基于医学图像属性的混合概率推断

田和坤[1)]

1)(西安交通大学)

摘 要 深度神经网络在医学图像诊断领域有着广泛的应用，在肿瘤检测、结节分类等方面已经有了显著的效果。然而，模型的性能与数据集的质量息息相关。没有大量、优质的训练数据，神经网络模型就很难通过训练表现出较好的性能。而医疗领域中最显著的问题就是高质量医疗数据集难以获取，导致目前纯神经网络仍难以克服泛化性差、缺乏可解释性、验证性不佳等问题。

针对上述问题，本文提出一种混合概率推断方法，将因果推理与神经网络相结合构建了强大的推理系统。具体而言，需要首先对病理图像的特征进行提取，将其表示为病理属性，之后利用图神经网络建立属性与属性之间、属性与疾病之间的关系图，并采用贝叶斯网络对因果关系进行推理建模，最后将二者结果紧密耦合，从而得到一个泛化能力强、验证效果好、具有可解释性的诊断方法。

本方法可以用于 CT 图像的肺结节良恶性分类，其准确度达到 95.36%；也可以利用胸部 x 射线图像来诊断结核病，准确度为 96.64%。实验证明，混合概率推断算法对于纯神经网络模型的性能有巨大改进，并且为未来因果推理与神经网络相结合的进一步研究提供了新的思路。

关键词 因果推理、神经网络、混合概率推断、医学图像诊断

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.01.2025.00001

Hybrid Probabilistic Inference Based on Medical Image Attributes

Tian Hekun[1)]

1)(Xi 'an Jiaotong University, China)

Abstract Deep neural networks have widespread applications in medical image diagnosis, and have achieved significant results in tumor detection and nodule classification. However, the performance of the model is closely related to the quality of the data set. Without a large and high-quality training data set, the neural network model is difficult to perform well through training. The most significant problem in the medical field is the difficulty of obtaining high-quality medical data sets, which makes it difficult for pure neural networks to overcome the problems of poor generalization, lack of explainability, and poor verification performance. In view of these problems, this paper proposes a hybrid probabilistic inference method, combining causal reasoning with neural networks to build a powerful inference system. Specifically, the features of the pathological images need to be extracted first, represented as pathological attributes, and then a graph neural network is used to establish relationships between attributes and diseases. A Bayesian network is used for causal inference modeling, and the results of the two are tightly coupled to obtain a diagnosis method with strong generalization ability, good verification effect, and explainability.This method can be used for CT image lung nodule classification, achieving an accuracy of 95.36 precent; it can also be used to diagnose tuberculosis using chest x-ray images, achieving an accuracy of 96.64 percent. The experimental results show that the hybrid probabilistic inference algorithm has a huge improvement in the performance of pure neural network models, and provides a new idea for further research on causal reasoning and neural networks.

Key words Causal reasoning; neural networks; mixed probability inference; medical image diagnosis

1 绪论

1.1 拟解决的问题

随着医学知识的快速更新与病情病因的日益复杂，病理图像中的医学信息已经远远超出了人工经验与简单推理所能处理的范畴。而深度神经网络因其能从海量数据提取大量有效信息的优势，已经被广泛应用到医学图像诊断的领域中，如肿瘤检测、病灶分析等。现有的图像诊断方法一般需要采用卷积神经网络等对图像中的特征进行提取、选择和分类，如 Kuo 等设计了一个全卷积神经网络对头部 CT 图像进行分割处理和联合分类[1]，用于定位

收稿日期：2025-01-11；修改日期：2025-01-11 田和坤，男，2002 年生. E-mail: 1229194407@qq.com.
第 1 作者手机号码：18109279541，E-mail: 1229194407@qq.com

人脑中颅内出血的具体位置。

然而，已有研究并未解决医学图像处理领域中高质量数据集缺乏的问题。没有大量的、广泛的数据集支持，模型就难以通过训练表现出较好的泛化能力，验证效果也可能不佳。且深度神经网络的结果并不具有可解释性，其输出结果和输入信息之间并没有很强的因果联系，导致其在实际应用中的说服力不足、可信度较低。因此，亟需一种兼具解释性、泛化性、验证性的医学图像诊断方法，用来改进已有的纯深度神经网络算法。

1.2 国内外研究现状

计算机视觉领域中主要采用三类方法对医学图像的表征进行学习和分析，分别是属性学习、贝叶斯神经网络、神经-符号学习。下面将分别对这三类方法进行调研和分析。

1.2.1 属性学习

属性学习一般需要从数据中提取中间语义表示来进行迁移学习，即让属性作为输入图片和预测结果之间的“桥梁”。早期的属性需要进行手动标记，由于这种方法很难覆盖到所有属性，之后的研究开始采用机器学习方法自动学习属性特征[2][3][4][5][6][7][8]。而近年来图神经网络 (GNN) 的出现使得科研人员能够对属性之间的关系进行学习，从而提高概率推断的准确性。

2009 年，美国华盛顿大学的 Kumar 等提出两种基于属性的图像处理方法用于人脸识别[4]，第一种方法采用经过训练的二元分类器来识别人脸部的属性是否存在，第二种分类器不再需要对属性类别进行手工标记，而是学习面部特定区域与参考人脸的相似度。这两种方法均不需要消耗大量资源对图片进行特征对齐，并且在一个包含大量公众人物的真实图像数据集 (PubFig) 中取得了很好的验证效果。而 2018 年，美国威斯康星大学的孟等证明了对属性之间的关系进行学习可以显著提升模型性能[9]。在训练时仅仅对部分属性进行排序，之后利用消息传递对图像表征、图像之间的关系和属性之间的关系进行端到端的学习，实验结果显示采用 GNN 对属性间依赖关系学习既可以明显减少参数数量，又不损失模型验证性能。

属性学习关注数据中最明显的特征信息，从而提高模型的泛化性能和验证性能，但大多数研究采用的端到端训练方法仍屏蔽了因果推理过程，模型的可解释性较差。

1.2.2 贝叶斯网络

贝叶斯网络是一种通过结构学习[10] 和参数学习[11] 进行因果关系建模和逻辑推理的自然方法，之后为了让贝叶斯网络能够直接地从数据中提取有效特征，研究人员将其嵌入神经网络结构中来构建更加强大的学习模型。而为了弥补神经网络可解释性差、过拟合和无法进行不确定性分析的缺点，不少研究提出将贝叶斯网络与深度学习相结合，构建贝叶斯神经网络 (BNN)，来对神经网络中的参数进行不确定性概率分析[12][13][14]。

2018 年，美国因特尔研究院的 Rohekar 等提出一种无监督的深度神经网络结构学习方法[15]，将神经网络结构学习问题转化为贝叶斯网络结构学习问题，通过学习生成图、构造随机逆的方式构造判别图，证明了生成图中变量之间的依赖关系在类条件判别图中仍然存在，在图像分类任务中验证了普通深度网络可以被更小的学习结构取代，显著降低了模型的训练成本。2018 年，中国合肥工业大学的石曾林等提出了一种概率深度体素扩张残差网络，用于 3D 核磁共振图像的心脏分割任务[16]。该方法基于暂退法 (Dropout) 来训练模型学习体素类标签的权值后验分布，并通过实验验证了该方法相比纯神经网络具有更好的可解释性。

综上所述，贝叶斯网络与神经网络二者具有互补的优势，二者结合可以获得更强可解释性的模型参数。然而仅仅将贝叶斯方法用于计算权值的后验分布并不能最大程度地利用其强大的因果推理能力，因此亟需一种方法能够将深度神经网络的表征能力和贝叶斯网络的因果关系建模能力嵌入到一个统一的框架中从而构建强大的学习模型。

1.2.3 神经-符号学习

对于神经-符号学习，早期的研究者主要关注如何将符号信息注入神经网络以提高模型性能[17][18]，近年来大多数研究希望将符号推理与深度神经网络相结合以提高模型的可解释性、泛化性和可验证性[19][20]。

2018 年，美国哈佛大学的 Yin 等设计了一个神经符号视觉问答系统[19]，首先从图像中恢复结构场景表示，并从问题中恢复系统跟踪信息，然后在场景上执行方法以获得答案。该方法既在 CLEVR 数据集上取得了 99.8% 的准确性，又为整个推理过程提供了充分的解释信息。

2020 年，香港大学李青等人引入语法模型作为连接神经感知和符号推理的符号模型[20]，并且提

出了一种新的反向搜索算法，模仿自顶向下的类人学习过程，通过符号推理模块有效地传播错误。此外，用马尔可夫链蒙特卡罗采样的最大似然估计来解释整个学习框架。神经-符号学习结合了两个强大的方法: 用于数据表征学习的深度神经网络和用于因果关系建模的符号推理学习。但由于大多数符号学习方法是不可微的，使得端到端的训练较为困难。

综上所述，已有的研究主要关注属性特征的学习，也不乏将神经网络与属性学习相结合的方法，然而目前的方法并不能将属性建模和神经网络相结合而得到一个强大的学习模型，使得方法兼具强大的因果推理能力和概率推断能力。同时由于因果推理网络的结构与传统神经网络的结构并不相同，因此目前的方法并没有找到一个效果良好的混合网络的训练方法。

2　创新思想与解决方法

2.1　创新思想

而本研究既将深度神经网络与贝叶斯网络相结合构建了一个对图像特征并行处理的混合概率推断方法，将二者优势很好地结合起来；又提出了用两个贝叶斯网络的反向梯度传播来对符号推理进行建模的方法，一定程度上解决了神经-符号学习难以训练的问题，最终得到的一个兼具较好的解释性、泛化性和验证性的医学图像诊断算法。

2.2　解决方法

本研究提出基于属性的混合神经概率推断方法，整体流程如图 1 所示。

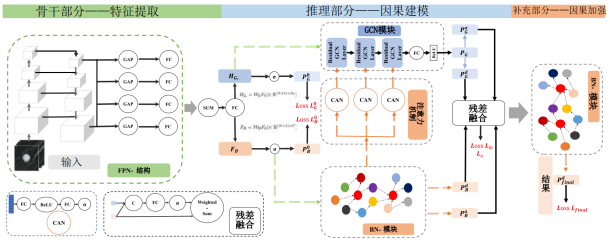

图 1　基于属性的混合神经概率推断方法

首先，该方法用卷积神经网络提取病理图像中的特征，之后用贝叶斯神经网络和图卷积神经网络来并行地对特征进行处理，由此建立属性和疾病的因果推理模型，并用无向图来表示属性与属性之间、属性与疾病之间的联系，之后采用交叉注意力机制和融合算法将二者紧密耦合，而为了加强因果推理的效果，该方法在最后增加了一个贝叶斯网络计算最终的预测结果。具体步骤将分为以下三个部分叙述:

2.2.1　骨干部分——特征提取

方法的第一步是要从图像块中提取有效的特征表示，首先采用卷积神经网络提取不同尺度的信息，之后引入特征金字塔网络和全局平均池化来优化特征表示，最终得到一个特征向量。这一部分的处理流程如图 2 所示。

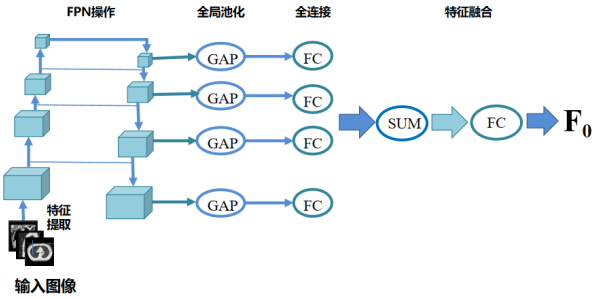

图 2　图像特征提取

该部分采用基于残差神经网络 (ResNet)[21] 或高效神经网络 (EfficientNet)[22] 作为主干网络，用来提取图像中不同层级的特征信息。之后在基础网络中加入特征金字塔网络 (FPN)，从而同时利用低层特征的高分辨率和高层特征的高语义信息，使得不同层级具有同样丰富的特征描述。

此外，本文在每个尺度中都采用全局平均池化 (GAP) 来减少参数数量。而为了不损失低级特征图中的有效信息，池化操作需保证每一级的特征图池化后得到的向量维度都相同。例如，当 FPN 中特征图是 64*64*256 时，通过将前两个维度中每个 2*4 相邻块的第三维连接到一起，重塑一个 32*16*2048 的特征图，之后在此基础上进行池化操作，从而得到一个 2048 维的特征向量。

最终，所有降维后的特征都会经过并行的全连接层 (FC)，求和后再次通过 FC 进行特征融合，得到总的特征向量 F_0。

2.2.2　推理部分——关系建模

从图像中提取特征表示后，贝叶斯网络和图卷积神经网络均会将其转化为属性，然后分别地对属性进行因果推理和关系建模，前者输出不同属性和疾病的边缘概率分布，后者输出属性和疾病的分类结果。二者并不是相互独立的，贝叶斯网络的计算结果会通过交叉注意力机制作用到图卷积神经网络中，使得图卷积神经网络更好地关注不同结点的特征值。而最终得到的推断结果和分类结果也会通过一个函数进行聚合，从而将因果推断和关系建模

相融合得到更精准的预测结果。

贝叶斯网络 (BN-1)。图 3 所示的贝叶斯网络将骨干部分得到的特征向量转化为属性特征的离散分布，之后利用属性描述进行因果推断，之后得到每个结点的边缘概率分布。这一部分主要关注三个研究内容：与骨干网络的连接、贝叶斯网络的设计与训练和最后的推理计算。

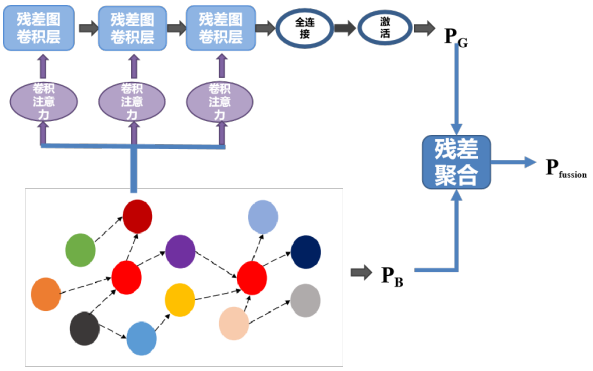

图 3 贝叶斯网络

与骨干网络的连接。第一步得到的特征向量 F_0 会再次作为 FC 层的输入，经给权重矩阵 W_B 的映射后转化为一个特征图 F_B，$F_B = W_B F_0 (\in R^{(N+1)\times C})$，其中 N 是属性的个数，C 是每个属性的评价等级个数，而额外加 1 是为了表示所要检测的某种目标疾病。此处的评价等级指的是对每个属性 (或疾病) 的存在与否、严重程度等评估维度的量化指标。也就是说，F_B 是 n 个 c 维向量的组合，每个向量是对不同属性（或疾病）的详细描述。之后 F_B 会经过 softmax 处理得到 P_B^0，作为属性 (或疾病) 不同评价指标的离散分布表示。

贝叶斯网络的设计。上文得到的 P_B^0 会作为贝叶斯网络的输入，用来对网络结构进行建模和分析。而贝叶斯网络 $B = \langle V_B, \xi_B, \Theta \rangle$，由两部分组成——一个有向无环图 (DAG)$\langle V_B, \xi_B \rangle$ 和一个联合概率分布表 (Θ)。其中，有向无环图的结点有 N+1 个，表示 N 种属性和 1 种疾病，结点 v_i 到 v_j 之间的有向边表示 v_j 的概率依赖于 v_i。每个结点有一张条件概率分布表 (CPT)，表示给定父节点情况下该子节点不同取值的概率分布。而 Θ 是所有的概率分布表的集合 (CPTs)，用来编码整个图中所有结点的联合概率分布。

贝叶斯网络的训练过程分为结构学习和参数学习两部分。首先采用基于贝叶斯准则 (BIC) 的动态规划方法对网络的拓扑结构进行学习，其次根据最大似然估计方法对每个参数进行更新，最终将属性特征分布转化为一个有向图和多张概率分布表。

概率估计结果。经过上述两步，便可以根据条件概率和属性的离散分布计算所有结点的边缘概率分布 $P_B(v_0)$，其中对疾病的概率估算结果会作为整个贝叶斯网络的输出。具体计算方式如下：

$$P_B(v_0) = \int \cdots \int_V P(v_0, v_1, \ldots v_n) dv_1 \ldots dv_n. \quad (1)$$

$$P(v_0, v_1, \ldots v_n) = \prod_{i=0}^{n} P(v_i | Parents(v_i)). \quad (2)$$

图卷积神经网络 (GCN)。同理，图卷积神经也需要对骨干网络的特征向量进行处理，将其转化为属性的特征表示。之后用无向图表示属性与属性之间的关系，并用图卷积的方式来增强每个属性的特征。这一部分也将分为三部分来阐述：与骨干网络的连接、属性间关系的建模以及图卷积处理。具体流程如图 4。

图 4 图卷积神经网络

与骨干网络的连接。图卷积神经网络与贝叶斯网络的第一步连接操作，同样是用一个权重矩阵 W_g 将特征向量 F_0 转化为一个特征图 $H_g = W_g F_0 (\in R^{(N+1)\times D_0})$。不同之处在于这里的 D_0 不再是对属性的评价指标个数，而是图卷积神经网络中每个结点的输入特征维度，并且此时暂且不需要对每个属性的特征描述进行归一化处理。

属性间关系建模。贝叶斯网络只对属性和疾病进行了由因至果的推理，但利用属性与属性之间的关系进行辅助推断可以显著提升方法的性能。因此本研究采用无向图 $g = (V_g, \xi_g)$ 来表示属性之间的联系。具体来说，每个属性和疾病都对应一个结点，结点之间的边表示二者之间存在某种联系。而边的权重表示这种联系的强弱，作为后续需要学习的参数。

图卷积处理。这里采用残差图卷积方法[23] 进行计算，即

$$H_{g_{l+1}} = F(g_l, W_l) + H_{g_l}. \quad (3)$$

$$F(F_l, W_g) = Update(Aggregate(g_l, W_l^{agg}), W_l^{update}). \quad (4)$$

其中 H_{g_l} 是 1 层中所有结点的特征图表示，$F(g_l, W_l)$ 是图卷积运算器，$g_l = (V_g, \xi_l)$ 是每一层的关系图，W_l^{agg} 和 W_l^{update} 分别是用于聚合和更新函数的权值。本研究采用最大池化节点特征聚合器来汇集节点 v_i 及其所有相邻结点的特征差异，并用一个以批量归一化和 ReLU 作为激活函数的多层感知器 (MLP) 作为节点特征更新器，从而将原始特征与聚合特征连接起来。

最后，完成更新的 H_{g_l} 会经过一个全连接层和一个 softmax 处理得到 P_G，作为属性和疾病的分类结果。

BN-1 和 GCN 的耦合。如图 5 所示，BN-1 和 GCN 并不是相互独立的，而是通过交叉注意力机制和预测结果的聚合紧密关联在一起。具体而言，BN-1 会提供注意力信息作用于 GCN 的每个结点，而 BN-1 的概率分布和 GCN 的分类信息会通过一个聚合函数进行融合。

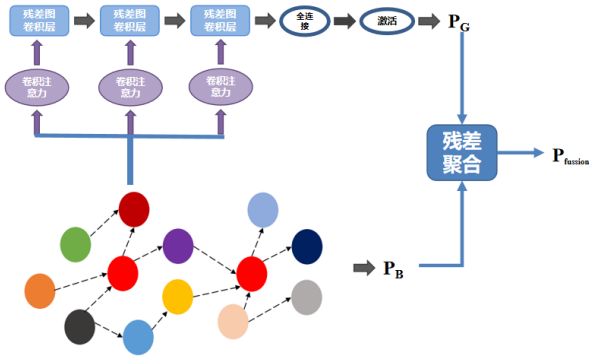

图 5　BN-1 和 GCN 的耦合

引入交叉注意力机制。BN-1 输出的边缘概率分布可以作为 GCN 中每一层的注意力模块，作为不同属性重要程度的考量。每一层的注意力值分为两部分，分别是空间注意力和通道注意力。每个结点的空间注意力值都不相同，计算方式如下：

$$M_{attn}^{l} = \sigma(W_{l_1} ReLU(W_{l_0} P_B(V))), l \in 1, 2, 3, ..., L. \tag{5}$$

其中，M_{attn}^{l} 是第 l 层的结点注意力向量，V 指代所有结点，$P_B(V) \in R^{N+1 \times C}$ 是 BN-1 中所有结点的概率分布。而为了增强 GCN 中不同层的特征，本研究在 GCN 中引入了挤压-激励模块 (SENet)[24]，即用 C_{attn}^{l} 表示每一层的注意力向量，如下：

$$C_{attn}^{l} = \sigma(W_{l_{ex}} ReLU(W_{l_{sq}} GAP(H_{g_l}))). \tag{6}$$

其中 $W_{l_{sq}}$ 和 $W_{l_{ex}}$ 分布是挤压和激励的参数值。之后，每一层的特征图都会乘以相应的通道注意力值，而每个结点会乘以各自的空间注意力值，更新后的特征值会作为 GCN 中下一层的输入。

预测结果的聚合。属性和疾病的分类结果均会采用如下的残差融合方案。

$$P_{fusion} = w_B \cdot P_B^{+}(1 - w_B) \cdot \sigma(W_0 Concat(P_G, P_B)). \tag{7}$$

其中，P_B 指 BN 中得到的边缘概率分布，P_G 指 GCN 中得到的分类结果，w_B 是可学习的权重系数，W_0 是权重矩阵，Concat 是连接函数。

2.2.3　补充部分——因果加强

BN-1 得到的因果推理结果可能会因与 GCN 的融合而受到损害，为了解决这一问题，在得到上文的聚合结果后再次引入一个贝叶斯网络（BN-2）。如图 6 所示，将 P_{fusion} 作为 BN-2 的输入，按照与 BN-1 类似的计算方法得到新的疾病预测结果，作为最终的输出。

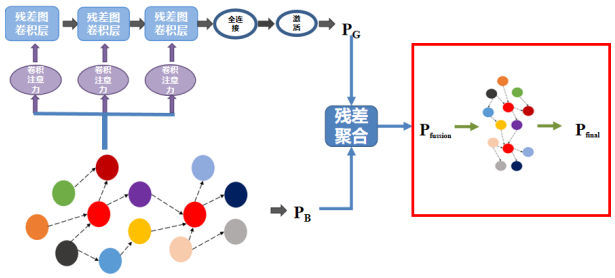

图 6　因果推理的加强

2.2.4　训练过程——交叉学习

除了 BN 模块，方法中的其他网络结构均可以采用随机梯度下降的方法来训练。然而 BN 模块的参数学习可以采用反向传播的方式进行更新，但结构部分无法采用相同方式进行学习。为了将贝叶斯网络的结构学习和参数学习紧密联系在一起，同时提高贝叶斯网络和其他网络结构的兼容度，本研究采用两阶段方式对整个网络进行交替训练。

第一阶段通过固定两个 BN 模块的结构和参数，利用反向梯度传播来更新 CNN 主干和 GCN 分支的权值；第二阶段则要以上一阶段得到的预测结果作为训练标签来更新 BN 模块的结构和条件概率表，前者采用基于贝叶斯准则 (BIC) 的动态规划方法对网络的拓扑结构进行学习，后者根据最大似然估计方法对每个参数进行更新。

其中，第一阶段采用深度监督策略，将多个监

督信号分配到混合网络的各个阶段，每个监控信号主要负责训练网络中模块的一个子集，整体损失函数是五个损失函数的加权组合, 如下所示:

$$L_{all} = w_1 * L_G^0 + w_2 * L_B^0 + w_3 * L_a + w_4 * L_d + w_5 * L_{final}. \tag{8}$$

具体而言，第一组监督信号 L_B^0、L_G^0 分别是贝叶斯模块和图卷积模块的损失值，计算方式如下:

$$L_B^0 = -\sum_{i=1}^{N+1}\sum_{j=1}^{C}(y_{ij} log(p_{ij}^{B,0}) + (1-y_{ij}) log(1-p_{ij}^{B,0})). \tag{9}$$

$$L_G^0 = -\sum_{i=1}^{N+1}\sum_{j=1}^{C}(y_{ij} log(p_{ij}^{G,0}) + (1-y_{ij}) log(1-p_{ij}^{G,0})). \tag{10}$$

第二组两个监督信号应用于 BN 和 GCN 的融合结果。对属性和疾病的分类结果分别分别计算两个交叉熵损失 L_d 和 L_a，计算方式如下:

$$L_d = -\sum_{j=1}^{C}(y_j^d log(p_j^d) + (1-y_j^d) log(1-p_j^d)). \tag{11}$$

$$L_a = -\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}^a log(p_{ij}^a) + (1-y_{ij}^a) log(1-p_{ij}^a)). \tag{12}$$

最后一个监督信号位于整个混合网络的末端，负责训练整个网络中的所有模块，计算对最终的疾病分类结果的损失，具体如下:

$$L_{final} = -\sum_{j=1}^{C}(y_j^d log(p_j^{final}) + (1-y_j^d) log(1-p_j^{final})). \tag{13}$$

上述五个监督信号会通过加权聚合的方式作用于整个网络，作为第一阶段的深度监督信号。

3 实验验证与分析

为证明本方法的有效性，需要将其分别应用于两个不同的医学图像诊断任务中——肺结节良恶性分类任务和胸部 x 射线图像诊断任务，并选择与疾病有高度因果关系的属性进行训练。同时希望通过消融实验来对模型性能进行分析，验证贝叶斯网络、图卷积网络等模块在网络中所发挥的作用。

3.1 肺结节良恶性分类任务

肺部结节的早期诊断对肺癌患者的长期治疗有着重要的意义，在临床实践中放射科医生会根据结节位置、形状等属性进行诊断。本研究将基于属性的诊断方法应用于 LIDC-IDRI 基准数据集中的 CT 图像进行肺结节的良恶性分类，并与现有的方法进行比较。

3.1.1 实验设置

实验数据。LIDC-IDRI 是世界上最大的肺癌数据集之一, 其数据来自 7 家医疗机构的 1018 例胸部 CT 扫描病例。每个病例都由一个 $512mm * 512mm$ 的切片和一个标注结节详细位置的 XML 文件组成。所有结节的直径在 3mm 到 30mm 之间，其恶性程度由 4 名经验丰富的放射科医生评估，最终评定出良性结节 1301 例，不确定结节 612 例，恶性结节 644 例。

除了对结节良恶性进行分类，还需要通过对 8 种属性进行描述来为每个结节进行分级，即微妙性、内部结构、钙化度、球形度、边缘度、分叶度、针状体和放射学坚固度。同时使用样条插值法将所有 CT 图像标准化为统一的大小 $1.0 * 1.0 * 1.0mm^3$。

评价指标。实验中采用 10 倍交叉验证，每个模型都独立训练和测试 5 次，每次都随机初始化权值。最后通过准确性、灵敏度、特异性、精度平均值和标准差等常用指标对模型性能进行评估。其中, 准确性、灵敏度、特异性、精度分别是模型正确分类的比例、正确预测正样本的比列、正确预测负样本的比例、正确预测的正样本数占所有预测为正的样本比例，它们取址范围均在 0 到 1 之间，数值越大表示模型性能越好。

实验环境和实施细节。所有模型都在 NVIDIA Titan X pascal 图形处理器上使用 PyTorch 训练 160 个周期，并采用 Adam 作为优化器。实验中将初始学习速率设置为 1e-3，每 30 个周期后减少 10 倍，重量衰减值设为 1e-4。输入图像的大小是 64mm * 64mm * 64mm，在单个 GPU 上的批处理大小是 6。

对比方法。本实验选取了目前表现较好的医学图像诊断方法作为对比，其训练数据来源均与本方法所用的数据一致。如表 1 所示，A-I 方法采用了纯神经网络架构来完成诊断任务，而 T1-M42 方法采用了属性学习、神经-符号学习等多种混合概率推断方法。其中 A 方法采用多作物卷积神经网络, 训练数据来自 LIDC 中的 825 个结节的子集；B 方法采用三维 CNN，训练数据为 1144 个结节子集 C。

C 到 H 的方法均使用相同数量的数据集，即 1945 个病例。其中，H 方法引入 1839 个未标记结节用于半监督学习。

3.1.2　实验结果

实验通过该方法与第一类方法比较来证明引入因果推理模块的有效性，而通过与第二类方法比较来证明将关系建模与概率推断耦合到一个统一框架的有效性。表 1 展示了方法与其他较先进的肺结节分类模型的比较结果，可以看到在应用数据增强和不增强的两种情况下，本方法均在各项评价指标中排名第一。

表 1　在 LIDC-IDRI 数据集上的肺结节分类模型的性能比较

	2* 方法	数量		评价指标					
		B	M	准确率	灵敏度	特异性	AUC	精度	F-指数
A	Multi-crop CNN[25]	528	297	87.14	77.00	93.00	93.00	未给出	未给出
B	3D CNN[26]	635	509	91.26	未给出	未给出	未给出	未给出	未给出
C	3D GLCM feature+SVM[27]	1301	644	85.38 ± 0.10	70.20 ± 0.15	92.80 ± 0.20	88.19 ± 0.16	82.85 ±0.38	75.99 ± 0.10
D	Multi-visual features[28]	1301	644	87.90 ± 0.17	84.50 ± 0.19	89.09 ± 0.25	93.77 ± 0.15	79.31 ± 0.37	81.82 ± 0.21
E	Deep + visual features[29]	1301	644	88.73 ±0.15	84.40 ± 0.20	90.88 ± 0.13	94.02 ± 0.20	82.09 ± 0.24	83.23 ± 0.21
F	TMME with Resnet-50[30]	1301	644	91.01 ± 0.10	83.83 ± 0.15	94.56 ± 0.13	95.35 ± 0.15	88.40 ± 0.24	86.07 ± 0.15
G	Knowledge-based[31]	1301	644	91.60 ± 0.15	86.52 ± 0.25	94.00 ± 0.30	95.70 ± 0.24	87.75 ± 0.52	87.13 ± 0.16
H	Semi-Supervised[12]	1301	644	92.53± 0.05	84.94 ± 0.17	96.28 ± 0.08	95.81 ± 0.19	未给出	未给出
I	Multi-Scale Cost-Sensitive[13]	1156	556	92.64 ± 0.12	85.58 ± 0.44	95.87 ± 1.26	94.00 ± 0.26	90.39 ± 0.48	87.91 ± 0.11
T1	Low-Level-Feature[3]	1301	644	88.73 ± 0.15	84.40 ± 0.20	90.88 ± 0.13	94.02 ± 0.20	82.09 ± 0.24	83.23 ± 0.21
T2	Basic-visual-Feature[4]	1301	644	91.01 ± 0.10	83.83 ± 0.15	94.56 ± 0.13	95.35 ± 0.15	88.40 ± 0.24	84.07 ± 0.15
M1	ResNet-50	1301	644	88.14 ± 0.23	82.17 ± 0.14	90.77 ± 0.15	91.21 ± 0.14	82.18 ± 0.11	82.36 ± 0.14
M2	Efficient-B4	1301	644	89.21 ± 0.12	83.83 ± 0.24	91.18 ± 0.22	92.05 ± 0.27	86.88 ± 0.19	83.04 ± 0.23
M3	ResNet-50-FPN	1301	644	90.01 ± 0.13	84.23 ± 0.21	91.54 ± 0.21	92.81 ±0.31	84.26 ± 0.12	84.71 ± 0.21
M4	Efficient-B4-FPN	1301	644	90.91 ± 0.22	85.74 ± 0.13	92.27 ± 0.15	93.23 ± 0.10	87.10 ± 0.24	86.97 ± 0.14
M31	ResNet-50-FPN-GCN-Relation[14]	1301	644	91.60 ± 0.15	86.52 ± 0.25	92.32 ±0.15	93.70 ± 0.24	86.75 ± 0.52	85.13 ± 0.16
M41	Efficient-B4-FPN-GCN-Relation[34]	1301	644	92.15 ± 0.12	86.97 ± 0.23	93.13 ± 0.11	93.89 ± 0.23	87.14 ± 0.21	85.57 ± 0.24
M32	ResNet-50-FPN-GRU[9]	1301	644	91.41 ± 0.11	86.12 ±0.14	92.92 ± 0.19	93.61 ± 0.23	87.44 ± 0.21	85.22 ± 0.15
M42	Efficient-B4-FPN-GRU[9]	1301	644	92.21 ± 0.10	86.94 ± 0.11	93.89 ± 0.24	93.72 ± 0.12	88.24± 0.13	86.63 ± 0.21
O1	本方法 ResNet-50-FPN	1301	644	93.74 ± 0.17	89.23 ± 0.21	95.76 ± 0.24	96.12 ± 0.12	94.21 ± 0.14	88.27± 0.31
O2	本方法 Our-Efficient-B4-FPN	1301	644	95.31 ± 0.15	90.51 ± 0.15	96.15 ± 0.22	96.47 ± 0.31	95.95 ± 0.24	88.83 ± 0.45
O2*	本方法 Efficient-B4-FPN*	1301	644	**95.36 ± 0.10**	**91.01 ± 0.16**	**96.47 ± 0.12**	**96.54 ± 0.32**	**96.96 ± 0.21**	**89.13 ± 0.15**

为了证明属性关系建模的有效性，将方法与 A-I 方法进行比较。具体而言，A 方法和 B 方法的准确率分别为 87.14% 和 91.26%，C 到 H 的方法使用相同大小的数据集，其中 H 方法准确率达到最高（92.53%），而方法 I 提出多尺度三维 ResNet，性能达到了 92.64%，相比之下，当采用相同的主干 (3D ResNet) 时，本方法（O1）获得了更好的准确率。此外，当使用高效 B4-FPN 骨干网络时，数据增强模型（O2*）达到了最高的 95.36%，而没有数据增强的模型（O2）比使用相同主干（M4）的基线增加了 4.41%（93.74%），证明了属性关系建模在提高肺结节分类性能的重要作用。

而为了证明将神经网络与概率推断相结合的有效性，需要进一步将方法与 M31、M41、M32、M42 等属性关系建模方法进行比较。M31、M41 中提出的关系网络模块和 M32、M42 中提出的 GRU 模块的在基准线上将准确率分别提高了 1.59% 和 1.30%，而本方法将相同基线的准确率提高了 3.63%。同理，为证明骨干网络的有效性。本文与两种经典的属性学习方法 T1 和 T2 进行了比较，其准确率较低，分别为 88.73% 和 91.01%，本方法的准确率分别高出 6.63%、4.35%。

综上所述，上述实验结果表明，本研究提出的混合神经-概率推理方法在 LIDC-IDRI 数据集上具有明显的优于现有方法的性能。

3.1.3　消融实验

为了验证不同网络模块的有效性，该实验在 LIDC 数据集上进行了消融实验，结果如表 2 所示。其中，“BN-1”和“BN-2”表示第一和第二个贝叶斯模块，“CNA-RES”表示跨网络注意和残差融合模块，“SEatt”表示在 GCN 中采用 SE-wise 注意力机制，“GradBN”表示通过 BN 模块的梯度反向传播，“交替训练”表示所提出的交替训练策略。“Relation”指在 GCN 模块之后删除 BN 模块并附加关系网络。

表 2　在 LIDC-IDRI 数据集上的分类方法的消融研究

GCN	BN-2	CNA-RES	SEatt	BN-1	GradBN	AlterTrain	Relation	准确率	灵敏度	特异性	AUC	精度	F-指数
✓	✓	✓	✓	✓	✓	✓	×	**95.31 ± 0.15**	**90.51 ± 0.15**	**96.15 ± 0.22**	**96.47 ± 0.31**	**95.95 ± 0.24**	**88.83 ± 0.45**
✓	×	✓	✓	✓	✓	✓	×	94.74 ± 0.09	88.92 ± 0.13	95.86 ± 0.41	95.83 ± 0.09	92.54 ± 0.06	88.04 ± 0.04
✓	×	×	✓	✓	✓	✓	×	92.11 ± 0.14	88.14 ± 0.22	93.74 ± 0.15	94.91 ± 0.26	88.11 ± 0.09	86.88 ± 0.12
×	×	×	×	✓	✓	✓	×	91.21 ± 0.22	87.26 ±0.14	93.21 ± 0.13	93.47 ± 0.54	88.01 ± 0.09	86.07 ± 0.56
✓	✓	×	✓	✓	✓	✓	×	93.64 ± 0.11	88.01 ± 0.15	95.03 ± 0.13	95.02 ± 0.13	91.42 ± 0.13	87.92 ± 0.04
✓	✓	✓	×	✓	✓	✓	×	94.01 ± 0.02	88.64 ± 0.08	95.23 ± 0.07	95.32 ± 0.11	91.11 ± 0.44	87.13 ± 0.09
✓	✓	×	✓	×	✓	✓	×	92.41 ± 0.14	88.14 ± 0.22	93.74 ± 0.15	94.91 ± 0.26	92.11 ± 0.09	86.88 ± 0.12
✓	✓	✓	✓	✓	×	✓	×	93.81 ± 0.10	88.04 ± 0.32	93.74 ± 0.15	94.91± 0.26	90.21 ± 0.07	86.96 ± 0.10
✓	✓	✓	✓	✓	✓	×	×	94.31 ± 0.10	88.28 ± 0.41	94.23 ± 0.12	94.87 ± 0.76	89.11 ± 0.22	87.03 ± 0.20
✓	×	×	×	×	×	×	✓	92.15 ± 0.12	86.97 ± 0.23	93.13 ± 0.11	93.89 ± 0.23	87.14 ± 0.21	85.57 ± 0.24
✓	×	×	×	×	×	×	×	91.31 ± 0.12	87.83 ± 0.24	93.15 ± 0.13	94.07 ± 0.12	86.65 ± 0.12	87.04 ± 0.13

可以看到，两个贝叶斯模块 (BN-1 和 BN-2)、交叉注意力机制和残差融合网络 (CNA-RES)、通过贝叶斯网络的反向传播等网络结构均对整体方法的性能有一定的贡献。其中，如果移除 CNA-RES，模型性能将下降 1.67%，如果在此基础上删除 BN-2，模型性能将显著下降 3.2%；而如果去除 BN-1 与 GCN 的耦合，性能会下降 2.9%，单独删除 BN-2 则会让模型性能下降 0.57%；此外如果删除两个 BN 模块和一个关系网络，性能会显著下降 3.16%。

综上所述，消融研究证明了方法的每个部分都发挥着重要作用，并且紧密耦合在一起，由此构成的混合神经概率推断方法才有较好的准确性。

3.2　胸部 x 射线图像诊断任务

结核病是全球死亡率最高的传染病之一，如果能在早期的胸片中发现可将其死亡率降低 70%。结核病在胸片中的放射学异常也可以用图像属性来描述，如弥漫性结节和纤维化表现。本研究为进一步评估方法的有效性，将其应用于胸部 x 涉嫌图像诊断结核病的任务中，并采用消融实验来来进一步

分析性能。

3.2.1 实验设置

实验数据。TB-Xatt 数据集共有 14200 张胸部 x 射线图像，每张图像有 3 处异常，而每种疾病有 1326 张图像。每张图像都由经验丰富的医生进行标注，对不同的放射学异常进行评估，包括“肺巩固”、“肺空化”、“弥漫性结节”、“纤维化条纹”、“肺不张”、“多发性结节”、“胸腔积液”。首先对原始 x 射线图像进行分割，并提取左右两侧肺的边界框，然后将这些边界框的大小调整为 512 x 256，并将每对左右边界框作为一个单一的图像样本并排放置。

评价指标。在该任务中同样采用 10 倍交叉验证，每个模型都独立训练和测试 5 次，每次随机初始化权重。模型的性能也使用与 LIDC 相同的一组指标进行评估，包括准确性、灵敏度/召回率、特异性、精确度等。

实验环境和实现细节。与 LIDC 不同，ResNet 和骨干网络是在 ImageNet 上预先训练的。优化器仍采用 Adam，初始学习速率设置为 1e-3，批处理大小为 32。每次训练需迭代 60 次，在 20 次和 40 次后，学习率分别降低原来的 10 倍。

对比方法。在同样的数据集上，将混合概率推断算法与现有的算法进行了比较，包括两种最先进的方法（S1 和 S2），两种属性学习方法（A1 和 A2）和两种关系建模方法（M31、M41 和 M32、M42）。

3.2.2 实验结果

与现有方法的性能比较结果如表 3 所示，可以看到无论是否使用数据增强方法，本方法在所有评价指标中均有最佳的表现。

表 3 TB-Xatt 数据集上结核病诊断模型的性能比较

2*	方法	评价指标					
		准确率	灵敏度	特异性	AUC	精度	F-指数
A1	Low-Level-Feature[3]	86.14±0.14	83.11± 0.26	89.17 ±0.10	90/11± 0.76	81.10±0.12	81.71 ±0.11
A2	Basic-visual-Feature[4]	87.23±0.12	84.16 ± 0.27	92.00 ± 0.09	91.27 ± 0.63	83.26 ± 0.71	83.44 ± 0.17
S1	Attention-Guide[35]	93.12 ± 0.10	86.17 ±0.16	91.22 ± 0.34	93.67 ± 0.43	87.21 ± 0.53	85.64 ± 0.29
S2	ADINet[36]	93.43 ± 0.19	87.24 ± 0.25	91.87 ± 0.11	94.11 ± 0.29	87.44 ± 0.54	84.21 ± 0.13
M1	ResNet-50	90.17 ±0.11	90.44 ± 0.12	89.16 ± 0.23	90.03 ± 0.21	82.09 ± 0.12	84.15 ± 0.23
M2	Efficient-B4	93.21 ± 0.25	91.54 ± 0.22	92.55 ± 0.11	92.67 ± 0.19	84.27 ± 0.98	87.23 ± 0.09
M3	ResNet-50-FPN	91.23 ± 0.28	91.96 ± 0.42	91.22 ± 0.33	91.54 ± 0.07	83.27 ± 0.08	85.26 ± 0.44
M4	Efficient-B4-FPN	94.19± 0.19	92.78 ± 0.09	92.97 ± 0.01	94.01 ± 0.10	86.11 ± 0.02	89.17 ± 0.08
M31	ResNet-50-FPN-GCN-Relation[34]	91.74 ± 0.29	92.16 ± 0.32	93.23 ± 0.25	92.26 ± 0.12	86.14 ± 0.11	86.54 ± 0.13
M41	Efficient-B4-FPN-GCN-Relation[34]	95.17 ± 0.11	92.29 ± 0.18	94.01 ± 0.12	94.83 ± 0.21	86.77 ± 0.14	90.21 ± 0.19
M32	ResNet-50-FPN-GRU[9]	91.66 ± 0.13	92.06 ± 0.34	94.26 ± 0.26	93.13 ± 0.12	86.01 ± 0.21	87.11 ± 0.98
M42	Efficient-B4-FPN-GCN[9]	94.65 ± 0.32	92.94 ± 0.18	95.19 ± 0.13	94.88 ± 0.23	86.98 ± 0.12	91.03 ± 0.24
O1	本方法 ResNet-50-FPN	94.67 ± 0.11	94.22 ± 0.33	96.25 ± 0.23	97.22 ± 0.43	90.29 ± 0.22	90.01± 0.10
O2	本方法 Our-Efficient-B4-FPN	97.13 ± 0.19	95.51 ± 0.10	97.11 ± 0.16	98.10 ± 0.34	92.44 ± 0.31	91.78 ± 0.27
O2*	本方法 Efficient-B4-FPN*	**98.41 ± 0.10**	**96.32 ± 0.11**	**97.91 ± 0.21**	**98.91 ± 0.12**	**93.24 ± 0.10**	**92.77 ± 0.14**

具体而言，A1 和 A2 是手工分析特征的经典属性学习方法，其准确率分别为 86.14%、87.23%，性能与目前最先进的疾病分类模型 S1、S2 相比略差。而本方法与 S1、S2 相比获得了更好的性能，准确率分别提高了 4.01%、3.70%。此外，原有的 ResNet-50 网络分类性能为 91.74% (M31) 和 91.66%，而本方法在基于 ResNet 的骨干网络上加入属性建模方法，将准确率提升了 3.44%。

3.2.3 消融实验

同理，本实验通过消融研究证明了方法框架中所有模块的有效性，结果如表 4 所示，这里本实验采用 EfficientNet-B4 作为主干网络。可以看到，移除 BN-2、CNA-RES、交叉训练等模块准确率分别会下降 0.35%、0.92%、0.6%。而移除 CNA-RES 和 GrandBN 的影响较大，模型性能会下降 4.68%。可以证明，所有所考虑的组件以及组件间的耦合都对最终性能有积极的影响。

表 4 在 TB-Xatt 数据集上的分类方法的消融研究

GCN	BN-2	CNA-RES	SEatt	BN-1	GradBN	AlterTrain	Relation	准确率	灵敏度	特异性	AUC	精度	F-指数
✓	✓	✓	✓	✓	✓	✓	×	**97.13 ± 0.19**	**95.51 ± 0.10**	**97.11 ± 0.16**	**98.10 ± 0.34**	**92.44 ± 0.31**	**91.78 ± 0.27**
✓	×	✓	✓	✓	✓	✓	×	96.78 ± 0.11	94.26 ± 0.13	96.78 ± 0.23	97.12 ± 0.11	90.93 ± 0.17	90.03 ± 0.07
✓	✓	×	✓	✓	✓	✓	×	96.21 ± 0.44	94.12 ± 0.31	96.22 ± 0.41	97.17 ± 0.41	90.61 ± 0.20	90.01 ± 0.20
×	×	×	×	✓	✓	✓	×	95.69 ± 0.12	93.58 ±0.17	95.44 ± 0.32	96.71 ± 0.21	90.01 ± 0.10	86.96 ± 0.24
✓	✓	✓	×	✓	✓	✓	×	96.89 ± 0.32	94.47 ± 0.55	96.79 ± 0.42	97.87 ± 0.31	91.02 ± 0.33	90.64 ± 0.25
✓	✓	×	✓	×	✓	✓	×	95.69 ± 0.14	94.55 ± 0.43	94.78 ± 0.11	95.94 ± 0.12	87.11 ± 0.23	91.21 ± 0.54
✓	✓	×	✓	×	✓	✓	×	92.41 ± 0.14	88.14 ± 0.22	93.74 ± 0.15	94.91 ± 0.26	92.11 ± 0.09	86.88 ± 0.12
✓	✓	✓	✓	✓	×	✓	×	96.12 ± 0.24	94.66 ± 0.31	94.51 ± 0.23	95.17± 0.21	90.14 ± 0.32	90.11 ± 0.23
✓	✓	✓	✓	✓	✓	×	×	96.53 ± 0.13	94.71 ± 0.24	94.62 ± 0.33	95.19 ± 0.47	90.28 ± 0.34	90.51 ± 0.33
✓	×	×	×	×	×	×	✓	94.67 ± 0.21	93.09 ± 0.08	93.46 ± 0.29	94.93 ± 0.21	86.77 ± 0.14	90.21 ± 0.19

此外，为证明本方法在数据集较少的情况下仍具有有效性，实验中对数据集进行了抽样，通过将越来越小的 TB-Xatt 数据集的子集作为训练集，同时保持验证和测试集的大小不变。抽样结果如表 5 所示，其中 P1 表示所有原始训练样本，P2、P3 和 P4 分别表示原始训练样本的 75%、50% 和 25% 。

表 5 用于性能比较的数据集规范

数据集	P1	P2	P3	P4
训练样本	11360	8530	5680	2840
验证样本	1420	1420	1420	1420
测试样本	1420	1420	1420	1420
训练过程 P/N 值	1/8	1/8	1/8	1/8
验证过程 P/N 值	1/3	1/3	1/3	1/3
测试过程 P/N 值	1/3	1/3	1/3	1/3

分层抽样确保在所有设置下阳性和阴性样本数量之间的比例相同。这个比率在训练集中为 1/8，在验证集和测试集为 1/3。所有结果如表 6 所示，这

表明在训练样本数量减少的情况下，所提的方法仍然优于基准方法。

表 6　使用越来越小的数据集训练的结核病诊断模型的性能

	2* 方法	评价指标					
		准确率	灵敏度	特异性	AUC	精度	F-指数
P1	Low-Level-Feature[3]	86.14 ± 0.14	83.11 ± 0.26	89.17 ± 0.10	90.11 ± 0.76	81.10 ± 0.12	81.71 ±0.11
	Basic-visual-Feature[4]	87.23 ± 0.12	84.16 ± 0.27	92.00 ± 0.09	91.27 ± 0.63	83.26 ±0.71	83.44 ± 0.17
	Efficient-B4-FPN-GCN-Relation[34]	94.97 ± 0.21	93.09 ± 0.08	93.46 ± 0.29	94.93 ± 0.21	86.77 ±0.14	90.21 ± 0.19
	Efficient-B4-FPN	94.19 ± 0.19	92.78 ± 0.09	92.97 ± 0.01	94.01 ± 0.10	86.11 ±0.02	89.17 ± 0.08
	本方法 Efficient-B4-FPN	**97.13 ± 0.19**	**95.51 ± 0.10**	**97.11 ± 0.16**	**98.10 ± 0.34**	**92.44 ±0.31**	**91.78 ± 0.27**
P2	Low-Level-Feature[3]	84.23 ± 0.12	80.07 ± 0.04	86.22 ± 0.20	86.78 ± 0.15	78.24 ± 0.09	78.16 ±0.27
	Basic-visual-Feature[4]	83.45 ± 0.22	81.07 ± 0.54	89.21 ± 0.12	87.13 ± 0.01	80.55 ±0.44	79.21 ± 0.03
	Efficient-B4-FPN-GCN-Relation[34]	91.26 ± 0.34	91.11 ± 0.25	90.24 ± 0.27	91.45 ± 0.67	84.12 ±0.03	87.64 ± 0.11
	Efficient-B4-FPN	91.21 ± 0.14	90.16 ± 0.22	89.98 ± 0.34	92.12 ± 0.22	82.56 ±0.13	86.26 ± 0.17
	本方法 Efficient-B4-FPN	**93.21 ± 0.44**	**91.02 ± 0.33**	**91.20 ± 0.21**	**93.56 ± 0.17**	**87.65 ±0.12**	**89.43 ± 0.14**
P3	Low-Level-Feature[3]	80.11 ± 0.09	75.67 ± 0.12	80.32 ± 0.55	81.26 ± 0.64	72.13 ± 0.07	72.51 ±0.62
	Basic-visual-Feature[4]	79.14 ± 0.24	78.27 ± 0.63	82.11 ± 0.10	80.65 ± 0.44	76.17 ±0.23	74.56 ± 0.35
	Efficient-B4-FPN-GCN-Relation[34]	87.24 ± 0.12	86.10 ± 0.65	87.21 ± 0.46	87.13 ± 0.55	81.21 ±0.46	83.27 ± 0.35
	Efficient-B4-FPN	86.25 ± 0.22	84.23 ± 0.27	84.22 ± 0.54	82.45 ±0.32	79.5 ± 0.21	80.24 ± 0.55
	本方法 Efficient-B4-FPN	**92.44 ± 0.65**	**90.01 ± 0.21**	**89.26 ± 0.17**	**89.22 ± 0.10**	**86.14 ±0.25**	**85.23 ± 0.17**
P4	Low-Level-Feature[3]	76.21 ± 0.09	79.29 ± 0.31	74.32 ± 0.45	77.26 ± 0.13	66.45 ± 0.16	67.84 ±0.17
	Basic-visual-Feature[4]	75.67 ± 0.54	77.15 ± 0.13	76.55 ± 0.43	76.21 ± 0.22	72.56 ±0.34	71.27 ± 0.21
	Efficient-B4-FPN-GCN-Relation[34]	80.11 ± 0.01	78.20 ± 0.15	81.31 ± 0.11	80.13 ± 0.24	75.56 ±0.31	76.27 ± 0.45
	Efficient-B4-FPN	78.11 ± 0.12	79.26 ± 0.36	78.23 ± 0.34	80.15 ± 0.24	76.27 ±0.15	73.62 ± 0.14
	本方法 Efficient-B4-FPN	**86.98 ± 0.27**	**87.23 ± 0.19**	**87.63 ± 0.24**	**88.11 ± 0.12**	**85.23 ±0.11**	**80.16 ± 0.47**

本实验进一步研究了来自 BN-1 和 GCN 模块的单个输出信号的分类性能，以显示当训练数据的大小变化时，这两种机制如何协同工作，如表 7 所示。可以看到 (1) 当训练数据集较小时，BN-1 的性能优于 GCN；(2) 当训练数据集较大时，GCN 的性能优于 BN-1。

表 7　在 TB-Xatt 数据集上 BN-1 和 GCN 的输入和输出信号的分类精度

方法	P1	P4
BN-1 输入	91.04	78.22
BN-1 输出	93.10	85.44
GCN 输入	92.12	78.23
残差融合 (GCN 和 BN-1)	97.07	86.10

针对上述现象的原因进行分析，当训练数据集较小时，BN-1 能够对属性与疾病之间的因果关系进行建模，以缓解数据不足的问题；而当训练数据集较大时，GCN 有能力从大规模注释数据集中学习强大的特征表示。此外，残差融合方案总是比单个网络 (BN-1 和 GCN) 获得更高的精度。显然，本模型的强泛化能力是由神经概率推理方法之间的互补性所实现的。

4　总结与展望

4.1　总结

本研究提出一种基于属性的混合神经概率推断方法用于医学图像诊断，该方法将神经网络模型和概率推断算法的优势相结合——前者擅长从大量数据中准确提取有效信息，后者擅长从少量数据中寻找因果关系并进行概率推断，从而训练得到一个泛化能力强、验证效果好、具有可解释性的学习模型。

具体而言，首先利用深度神经网络从输入图像中提取其显著特征，其次用贝叶斯神经网络（BN）和图卷积神经网络（GCN）来并行地对特征进行处理，从而建立属性和疾病的因果推理模型，并用无向图来表示属性与属性之间、属性与疾病之间的关系，最后用交叉注意力机制将二者的分类结果进行融合，从而得到准确的结果。

本方法在两个具有代表性的医学图像诊断任

务上均表现出了最佳的结果。第一个任务是对 LIDC-IDRI 基准数据集中的 CT 图像进行肺结节的良恶性分类，准确度达到 95.36%；第二个任务是利用胸部 x 射线图像诊断结核病, 准确度为 96.64%。此外，采用消融实验证明了每个模块的有效性，同时表明在训练数据有限的情况下，混合神经概率推断算法比纯神经网络结构具有更好的泛化性能。

4.2 展望

本方法从医生的诊断过程入手，建立病例特征和疾病检测结果的因果推理模型，以提升方法的可验证性、可泛化性和可解释性。实验数据证明了方法的有效性，但是并未对方法的详细推理过程进行可视化解释分析。在医学图像诊断领域，医生和研究人员不仅需要知道模型的最终诊断结果，更需要理解模型是如何通过观察图像特征和属性来做出这一诊断的。缺乏这种可视化解释可能会限制医生对诊断方法的信任度，进而影响模型在临床实践中的应用。未来可能的研究可以集中在模型可解释性的验证、泛化性的证明等，包括开发更先进的可视化工具，以及通过临床反馈来验证和改进模型的解释性等。

参考文献

[1] KUO W, HäneChristian, MUKHERJEE P, идр. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning[J]. Proceedings of the National Academy of Sciences, 2019, 116(45):22737-22745.

[2] AKATA Z, PERRONNIN F, HARCHAOUI Z, et al. Label-embedding for attribute-based classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2013: 819-826.

[3] FERRARI V, ZISSERMAN A. Learning visual attributes[J]. Advances in neural information processing systems, 2007, 20.

[4] KUMAR N, BERG A C, BELHUMEUR P N, et al. Attribute and simile classifiers for face verification[C]//2009 IEEE 12th international conference on computer vision. [S.l.]: IEEE, 2009: 365-372.

[5] LAMPERT C H, NICKISCH H, HARMELING S. Attribute-based classification for zero-shot visual object categorization[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(3):453-465.

[6] LIANG K, CHANG H, MA B, et al. Unifying visual attribute learning with object recognition in a multiplicative framework[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(7):1747-1760.

[7] LIANG K, GUO Y, CHANG H, et al. Incomplete attribute learning with auxiliary labels.[C]//IJCAI. [S.l.: s.n.], 2017: 2252-2258.

[8] MIN W, MEI S, LIU L, et al. Multi-task deep relative attribute learning for visual urban perception[J]. IEEE Transactions on Image Processing, 2019, 29:657-669.

[9] MENG Z, ADLURU N, KIM H J, et al. Efficient relative attribute learning using graph neural networks[C]//Proceedings of the European conference on computer vision (ECCV). [S.l.: s.n.], 2018: 552-567.

[10] WIT E, HEUVEL E V D, ROMEIJN J W. ‘all models are wrong...’: an introduction to model uncertainty[J]. Statistica Neerlandica, 2012, 66(3):217-236.

[11] PEARL J. Bayesian networks[M]//The Handbook of Brain Theory and Neural Networks. Cambridge: MIT Press, 1998.

[12] GAL Y, GHAHRAMANI Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//international conference on machine learning. [S.l.]: PMLR, 2016: 1050-1059.

[13] KINGMA D P, SALIMANS T, WELLING M. Variational dropout and the local reparameterization trick[J]. Advances in neural information processing systems, 2015, 28.

[14] KIPF T F, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[15] ROHEKAR R Y, NISIMOV S, GURWICZ Y, et al. Constructing deep neural networks by bayesian network structure learning[J]. Advances in Neural Information Processing Systems, 2018, 31.

[16] SHI Z, ZENG G, ZHANG L, et al. Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images[C]//Medical Image Computing and Computer

Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. [S.l.]: Springer, 2018: 569-577.

[17] SHAVLIK J W. Combining symbolic and neural learning[J]. Machine Learning, 1994, 14:321-331.

[18] SHAVLIK J W, MOONEY R J, TOWELL G G. Symbolic and neural learning algorithms: An experimental comparison[J]. Machine learning, 1991, 6:111-143.

[19] YI K, WU J, GAN C, et al. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding[J]. Advances in neural information processing systems, 2018, 31.

[20] LI Q, HUANG S, HONG Y, et al. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2020: 5884-5894.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 770-778.

[22] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. [S.l.]: PMLR, 2019: 6105-6114.

[23] LI G, MULLER M, THABET A, et al. Deepgcns: Can gcns go as deep as cnns?[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2019: 9267-9276.

[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 7132-7141.

[25] SHEN W, ZHOU M, YANG F, et al. Multi-scale convolutional neural networks for lung nodule classification[C]//Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24. [S.l.]: Springer, 2015: 588-599.

[26] HUSSEIN S, CAO K, SONG Q, et al. Risk stratification of lung nodules using 3d cnn-based multi-task learning[C]//Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25. [S.l.]: Springer, 2017: 249-260.

[27] HAN F, WANG H, ZHANG G, et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules[J]. Journal of digital imaging, 2015, 28:99-115.

[28] DHARA A K, MUKHOPADHYAY S, DUTTA A, et al. A combination of shape and texture features for classification of pulmonary nodules in lung ct images[J]. Journal of digital imaging, 2016, 29:466-475.

[29] XIE Y, ZHANG J, XIA Y, et al. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct[J]. Information Fusion, 2018, 42:102-110.

[30] XIE Y, XIA Y, ZHANG J, et al. Transferable multi-model ensemble for benign-malignant lung nodule classification on chest ct[C]//Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. [S.l.]: Springer, 2017: 656-664.

[31] XIE Y, XIA Y, ZHANG J, et al. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct[J]. IEEE transactions on medical imaging, 2018, 38(4):991-1004.

[32] XIE Y, ZHANG J, XIA Y. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest ct[J]. Medical image analysis, 2019, 57:237-248.

[33] XU X, WANG C, GUO J, et al. Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks[J]. Medical Image Analysis, 2020, 65:101772.

[34] SANTORO A, RAPOSO D, BARRETT D G, et al. A simple neural network module for relational reasoning[J]. Advances in neural information processing systems, 2017, 30.

[35] GUAN Q, HUANG Y, ZHONG Z, et al. Thorax disease classification with attention guided convolutional neural network[J]. Pattern Recognition Letters, 2020, 131:38-45.

[36] MENG Q, SHIN'ICHI S. Adinet: Attribute driven incremental network for retinal image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2020: 4033-4042.