1995), the causality from X to Y unfolds as follows:

$$\begin{aligned} P(Y|do(X)) &= \sum_{e} P(Y|do(E))P(E=e|do(X)) \\ &= \sum_{e} P(E=e|x) \sum_{x'} P(Y|E=e,x')P(x'), \end{aligned} \tag{2}$$

where x' is an index of summation at $P(Y|do(E))$.

4. Method

The proposed framework illustrated in Fig. 2 encompasses crucial components: NLRM, PMRM, QMIA, and KVMIA. Acting as the foundational model, the NLRM plays a pivotal role in extracting mediators as $P(E|do(X))$ in Eq. (1) and conducting predictions for medical images. Mediators refer to invariant features in images, such as texture characteristics, that can effectively reflect lesions. The NLRM employs a self-attention-based convolutional neural network to extract these features from images. PMRM is a mechanism that ranks the importance of image sub-blocks by assessing the distance between the sub-blocks and global features. It aids in the identification of lesion and non-lesion areas within the image for QMIA. Front-door adjustment assumes no unmeasured confounders between the input X and the mediator (Yang et al., 2023). To mitigate this impact, QMIA employs an intra-class sample exchange mechanism for lesion sub-blocks, enabling the NLRM module to focus on learning robust, lesion-specific features associated with the disease process. This reduces the interference of unmeasured confounders and ensures the mediator captures diagnostically relevant features rather than spurious correlations. Concurrently, KVMIA facilitates causal intervention on the mediator, which corresponds to the process of $P(Y|do(E))$ as Eq. (2). KVMIA implements causal intervention on mediators through an exchange mechanism of non-lesion sub-blocks between samples. The whole process is designed to implement the front-door adjustment, effectively diminishing the effects of unobserved confounders, thus improving generalization ability.

4.1. Non-local reasoning module

NLRM serves as the mediator extractor by abstracting disease features in images into highly correlated semantic features that are less affected by background information. Non-local reasoning module aims to capture long-range dependencies and relationships across the input space to fundamentally enhance CNN attention computation, as non-local reasoning in neural networks.

Given an input image x, the mediator of feature map $\mathbf{E}$ is computed using a convolutional neural network. Specifically, we utilize the first

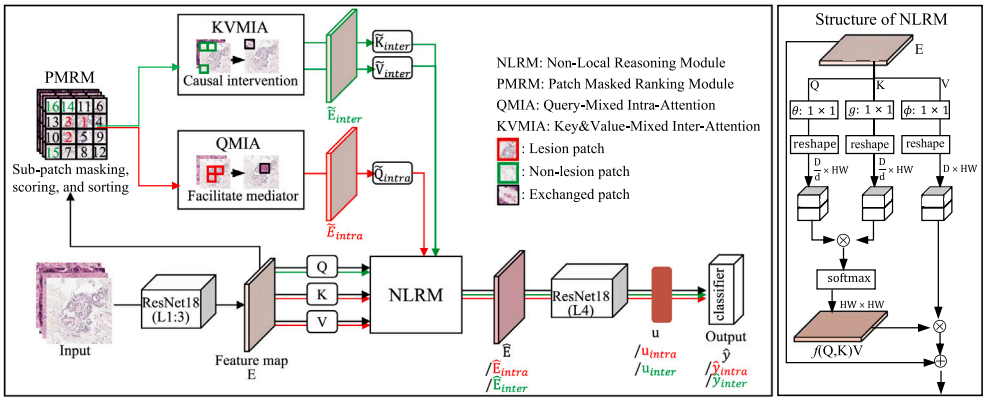

Fig. 2. The framework of CausalMixNet, which consists of NLRM, PMRM, QMIA, and KVMIA. NLRM is utilized to generate mediator **E** by capturing global dependencies within the image. PMRM dissects the image into multiple sub-patches and systematically evaluates the significance of each sub-patch through masking, scoring, and sorting. QMIA marked with red lines can enhance the learning of the mediator. KVMIA marked with green lines is developed to perform causal interventions on the mediator. Collectively, these processes constitute the implementation of front-door adjustment for deconfounding.

three layers of ResNet18 (He et al., 2016) for this purpose, as expressed in the following equation:

$$\mathbf{E} = ResNet^{1:3}(x) \in \mathbb{R}^{D\times H\times W}, \tag{3}$$

where D, H, and W are the number of channels, height and width of **E**, respectively. The mediator extractor encompasses low-level contextual features, such as colors and edges, that are essential for discerning subtle visual distinctions in lesions. To effectively capture pertinent information from surrounding pixels, this module employs a self-attention mechanism based on query (**Q**), key (**K**), and value (**V**). Query is a representation or feature extracted from a specific location on the feature map, probing other positions. Key encodes descriptors of the feature map, guiding the queries to find related content. Values represent the substantive information linked to the keys, which are harnessed to refine the feature map. When a query aligns with a key, the corresponding value will be aggregated to calculate the final output. The calculations of **Q**, **K**, and **V** for the input tensor **E** are expressed as follows:

$$\begin{aligned}\mathbf{Q} &= \theta(\mathbf{E}) = \mathbf{E}\mathbf{W}_\theta \in \mathbb{R}^{\frac{D}{d}\times HW},\\ \mathbf{K} &= \phi(\mathbf{E}) = \mathbf{E}\mathbf{W}_\phi \in \mathbb{R}^{\frac{D}{d}\times HW},\\ \mathbf{V} &= g(\mathbf{E}) = \mathbf{E}\mathbf{W}_g \in \mathbb{R}^{D\times HW},\end{aligned} \tag{4}$$

where $\theta(\cdot)$, $\phi(\cdot)$, and $g(\cdot)$ represent the learnable transformation functions with parameters $\mathbf{W}_\theta$, $\mathbf{W}_\phi$, and $\mathbf{W}_g$, respectively. The parameter d serves as a shrinkage parameter designed to reduce computational costs. The affinity between all positions is calculated by the pairwise function $f(\cdot,\cdot)$ as:

$$f(\mathbf{Q},\mathbf{K}) = softmax(\mathbf{Q}^T\mathbf{K}). \tag{5}$$

Subsequently, the output $\hat{\mathbf{E}}$, which integrates surrounding information, is expressed through the aggregation and residual structure as given by Eq. (6):

$$\hat{\mathbf{E}} = \mathbf{E} + f(\mathbf{Q},\mathbf{K})\mathbf{V}. \tag{6}$$

Next, $\hat{\mathbf{E}}$ is further fed into the subsequent layer of ResNet18 to obtain the one-dimensional feature vector $\mathbf{u} \in \mathbb{R}^{D^4}$, where D^4 represents the dimension of **u**. The prediction $\hat{\mathbf{y}}$ for x is then computed by applying a classifier to **u**. To optimize the model, we minimize the negative log-likelihood of the cross-entropy loss as follows:

$$\mathcal{L}_o = -argmin \sum \mathbf{y} log\hat{\mathbf{y}}, \tag{7}$$

where **y** is the ground-truth label of x.

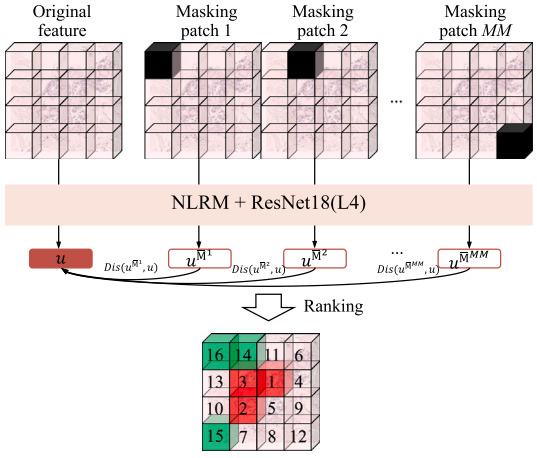

Fig. 3. The PMRM process divides an image into $M \times M$ sub-patches and assesses the importance of each sub-patch by contrasting the distance between the patch-masked sample and the original sample. The red sub-patches represent lesion areas, while the green sub-patches signify non-lesion regions.

4.2. Patch masked ranking module

PMRM is designed to capture pivotal cues within the image associated with corresponding labels, as shown in Fig. 3. This module dissects the image into multiple sub-patches and systematically evaluates the significance of each sub-patch through a crafted process involving masking, scoring, and sorting. Given the feature matrix **E**, it is divided into $M \times M$ (abbreviated as MM) sub-patches, where the ith sub-patch is denoted as $\mathbf{M}^i \in \mathbb{R}^{D\times\frac{H}{M}\times\frac{W}{M}}$. To gauge the significance of each sub-patch, MM patch-masked samples are generated by applying masks to the sub-patches within the completed feature matrix. The ith patch-masked sample is expressed as $\mathbf{E}^{\overline{\mathbf{M}}^i}$:

$$\mathbf{E}^{\overline{\mathbf{M}}^i} = \{\mathbf{E}|\mathbf{M}^i = 0\}. \tag{8}$$

The resultant $\mathbf{E}^{\overline{\mathbf{M}}^i}$ is then fed into the last convolutional layer to derive the corresponding feature vector $\mathbf{u}^{\overline{\mathbf{M}}^i}$. Subsequently, a ranking process is adopted to appraise the importance of sub-patches, determined by calculating the Euclidean distance between $\{\mathbf{u}^{\overline{\mathbf{M}}^1},\ldots,\mathbf{u}^{\overline{\mathbf{M}}^{MM}}\}$ and the feature vector of the complete image **u**. Notably, when the distance between $\mathbf{u}^{\overline{\mathbf{M}}^i}$ and **u** is larger, the ith sub-patch is considered more important, as the masked area loses more crucial information. Through

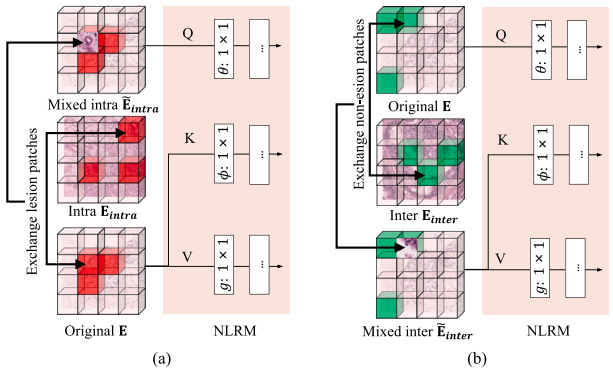

Fig. 4. (a) Showcase of the QMIA process. QMIA enhances the learning of mediator by enhancing the diversity of the query feature. Lesion patches are randomly mixed between samples of the same category. (b) Showcase of the KVMIA process. Causal intervention is conducted on the mediator through randomized mixing of non-lesion patches across samples, altering key and value features dynamically.

this process, the k_ith sub-patch $\mathbf{R}^{k_i}$ is sorted as follows:

$$\mathbf{R}^{k_i} = \{\mathbf{M}^{k_i} | Dis^{k_1} \leq \cdots \leq Dis^{k_i} \leq \cdots \leq Dis^{k_{MM}}\}, \tag{9}$$

where $k_i \in \{1, \ldots, MM\}$ denotes the ranking index of a sub-patch, and Dis^{k_i} represents the Euclidean distance between $\mathbf{u}$ and $\mathbf{u}^{\overline{\mathbf{M}}^{k_i}}$.

4.3. Query-mixed intra-attention

QMIA is designed to facilitate the learning of mediators, focusing on capturing lesion features effectively. Inspired by the attention mechanism within the global contextual network, the query feature plays a crucial role in identifying disease-related features. Therefore, the effective integration of lesion information through the query feature is paramount. To address this issue, this module augments the diversity of the query feature by randomly mixing lesion patches of samples within the same class. The lesion patches are selected based on their ranking in PMRM. The diversified query features enable the retrieval of lesion information from multiple perspectives, thereby improving the learning of the mediator for ResNet18.

The process of QMIA is illustrated in Fig. 4.(a). Firstly, we randomly select a sample x_{intra} from the same category as sample x within a mini-batch, and the ranked sub-patches of x_{intra} are denoted as $\{\mathbf{R}^1_{intra}, \ldots, \mathbf{R}^{MM}_{intra}\}$. Next, a mixed intra-sample $\tilde{x}_{intra}$ is generated by randomly exchanging a lesion patch of x with a lesion patch of x_{intra}. This process can be formulated as:

$$\tilde{\mathbf{E}}_{intra} = \{\mathbf{E} | \mathbf{R}^{k_1} \leftrightarrows \mathbf{R}^{k_2}_{intra}\}, \tag{10}$$

where $k_1 \in \{1, \ldots, \mathcal{K}\}$, $k_2 \in \{1, \ldots, \mathcal{K}\}$, and $\mathcal{K}$ represents the range of the selected top-$\mathcal{K}$ patches. With the obtained mixed intra-sample, the mixed intra-query is generated as:

$$\tilde{\mathbf{Q}}_{intra} = \theta(\tilde{\mathbf{E}}_{intra}) = \tilde{\mathbf{E}}_{intra}\mathbf{W}_\theta \in \mathbb{R}^{\frac{D}{d} \times HW}. \tag{11}$$

Subsequently, we utilize the mixed intra-query $\tilde{\mathbf{Q}}_{intra}$, original key $\mathbf{K}$, and original value $\mathbf{V}$ to summarize disease-related features in an augmented manner by:

$$\hat{\mathbf{E}}_{intra} = \mathbf{E} + f(\tilde{\mathbf{Q}}_{intra}, \mathbf{K})\mathbf{V}. \tag{12}$$

Finally, we minimize the negative cross-entropy loss to optimize the predicted result $\hat{\mathbf{y}}_{intra}$ of mixed intra-sample by:

$$\mathcal{L}_{intra} = -argmin \sum \mathbf{y} log \hat{\mathbf{y}}_{intra}. \tag{13}$$

4.4. Key&value-mixed inter-attention

KVMIA is a key module for causal intervention of $P(Y|do(E))$, thereby diminishing the impact of unobserved confounders. In light of Eq. (1), $P(Y|do(E))$ can be expressed as $\sum P(Y|E,x)P(x)$. This formulation facilitates the coexistence of the mediator $\mathbf{E}$ with confounding factors across different stratums, ensuring equal probability of coexistence and computing the average causal effect. Therefore, two essential elements underline causal intervention on the mediator: (1) fostering the coexistence of the mediator and confounding factors from different stratums, and (2) maintaining an equal probability of coexistence. To achieve these goals, KVMIA constructs the mixed key&value features for sample x by randomly exchanging non-lesion patches with another randomly selected sample. Swapping non-lesion patches actively facilitates the coexistence of the mediator with confounding factors of different stratums, while random operations ensure an equal probability of coexistence, facilitating causal intervention on the mediator.

As illustrated in Fig. 4.(b), firstly, a data x_{inter} is randomly selected within a mini-batch, and its ranked sub-patches are denoted as $\{\mathbf{R}^1_{inter}, \ldots, \mathbf{R}^{MM}_{inter}\}$. Next, the mixed inter-sample $\tilde{x}_{inter}$ is generated by randomly exchanging a non-lesion patch of x with a non-lesion patch of x_{inter}. Similarly, this process can be formulated as:

$$\tilde{\mathbf{E}}_{inter} = \{\mathbf{E} | \mathbf{R}^{k_1} \leftrightarrows \mathbf{R}^{k_2}_{inter}\}, \tag{14}$$

where $k_1 \in \{MM - \mathcal{K}, \ldots, MM\}$ and $k_2 \in \{MM - \mathcal{K}, \ldots, MM\}$. With the obtained mixed inter-sample, the mixed inter-key and inter-value are generated as:

$$\begin{aligned} \tilde{\mathbf{K}}_{inter} &= \phi(\tilde{\mathbf{E}}_{inter}) = \tilde{\mathbf{E}}_{inter}\mathbf{W}_\phi \in \mathbb{R}^{\frac{D}{d} \times HW}, \\ \tilde{\mathbf{V}}_{inter} &= g(\tilde{\mathbf{E}}_{inter}) = \tilde{\mathbf{E}}_{inter}\mathbf{W}_g \in \mathbb{R}^{D \times HW}. \end{aligned} \tag{15}$$

Then, we can utilize the original $\mathbf{Q}$, mixed inter-key $\tilde{\mathbf{K}}_{inter}$, and mixed inter-value $\tilde{\mathbf{V}}_{inter}$ to perform causal intervention by:

$$\hat{\mathbf{E}}_{inter} = \mathbf{E} + f(\mathbf{Q}, \tilde{\mathbf{K}}_{inter})\tilde{\mathbf{V}}_{inter}. \tag{16}$$

QMIA and KVMIA employ different mixing mechanisms due to their distinct objectives: QMIA is designed to enhance the model's comprehension of lesion features in the mediator, necessitating modification solely in the query feature. A targeted adjustment in the query allows the streamlining of the learning process. On the other hand, KVMIA operates with the objective of introducing various new confounding factors to intervene in the mediator. This intervention requires to detect the nuanced alternations in both the key and value features. Finally, akin to QMIA, the negative cross-entropy loss is adopted to optimize the predicted result $\hat{\mathbf{y}}_{inter}$ of mixed inter-sample as follows:

$$\mathcal{L}_{inter} = -argmin \sum \mathbf{y} log \hat{\mathbf{y}}_{inter}. \tag{17}$$

4.5. Overall objective function

Combining the contributions of NLRM, PMRM, QMIA, and KVMIA, we construct the overall objective function for the proposed framework as:

$$\mathcal{L} = \mathcal{L}_o + \alpha \mathcal{L}_{intra} + \beta \mathcal{L}_{inter}, \tag{18}$$

where α and β represent the hyper-parameters, finely tuned to harmonize the interplay among the three fundamental components.

5. Experiment

In this section, we conduct comprehensive experiments to assess the performance of CausalMixNet. Initially, we introduce the datasets used, evaluation protocols, compared methods, and implementation details. Subsequently, we disclose the quantitative results for both in-domain and out-of-domain (OOD) experiments spanning four established medical datasets. Further experimental analysis is conducted

to evaluate the capabilities of the proposed methods, encompassing ablation studies, gender bias analysis, attribute bias analysis, parameter sensitivity examination, robustness checks against noise using the CUB-200-2011 dataset (Wah et al., 2011), and the effectiveness assessment of PMRM. Finally, we present a qualitative analysis to elucidate the interpretability of the methods used in comparison.

5.1. Datasets and evaluation protocols

Our assessment of the proposed method spans seven extensively recognized datasets: BRACS (Brancati et al., 2022), DDR (Li et al., 2019), APTOS,[1] FGADR (Zhou et al., 2021), NIH dataset (Wang et al., 2017), FJ-Thyroid, and CUB-200-2011 (Wah et al., 2011).

(1). In-domain evaluations: the BRACS and DDR datasets. The BRACS dataset, which is publicly available,[2] is pertinent to breast carcinoma subtyping in hematoxylin and eosin (H&E) histopathology images, comprising 4539 regions of interest (ROIs) extracted from whole slide images (WSIs). This dataset includes three primary lesion types: benign, malignant, and atypical, which are further categorized into seven subtypes. The benign category consists of normal (484 ROIs), pathological benign (836 ROIs), and usual ductal hyperplasia (517 ROIs). The malignant category compasses ductal carcinoma in situ (790 ROIs) and invasive carcinoma (649 ROIs). The atypical category contains flat epithelial atypia (756 ROIs) and atypical ductal hyperplasia (ADH, 507 ROIs). For diagnostic, benign and malignant ROI images from BRACS are utilized following the official partition, with 2646/222/408 ROIs allocated for training, validation, and testing, respectively.

The DDR dataset is a publicly available dataset of diabetic retinopathy,[3] containing 12,522 optical fundus images distributed among five classes. These classes contain 6266, 630, 4477, 236, and 913 samples, respectively. We randomly partition this dataset into training, validation, and test sets with a 7:1:2 ratio.

(2). Out-of-domain evaluations: the APTOS and FGADR datasets. The source model is trained using the DDR training subset and vetted against the DDR validation subset. The APTOS dataset, curated by Aravind Eye Hospital for the APTOS 2019 Blindness Detection Competition, includes 3657 publicly available[1] optical fundus images with official diabetic retinopathy (DR) grading. The dataset is distributed across five classes with the following sample counts: 1801, 370, 998, 193, and 295. All samples are employed to assess the out-of-domain generalization capabilities.

The FGADR dataset is a publicly available dataset,[4] offering a comprehensive set of fine-grained annotations for DR. It comprises two distinct subsets: Seg-set and Grade-set. Our analysis incorporates 1841 images with image-level DR labels from this dataset. The sample distribution across the classes is 101, 212, 594, 647, and 287, respectively. All samples are used to assess the generalization ability in out-of-domain scenarios.

(3). Gender bias analysis: the NIH dataset. Following Luo et al. (2022), we select data from the NIH dataset[5] involving patients with no findings and pneumothorax, constructing three experimental setups with varying gender biases. The sample selection for the first two setups strictly follows this work (Luo et al., 2022). Specifically: (1) GbP-Tr1 includes 800 male pneumothorax samples, 100 male no finding samples, 800 female no finding samples, and 100 female pneumothorax samples; (2) GbP-Tr2 consists of 800 female pneumothorax samples, 100 female no finding samples, 800 male no finding samples, and 100 male pneumothorax samples. Additionally, we establish a gender-balanced experimental setup (GbP-B) as a control, comprising 450 female pneumothorax samples, 450 female no finding samples, 450 male no finding samples, and 450 male pneumothorax samples. For the validation and testing sets, we evenly collect 150 and 250 samples from each group (with or without pneumothorax; male or female) according to the settings in Luo et al. (2022).

(4). Attribute bias analysis: the FJ-Thyroid dataset. To evaluate the capacity to address attribute bias, we employed the privately held FJ-Thyroid dataset, which was collected from Fujian Provincial Hospital and consists of 577 thyroid ultrasound images from 289 patients, most of whom contributed two images. This dataset facilitates a binary classification task to differentiate between malignant and benign thyroid tumors, with class labels derived from biopsy results. Each lesion is annotated by experienced clinicians according to the TI-RADS criteria (Tessler et al., 2018), encompassing eight attributes: X1: shape taller than wide; X2: perinodular halo; X3: well circumscribed; X4: microlobulation; X5: hypoechoic; X6: homogeneous echotexture; X7: mainly cystic; and X8: microcalcification. Notably, attributes X3, X5, and X7 exhibit the highest variances at 0.4227, 0.4967, and 0.5001, respectively. For the attribute bias analysis, we established three dataset settings based on these attributes. Setting 1 (FJ-Tr1) features a test set with no occurrences of the three attributes, accounting for 20% of the total samples. The remaining data is randomly divided into training and validation sets at a 7:1 ratio, resulting in a training and validation set with a greater prevalence of significant attributes than the test set. Setting 2 (FJ-Tr2) includes a test set with all three attributes, also making up 20% of the total sample size. The remaining samples are again divided into training and validation sets at a 7:1 ratio, leading to a training and validation set with fewer significant attributes than the test set. Setting 3 (FJ-B) involves a random division of the dataset into training, validation, and test sets at a ratio of 7:1:2, ensuring balanced representation of attributes. All partitions are designed to maintain patient non-overlapping conditions.

(5). Noise tesing: the CUB-200-2011 dataset. The CUB-200-2011 dataset provides 11,788 images of 200 bird species, presenting an ideal challenge due to the species' shared global features but different attributes. This dataset is a publicly available natural image dataset.[6] To simulate noise, we use images from the first 100 classes as base categories, adding images from the remaining 100 species as noise (Dubey et al., 2018). The resulting noisy training set and the pristine test set, taken from base categories, are divided in an 80:20 ratio.

Our evaluation protocol adopts the widely-used medical diagnostic metrics: accuracy (Acc), precision (Precision), recall (Recall), F1-score (F1), and the mean area under the ROC curve (AUC), to quantitatively evaluate the proposed methods across various datasets.

5.2. Compared methods and implementation details

Compared Methods. In order to comprehensively evaluate our proposed method, we benchmark against nine state-of-the-art comparative methods. These methods, known for their efficiency in various disease diagnostics and causal inference methods, include GDRNet (Che et al., 2023), GREEN (Liu et al., 2020), CABNet (He et al., 2021), Mixup-Net (Zhang et al., 2018), MixStyleNet (Zhou et al., 2020), Fishr (Rame et al., 2022), PBBL (Luo et al., 2022), Meta-Causal (Chen et al., 2023), and FAGT (Nguyen et al., 2023). These methods are designed to extract domain-invariant features through supervised learning (Liu et al., 2020; He et al., 2021), domain generalization techniques (Che et al., 2023; Zhang et al., 2018; Zhou et al., 2020; Rame et al., 2022), data augmentation strategies (Che et al., 2023; Zhang et al., 2018; Zhou et al., 2020), or causal inference strategies (Chen et al., 2023; Nguyen et al., 2023). Comparisons with these methods enable a reliable validation of our proposed method's performance in both in-domain, out-of-domain scenarios and bias analysis.

[1] https://www.kaggle.com/competitions/aptos2019-blindness-detection.
[2] https://www.bracs.icar.cnr.it/.
[3] https://www.kaggle.com/datasets/mariaherrerot/ddrdataset.
[4] https://csyizhou.github.io/FGADR/.
[5] https://www.kaggle.com/datasets/nih-chest-xrays/data.
[6] https://www.kaggle.com/datasets/wenewone/cub2002011.

Table 1
The experiment results of in-domain and out-of-domain tests in DDR, APTOS, and FGADR datasets. The **best results** and second-best results are marked with corresponding formats.

Method	In-domain (DDR)			OOD (APTOS)			OOD (FGADR)		
	Acc↑	AUC↑	F1↑	Acc↑	AUC↑	F1↑	Acc↑	AUC↑	F1↑
GDRNet	$59.40_{\pm 0.62}$	$77.87_{\pm 0.91}$	$33.26_{\pm 1.71}$	$62.66_{\pm 0.59}$	$76.55_{\pm 0.90}$	$34.55_{\pm 0.80}$	$30.74_{\pm 0.32}$	$72.49_{\pm 0.68}$	$25.60_{\pm 0.66}$
GREEN	$76.69_{\pm 0.87}$	$87.11_{\pm 0.58}$	$51.26_{\pm 1.00}$	$65.26_{\pm 0.98}$	$75.47_{\pm 0.39}$	$35.44_{\pm 1.09}$	$37.47_{\pm 0.84}$	$74.13_{\pm 0.74}$	$27.47_{\pm 0.69}$
CABNet	$71.60_{\pm 0.97}$	$86.87_{\pm 0.78}$	$53.40_{\pm 0.72}$	$46.94_{\pm 1.19}$	$75.46_{\pm 0.38}$	$37.78_{\pm 1.18}$	$42.14_{\pm 0.50}$	$74.27_{\pm 0.62}$	$27.60_{\pm 0.46}$
MixupNet	$71.69_{\pm 1.45}$	$84.99_{\pm 2.67}$	$51.67_{\pm 0.91}$	$50.88_{\pm 1.91}$	$71.91_{\pm 2.70}$	$33.70_{\pm 0.88}$	$32.40_{\pm 0.16}$	$68.10_{\pm 2.86}$	$20.46_{\pm 2.70}$
MixStyleNet	$60.04_{\pm 3.95}$	$75.80_{\pm 0.91}$	$36.61_{\pm 3.35}$	$40.73_{\pm 0.27}$	$67.50_{\pm 0.67}$	$25.08_{\pm 1.55}$	$20.59_{\pm 3.49}$	$47.30_{\pm 1.28}$	$14.77_{\pm 2.34}$
Fishr	$74.10_{\pm 1.02}$	$87.22_{\pm 0.89}$	$51.01_{\pm 0.39}$	$57.73_{\pm 0.38}$	$76.68_{\pm 0.63}$	$38.73_{\pm 0.51}$	$35.17_{\pm 1.10}$	$74.29_{\pm 0.97}$	$20.84_{\pm 1.58}$
Meta-causal	$78.52_{\pm 0.23}$	$87.89_{\pm 0.63}$	$52.37_{\pm 1.87}$	$65.30_{\pm 0.47}$	$75.60_{\pm 1.23}$	$41.01_{\pm 0.23}$	$41.84_{\pm 1.86}$	$73.39_{\pm 0.52}$	$25.46_{\pm 0.90}$
FAGT	$77.41_{\pm 0.70}$	$87.64_{\pm 0.43}$	$51.55_{\pm 0.77}$	$65.27_{\pm 1.03}$	$77.31_{\pm 0.25}$	$40.33_{\pm 0.39}$	$37.45_{\pm 0.85}$	$72.90_{\pm 0.28}$	$27.56_{\pm 1.02}$
Ours	$\mathbf{79.62}_{\pm 1.06}$	$\mathbf{89.50}_{\pm 0.13}$	$\mathbf{56.97}_{\pm 0.43}$	$\mathbf{68.16}_{\pm 2.09}$	$\mathbf{79.49}_{\pm 0.11}$	$\mathbf{41.62}_{\pm 0.31}$	$\mathbf{45.03}_{\pm 1.70}$	$\mathbf{75.63}_{\pm 0.19}$	$\mathbf{30.28}_{\pm 2.18}$

Table 2
The experiment results of in-domain testing for multi-class classification task and binary classification task in BRACS dataset. The **best results** and second-best results are marked with corresponding formats.

Method	Multi class (5 classes)		Binary class			
	Acc↑	F1↑	Acc↑	Precision↑	Recall↑	F1↑
GREEN	$62.14_{\pm 1.53}$	$62.35_{\pm 1.27}$	$83.85_{\pm 1.53}$	$\mathbf{84.76}_{\pm 0.80}$	$81.63_{\pm 1.61}$	$81.25_{\pm 0.92}$
CABNet	$63.27_{\pm 0.83}$	$63.75_{\pm 0.42}$	$84.91_{\pm 1.38}$	$84.26_{\pm 1.16}$	$79.02_{\pm 1.17}$	$79.85_{\pm 1.25}$
MixupNet	$63.71_{\pm 0.72}$	$63.61_{\pm 0.68}$	$84.57_{\pm 0.75}$	$80.62_{\pm 2.90}$	$80.30_{\pm 1.67}$	$80.32_{\pm 0.57}$
MixStyleNet	$65.97_{\pm 1.02}$	$65.22_{\pm 0.86}$	$85.10_{\pm 0.12}$	$81.28_{\pm 0.21}$	$81.34_{\pm 0.36}$	$81.41_{\pm 0.53}$
Fishr	$65.11_{\pm 0.13}$	$65.13_{\pm 0.89}$	$84.20_{\pm 1.44}$	$80.81_{\pm 1.23}$	$80.87_{\pm 0.94}$	$80.33_{\pm 0.66}$
Meta-causal	$65.75_{\pm 0.20}$	$65.62_{\pm 0.41}$	$85.29_{\pm 1.24}$	$83.73_{\pm 1.41}$	$82.56_{\pm 1.29}$	$82.05_{\pm 1.13}$
FAGT	$65.30_{\pm 0.24}$	$\mathbf{66.14}_{\pm 0.36}$	$85.13_{\pm 0.31}$	$80.26_{\pm 0.98}$	$\mathbf{86.18}_{\pm 0.74}$	$\mathbf{83.77}_{\pm 0.76}$
Ours	$\mathbf{65.98}_{\pm 0.46}$	$65.75_{\pm 0.43}$	$\mathbf{86.13}_{\pm 0.78}$	$83.16_{\pm 1.43}$	$83.92_{\pm 4.63}$	$83.76_{\pm 1.54}$

Table 3
The ablation study of in-domain and out-of-domain tests in DDR, APTOS, and FGADR datasets.

Method	In-domain (DDR)			OOD (APTOS)			OOD (FGADR)		
	Acc↑	AUC↑	F1↑	Acc↑	AUC↑	F1↑	Acc↑	AUC↑	F1↑
Baseline	$79.00_{\pm 0.42}$	$88.57_{\pm 0.02}$	$50.74_{\pm 1.35}$	$66.67_{\pm 0.40}$	$\mathbf{81.09}_{\pm 0.07}$	$39.73_{\pm 0.99}$	$36.95_{\pm 1.78}$	$75.41_{\pm 0.14}$	$24.37_{\pm 2.26}$
QMIA	$79.50_{\pm 0.73}$	$89.13_{\pm 0.49}$	$56.94_{\pm 1.02}$	$66.81_{\pm 1.07}$	$79.09_{\pm 0.71}$	$41.45_{\pm 0.80}$	$43.16_{\pm 1.78}$	$75.41_{\pm 1.74}$	$28.51_{\pm 2.03}$
KVMIA	$79.53_{\pm 0.84}$	$89.11_{\pm 0.13}$	$56.57_{\pm 0.77}$	$65.66_{\pm 3.40}$	$78.97_{\pm 0.84}$	$40.41_{\pm 0.88}$	$43.56_{\pm 2.85}$	$75.61_{\pm 2.13}$	$28.84_{\pm 2.44}$
CausalMixNet	$\mathbf{79.62}_{\pm 1.06}$	$\mathbf{89.50}_{\pm 0.13}$	$\mathbf{56.97}_{\pm 0.43}$	$\mathbf{68.16}_{\pm 2.09}$	$79.49_{\pm 0.11}$	$\mathbf{41.62}_{\pm 0.31}$	$\mathbf{45.03}_{\pm 1.70}$	$\mathbf{75.63}_{\pm 0.19}$	$\mathbf{30.28}_{\pm 2.18}$

Table 4
The ablation study for BRACS dataset.

Method	Multi class (5 classes)		Binary class			
	Acc↑	F1↑	Acc↑	Precision↑	Recall↑	F1↑
Baseline	$63.64_{\pm 0.27}$	$63.47_{\pm 0.76}$	$83.98_{\pm 1.31}$	$79.36_{\pm 2.74}$	$83.54_{\pm 1.53}$	$82.87_{\pm 0.39}$
QMIA	$\mathbf{66.03}_{\pm 0.10}$	$65.47_{\pm 0.22}$	$85.13_{\pm 0.52}$	$\mathbf{83.27}_{\pm 0.78}$	$81.30_{\pm 0.69}$	$82.17_{\pm 0.72}$
KVMIA	$63.89_{\pm 0.81}$	$63.48_{\pm 0.86}$	$84.91_{\pm 0.11}$	$80.36_{\pm 1.74}$	$\mathbf{85.95}_{\pm 1.63}$	$83.64_{\pm 1.83}$
CausalMixNet	$65.98_{\pm 0.46}$	$\mathbf{65.75}_{\pm 0.43}$	$\mathbf{86.13}_{\pm 0.78}$	$83.16_{\pm 1.43}$	$83.92_{\pm 4.63}$	$\mathbf{83.76}_{\pm 1.54}$

Implementation Details. In our experimental setup, ResNet18 serves as the baseline model for the in-domain, out-of-domain, attribute bias analysis, and noise testing. For gender bias analysis, we employed DenseNet121 (Huang et al., 2017) as the backbone model, in line with the approach used in Luo et al. (2022). All experiments are executed on an RTX 3090 GPU. Images are resized to a universal dimension of 256 × 256 pixels. We employ the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.05 to refine our model. For the experiments involving the BRACS dataset, we set the learning rate to 10e−4, the batch size to 16, and limit the training to 30 epochs. The hyper-parameters α and β are specifically tuned to 5.0 and 3.0, respectively. For the DDR and FJ-Thyroid datasets, the learning rate is fixed at 10e−4, with a larger batch size of 64, and the training duration capped at 30 and 100 epochs, respectively. The α and β are adjusted to 0.5 and 1.0, respectively. For the experiments on the NIH dataset, the learning rate is set to 10^{-4}, the batch size to 128, and both α and β are set to 5.0. In tests with the CUB-200-2011 dataset, we use a learning rate of 10e−3, a batch size of 128, and a shorter training span of 20 epochs, setting both α and β to 5.0. The $\mathcal{K}$ for the top-$\mathcal{K}$ sub-patches and M are set as 3 and 4, respectively, across all datasets. We use the validation set to implement an early stopping training strategy. All experimental results are based on the average of five trials conducted with different random seeds. Code is available at https://github.com/Yajie-Zhang/CausalMixNet.

5.3. Quantitative results

In-Domain Results. In-domain experiments evaluate model performance on specific tasks under independent and identically distributed (i.i.d.) conditions, providing a baseline measure of predictive accuracy. Tables 1 and 2 provide results from in-domain testing on the DDR and BRACS datasets, respectively. Table 1 indicates that CausalMixNet achieves commendable results on the DDR dataset. In the F1-score metric, for instance, CausalMixNet outshines the runner-up method by an impressive 3.57%. These findings suggest that CausalMixNet is proficient at discerning both positive and negative instances, showcasing robust overall capabilities. Similarly, our method CausalMixNet is assessed on the BRACS dataset through a multi-class classification task across five categories, as well as a binary classification task targeting benign versus malignant differentiation. Table 2 reveals that CausalMixNet outperforms competing approaches in both multi-class

and binary classification tasks. The in-domain experiments demonstrate that CausalMixNet achieves strong predictive performance across a variety of modalities, suggesting its effectiveness in learning relevant features for the classification tasks.

Out-of-Domain Results. We test the generalization performance of the compared methods using the APTOS and FGADR datasets. Strong generalization suggests that a model has likely learned the true causal factors of diseases, as opposed to simply memorizing dataset-specific noise. As shown in Table 1, CausalMixNet stands out as the leader in generalization across both datasets, boasting an average improvement of 3% over the closest rival. On the FGADR dataset, CausalMixNet achieves an 3% increase in accuracy and F1-score, respectively, over the second-best method. Specifically, our proposed approach, which utilizes a sub-patch mixing strategy for causal intervention, demonstrates superior generalization compared to the MixUp technique used in MixupNet and MixStyleNet. This superiority can be contributed to our method's emphasis on exploring authentic causal relationships, rather than depending primarily on data augmentation for model generalization. Furthermore, the integration of PMRM steers an importance-weighted mixing scheme that enables the model to learn more precise disease markers and execute effective causal interventions.

5.4. Exploration experiments

Ablation Study. To evaluate the efficacy of the key modules in our methodology, QMIA and KVMIA, we conduct ablation studies in both in-domain and out-of-domain scenarios using the DDR, BRACS, APTOS, and FAGDR datasets. Tables 3 and 4 exhibit a performance increase ranging from 1% to 3% over the baseline in in-domain experiments on the DDR and BRACS datasets. Specifically, KVMIA demonstrates a marked impact on the BRACS dataset, while both QMIA and KVMIA are comparable on the DDR dataset. Out-of-domain ablation results, as shown in Table 3, underscore the benefits of QMIA for both APTOS and FAGDR datasets. The synergy between QMIA and KVMIA, where the latter builds on the findings of the former, facilitates enhanced model generalization, underscoring the complementary nature between these two modules.

Gender-Biased Analysis. Fig. 5 presents the AUC performance and, crucially, the variance in AUC scores for various algorithms across three gender-biased training scenarios (GbP-Tr1, GbP-Tr2, GbP-B). A lower variance in AUC across these settings indicates greater robustness to gender bias. CausalMixNet exhibits lower variance than most other methods, demonstrating its stability in the presence of this potential confounder. While MixStyleNet shows slightly lower variance (approximately 0.003 difference), CausalMixNet generally achieves higher AUC scores across the different gender ratio settings, suggesting a better balance between robustness and overall performance.

Attribute Bias Analysis. As illustrated in Fig. 6, we evaluate multiple algorithms on AUC metric and the variance of AUC results across three attribute bias scenarios: FJ-Tr1, FJ-Tr2, and FJ-B. Vertically, CausalMixNet consistently exhibits the lowest variance in three scenarios, indicating its robust capacity to effectively mitigate attribute bias. Horizontally, CausalMixNet outperforms other algorithms in terms of AUC, demonstrating its effectiveness in both independent and identically distributed (FJ-B) and out-of-distribution (FJ-Tr1 and FJ-Tr2) settings.

Parameter Analysis. We conduct an analysis of the hyperparameters α and β on the DDR dataset. The selection of these hyperparameters is based on the validation performance. As shown in Table 5, when varying these two parameters from 0.5 to 5.0, their impact on the overall performance is moderate. Optimal results on the DDR dataset are achieved with $\alpha = 0.5$ and $\beta = 1.0$. A similar method was applied to determine the optimal hyperparameters for other datasets. On the BRACS dataset, the best performance is obtained with $\alpha = 5.0$ and $\beta = 1.0$. On the FJ-Thyroid dataset, the α and β are set as 0.5 and 1.0, respectively. For the NIH and CUB-200-2011 dataset, the optimal settings are $\alpha = 5.0$ and $\beta = 5.0$.

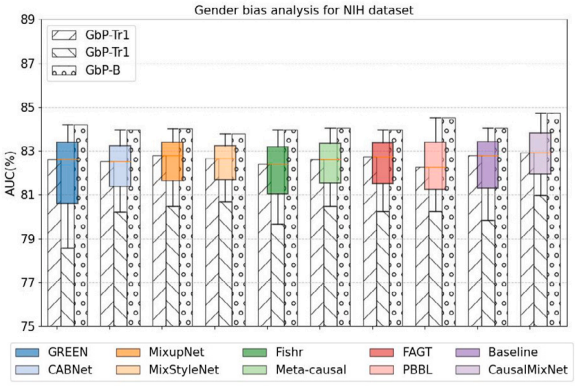

Fig. 5. The gender bias analysis for NIH dataset. The bar charts without colored-band textures represent the AUC results of various algorithms under the three experimental settings of GbP-Tr1, GbP-Tr2, and GbP-B. The transparent-colored box plots represent the variance of the AUC results of each algorithm in the three settings. Lower variance (smaller box and whisker spread) indicates greater robustness to gender bias.

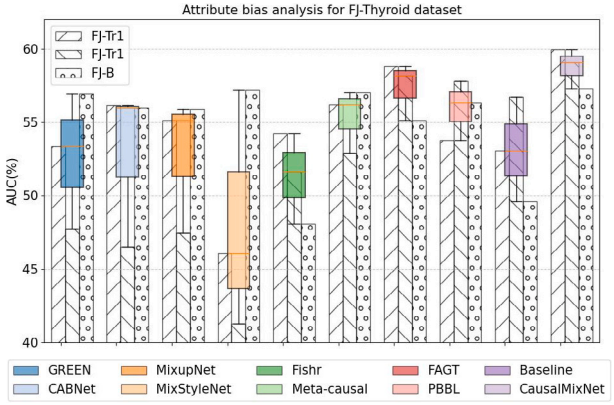

Fig. 6. The attribute bias analysis for FJ-Thyroid dataset. The bar charts without colored-band textures represent the AUC results of various algorithms under the three experimental settings of FJ-Tr1, FJ-Tr2, and FJ-B. The transparent-colored box plots represent the variance of the AUC results of each algorithm in the three settings. Lower variance (smaller box and whisker spread) indicates greater robustness to attribute bias.

Table 5
The parameter analysis of α and β in DDR dataset for in-domain testing on the validation set.

Parameters	In-domain (DDR)		
	Acc↑	AUC↑	F1↑
$\alpha = 0.5, \beta = 0.5$	**79.81**$_{\pm 0.64}$	89.63$_{\pm 0.26}$	55.11$_{\pm 1.04}$
$\alpha = 0.5, \beta = 1.0$	79.77$_{\pm 0.97}$	89.82$_{\pm 0.20}$	**56.86**$_{\pm 0.45}$
$\alpha = 0.5, \beta = 2.0$	79.62$_{\pm 0.34}$	**89.80**$_{\pm 0.50}$	56.64$_{\pm 0.62}$
$\alpha = 0.5, \beta = 3.0$	79.22$_{\pm 0.12}$	89.31$_{\pm 0.28}$	54.20$_{\pm 1.35}$
$\alpha = 0.5, \beta = 4.0$	78.66$_{\pm 0.43}$	88.94$_{\pm 0.36}$	52.03$_{\pm 1.65}$
$\alpha = 0.5, \beta = 5.0$	78.56$_{\pm 0.31}$	88.79$_{\pm 0.61}$	55.64$_{\pm 0.37}$
$\alpha = 1.0, \beta = 1.0$	79.42$_{\pm 1.22}$	89.50$_{\pm 0.34}$	55.12$_{\pm 0.65}$
$\alpha = 2.0, \beta = 1.0$	78.60$_{\pm 0.65}$	89.15$_{\pm 0.62}$	56.19$_{\pm 0.42}$
$\alpha = 3.0, \beta = 1.0$	78.96$_{\pm 0.52}$	89.24$_{\pm 0.34}$	55.43$_{\pm 0.85}$
$\alpha = 4.0, \beta = 1.0$	79.50$_{\pm 0.35}$	89.62$_{\pm 0.35}$	56.52$_{\pm 0.87}$
$\alpha = 5.0, \beta = 1.0$	78.49$_{\pm 0.98}$	89.10$_{\pm 0.55}$	56.77$_{\pm 1.44}$

Noise Testing. To evaluate the model's resilience against unobservable confounding factors, we introduce simulated noise into the CUB-200-2011 dataset. According to the noise settings in the Ref. Dubey et al. (2018), the model is trained on a dataset consisting of (1-a%) samples from the base classes and a% samples from other classes, with the latter serving as noise. For evaluation, a noise-free test set

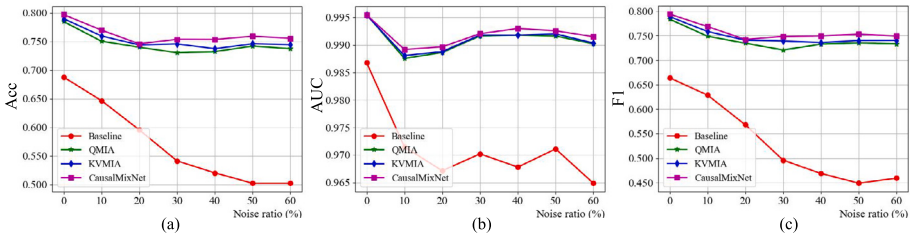

Fig. 7. The noise testing on CUB-200-2011 dataset.

Table 6
The results in DDR dataset for evaluating the effectiveness of PMRM.

Method	In-domain (DDR)			OOD (APTOS)			OOD (FGADR)		
	Acc↑	AUC↑	F1↑	Acc↑	AUC↑	F1↑	Acc↑	AUC↑	F1↑
Random	**80.35**$_{\pm 0.03}$	88.94$_{\pm 0.48}$	56.08$_{\pm 0.64}$	67.21$_{\pm 0.59}$	78.86$_{\pm 0.50}$	40.53$_{\pm 1.28}$	42.42$_{\pm 0.73}$	75.45$_{\pm 1.47}$	27.74$_{\pm 0.63}$
Intra-sample	79.12$_{\pm 1.08}$	88.76$_{\pm 0.22}$	56.88$_{\pm 0.64}$	67.62$_{\pm 0.87}$	78.52$_{\pm 0.14}$	41.27$_{\pm 0.40}$	41.87$_{\pm 1.17}$	73.40$_{\pm 1.11}$	26.97$_{\pm 1.52}$
Inter-sample	79.04$_{\pm 1.00}$	89.39$_{\pm 0.02}$	55.86$_{\pm 0.39}$	65.60$_{\pm 0.96}$	**79.88**$_{\pm 0.11}$	41.55$_{\pm 0.56}$	**45.19**$_{\pm 0.86}$	75.39$_{\pm 0.54}$	29.81$_{\pm 1.58}$
PMRM	79.62$_{\pm 1.06}$	**89.50**$_{\pm 0.13}$	**56.97**$_{\pm 0.43}$	**68.16**$_{\pm 2.09}$	79.49$_{\pm 0.11}$	**41.62**$_{\pm 0.31}$	45.03$_{\pm 1.70}$	**75.63**$_{\pm 0.19}$	**30.28**$_{\pm 2.18}$

Table 7
Model efficiency and training analysis for the all methods on the DDR dataset.

Method	FPS	GPU memory (MB)	Parameter memory (MB)	Training time (Hours)
GDRNet	35.75	12 532	43.25	0.3385
GREEN	36.19	5862	43.16	0.3620
CABNet	35.45	**2574**	43.29	0.3101
MixupNet	35.73	3590	42.64	**0.2662**
MixStyleNet	35.89	18 610	56.16	0.3046
Fishr	**36.59**	8808	46.67	0.3265
Meta-causal	36.29	15 644	89.68	0.6737
FAGT	35.94	12 536	86.78	0.8962
baseline	35.86	2954	**43.02**	0.3041
Ours	34.39	8066	43.15	0.4696

containing only base class images is used. In this setting, we varied the noise ratio from 0% to 60%. Fig. 7 displays how the baseline performance deteriorates with increasing noise levels, showing a drop of up to 20% in accuracy and F1-score. However, incorporating QMIA and KVMIA modules stabilizes performance, exhibiting a steady trend despite escalating noise. These experiments highlight the distinct advantage of CausalMixNet in mitigating the impact of unobservable confounding factors.

Effectiveness of PMRM. To assess the significance of PMRM, we employ various intervention strategies on the mediator. These include: (1) a random approach, indiscriminately swapping sub-patches without considering their relevance; (2) intra-sample, leveraging $\mathbf{E}_{intra}$ within the same category to modify $\mathbf{Q}/\mathbf{K}/\mathbf{V}$; and (3) inter-sample, leveraging $\mathbf{E}_{inter}$ across different categories to modify $\mathbf{Q}/\mathbf{K}/\mathbf{V}$. Table 6 presents the outcomes, revealing that the proposed PMRM attains superior performance. Conversely, disregarding the significance of sub-patches or simply exchanging full $\mathbf{Q}/\mathbf{K}/\mathbf{V}$ sets leads to a decline in results.

Model Efficiency and Training Analysis. We assess the frames per second (FPS), GPU memory usage, parameter memory, and training time on the DDR dataset, as shown in Table 7. The FPS results indicate the adaptability of various methods for processing real-time data (Čížek and Faigl, 2018). CausalMixNet achieved an FPS of 34.39, which, while slightly lower than some of the competing methods, demonstrates a commendable level of performance suitable for real-time applications. The overall similarity in FPS among the different methods indicates that they are well-optimized for handling real-time data streams, with only slight variations in processing speed. In terms of GPU memory usage, CausalMixNet consumes more memory compared to GREEN, CABNet, MixupNet, and the Baseline. This increased usage is primarily due to the dual data augmentation performed during QMIA and KVMIA. Regarding parameter memory, CausalMixNet exhibits a smaller parameter memory footprint than other compared methods, with only a marginal increase in parameters over the baseline method. Furthermore, CausalMixNet requires the shorter training time compared to other causal inference methods (Meta-causal and FAGT). In conclusion, CausalMixNet demonstrates a balanced performance with commendable FPS and parameter memory usage, making it suitable for real-time applications.

5.5. Visualization analysis

We perform a visualization analysis of DenseNet121 using Grad-CAM (Selvaraju et al., 2017) on the NIH dataset for pneumothorax. The affected side in a pneumothorax typically displays a distinct black (radiolucent) area, with the lung's edge shifting towards the center due to the presence of gas (Reid et al., 2024). As shown in Fig. 8, the lesions in the original images are outlined in red by a professional physician. Since gas and the lung border can affect the entire lobe, most annotated regions encompass the whole lobe. The visualization results demonstrate that the CausalMixNet method is more accurate in lesion localization compared to other methods. In some cases, the localized regions partially miss the lesions, particularly under suboptimal image contrast conditions. In contrast, other approaches either fail to identify the lesions or focus on irrelevant areas, such as the shoulders. This highlights the effectiveness of CausalMixNet in addressing gender bias and improving diagnostic accuracy. By accurately localizing the lesions, CausalMixNet not only enhances the identification rate of pneumothorax cases but also provides more reliable support for clinical decision-making, further emphasizing the importance of selecting appropriate models to tackle gender bias.

6. Conclusion

This study presents CausalMixNet as a sophisticated deconfounding method for investigating causality within medical image diagnosis, effectively integrating NLRM, PMRM, QMIA, and KVMIA modules to

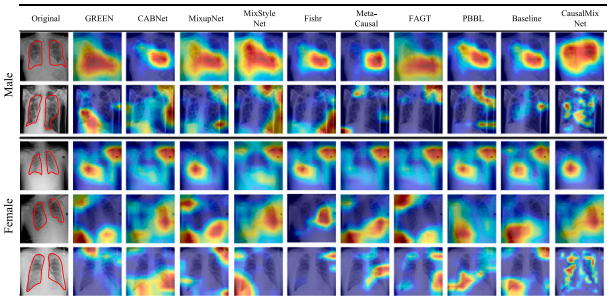

Fig. 8. The visualization analysis on the NIH dataset. Professional doctors have annotated the lesions in the original images, which are delineated in red for clarity.

mitigate the influence of unobservable confounding factors. NLRM is specifically crafted to extract the mediator in front-door adjustment, establishing a causal link from image features to diagnostic labels. PMRM is charged with the identification of lesion and non-lesion patches, refining the focus of the analysis. QMIA enhances the discernment of the mediator by diversifying the query feature and deepening the learning of lesion characteristics. KVMIA applies causal intervention to the mediator by integrating key and value features to compute average causal effects. Empirical evaluation across four datasets in both in-domain and out-of-domain testing scenarios demonstrate that CausalMixNet can be consistently effective, scalable, and reliable in improving image classification accuracy. Additionally, our analysis of gender bias illustrates CausalMixNet's effectiveness in reducing the influence of unobserved confounders. Through comprehensive experimentation, we have validated the contribution of each individual component within CausalMixNet, the system's resilience in noisy conditions, and its exceptional interpretability.

Although CausalMixNet demonstrates capabilities in causal inference, several limitations warrant discussion. While inference latency and memory usage during deployment are competitive with baseline methods, CausalMixNet requires greater GPU memory and a longer convergence period during training. Furthermore, the attention mechanism can sometimes produce fragmented or incomplete heatmaps, particularly when analyzing complex or low-contrast lesion areas. To address these limitations, our future efforts will focus on several key areas: First, we will explore optimizing the model architecture and training procedures to reduce GPU memory consumption and convergence period. Second, investigating more sophisticated attention mechanisms, potentially incorporating techniques from transformer models or recurrent networks, could improve the continuity and completeness of lesion localization, especially in challenging imaging scenarios. We aim to extend our framework to tackle both observable and unobservable confounders within multimodal contexts and larger-scale model settings.

CRediT authorship contribution statement

Yajie Zhang: Writing – original draft, Methodology. **Yu-An Huang:** Methodology. **Yao Hu:** Writing – review & editing, Methodology. **Rui Liu:** Validation, Methodology. **Jibin Wu:** Writing – review & editing, Supervision, Resources. **Zhi-An Huang:** Writing – review & editing, Supervision, Resources. **Kay Chen Tan:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially supported by the National Nature Science Foundation of China (Grant No. 62202399 and U21A20512), the Research Grants Council of the Hong Kong SAR (Grant No. PolyU15218622, PolyU15215623, and C5052-23G), and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515012944 and 2024A1515011984), the Hong Kong Polytechnic University (Project IDs: P0043563, P0046094), the Fundamental Research Funds for the Central Universities, China (Grant No. G2023KY05102), and the General Program of National Natural Science Foundation of China (Grant No. 62472353).

Data availability

Data will be made available on request.

References

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D., 2019. Invariant risk minimization. arXiv preprint arXiv:1907.02893.

Bareinboim, E., Forney, A., Pearl, J., 2015. Bandits with unobserved confounders: A causal approach. In: Advances in Neural Information Processing Systems. vol. 28.

Bissoto, A., Valle, E., Avila, S., 2020. Debiasing skin lesion datasets and models? Not so fast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 3192–3201. http://dx.doi.org/10.1109/CVPRW50498.2020.00378.

Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al., 2022. Bracs: A dataset for breast carcinoma subtyping in H&E histology images. Database 2022, baac093. http://dx.doi.org/10.1093/database/baac093.

Che, H., Cheng, Y., Jin, H., Chen, H., 2023. Towards generalizable diabetic retinopathy grading in unseen domains. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 430–440. http://dx.doi.org/10.1007/978-3-031-43904-9_42.

Chen, J., Gao, Z., Wu, X., Luo, J., 2023. Meta-causal learning for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7683–7692. http://dx.doi.org/10.1109/CVPR52729.2023.00742.

Chen, Y., Guo, X., Xia, Y., Yuan, Y., 2022a. Disentangle then calibrate: Selective treasure sharing for generalized rare disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 512–522. http://dx.doi.org/10.1007/978-3-031-16437-8_49.

Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S., 2022b. C-CAM: Causal CAM for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11666–11675. http://dx.doi.org/10.1109/CVPR52688.2022.01138.

Chen, X., You, S., Tezcan, K.C., Konukoglu, E., 2020. Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. 64, 101713. http://dx.doi.org/10.1016/j.media.2020.101713.

Čížek, P., Faigl, J., 2018. Real-time FPGA-based detection of speeded-up robust features using separable convolution. IEEE Trans. Ind. Inform. 14 (3), 1155–1163. http://dx.doi.org/10.1109/TII.2017.2764485.

Deng, W., Zhong, Y., Dou, Q., Li, X., 2023. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 158–169. http://dx.doi.org/10.1007/978-3-031-34048-2_13.

Dubey, A., Gupta, O., Raskar, R., Naik, N., 2018. Maximum-entropy fine grained classification. Adv. Neural Inf. Process. Syst. 31.

Duboudin, T., Dellandréa, E., Abgrall, C., Hénaff, G., Chen, L., 2021. Encouraging intra-class diversity through a reverse contrastive loss for single-source domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 51–60. http://dx.doi.org/10.1109/ICCVW54120.2021.00012.

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., Kompatsiaris, I., 2022. A survey on bias in visual datasets. Comput. Vis. Image Underst. 223, 103552. http://dx.doi.org/10.1016/j.cviu.2022.103552.

Gao, R., Tang, Y., Xu, K., Huo, Y., Bao, S., Antic, S.L., Epstein, E.S., Deppen, S., Paulson, A.B., Sandler, K.L., et al., 2020. Time-distanced gates in long short-term memory networks. Med. Image Anal. 65, 101785. http://dx.doi.org/10.1016/j.media.2020.101785.

Gokhale, T., Anirudh, R., Thiagarajan, J.J., Kailkhura, B., Baral, C., Yang, Y., 2023. Improving diversity with adversarially learned transformations for domain generalization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 434–443. http://dx.doi.org/10.1109/WACV56688.2023.00051.

He, A., Li, T., Li, N., Wang, K., Fu, H., 2021. CABNet: Category attention block for imbalanced diabetic retinopathy grading. IEEE Trans. Med. Imaging 40 (1), 143–153. http://dx.doi.org/10.1109/TMI.2020.3023463.

He, X., Tan, E.-L., Bi, H., Zhang, X., Zhao, S., Lei, B., 2022. Fully transformer network for skin lesion analysis. Med. Image Anal. 77, 102357. http://dx.doi.org/10.1016/j.media.2022.102357.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Hu, Y., Huang, Z.-A., Liu, R., Xue, X., Sun, X., Song, L., Tan, K.C., 2023. Source free semi-supervised transfer learning for diagnosis of mental disorders on fMRI scans. IEEE Trans. Pattern Anal. Mach. Intell. 45 (11), 13778–13795. http://dx.doi.org/10.1109/TPAMI.2023.3298332.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2261–2269. http://dx.doi.org/10.1109/CVPR.2017.243.

Huang, Z.-A., Zhu, Z., Yau, C.H., Tan, K.C., 2021. Identifying autism spectrum disorder from resting-state fMRI using deep belief network. IEEE Trans. Neural Netw. Learn. Syst. 32 (7), 2847–2861. http://dx.doi.org/10.1109/TNNLS.2020.3007943.

Jiang, X., Zhang, D., Li, X., Liu, K., Cheng, K.-T., Yang, X., 2025. Labeled-to-unlabeled distribution alignment for partially-supervised multi-organ medical image segmentation. Med. Image Anal. 99, 103333. http://dx.doi.org/10.1016/j.media.2024.103333.

Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.-A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al., 2024. Generative models improve fairness of medical classifiers under distribution shifts. Nat. Med. 30, 1–8. http://dx.doi.org/10.1038/s41591-024-02838-6.

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H., 2019. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf. Sci. 501, 511–522. http://dx.doi.org/10.1016/j.ins.2019.06.011.

Li, Y., Lao, Q., Kang, Q., Jiang, Z., Du, S., Zhang, S., Li, K., 2023. Self-supervised anomaly detection, staging and segmentation for retinal images. Med. Image Anal. 87, 102805. http://dx.doi.org/10.1016/j.media.2023.102805.

Liu, S., Gong, L., Ma, K., Zheng, Y., 2020. Green: A graph residual re-ranking network for grading diabetic retinopathy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 585–594. http://dx.doi.org/10.1007/978-3-030-59722-1_56.

Liu, R., Huang, Z.-A., Hu, Y., Huang, L., Wong, K.-C., Tan, K.C., 2024a. Spatio-temporal hybrid attentive graph network for diagnosis of mental disorders on fMRI time-series data. IEEE Trans. Emerg. Top. Comput. Intell. 1–13. http://dx.doi.org/10.1109/TETCI.2024.3386612.

Liu, R., Huang, Z.-A., Hu, Y., Zhu, Z., Wong, K.-C., Tan, K.C., 2024b. Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI. IEEE Trans. Neural Netw. Learn. Syst. 35 (6), 7627–7641. http://dx.doi.org/10.1109/TNNLS.2022.3219551.

Liu, R., Huang, Z.-A., Hu, Y., Zhu, Z., Wong, K.-C., Tan, K.C., 2024c. Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fMRI data. IEEE Trans. Neural Netw. Learn. Syst. 35 (8), 10591–10605. http://dx.doi.org/10.1109/TNNLS.2023.3243000.

Luo, L., Xu, D., Chen, H., Wong, T.-T., Heng, P.-A., 2022. Pseudo bias-balanced learning for debiased chest X-ray classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 621–631. http://dx.doi.org/10.1007/978-3-031-16452-1_59.

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., D'Amour, A., 2022. Causally motivated shortcut removal using auxiliary labels. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 739–766.

Mao, H., Liu, H., Dou, J.X., Benos, P.V., 2022. Towards cross-modal causal structure and representation learning. In: Machine Learning for Health. PMLR, pp. 120–140.

Miao, J., Chen, C., Liu, F., Wei, H., Heng, P.-A., 2023. CauSSL: Causality-inspired semi-supervised learning for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21369–21380. http://dx.doi.org/10.1109/ICCV51070.2023.01959.

Nguyen, T., Do, K., Nguyen, D.T., Duong, B., Nguyen, T., 2023. Causal inference via style transfer for out-of-distribution generalisation. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1746–1757. http://dx.doi.org/10.1145/3580305.359927.

Nie, W., Zhang, C., Song, D., Bai, Y., Xie, K., Liu, A.-A., 2023. Chest X-ray image classification: A causal perspective. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 25–35. http://dx.doi.org/10.1007/978-3-031-43898-1_3.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82 (4), 669–688.

Pölsterl, S., Wachinger, C., 2021. Estimation of causal effects in the presence of unobserved confounding in the Alzheimer's continuum. In: Information Processing in Medical Imaging. Springer, pp. 45–57. http://dx.doi.org/10.1007/978-3-030-78191-0_4.

Qu, J., Xiao, X., Wei, X., Qian, X., 2024. A causality-inspired generalized model for automated pancreatic cancer diagnosis. Med. Image Anal. 94, 103154. http://dx.doi.org/10.1016/j.media.2024.103154.

Rame, A., Dancette, C., Cord, M., 2022. Fishr: Invariant gradient variances for out-of-distribution generalization. In: International Conference on Machine Learning. PMLR, pp. 18347–18377.

Reid, W., Chung, F., Hill, K., 2024. Chest X-rays. Cardiopulm. Phys. Ther.: Manag. Case Stud. 141.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 618–626. http://dx.doi.org/10.1109/ICCV.2017.74.

Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K., 2023. Explaining the black-box smoothly—A counterfactual approach. Med. Image Anal. 84, 102721. http://dx.doi.org/10.1016/j.media.2022.102721.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312. http://dx.doi.org/10.1109/TMI.2016.2535302.

Tessler, F.N., Middleton, W.D., Grant, E.G., 2018. Thyroid imaging reporting and data system (TI-RADS): a user's guide. Radiology 287 (1), 29–36. http://dx.doi.org/10.1148/radiol.2017171240.

Thiagarajan, J.J., Thopalli, K., Rajan, D., Turaga, P., 2022. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. Sci. Rep. 12 (1), 597. http://dx.doi.org/10.1038/s41598-021-04529-5.

Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A., 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. 79, 102470. http://dx.doi.org/10.1016/j.media.2022.102470.

Wachinger, C., Becker, B.G., Rieckmann, A., Pölsterl, S., 2019. Quantifying confounding bias in neuroimaging datasets with causal inference. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 484–492. http://dx.doi.org/10.1007/978-3-030-32251-9_53.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The caltech-ucsd birds-200–2011 dataset. CNS-TR-2011-001.

Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D., 2022. Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Trans. Med. Imaging 41 (7), 1688–1698. http://dx.doi.org/10.1109/TMI.2022.3146973.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3462–3471. http://dx.doi.org/10.1109/CVPR.2017.369.

Wang, J., Zhong, C., Feng, C., Zhang, Y., Sun, J., Yokota, Y., 2023. Disentangled representation for cross-domain medical image segmentation. IEEE Trans. Instrum. Meas. 72, 1–15. http://dx.doi.org/10.1109/TIM.2022.3221131.

Xu, C., Xu, L., Ohorodnyk, P., Roth, M., Chen, B., Li, S., 2020. Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal GANs. Med. Image Anal. 62, 101668. http://dx.doi.org/10.1016/j.media.2020.101668.

Yang, X., Zhang, H., Cai, J., 2023. Deconfounded image captioning: A causal retrospect. IEEE Trans. Pattern Anal. Mach. Intell. 45 (11), 12996–13010. http://dx.doi.org/10.1109/TPAMI.2021.3121705.

Zardavas, D., Irrthum, A., Swanton, C., Piccart, M., 2015. Clinical management of breast cancer heterogeneity. Nat. Rev. Clin. Oncol. 12 (7), 381–394. http://dx.doi.org/10.1038/nrclinonc.2015.73.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations.

Zhang, Y., Huang, Z.-A., Hong, Z., Wu, S., Wu, J., Tan, K.C., 2024. Mixed prototype correction for causal inference in medical image classification. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4377–4386. http://dx.doi.org/10.1145/3664647.368139.

Zhong, S., Wang, W., Feng, Q., Zhang, Y., Ning, Z., 2025. Cross-view discrepancy-dependency network for volumetric medical image segmentation. Med. Image Anal. 99, 103329. http://dx.doi.org/10.1016/j.media.2024.103329.

Zhou, Y., Wang, B., Huang, L., Cui, S., Shao, L., 2021. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. IEEE Trans. Med. Imaging 40 (3), 818–828. http://dx.doi.org/10.1109/TMI.2020.3037771.

Zhou, K., Yang, Y., Qiao, Y., Xiang, T., 2020. Domain generalization with MixStyle. In: International Conference on Learning Representations.