Recognizing Non-Small Cell Lung Cancer Subtypes by a Constraint-Based Causal Network from CT Images

Zhengqiao Deng[1], Shuang Qian[1], Jing Qi[2], Li Liu(✉)[1], and Bo Xu[2,3]

[1] School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China
{20161630,202024131068,dcsliuli}@cqu.edu.cn
[2] Department of Biochemistry and Molecular Biology, Tianjin Medical University Cancer Institute and Hospital, Key Laboratory of Cancer Prevention and Therapy, National Clinical Research Center for Cancer, Tianjin's Clinical Research Center for Cancer, Tianjin 300060, China
[3] Center for Intelligent Oncology, Chongqing University Cancer Hospital, Chongqing University School of Medicine, Chongqing 401331, China

Abstract. The primary goal of non-small cell lung cancer (NSCLC) recognition from CT images is to discover representative features, with each being responsible for NSCLC diagnosis. A key challenge in CT image feature selection is the fact that rich causal dependencies are often neglected among either radiomics or deep learning-based features. This leads us to present a constraint-based model to construct a causal network that explicitly discovers and leverages the inherent local causal variability of these deep and radiomics features under a global view. In particular, an identified network skeleton is generated to characterize a unique causal configuration of a particular NSCLC subtype as a variable number of nodes and links, and as a result, the resulting causal network satisfies the causal Markov property and all local cause-effect dependencies are globally consistent. Furthermore, a representative node selector is devised to select the most representative causal features from the causal network for NSCLC subtype recognition. Empirical evaluations on one benchmark dataset and one in-house dataset suggest our model significantly outperforms the state-of-the-art methods.

Keywords: Non-small cell lung cancer recognition · Constraint-based network · Cause-effect dependency · Feature selection

1 Introduction

Non-small cell lung cancer (NSCLC), a leading cause of cancer deaths all over the world, has different characteristics such as adenocarcinoma (ADC) and squamous cell carcinoma (SCC), and thus its subtype recognition has become an important research field, given its role in guiding the subsequent treatment for patients with lung cancer. The golden standard for NSCLC diagnosis is pathological diagnosis,

which has not yet been fully elucidated and are commonly expensive and time-consuming. Although experienced doctors can make an initial diagnosis from radiographic data, there are still urgent needs for data-driven models that can detect different subtypes from CT images. Currently, these models can be divided into two categories: conventional models which are required to manually encode radiological features, and deep models which can automatically discover features from images. However, these image data generally has the properties of high-dimensional but small samples, which may bring about degradation in accuracy and efficiency of recognition model by curse of dimensionality and overfitting [8]. In addition, most features of CT images are unrelated to NSCLC subtypes and have no effect on their diagnosis or even have negative impacts. Therefore, feature selection is especially significant for the recognition of NSCLC subtypes [9].

Current techniques are becoming mature to select features. Here, a review [17] reports a repository of near 40 representative feature selection algorithms, which have been used in the field of radiomics, such as LASSO [2], PCA [2], RFE [27], mutual information [23] and other deep-based features. It is worth noting that these approaches commonly assume that features are independent without consideration of their causal relationships [10]. However, rich causal relations exist among radiological features in CT images with their unique values in cancer diagnosis [7, 18]. For example, as illustrated in Fig. 1, it is known that *pleural tag* or *air bronchogram* is the cause of the NSCLC subtype *adenocarcinoma* in CT image-based detection. In fact, most of the existing data-driven models may find that there is a heavy correlation between *pleural tag* and *air bronchogram* but unfortunately cannot discover the further interpretation that the *adenocarcinoma* is the common effect of these two symptoms, which leads to their extrinsic association. As a result, it could be rather difficult to determine the significant factors, which is extremely important to NSCLC diagnosis because a wrong release of a patient can have bad consequences. The main focus of this paper is on causal discovery of features in CT images, since an important assumption for lung cancer diagnosis is the causal relationships between the radiological imaging data and cancer types [4].

Despite being a very challenging problem, there has been a rapid growth of interest in selecting causal features in recent years. The most popular modeling paradigm might be that of the graphical causal modeling, where the Bayesian network is the most commonly used structure in causal discovery that calculates the relationships of all features by constructing nodes (e.g., imaging features or NSCLC subtypes) and edges (i.e. their causal relations) as well as their joint probability distributions under certain constraints, often entailed by Markov blanket property. In a typical manner, they can be divided into two main categories, *score-based* models which maximize a score criterion to learn a causal network, and *constraint-based* models which use conditional independence and dependence constraints to discovery causal structures of observed variables [11]. However, imaging features are often not causally sufficient (i.e. there exist unobserved causes for two observed variables [26]). If without assuming causal sufficiency, score-based models must require a predeterminate number of such latent

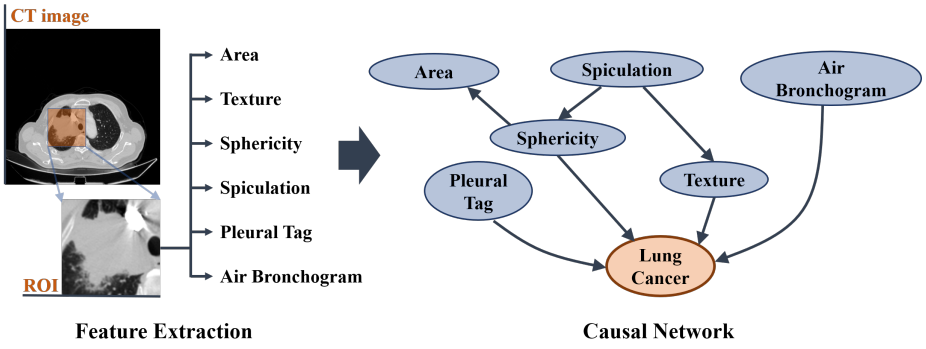


Fig. 1. An example of a causal network representing lung cancer and features of CT images.

variables, which is almost unavailable in CT images. In contrast, the constraint-based algorithms are more capable of handling the awkward situation of the lack of such prior knowledge. Moreover, constraint-based algorithms are more computationally efficient than score-based methods, which are NP-hard in terms of complexity [6].

In this work, we adopt a constraint-based model to discover causal features in CT images for NSCLC recognition. Normally, most of such models construct causal structures under Markov blanket property among features by leveraging exhaustively global search to learn from a complete graphical network. Unfortunately, the main challenge is their computational cost as the number of such relations and possible causal networks is super-exponential to the number of variables (i.e., node) [11]. In fact, checking Markov blanket consistency becomes intractable with the growth of network size. Besides, since Bayesian network structure is a directed acyclic graph, some highly correlated relations has to be removed from the the network in order to maintain causal consistency, which nevertheless would result in information loss. For example, *air bronchogram* is the most relevant feature of *adenocarcinoma*, which is important in NSCLC diagnosis, but it cannot be identified as a causal variable by existing constraint-based approaches due to the Markov blanket inconsistency. Subsequently, these approaches are rather limited in identifying feature variables with meaningful cause-effect relationships between them in the network [5].

To address these aforementioned issues, we present a constraint-based causal network model for NSCLC subtypes recognition. Specifically, our model considers a principled way of discovering and applying causal relations of CT imaging features associated with NSCLC subtypes. In short, by generating from an identified network skeleton with highly correlated nodes (features) under the constraint of causal Markov property, a causal network is constructed to discover representative causal features from CT images. Note that the node set in the identified network is composed of both deep-based features and representative

radiomics features. Now each resulting causal network contains its unique set of directed links that represent cause-effect relations, together with other commonly used radiomics features such as shape, texture and statistical information of the tumor lesion. Moreover, we design a *representative node selector* to choose the most representative causal features from the causal network for NSCLC subtype recognition. In this way, our causal network-based method is more capable of characterizing the inherit cause-effect dependencies of CT imaging features in a non-invasive NSCLC diagnosis when compared to existing approaches, which is also verified during empirical evaluations to be detailed in later sections.

2 Related Work

Existing approaches for NSCLC recognition can be divided into two categories.

2.1 Conventional Models with Radiomics Features

It is commonly known that conventional models can be adopted for diagnosis of lung cancer with radiomics features, which include high-throughput quantitative metrics from medical images related to tumor pathobiology and the creation of minable high dimensional database [13, 14, 19]. Since radiomics data contains a large number of features describe intensity distribution, spatial relationships between the various intensity levels, texture heterogeneity patterns, shape and the relations of the tumour with the surrounding tissues, it is necessary to apply feature selection to eliminate redundant features that are not relevant to the label [15]. Wang et al. [22] compared the performance of several conventional machine learning methods in predicting the prognostic recurrence of NSCLC using PCA to select features. Zhu et al. [28] successfully performed a radiomics analysis with LASSO logistic model to distinguish ADCs from SCCs. Han et al. [12] evaluated ten feature selection techniques as well as ten conventional models for NSCLC classification. These studies demonstrated that the manually encoded radiological features are capable of characterizing properties as potential biomarkers for recognizing NSCLC subtypes. However, such features are normally hand-coded or defined based on domain knowledge, which would be not practicable since tumor pathobiology in NSCLC is not completely elucidated and its corresponding radiomics images are often intricate. Moreover, the causal relationships between features cannot be exploited by these conventional methods, leading to the significant information loss.

2.2 Deep Model-Based Recognition from Raw CT Images

Different from conventional models, deep models have been at the forefront of this research field, which can automatically quantify radiographic characteristics of tumor and its surroundings without handcraft from CT images for NSCLC diagnosis. Among them, CNNs have achieved excellent performance, and not surprisingly an increasing amount of CNN-based variants are presented in CT

image-based classification. These algorithms provide a great aid in the diagnosis of lung cancer, including segmentation of the lung and tumor area, prediction of invasiveness and survival analysis, etc. [20, 21, 24]. Aonpong et al. [3] found that the deep models such as ResNet can achieve better result in the diagnosis of NSCLC compared to conventional models using radiomics features. However, deep models only take raw CT images as input, which are limited to manage radiomics information such as texture, shape and density. Several recent studies attempted to merge such radiomics features in deep models instead of merely using CT images. Han et al. [12] designed a fusion algorithm that can combine radiomic and deep-based features to help radiologists to differentiate the subtypes of NSCLC via PET/CT images. Aonpong et al. [2] embedded selected radiomics features in a deep network for recurrence prediction of NSCLC. It was shown that these fusion models can achieve better performance than deep models that use raw images alone. However, a major limitation of these deep models concerns that the relationships that are learned from raw images are often hard to understand by human beings, which are extremely crucial in cancer diagnosis. In addition, they usually lack the expressive power to characterize and propagate rich causal dependencies in NSCLC recognition, and thus they are limited to capture the inherent causal variability of radiological image features in a global view. To address the issues, we present a constraint-based causal network to discover and utilize the cause-effect relations of both radiomics features and deep features extracted from CT images to discriminate NSCLC subtypes.

3 Preliminaries

3.1 Data Acquisition

In this study two NSCLC datasets are considered, including one in-house dataset and one publicly-available benchmark dataset.

A public dataset named NSCLC-Radiomics-Lung (P-NSCLC) [1], which includes 422 NSCLC patients is used in this study. It contains CT images with manually segmented gross tumor volumes and contour annotations for each patient. A number of 203 patients were eventually selected in our study (51 patients were diagnosed with ADC, and 152 patients were diagnosed with SCC).

To our best knowledge, the above mentioned dataset is so far the only ones publicly available for the field of NSCLC recognition. To this end, we propose a new NSCLC dataset (named I-NSCLC) collected from a hospital between May 2018 and September 2019, which includes 466 NSCLC patients' computed tomography scans. The inclusion criteria were as follows: (1) Patients were diagnosed with a primary NSCLC subtype, i.e., ADC or SCC; (2) Patients received no treatment before pathological diagnosis; (3) Patients were considered with available thoracic enhanced CT images. In the entire dataset, 368 patients were diagnosed with ADC, and 98 patients were diagnosed with SCC. The collected CT images are consecutive thoracic series in *digital imaging and communications in medicine* (DICOM) format. The corresponding CT system we used in this study is a 64-channel multi-detector CT scanning system (64-slice LightSpeed

VCT, GE Medical Systems, Milwaukee, WI, USA), with the same scanning parameters (120kV; 400mAs; detector coverage: 40mm; rotation time: 0.6s; matrix size: 512×512). A subset of samples are provided in the supplementary material, and once ready we plan to share the entire dataset in the community.

3.2 CT Image Preprocessing

All the CT images of the enrolled patients in both the in-house and public-available NSCLC dataset were manually annotated by an experienced radiologist using ITK-snap software using a standard clinical delineation protocol. Each single CT image was checked for delineating the corresponding tumor solid lesions. Then regions of interest (ROI) of those patients were stored separately as the mask information of the original CT image.

For each raw CT image, an intensity normalization was applied to rescale the pixel intensity to $[0, 255]$. Combined with the mask information, each CT image with tumor lesion was center cropped to 128×128 pixels. Radiomics features were extracted from the ROIs of the CT images. These features were divided into the following seven categories: 19 first order statistics features, 16 shape-based (3D) features, 10 shaped-based (2d) features, 34 gray level co-occurrence matrix (GLCM) features, 16 gray level run length matrix (GLRLM) features, 16 gray level size zone matrix (GLSZM) features and 14 gray level dependence matrix (GLDM) features. All the features were extracted from the raw CT images except for shape that are independent of gray value. For each raw image, a set of filters were leveraged to generate derived images, which could also be used for feature extraction. In particular, most of the features are consistent with the standard definitions as described by the *Imaging Biomarker Standardization Initiative* (IBSI) [29].

3.3 Problem Formulation

Given a NSCLC dataset $\mathcal{D}$ of N samples from a set of C subtypes, with each sample representing a patient consisting of a series of CT images of size $C \times W \times H$ (indicating channels, width and height, respectively) and a mask of tumor lesion. Each sample consists of M features extracted from patients' CT images and a corresponding NSCLC subtype. A causal network $\mathbf{G} = (\mathbf{X}, \mathbf{E})$ indicates the causal dependencies of a variable set (including NSCLC feature variable and label variable), where a node in $\mathbf{X}$ represents a variable and a link in $\mathbf{E}$ between any two nodes represents their causal relationship [25]. There are two types of link in $\mathbf{E}$: directed links ($\rightarrow$) and undirected links ($\leftrightarrow$) (can also be seen as two-way links). A directed link ($\rightarrow$) in $\mathbf{E}$ describes that the head node x_i is a direct cause of the tail one, denoted by $x_i \rightarrow x_j$, where $x_i, x_j \in \mathbf{X}$. An undirected link $x_i \leftrightarrow x_j$ means that there exist two Markov equivalence class of $\mathbf{G}$ containing $x_i \rightarrow x_j$ and $x_j \rightarrow x_i$ respectively, indicating a cause-effect relation between x_i and x_j with uncertain direction.

A causal network is a fully *completed partially directed acyclic graph* (i.e., its directed subgraphs does not contain a directed cycle) because causality is

transitive, irreflexive and anti-symmetric. For any $\{x_i, x_j\} \in \mathbf{V}$, x_i and x_j are conditionally independent if there is a set of variables $\mathbf{X}' \subseteq \mathbf{X} \backslash \{x_i, x_j\}$ over the dataset $\mathcal{D}$ satisfying $P(x_i, x_j \mid \mathbf{X}', \mathcal{D}) = P(x_i \mid \mathbf{X}', \mathcal{D}) P(x_j \mid \mathbf{X}', \mathcal{D})$, denoted by $x_i \perp\!\!\!\perp x_j \mid (\mathbf{X}', \mathcal{D})$.

A causal network $\mathbf{G}$ must satisfy *causal Markov condition*. That is, give a node $x_i \in \mathbf{X}$ and its parents $\mathbf{Pa}(x_i)$, if x_i is not a cause of x_j, then x_i is conditionally independent of x_j given $\mathbf{Pa}(x_i)$, i.e., $x_i \perp\!\!\!\perp x_j \mid \mathbf{Pa}(x_i)$, where $x_j \in \mathbf{V} \backslash \{x_i, \mathbf{Pa}(x_i)\}$. There are three important structures on a triplet $\langle x_i, x_j, x_k \rangle$ in the causal network, *v-structure*, *chain* and *fork*, as illustrated in Table. 1. The *d-separation* criterion captures exactly the conditional independence relationships that are implied by the Markov condition. Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be disjoint subsets of the nodes of $\mathbf{X}$. $\mathcal{P}$ is a acyclic path between node x_i and x_j, where $x_i \in \mathbf{A}$, $x_j \in \mathbf{B}$. We say $\mathcal{P}$ is *blocked* by the subset $\mathbf{C}$ if and only if (1) there is a *chain* $x_i \rightarrow x_k \rightarrow x_j$ or a *fork* $x_i \leftarrow x_k \rightarrow x_j$ such that $x_k \in \mathbf{C}$; or (2) $\mathcal{P}$ contains a *v-structure* $x_i \rightarrow x_k \leftarrow x_j$ and neither x_k nor any of its descendants are in $\mathbf{C}$. The causal networks constructed according to the conditional independence determined by *d-separation* are not unique, they are Markov equivalence classes of the real Bayesian network with the same skeleton and *v-structures*.

Table 1. Orientation rules in *d-separation*. Adj(nonAdj) means two nodes do(not) have a link

Structure	Probability	Definition
v-structure $x_i \rightarrow x_k \leftarrow x_j$	$P(x_k \mid x_i, x_j) =$ $P(x_k \mid x_i) P(x_k \mid x_j)$	x_k is a common effect of x_i and x_j. x_k is called a *collider*
chain $x_i \rightarrow x_k \rightarrow x_j$	$P(x_k \mid x_i, x_j)$ $= P(x_k \mid x_i)$	x_i is an indirected cause of x_k
fork $x_i \leftarrow x_k \rightarrow x_j$	$P(x_j \mid x_k, x_i)$ $= P(x_j \mid x_i)$	x_k is a common cause of x_i and x_j

Generally, the constraint-based algorithms consist of two key steps to determine a causal network: (1) identifying the network skeleton (the candidate nodes and links); (2) orienting links as many as possible. Notice that we assume the node set $\mathbf{X}$ (observed variables) is causal sufficient, i.e., all the relevant features in the network have been observed and there is no unobserved common cause. A causal network characterizes the relationships between variable features and labels. This inspires us to present in what follows a constraint-based method where these networks can be systematically discovered to construct the final causal network characterizes the causal relationships among various radiomics and deep features for the NSCLC subtypes.

4 Our Model

To generate a causal network for histologic subtypes of NSCLC, two types of features (nodes) are considered in our model. That is, deep features are learned from a deep model while representative radiomics are selected as the nodes of the skeleton. A causal network is then generated by causal link orientation. Finally, the most important causal features are selected for the classification of NSCLC histologic subtypes. The main procedure of our approach is illustrated in Fig. 2.

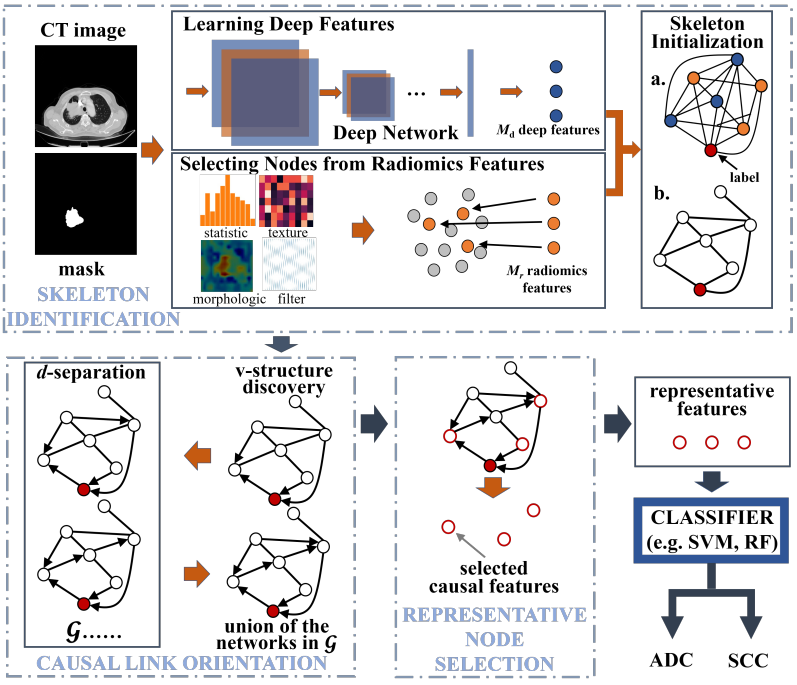


Fig. 2. The framework of our approach.

4.1 Skeleton Identification

To discover causal relationships between features using a constraint-based approach, a network skeleton consist of candidate nodes and links of the final causal network needs to be extracted first.

Learning Deep Features. A homogeneous deep model is constructed to extract deep features related to NSCLC subtypes from CT images automatically.

It consists of two main components: convolution layer and residual block. First, a convolution layer is defined as:

$$\begin{aligned} \text{ConvLayer: } \boldsymbol{Y} = \mathsf{F}_c(\boldsymbol{A}) = \text{ReLu}(\mathbf{W} \cdot \text{BN}_{\gamma,\beta}(\text{Conv}(\boldsymbol{A}))), \\ \text{with } \text{Conv}(\boldsymbol{A}) : \boldsymbol{A} = (\mathbf{a}_1, ..., \mathbf{a}_{in}) \mapsto \boldsymbol{Y} = (\mathbf{y}_1, ..., \mathbf{y}_{out}), \end{aligned} \tag{1}$$

where $\mathbf{W}$ is the corresponding weight vector of the ReLu function, γ and β are the internal parameters in the BatchNorm function. Conv is the convolution function with the sizes of the input channels *in* and output channels *out*. Note that $\mathbf{y}_c = \mathcal{K}_c^s * \boldsymbol{A}$ $(c = 1, ..., out)$, where $\mathcal{K}_c^s$ is the c-th convolution kernel of the size s and $*$ is the convolution operator. Subsequently, a residual block is constructed as follows:

$$\begin{aligned} \text{ResBlk: } \boldsymbol{Y} = \mathsf{F}_r(\boldsymbol{A}) = \boldsymbol{A} + \sum_{i=1}^{Cad} \mathsf{F}_{\mathcal{T}}(\boldsymbol{A}), \\ \text{with } \mathsf{F}_{\mathcal{T}}(\boldsymbol{A}) = \mathsf{F}_c^3(\boldsymbol{A}), \end{aligned} \tag{2}$$

where $\mathsf{F}_{\mathcal{T}}$ is a transformation including 3 *ConvLayers* with the kernel sizes of 1×1, 3×3 and 1×1 consecutively. Cad is the cardinality that is introduced to control the number of complex transformations. In this way, the split-transform-merge strategy is exploited in the block to reduce the number of parameters in an easy and extensive way.

Given a CT image $\mathcal{I} \in \mathbb{R}^{C \times W \times H}$ and its corresponding label c, our deep model aims to discover the deep features by leveraging the following structures:

$$\begin{aligned} \mathcal{I}_1 &= \text{MaxPool}(\mathsf{F}_c(\mathcal{I})), \\ \mathcal{I}_2 &= \mathsf{F}_r^4(\mathcal{I}_1), \\ \mathbf{V}_{M_d} &= \mathsf{FC}(\text{GAP}(\mathcal{I}_2)), \\ c &= \text{softmax}(\mathsf{FC}(\mathbf{V}_{M_d})), \end{aligned} \tag{3}$$

where GAP is the global average pooling and FC is a fully connected layer. Finally, $\mathbf{V}_{M_d}$ represents the vector of M_d deep features learned from our model.

Selecting Nodes from Radiomics Features. Now let us consider a radiomics feature set $\mathcal{R} = \{(s_1, c_1), ..., (s_i, c_i), ..., (s_N, c_N)\}$ where s_i is the i-th sample composed of R radiomics features denoted by $[s_{i1}, ..., s_{iR}]$, $c_i \in \{1, ..., C\}$ is its corresponding NSCLC subtype and N is the number of samples. Our goal is to find a weighting vector $\mathbf{w}$ that reflects the importance of each feature to the corresponding label. The weighting vector could be denoted as $\mathbf{w} = [w_1, \ldots, w_R]$, where w_r represent the importance weight of the r-th feature.

Here, the probability P_{ij} of s_i and s_j with the same NSCLC subtype can be defined as follows:

$$p_{ij} = \begin{cases} \frac{\mathcal{K}(\mathsf{d}_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{k \neq i} \mathcal{K}(\mathsf{d}_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_k))}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{4}$$

where $\mathcal{K}(z) = \exp\left(-\frac{z}{\sigma}\right)$ is a kernel function with a kernel width of σ and $\mathsf{d}_{\mathbf{w}}(s_i, s_j) = \sum_{r=1}^{R} w_r^2 \left|s_{ir} - s_{jr}\right|$ is a weighting distance function. To this end, the optimal radiomics feature weight vector $\hat{w}$ over dataset $\mathcal{R}$ can be computed:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\{\mathcal{L}(\mathbf{w}, \mathcal{R})\} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} P_{ij} \left(1 - c_{ij}\right) + \lambda \sum_{r=1}^{R} w_r^2 \right\}, \tag{5}$$

where $\lambda > 0$ is regularization parameter to alleviate overfitting, and $c_{ij} = 1$ when s_i and s_j are of the same class, otherwise $c_{ij} = 0$. A gradient based optimizer can be used to optimize the above objective function, such as delta-bardelta or conjugate gradients. It is worth noting that the larger the weight value in $\hat{\mathbf{w}}$, the greater the importance it has in NSCLC classification task. We select top M_r features with the largest weight values as the skeleton nodes of radiomics features.

Skeleton Initialization. A complete undirected graph $\mathbf{G}_1 = (\mathbf{X}, \mathbf{E}_1)$is constructed to initial the skeleton for causal link orientation, where $\mathbf{X}$ composed of M_d deep features nodes, M_r radiomics features nodes and their corresponding NSCLC subtype label, and $K = |\mathbf{V}| = M_d + M_r + 1$. Then, conditional independence test is performed for each edge $x_i \leftrightarrow x_j$ in the graph. x_i and x_j are not independent given any subset of nodes in $\mathbf{X}$ except x_i and x_j, i.e., $\mathbf{E}_1 = \{x_i \leftrightarrow x_j : \forall \mathbf{U} \subseteq \mathbf{V} \setminus \{x_i, x_j\}, x_i \not\perp\!\!\!\perp x_j \mid \mathbf{U}\}$. In consequence, all links that with two conditionally independent nodes are removed from the graph to satisfy causal Markov condition.

4.2 Causal Link Orientation

Now we have a causal network skeleton $\mathbf{G}_1$ which is an undirected graph without any determined causal relation. We continue to direct the links in $\mathbf{G}_1$. First we construct a network $\mathbf{G}_2 = (\mathbf{X}, \mathbf{E}_2)$ by analyse the *v-structures* in $\mathbf{G}_1$. For any triplets $\langle x_i, x_j, x_k \rangle$ with the structure of $x_i \leftrightarrow x_k \leftrightarrow x_j$ in $\mathbf{E}_1$, if $\mathbf{E}_1$ dose not contain such a link $x_i \leftrightarrow x_j$ and x_k dose not belong to any subset $\mathbf{U} \subseteq \mathbf{V} \setminus \{x_i, x_j\}$ so that $x_i \perp\!\!\!\perp x_j \mid \mathbf{U}$, then we add the link $x_i \rightarrow x_k \leftarrow x_j$ to $\mathbf{E}_2$; otherwise, we still keep the undirected link $x_i \leftrightarrow x_k \leftrightarrow x_j$ in $\mathbf{E}_2$. After analysing all *v-structures* in $\mathbf{E}_1$, we get a partially directed acyclic graph $\mathbf{G}_1$, which satisfies causal Markov condition.

Next, we further orientate other undirected links in $\mathbf{G}_2$ according to the *d-separation* criterion, where if x_i and x_j are *d-separated* by x_k. then x_i and x_j are independent given x_k; otherwise, x_i and x_j are interdependent given x_k. Four orientation rules of the corresponding sub-graph are illustrated in Fig. 3. With these rules, we get a set of different networks $\mathcal{G}$ by orientating specific undirected links. These networks in $\mathcal{G}$ are Markov equivalent due to their same skeleton and same *v-structures*.

The final causal network $\mathbf{G} = (\mathbf{X}, \mathbf{E})$ is the union of the networks in $\mathcal{G}$, where a directed link $x_i \rightarrow x_j$ exists in $\mathbf{E}$ if and only if it exists in every network in $\mathcal{G}$,

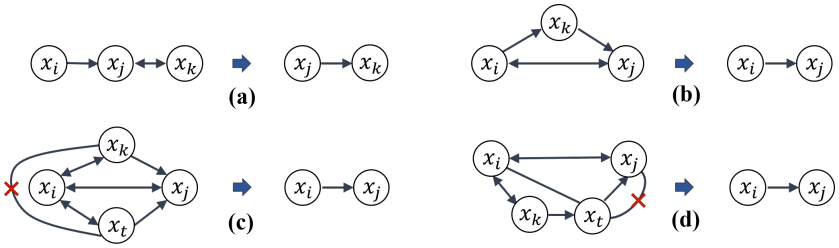


Fig. 3. Four types of orientation rules in *d-separation*. A red cross means there does not exist any link between the two nodes.

otherwise $x_i \leftrightarrow x_j$ remains in $\mathbf{E}$. Note that $\mathbf{G}$ is a completed partially directed acyclic graph representing a Markov equivalence class of $\mathcal{G}$ and thus satisfies causal Markov condition. Besides, the causal probability ξ_{ij} of any two nodes x_i and x_j can be calculated in term of standard deviations:

$$\xi_{ij} = \frac{\sigma^2_{a\mathbf{x}_i + b\mathbf{x}_j} - \sigma^2_{a\mathbf{x}_i - b\mathbf{x}_j}}{\sigma^2_{a\mathbf{x}_i + b\mathbf{x}_j} + \sigma^2_{a\mathbf{x}_i - b\mathbf{x}_j}}, \tag{6}$$

where $\sigma_{\mathbf{X}} = \sqrt{\mathrm{Var}(\mathbf{X})}$.

4.3 Representative Node Selection

To estimate the cause-effect level of each feature node on label node in the causal network $\mathbf{G}$, we introduce a representative node selector that estimates the regression coefficients [16] defined as follows:

$$\mathcal{L}_s\left(\mathbf{w}_s, \lambda_1, \lambda_2\right) = \left\|\mathbf{C} - \mathbf{S}\mathbf{w}_s\right\|_2^2 + \lambda_1 \left\|\mathbf{w}_s\right\|_1 + \lambda_2 \mathbf{w}_s^T \Phi \mathbf{w}_s, \tag{7}$$

where $\mathbf{w}_s = [w_{s1}, \ldots, w_{sK}]$ is a vector, referred as to causal weights, representing the cause-effect level of each feature node in $\mathbf{G}$. $\mathbf{S} \in \mathbb{R}^{N \times K}$ is a sample matrix from dataset $\mathcal{D}$ where its (i, m)-th entry is the m-th feature of i-th sample s_{im} and $\mathbf{C} \in \mathbb{R}^N$ is its corresponding vector of labels of the NSCLC subtypes. $\Phi \in \mathbb{R}^{K \times K}$ is a Laplacian matrix for the network $\mathbf{G}$ with the (i, j)-th element defined by:

$$\Phi_{ij} = \begin{cases} 1/\deg_i, & \text{if } i = j \text{ and } \deg_i \neq 0 \\ -\xi_{ij}/\sqrt{\deg_i \deg_j}, & \text{if } \mathrm{Adj}\left(x_i, x_j\right) \in \mathbf{E} \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $\deg_i = \sum_{\mathrm{Adj}(x_i, x_k) \in \mathbf{E}} \xi_{ik}$ and $\mathrm{Adj}\left(x_i, x_k\right)$ means that there exists a link between the two nodes. The first term in $\mathcal{L}_s$ only seeks to minimize regression errors by regrading each node (variable) individually. The last term considers the coefficient and correlation of two neighboring variables having cause-effect

relations. The tuning parameters λ_1, λ_2 control the amount of regularization for sparsity and smoothness, respectively.

Now we are ready to evaluate the causal weights by minimizing the following function using the sparse Laplacian shrinkage with the graphical Lasso estimator:

$$\widehat{\mathbf{w}}_s = \underset{\mathbf{w}_s}{\operatorname{argmin}} \mathcal{L}_s\left(\mathbf{w}_s, \lambda_1, \lambda_2\right) = \left(\mathbf{C} - \mathbf{S}\mathbf{w}_s\right)^T \left(\mathbf{C} - \mathbf{S}\mathbf{w}_s\right) + \lambda_1 \sum_i |w_{si}| + \lambda_2 \sum_{\mathrm{Adj}(x_i, x_j) \in \mathbf{E}} \left(\frac{w_{si}}{\sqrt{\deg_i}} - \frac{w_{sj}}{\sqrt{\deg_j}} \right)^2 \xi_{ij}. \tag{9}$$

To recognize NSCLC subtypes, we select the features whose causal weight $\widehat{w}_s$ is greater than a threshold θ_1 as well as the features whose important weight $\widehat{w}$ is greater than a threshold θ_2.

5 Experiments

In our experiments, all the competing models for NSCLC histologic subtypes are evaluated over accuracy, sensitivity, specificity and area under the curve (AUC) of receiver operating characteristic (ROC). The accuracy measures the proportion of the correctly classified samples among the total tested samples. The sensitivity and specificity show the ability to correctly identify samples with ADC or with SCC. The AUC is employed to measure the quality of the model's predictions.

5.1 Comparison Results Against Other Competing Models

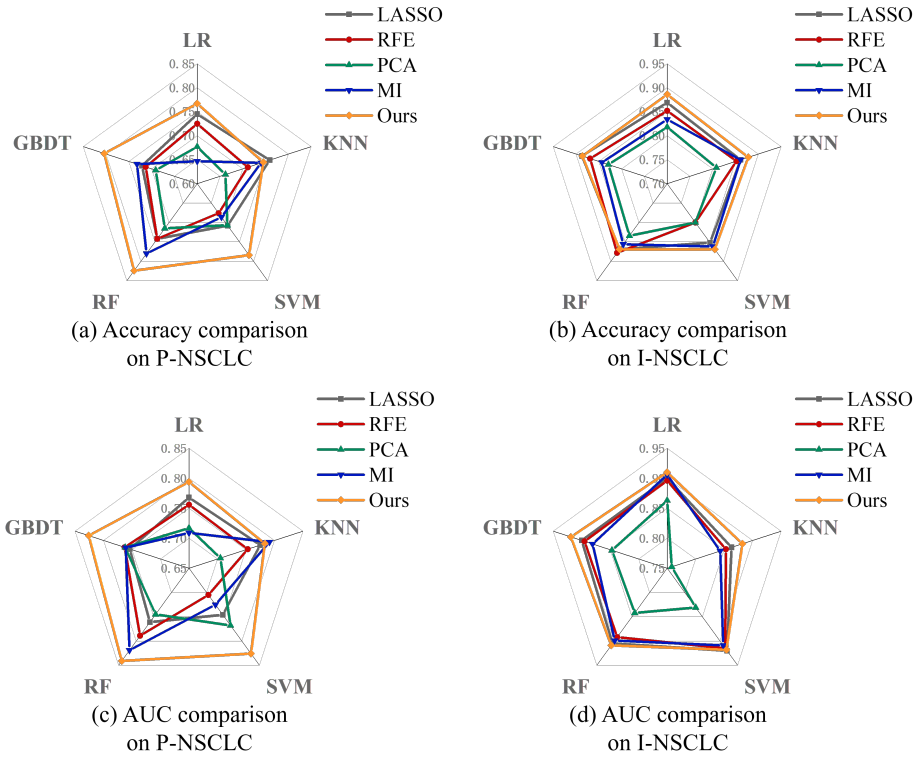
Three types of models for NSCLC classification have been taken into account in this part, i.e., five conventional classifiers merely using radiomics (LR, KNN, SVM, RF and GBDT), two deep models directly learning from raw CT images (VGG16 and ResNet) and two fusion models that combine the radiomics and deep features (ResNet_Fusion [2] and VGG16_Fusion [12]). Table. 2 shows the accuracy, sensitivity, specificity and AUC performance. Our approach clearly outperforms other NSCLC classification methods on both datasets with a more stable performance with around $2\% - 20\%$ boost. This is mainly due to our approach utilized a combination of deep and radiomics features to generate the causal skeleton and subsequently take advantage of the causal-effect dependency information. Also, it is obvious that the two fusion models that merge deep network (for image features) and conventional models (for radiomics features) have better performance than either of them. This might explain why our approach achieves the best performance by a large margin with nearly 20% boost at most compared with LR on I-NSCLC dataset where a large number of features exist but only a few causal features are discovered, which are extremely significant for NSCLC recognition.

Table 2. Performance of different models in NSCLC Classification. The value in the bracket shows the metric change taken our model as a baseline.

Dataset	Method	Metrics (%)			
		Accuracy	Sensitivity	Specificity	AUC
P-NSCLC	**LR**	74.5(-9.1)	76.9(-7.7)	70.6(-12.6)	76.8(-8.9)
	KNN	75.9(-7.7)	78.3(-6.3)	68.6(-14.6)	77.5(-8.2)
	SVM	70.8(-12.8)	72.4(-12.2)	66.7(-16.5)	74.6(-11.1)
	RF	74.2(-9.4)	76.3(-8.3)	68.8(-14.4)	76.1(-9.6)
	GBDT	72.1(-11.5)	73.7(-10.9)	70.6(-12.6)	75.4(-10.3)
	VGG16	76.5(-7.1)	80.3(-4.3)	65.9(-17.3)	78.3(-7.4)
	ResNet	79.4(-4.2)	81.4(-3.2)	73.8(-9.4)	82.5(-3.2)
	VGG16_Fusion	78.8(-4.8)	80.2(-4.4)	75.1(-8.1)	80.2(-5.5)
	ResNet_Fusion	81.7(-1.9)	82.5(-2.1)	79.7(-3.5)	83.3(-2.4)
	Ours	**83.6**	**84.6**	**83.2**	**85.7**
I-NSCLC	**LR**	86.9(-3.5)	70.1(-19.9)	91.3(+0.8)	90.0(-2.1)
	KNN	86.1(-4.3)	84.2(-5.8)	86.4(-4.1)	86.3(-5.8)
	SVM	85.2(-5.2)	82.7(-7.3)	86.0(-4.5)	87.5(-4.6)
	RF	86.9(-3.5)	86.5(-3.5)	88.7(-1.8)	90.9(-1.2)
	GBDT	88.7(-1.7)	73.9(-16.1)	92.3(+1.8)	90.0(-2.1)
	VGG16	85.9(-4.5)	77.3(-12.7)	89.1(-1.4)	88.6(-3.5)
	ResNet	87.1(-3.3)	73.9(-16.1)	92.3(+1.8)	90.6(-1.5)
	VGG16_Fusion	88.9(-1.5)	72.2(-17.8)	96.5(+6.0)	90.1(-2.0)
	ResNet_Fusion	87.7(-2.7)	78.9(-11.1)	90.1(-0.4)	90.9(-1.2)
	Ours	**90.4**	**90.0**	**90.5**	**92.1**

5.2 Comparison Results Against Other Radiomics Feature Selection Approaches

To further explore the role of constraint-based method on feature selection, we compared the performance of four commonly used feature selection approaches in radiomics (i.e., LASSO, RFE, PCA, MI) [2,23,27]. Since the features selected by these approaches including ours can be employed in arbitrary classifiers. They were evaluated in several individual classifiers. The results are reported in Fig. 4. Overall, the constraint-based model outperforms other conventional approaches as being capable to capture causal features. This is mainly because the other approaches are greatly affected by outliers, including confounding factors and hidden variables, which often appear in medical datasets. Note that other approaches mainly consider the correlation between features, while ignoring those causal relationships. In addition, it can be seen that the samples generated by our approach are more classifier-agnostic than those of other approaches, which can be adopted by various classifiers. Theoretically, the time complexity of our model is $O(K^3 + NK^2) + O(N^2RM_r) + O_{df}$, where the first value represents the complexity of causal network construction, while the latter two values indicate the complexity of radiomics feature selection and deep feature learning, respectively, and thus it is affordable for practical usage in NSCLC recognition.



(a) Accuracy comparison on P-NSCLC

(b) Accuracy comparison on I-NSCLC

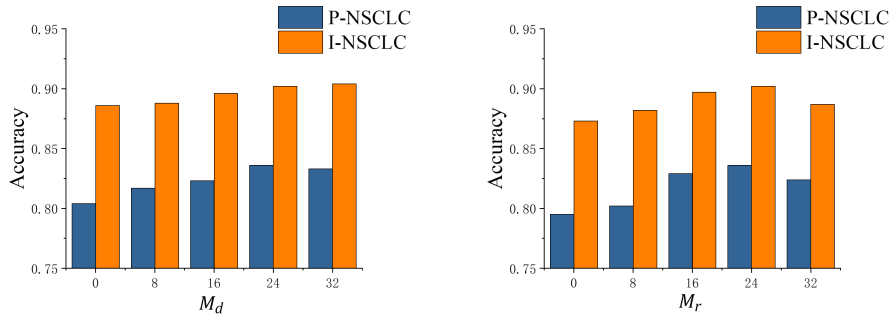
(c) AUC comparison on P-NSCLC

(d) AUC comparison on I-NSCLC

Fig. 4. Accuracy and AUC comparison against other feature selection methods.

5.3 Ablation Study

In this section, we conducted ablation studies to measure the effectiveness of the modules in our model. We evaluate the impact of the respective sizes of deep and radiomics feature nodes in the network skeleton on causal relation discovery and final classification performance. As shown in Fig. 5, it is clear that the combination of both types of features can achieve better classification performance than merely the usage of either of them. It is worth noting that $M_d = 0$ (*resp.* $M_r = 0$) represents that the skeleton is completely composed of radiomics features (*resp.* deep features). There is an improvement in accuracy when M_d and M_r grow to 32 and 24 on I-NSCLC, respectively. Similarly, it can be seen that the best setting on P-NSCLC is $M_d = 24$ and $M_r = 24$. In addition, a large value of M_d and M_r will bring about a very complicated network structure with a great number of nodes (features), which results in a huge computational complexity.



(a) Changes on different M_d ($M_r = 24$) (b) Changes on different M_r ($M_d = 24$)

Fig. 5. Accuracy comparison of different settings of our model.

6 Conclusion

In this paper, we present a constraint-based network approach where causal Markov condition is incorporated to exploit the causal-effect dependencies among a combination of deep and radiomics features for NSCLC recognition from CT images. It is more reliable and flexible than existing methods on NSCLC subtype classification by explicitly discovering representative features under a causal view. As for future work, we will further consider utilizing not only the selected variables (nodes) but also their causal relations (links) as features to train our model, and we will explore the applications of our approach on other types of cancers.

Acknowledgements This work was supported by grants from the National Major Science and Technology Projects of China (grant no. 2018AAA0100703), the National Natural Science Foundation of China (grant nos. 61977012, 61977054), the Central Universities in China (grant no. 2021CDJYGRH011).

References

1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications **5**(1), 1–9 (2014)
2. Aonpong, P., Iwamoto, Y., Han, X.H., Lin, L., Chen, Y.W.: Genotype-guided radiomics signatures for recurrence prediction of non-small cell lung cancer. IEEE Access **9**, 90244–90254 (2021)
3. Aonpong, P., Iwamoto, Y., Wang, W., Lin, L., Chen, Y.W.: Hand-crafted and deep learning-based radiomics models for recurrence prediction of non-small cells lung cancers. In: Innovation in Medicine and Healthcare, pp. 135–144. Springer (2020)
4. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 1–10 (2020)

5. Chaudhary, M.S., Gonzalez, D.L., Bello, G.A., Angus, M.P., Desai, D., Harenberg, S., Doraiswamy, P.M., Semazzi, F.H., Kumar, V., Samatova, N.F., et al.: Causality-guided feature selection. In: International Conference on Advanced Data Mining and Applications. pp. 391–405. Springer (2016)
6. Coumans, V., Claassen, T., Terwijn, S.: Causal discovery algorithms and real world systems. Ph.D. thesis, Masters thesis (2017)
7. Debbi, H.: Causal explanation of convolutional neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 633–649. Springer (2021)
8. Duangsoithong, R., Phukpattaranont, P., Windeatt, T.: Bootstrap causal feature selection for irrelevant feature elimination. In: The 6th 2013 Biomedical Engineering International Conference. pp. 1–5. IEEE (2013)
9. Fan, X., Wang, Y., Tang, X.Q.: Extracting predictors for lung adenocarcinoma based on granger causality test and stepwise character selection. BMC bioinformatics **20**(7), 83–96 (2019)
10. Feng, X., Wang, S., Liu, Q., Li, H., Liu, J., Xu, C., Yang, W., Shu, Y., Zheng, W., Yu, B., et al.: Selecting multiple biomarker subsets with similarly effective binary classification performances. JoVE (Journal of Visualized Experiments) (140), e57738 (2018)
11. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: Problems and methods. ACM Computing Surveys (CSUR) **53**(4), 1–37 (2020)
12. Han, Y., Ma, Y., Wu, Z., Zhang, F., Zheng, D., Liu, X., Tao, L., Liang, Z., Yang, Z., Li, X., et al.: Histologic subtype classification of non-small cell lung cancer using pet/ct images. European journal of nuclear medicine and molecular imaging **48**(2), 350–360 (2021)
13. Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S.A., Schabath, M.B., Forster, K., Aerts, H.J., Dekker, A., Fenstermacher, D., et al.: Radiomics: the process and the challenges. Magnetic resonance imaging **30**(9), 1234–1248 (2012)
14. Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., De Jong, E.E., Van Timmeren, J., Sanduleanu, S., Larue, R.T., Even, A.J., Jochems, A., et al.: Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology **14**(12), 749–762 (2017)
15. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer **48**(4), 441–446 (2012)
16. Li, C., Li, H.: Variable selection and regression analysis for graph-structured covariates with an application to genomics. The annals of applied statistics **4**(3), 1498 (2010)
17. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM computing surveys (CSUR) **50**(6), 1–45 (2017)
18. Raghu, V.K., Zhao, W., Pu, J., Leader, J.K., Wang, R., Herman, J., Yuan, J.M., Benos, P.V., Wilson, D.O.: Feasibility of lung cancer prediction from low-dose ct scan and smoking factors using causal models. Thorax **74**(7), 643–649 (2019)
19. Shayesteh, S., Nazari, M., Salahshour, A., Sandoughdaran, S., Hajianfar, G., Khateri, M., Yaghobi Joybari, A., Jozian, F., Fatehi Feyzabad, S.H., Arabi, H., et al.: Treatment response prediction using mri-based pre-, post-, and delta-radiomic features and machine learning algorithms in colorectal cancer. Medical physics **48**(7), 3691–3701 (2021)

20. Shaziya, H., Shyamala, K., Zaheer, R.: Automatic lung segmentation on thoracic ct scans using u-net convolutional network. In: 2018 International conference on communication and signal processing (ICCSP). pp. 0643–0647. IEEE (2018)
21. Wang, X., Li, Q., Cai, J., Wang, W., Xu, P., Zhang, Y., Fang, Q., Fu, C., Fan, L., Xiao, Y., et al.: Predicting the invasiveness of lung adenocarcinomas appearing as ground-glass nodule on ct scan using multi-task learning and deep radiomics. Translational Lung Cancer Research **9**(4), 1397 (2020)
22. Wang, X., Duan, H.h., Nie, S.d.: Prognostic recurrence analysis method for non-small cell lung cancer based on ct imaging. In: 2019 International Conference on Image and Video Processing, and Artificial Intelligence. vol. 11321, p. 113211T. International Society for Optics and Photonics (2019)
23. Wang, Y., Yue, W., Li, X., Liu, S., Guo, L., Xu, H., Zhang, H., Yang, G.: Comparison study of radiomics and deep learning-based methods for thyroid nodules classification using ultrasound images. Ieee Access **8**, 52010–52017 (2020)
24. Wu, Y., Ma, J., Huang, X., Ling, S.H., Su, S.W.: Deepmmsa: A novel multimodal deep learning method for non-small cell lung cancer survival analysis. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 1468–1472. IEEE (2021)
25. Yan, X., Liao, J., Luo, H., Zhang, Y., Liu, L.: Predicting cancer risks by a constraint-based causal network. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
26. van der Zander, B., Liśkiewicz, M., Textor, J.: Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. Artificial Intelligence **270**, 1–40 (2019)
27. Zhang, Y., Cheng, C., Liu, Z., Wang, L., Pan, G., Sun, G., Chang, Y., Zuo, C., Yang, X.: Radiomics analysis for the differentiation of autoimmune pancreatitis and pancreatic ductal adenocarcinoma in 18f-fdg pet/ct. Medical Physics **46**(10), 4520–4530 (2019)
28. Zhu, X., Dong, D., Chen, Z., Fang, M., Zhang, L., Song, J., Yu, D., Zang, Y., Liu, Z., Shi, J., et al.: Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. European radiology **28**(7), 2772–2778 (2018)
29. Zwanenburg, A., Leger, S., Vallières, M., Löck, S.: Image biomarker standardisation initiative. arXiv preprint arXiv:1612.07003 (2016)