Causally-Aware Intraoperative Imputation for Overall Survival Time Prediction

Xiang Li[1,*], Xuelin Qian[1,*], Litian Liang[1,*], Lingjie Kong[1], Qiaole Dong[1], Jiejun Chen[2]
Dingxia Liu[2], Xiuzhong Yao[2,†], Yanwei Fu[1,†]

[1]School of Data Science, Fudan University
[2]Department of Radiology, Zhongshan Hospital, Fudan University
{li_x20,xlqian,ltliang19,qldong18,jjchen20,yanweifu}@fudan.edu.cn
{jkong22,dxliu21}@m.fudan.edu.cn   yao.xiuzhong@zs-hospital.sh.cn

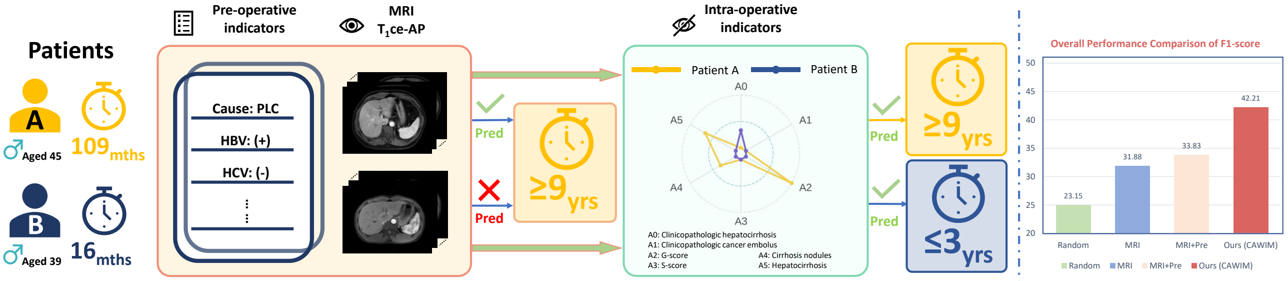

Figure 1. Our idea illustration. Patient A and B have different OS time (109 months and 16 months, respectively) but similar pre-operative image-based patterns and indexes; thus, preoperative based model fails to distinguish A and B (their predicted OS time are both larger than 9 years). However, their intra-operative attributes, which can describe the severity of disease in a more informative way, are very different from each other (*e.g.*, G-score, S-score, Clinicopathologic hepatocirrhosis). Therefore, it can help the model discriminate A and B. In the right image, we show that our CAWIM – that leverages intra-operative indexes in the training stage – can largely improve the prediction power than other methods. Our method managed to correctly classify these two cases, and the overall performance of our CAWIM surpasses other baseline models by approximately 10 points on 4-category classification task.

Abstract

Previous efforts in vision community are mostly made on learning good representations from visual patterns. Beyond this, this paper emphasizes the high-level ability of causal reasoning. We thus present a case study of solving the challenging task of Overall Survival (OS) time in primary liver cancers. Critically, the prediction of OS time at the early stage remains challenging, due to the unobvious image patterns of reflecting the OS. To this end, we propose a causal inference system by leveraging the intraoperative attributes and the correlation among them, as an intermediate supervision to bridge the gap between the images and the final OS. Particularly, we build a causal graph, and train the images to estimate the intraoperative attributes for final OS prediction. We present a novel Causally-aware Intraoperative Imputation Model (CAWIM) that can sequentially predict each attribute using its parent nodes in the estimated causal graph. To determine the causal directions, we propose a splitting-voting *mechanism, which votes for the direction for each pair of adjacent nodes among multiple predictions obtained via causal discovery from heterogeneity. The practicability and effectiveness of our method are demonstrated by the promising results on liver cancer dataset of 361 patients with long-term observations.*

1. Introduction

The success of recent deep learning model is largely attributed to learning the good representations for visual patterns. Such representations essentially facilitate various vision task, such as recognition and synthesis [15, 25, 33]. Nevertheless, one important goal for the vision community is to model and summarize the relationships of observed variables of a system, in order to enable well predictions on similar data. Essentially, it is desirable to understand how the system is changed if one modifies these relationships under certain conditions, *e.g.*, the effects of a treatment in healthcare. Thus this demands the *high-*

[*] Equal contribution
[†] Corresponding author

level ability of causal reasoning beyond the previous efforts of only learning good representations for visual patterns [1, 5, 16, 21, 35, 37]. This naturally leads into our task of causal inference.

This paper presents a case study of solving the challenging task of Overall Survival (OS) time estimation in Primary Liver Cancers (PLC). Generally, the liver cancer remains one of the most common malignancies worldwide in the 21st century, as there are about one million new cases every year [36]. The five-year survival rate for advanced PLC is only about 5% [7]. Therefore, early and accurate prediction of OS time estimation can provide informative guidance for individualized treatment planning and reducing burden of medical resources [6, 27, 32, 34]. On the other hand, one shall easily notice that with the renaissance of deep learning, great achievements have been made on medical imaging analysis, such as diagnosis, segmentation, and classification [16, 21, 35, 37]. Unfortunately, it still remains challenging for experienced clinicians to predict OS time at early stage, even with advanced modern diagnostic tools such as Magnetic Resonance Imaging (MRI) or tumor marker tests [3, 11], and the deep learning tools [1, 5].

Some studies propose to leverage deep neural networks for OS time prediction. They focus on fusion learning of multi-modal image features and some basic information (*i.e.*, age and gender) [5, 8, 14, 22, 24, 30, 39]. Nevertheless, the accurate prediction based on only preoperative information (such as image and tumor marker indexes) is still challenging, possibly due to missing information from early diagnosis stage to the final stage. This missing information includes the texture and pathological attributes of the liver, health level of the patient, and post-operative treatment [9, 17, 20]. For example, as in Fig. 1, patient A and B with different OS time have almost identical preoperative indicators *cause, history of disease,* etc,, making the preoperative-based model hard for discrimination.

To amend this problem, we present a causal inference system that can well utilize the intraoperative information, which is pretty easy to be accessed in training data. According to medical priories, such information records pathological attributes, which can be more reflective about the patient's health level and the postoperative recovery. Inspired by this, we propose to leverage this auxiliary information to help build our causal inference system. Again, we take the example in Fig. 1. Although preoperative information cannot differentiate patient A from B, their intraoperative index shows great difference in distribution, which can thus be employed for the discrimination. Additionally, there are many indicators from medical experts that the intraoperative indexes are related to each other. For instance, the *clinicopathologic hepatocirrhosis* is dependent on the *hepatocirrhosis*; the *sum of tumor diameter* is affected by the *number of tumors*; the fibrosis (reflected on S-Score) can probably deteriorate to cirrhosis [4], *etc*. By leveraging these relationships, we can better understand the causal inference concrete from both observed data modelling and medical expert-level knowledge of these variables.

To this end, we encapsulate these priors and the inspired proposals into a new method, dubbed as *Causally-Aware Intraoperative Imputation Model* (CAWIM). It incorporates causal discovery module to sequentially estimate intraoperative indexes as an intermediate stage towards final OS time prediction. Specifically, our model is composed of two key steps: **i)** estimating the intraoperative indexes using preoperative information, *i.e.*, image and indexes; **ii)** followed by OS time prediction using estimated intraoperative indexes and preoperative information. To achieve more accurate prediction of intraoperative indexes that is determinant to the prediction power of the whole method, we propose a ***C****ausally-****a****ware* ***D****irected* ***A****cyclic* ***G****raph* (CaDAG) module. It learns the causal structure represented as a DAG over intraoperative features. To identify the causal relations beyond the traditional PC algorithm [26], we propose a *splitting-voting* mechanism, which is inspired by the recent work [18] that learn the causal structure with the assistance of an auxiliary domain index variable. Our proposed mechanism can not only identify the causal relations even when this domain index variable is not available, but also can be theoretically guaranteed that the learned graph is not acyclic. During test stage, we sequentially estimate each intraoperative index with preoperative information and additionally, its parent set among other intraoperative indexes. The utility of our method can be demonstrated by a significant improvement of OS time prediction on an inhouse liver cancer dataset, as well as better interpretability of learned causal structure, more accurate estimation of intraoperative indexes and more interpretable visualization results.

In a nutshell, we for the first time present a case study of building a causally-aware intraoperative imputation system for the challenging task of overall survival time prediction. The proposed method of building the casual inference system, can be naturally extended to other similar medical tasks. Our key contributions are listed as follows. **(1) New Paradigm for OS time Prediction.** We propose to leverage intraoperative indexes as an intermediate stage during training. The leverage of this information can significantly alleviate the "missing information" issue. To the best of our knowledge, we are the first to leverage auxiliary information (in addition to preopearative features) for OS time prediction. **(2) Causal Structure Learning.** We propose a novel *splitting-voting* mechanism that can identify the causal structure even when the domain index variable is missing. **(3) Better Prediction Power.** Our method can significantly improve the prediction power for liver cancer over the competitors. The methods are evaluated on the medical dataset, which, to the best of our knowledge, is the largest primary live cancer dataset.

2. Related Work and Preliminaries

Previous efforts [23,39] directly take the OS time prediction as the vanilla classification task by employing various deep models, whilst we reformulate it by causal discovery method for the first time. We will review the most related literature here.

MRI based deep models. OS time is defined as the duration from patient's first scan to cancer-related death [6]. Since Magnetic Resonance Image (MRI) plays an important role in tumor-related studies, various MRI based deep learning model have been proposed for feature extraction and representing [28], multi-modal feature fusion [10, 30, 38, 39], and improvement on model structure [5]. However, since missing information during disease progression can mediate the correlation from the preoperative stage to the final stage, the results of these image based methods are typically unsatisfied. In contrast, our method can amend this problem by leveraging intraoperative indexes as a bridge.

Causal discovery. Causal discovery is to learn the causal graph over endogenous variables. Typical methods include PC algorithm [19], a constraint-based method that implement iterative conditional independence test to identify the skeleton and directions via v-structure. Under Markovian and faithfulness assumptions [26, 31], this method is provable to identify an equivalent class of the causal graph. To further identify more directions, recent work [18] leveraged a domain index variable that splits the dataset into multiple domains. With these heterogeneous data, they can identify mutable variables that change across the domain index variable. On the basis of this, they further identified edges among these mutable variable set by implementing Hilbert Schmidt Independence Criterion (HSIC) norm [13]. However, such domain index may not be available in many scenarios, making the identification of these directions difficult. Besides, this method may not ensure that the learned graph is acylic. To address these issues, we in this paper propose a *splitting-voting* mechanism to mine the causal relations among intraoperative indexes so as for leveraging each other to better estimate these indexes, which is provable to return a directed acyclic graph.

Preliminaries. (1) Causal Discovery. By faithfuness and Markovian assumptions [26], we can use conditional independent tests for causal discovery. That is learning the causal graph over $\mathbf{V}$. **(2) Structural causal model (SCM).** Before fully developing our contributions in methodology, we give as preliminary the definition of SCM, which is defined as $\langle G, \mathcal{F}, P\langle\epsilon\rangle\rangle$: **i)** directed acycle graph $G = (\mathbf{V}, \mathbf{E})$ is as causal structure, with node set $\mathbf{V}$ and edge set $\mathbf{E}$; **ii)** the autonomous structural functions $\mathcal{F} = \{f_k\}_{V_k \in V}$, where disturbance on V_k has no effect on others; **iii)** probability measure $P(\epsilon)$ for exogenous variables $\{\epsilon_k\}_k$. Given the assumption that $\{\epsilon_k\}_k$ are independent, each P can be obtained by *Causal Markov Condition* with G as $\mathrm{P}(\{V_K = v_k\}_{V_k \in \mathbf{V}}) = \prod_k \mathrm{P}(V_k = v_k | Pa(k) = pa(k))$. **(3) Causal discovery from heterogeneity.** In [18], the authors proposed to leverage a domain index variable C to identify some causal directions. This method is built upon the PC and faithfulness assumption at a distributional level. It first learns the skeleton of the graph and identifies a variable set that is affected by E, followed by identifying the directions between E and its neighbors. Specifically, if $C \to V_i, C \to V_j$, and $V_i \perp V_j | X, C$ for some deconfounding set X between V_i and V_j, $\{P(V_i|X, C = c\}$ is independent of $\{P(V_j|V_i, C = c, X)\}$, while $\{P(V_j|X, C = c)\}$ and $\{P(V_j|V_i, C = c, X)\}$ are dependent. We can test these Independence using *Hilbert Schmidt Independence Criterion* (HSIC) [12] norms $\delta_{v_i \to v_j | X}$ and $\delta_{v_j \to v_i | X}$. One can obtain that $V_i \to V_j$ if $\delta_{v_i \to V_j} < \alpha$ with pre-set significance level α. However, in many applications such domain index variable is missing, *i.e.*, all data are pooled together. In this paper, we propose a *splitting-voting* mechanism (in the CaDAG module) to identify causal relations. In the following, we first introduce some basic assumptions that our method is built upon.

3. Methodology

Problem setting. For the OS time classification task, suppose we have $\{X_i, \mathbf{B}_i, \mathbf{A}_i, y_i\} \sim_{i.i.d} P(X, \mathbf{B}, \mathbf{A}, y)$, where X denotes the image acquired from structural Magnetic Resonance Image (sMRI); $\mathbf{B}, \mathbf{A}$ denote the preoperative and intraoperative attributes, respectively. We have the final label y. Our goal is to predict y from the image X and preoperative attributes $\mathbf{B}$ that are recorded before surgery. Our method is built upon the following three assumptions, commonly utilized in the causal inference works [18, 26].

Definition 1 (Causal Graph) *We assume the causal graph over $\mathbf{A}$ is a directed acyclic graph (DAG) and denote the corresponding SCM as $M := \langle G := (\mathbf{A}, \mathbf{E}), \mathcal{F}, P(\varepsilon)\rangle$.*

Definition 2 (Markovian and Faithfulness) *For triplets of disjoint sets $\mathbf{V}_i, \mathbf{V}_j, \mathbf{V}_k$, it holds that $\mathbf{V}_i \perp_d \mathbf{V}_j | \mathbf{V}_k \leftrightarrow \mathbf{V}_i \perp \mathbf{V}_j | \mathbf{V}_k$, where $\perp_d$ and $\perp$ respectively mean d-separation and probability independence. This is the common property in DAG.*

Definition 3 (Distributional Faithfulness) *If $X_i \to X_j$ and at least $E \to V_i$ or $E \to V_j$ holds, $\{P^e(V_i|V_j, \mathbf{Z})\}$ is dependent to $\{P^e(V_j|\mathbf{Z})\}$, where $\mathbf{Z}$ denotes the minimal deconfounding set. Particularly, $\mathbf{Z}$ is a deconfounding set between V_i and V_j if we have $V_i \perp V_j | \mathbf{Z}$ and $\mathbf{Z} \cap (\mathrm{De}(V_i) \cup \mathrm{De}(V_j)) = \emptyset$.*

In term of this assumption, we further have the following result for edge orientation [18].

Theorem 4 (Theorem 2 in [18]) *Denote E as domain index variable. Under assumptions 1, 2, 3, for each adjacent pair (V_i, V_j) such that $E \to V_i$ or $E \to V_j$ holds, we have $V_i \to V_j$ if $\{P^e(V_i|\mathbf{Z})\}$ and $\{P^e(V_j|V_i, \mathbf{Z})\}$ are independent while $\{P^e(V_j|\mathbf{Z})\}$ and $\{P^e(V_i|V_j, \mathbf{Z})\}$ are dependent.*

3.1. Causally-aware Directed Acyclic Graph

The Causally-aware DAG. We propose a novel Causally-Aware Intraoperative Imputation Model (CAWIM). The pipeline of our method is shown in Fig. 2. It is composed of *Causally-aware Directed Acycled Graph* (CaDAG), and a classification model with MRI imaging encoder. The CaDAG enables the causally-aware intraoperative reasoning, by managing to better represent and encode the correlation of intraoperative information. Specifically, we first use preoperative information, *i.e.*, MRI images and indexes to estimate the intraoperative features. The estimated intraoperative features $\tilde{\mathbf{A}}$, together with the MRI X and preoperative features $\mathbf{B}$ are then utilized for OS classification. To accurately estimate intraoperative indexes $\mathbf{A}$, we propose CaDAG to learn causal structure over $\mathbf{A}$, and then leverage the parent nodes of each index for sequentially imputation during test stage. We give the details about the CaDAG and the intraoperative imputation model for classification, while some theoretical proofs, and the specification of our networks are in the appendix.

We aim at learning the causal structure over $\mathbf{A}$. However, without the domain index, we actually have the difficulty of dividing the whole dataset into multiple domains. To this end, we propose a *splitting-voting* mechanism for this problem. Particularly, this mechanism first generates the domain-index variable E by randomly splitting the whole dataset into multiple domains for $m > 0$ times. Then for each time we identify the mutable variable set and generate a prediction for each identifiable causal directions, with the assistance of domain index variable E. Finally, we vote for a final direction among m predictions for each pair of variables. As shown in Fig. 3, the whole procedure is a sequence of the four steps. Step **i)**, **ii)** and **iii)** provide preliminary causal relationship between each pair of nodes. Step **iv)** is our proposed *splitting-voting*, aiming to finally determine all the causal directions and output a causal DAG.

i) Learning skeleton of the causal graph. We first implement PC algorithm to learn the skeleton [26]. That is an undirected graph over all attributes. This can be achieved by iterative conditional independence test. Then we determine A_i and A_j to be adjacent if they are not independent conditional on any subsets. Finally, we utilize the v-structure to learn the equivalence class of DAG.

ii) Random splitting and identification of mutable variable sets. To determine more directions, we randomly split the dataset into multiple domains for m times. For each time we have a domain index variable E_i and multiple domains $\{\mathcal{D}_e^i\}$ [18]; and we identify the mutable variable set $\mathbf{M}_i$ such that each $X \in \mathbf{M}_i$ has $X \not\perp E_i|\mathbf{A} - X$ is conditionally dependent to E_i. We thus can implement conditional independent tests to identify $\mathbf{M}_i$. Steps i) and ii) are summarized in Alg. 1.

Algorithm 1 Identify the skeleton and mutable set.

INPUT: $\{\mathcal{D}_e^i := \{\mathbf{A}_i^e\}_{i=1}^{n_e}|e\}$, domain index variable E_i.
OUTPUT: Equivalent class of DAG and $\mathbf{M}_i$.

```
Implement the PC algorithm to learn the equivalence class of DAG via v-structure.
For each $A \in \mathbf{A}$, add the edge $E_i \to A$ in the graph iff $E_i \not\perp A$ given any subsets.
```

iii) Identifying directions via changed causal models. For each adjacent pair A_i and A_j such that at least one of them is adjacent to E, we identify the causal directions among $\mathbf{M}_i$ via HSIC Norm [12] by testing the independence between $\{P^e(A_i|A_j, \mathbf{Z})\}$ and $\{P^e(A_j|\mathbf{Z})\}$ if $A_i \leftarrow A_j$, and that between $\{P^e(A_j|A_i, C)\}$ and $\{P^e(A_i|\mathbf{Z})\}$ for determining the direction between A_i and A_j. Here, $\mathbf{Z}$ denotes the deconfounding set. In this regard, the direction between A_i and A_j is then determined according to the following rule: *if $\hat{\Delta}_{A_i \to A_j|\mathbf{Z},E} < \hat{\Delta}_{A_j \to A_i|C,E}$, output $A_i \to A_j$; if $\hat{\Delta}_{A_i \to A_j|\mathbf{Z},E} > \hat{\Delta}_{A_j \to A_i|\mathbf{Z},E}$, output $A_i \leftarrow A_j$; if $\hat{\Delta}_{A_i \to A_j|\mathbf{Z},E} > \alpha$ and $\hat{\Delta}_{A_j \to A_i|\mathbf{Z},E} > \alpha$, where α is a threshold, making the direction undetermined.*

Algorithm 2 Identify causal directions among $\mathbf{M}$.

INPUT: E; skeleton and M via Alg.1.
OUTPUT: Directed graph among $\mathbf{M}$.

```
For each adjacent $(A_i, A_j)$ such that one of $V_i \in \mathbf{M}$,
    Detect deconfounding set $\mathbf{Z}$.
    Calculate $\hat{\Delta}_{A_i \to A_j|\mathbf{Z},E} > \alpha$ and $\hat{\Delta}_{A_j \to A_i|\mathbf{Z},E} > \alpha$.
    Determine $A_i \to A_j$ or $A_j \to A_i$.
```

iv) Voting for the direction of each edge. For each time i we implement Alg. 2 to predict an direction of the edge (A_i, A_j). Then we vote for $A_i \to A_j$ if the frequency $f_{i \to j}$ of $A_i \to A_j$ surpasses a pre-set threshold α $(0 < \alpha < 1)$. Finally, if there exists a cycle of N variables: $A_{i_1} \to A_{i_2} \to ... \to A_{i_1}$, we orient $(A_{i_k}, A_{i_{k+1}})$ with the minimum frequency among $\{f_{i_k,i_{k+1}}, f_{i_N,i_1}\}$.

We show that this procedure can generate a DAG over $\mathbf{A}$, with the proof in appendix.

Theorem 5 *Under definitions and assumptions 1, 2, 3, the learned graph via our CaDAG is a directed acyclic graph.*

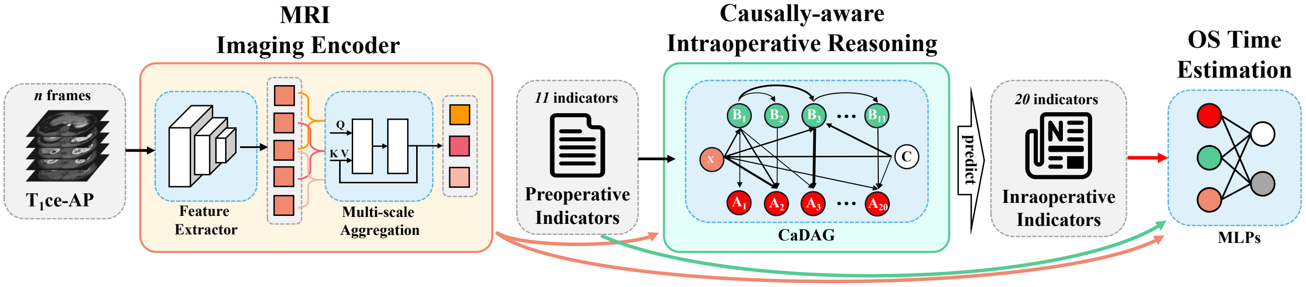

Figure 2. Overview of our CAWIM model. We concatenate the MRI features encoded by the *MRI Imaging Encoder*, preoperative indicators, and the estimated intraoperative indicators obtained from CaDAG to perform the final prediction on OS time.

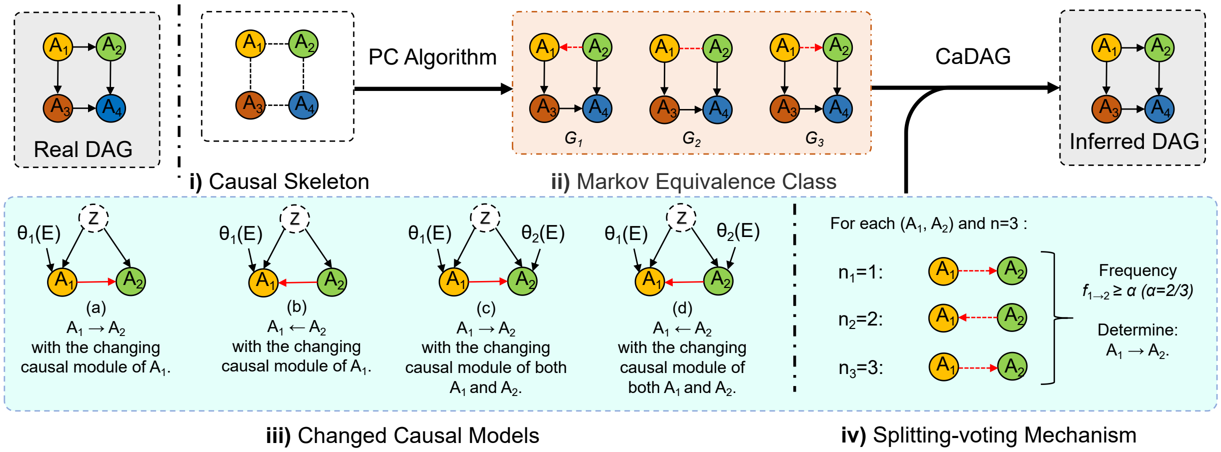
A$_1$
A$_2$
A$_3$
A$_4$
Real DAG
i) Causal Skeleton
PC Algorithm
G$_1$
G$_2$
G$_3$
ii) Markov Equivalence Class
CaDAG
Inferred DAG
θ$_1$(E)
θ$_2$(E)
Z
(a)
A$_1$ → A$_2$
with the changing causal module of A$_1$.
(b)
A$_1$ ← A$_2$
with the changing causal module of A$_1$.
(c)
A$_1$ → A$_2$
with the changing causal module of both A$_1$ and A$_2$.
(d)
A$_1$ ← A$_2$
with the changing causal module of both A$_1$ and A$_2$.
For each (A$_1$, A$_2$) and n=3 :
n$_1$=1:
n$_2$=2:
n$_3$=3:
Frequency
$f_{1\to2} \geq \alpha$ (α=2/3)
Determine:
A$_1$ → A$_2$.
iii) Changed Causal Models
iv) Splitting-voting Mechanism

Figure 3. Schematic diagram of a four-variable case A_1, A_2, A_3, and A_4, with the real DAG shown in the top left corner. Through step **i)** and **ii)** in the upper part, a Markov Equivalence Class of 3 causal graphs is observed with the direction between A_1 and A_2 undetermined. For step **iii)**, we split the dataset into ($n = 3$) domains and consider the 4 circumstances shown fig(a) to (d) in the lower left part. For step **iv)**, we vote for the direction between A_1 and A_2 as $A_1 \to A_2$ according to the number of splitting ($n = 3$), frequency $f_{1\to2} = 2/3$ and threshold $\alpha = 2/3$, as shown in the lower right part. Finally, we obtain an inferred DAG with each direction determined.

3.2. Intraoperative Imputation for Classification

Given MRI image $X \in \mathbb{R}^{H\times W}$ of a patient that contains multiple slides, we first resample these slides into a fixed K slides for each patient. As each slice is a gray images, we repeat the images for 3 times along the channel dimension; and then we employ a neural network to extract features from each slice. The extracted features are denoted as $F_1, ..., F_K$. In order to aggregate the information of all slices, we divide $F_1, ..., F_K$ into g groups according to the acquired position. Each group contains consecutive $\lfloor K/g \rfloor$ slices. Then, we implement a Multi-head Cross-Attention (MCA) mechanism inside each group to aggregate features:

$$\text{head}_j = \text{softmax}(Q^j(F\mathbf{W}_K^j)^T)(F\mathbf{W}_V^j) \qquad (1)$$

$$\text{MCA} = [\text{head}_1; \cdots ; \text{head}_h]\mathbf{W}_O \qquad (2)$$

where F is the slice's feature within a group, the query vector Q^j and linear projection layers $\mathbf{W}_K^j, \mathbf{W}_V^j, \mathbf{W}_O, 1 \leq j \leq h$ are all learnable parameter. Finally, we concatenate $\{F_k\}$ from g groups as our final image feature $I \in \mathbb{R}^{g\times d}$ with d denoting the dimension of each F_k with $k = 1, ..., K$.

Afterwards, we sequentially estimate each A_j by preoperative information, as our learned causal graph over $\mathbf{A}$. So for each A_j, we predict

$$\hat{A}_j = f^j_{\text{intro}}(\mathbf{B}, I, \text{Pa}(A_j)) \qquad (3)$$

where $\text{Pa}(A_j)$ denotes the parent node set of A_j. We use cross entropy loss to optimize f^j_{intro} if A_j is categorical; and optimize f^j_{intro} if A_j is continuous. After obtaining all predicted $\{\hat{A}_i\}_{i=1}^d$, we use $(I, \mathbf{B}, \hat{\mathbf{A}})$ to predict Y via $f_\theta(I, \mathbf{B}, \hat{\mathbf{A}})$ and obtain θ by optimizing the cross-entropy $\mathcal{L}(y, f_\theta(I, \mathbf{B}, \hat{\mathbf{A}}))$. During test stage, given a new sample $(x, \mathbf{b})$, we first extract I and then sequentially estimate $\mathbf{a}_i$ and finally predict y using $f_\theta(x, \mathbf{b}, \hat{\mathbf{a}})$.

4. Experiments

Dataset overview. Ethics Committee of Zhongshan Hospital, Fudan University approved the protocol of this study and waived the requirement for patient-informed consent (B2021-325R). Because open source data sets generally lack sufficient intraoperative information, it takes us around ten years to collect and build our own medical dataset. Particularly, we conduct a search through the medical records in the hospital information system, and build a dataset with

439 patients infected by primary liver cancer. Long-term medical data of every patient has been recorded from the first scan to the cancer-related death. Till the submission of this manuscript, around 90% patients have unfortunately passed away. This gives us the ground-truth OS labels. Information like MRI scanning before the surgery, blood test and operation reports, follow-up records are also collected. After removing samples with missing information, 361 patients were enrolled, with 306 (84.8%) men, and the average age was 53.7 (±11.7) years. Specifically, each patient has MRI scanning with four modalities, including T$_1$ weighted images (T$_1$-WI), T$_2$ weighted images (T$_2$-WI), T$_1$ in-phase (T$_1$-IP) and contrast–enhanced T$_1$ at the arterial phase (T$_1$ce-AP). Following the guidance of medical experts, we select 11 preoperative and 20 intraoperative indicators that are related to OS time for our study. Please refer to the appendix for details.

Data preprocessing. In our dataset, the survival time ranges from 0 to 130 months, which are cauterized into four classes: short-term survival ($\leq$ 36 months), middle-short-term survival (between 36 and 72 months), middle-long-term survival (between 72 and 108 months), and long-term survival ($\geq$ 108 months). In our paper, we consider the classification task: given a new sample, our goal is to classify which class of OS time this sample belongs to. To extract I from MRI image, we resample $K = 20$ slices and employ Resnet-34 for feature extraction. We set $g = 4$ and the dimension F_k as 512. For f^j_{intro} and f_θ, we employ a multi-layer perceptron (MLP) for parameterization.

We encode categorical covariates into dummy variables and implement zero-mean normalization for continuous variables. For MRI images, we resize them into 256 × 256 as input. We split the whole dataset into 5 folds. To remove the effect of imbalance across classes and randomness, we adopt average precision, recall, and F$_1$-Score measurements over five folds for classification evaluation.

Implementation details. We use the ImageNet pretrained Resnet-34 to initialize our CNN encoder. We train 150 epochs for all methods with T$_1$ce-AP modality, using SGD as the optimizer with learning rate $3e{-}3$ and $3e{-}2$ for CNN encoder and later MLP, respectively. We set weight decay to $5e - 4$ and nesterov momentum factor to 0.9. Besides, we decrease the learning rate every 60 epochs by a factor of $1/10$. We set the Batch size to 8. In our voting mechanism of CaDAG, we in total implement Alg. 2 for seven times and set the voting threshold as 4/7. Our model is trained with Pytorch on NVIDIA GeForce RTX 3090 GPUs.

Compared Methods. We compare our method with four baseline methods: **i)** *Random*: random guess among four classes as a trivial baseline, of which the expected F$_1$ score is 25%; **ii)** *MRI*: train the end-to-end method with T$_1$ce-AP only as input for classification; **iii)** *MRI+pre*: additionally take preoperative features as input, on the basis of the method *MRI*; **iv)** *ground-truth (GT) Intra*: only take the ground-truth intraoperative features as input. It is worth to mention that although the intraoperative features are not allowed to use during test stage, the comparision with method **iv** (*i.e.*, GT Intra) can provide information that to which extent, the intraoperative features can help OS time prediction.

Table 1. Comparisons of Our CAWIM with Other Baselines.

	Model	Precision	Recall	F$_1$-Score
(a)	Random	26.04±2.10	24.54±0.80	23.15±0.64
(b)	MRI	40.79±7.97	33.45±2.66	31.88±3.92
(c)	MRI+Pre	36.99±5.30	35.62±3.50	33.83±3.64
(d)	gt Intra	41.26±2.74	41.62±1.95	39.77±2.47
(e)	CAWIM	**45.58**±5.86	**43.70**±5.89	**42.21**±4.92

4.1. Results and Analysis

Table 1 records comparisons of our proposed CAWIM (Tab. 1(d)) with other baseline models. (Tab. 1(a), Tab. 1(b), and Tab. 1(c)). Obviously, our CAWIM constantly outperform others in all metrics. Specifically, CAWIM improved *MRI+Pre* by nearly 10% in terms of F$_1$-Score, which can be contributed to the prediction of intraoperative features equipped with causally aware prediction mechanism. Without this mechanism, using MRI and preoperative features only can outperform random guess by only 8.33%, which demonstrates the limited information of the image for OS time. In contrast, it is also interesting to note from Tab. 1(d) that using intraoperative features can be more informative than MRI features, in terms of OS prediction, which is natural as the attributes that described the functions and textures of the liver could affect the postoperative recovery and thus the OS time. Finally, our CAWIM (Tab. 1(e)) that combines the information of intraoperative features and the information from images, can achieve further improvement.

4.2. Ablation Study

To test the effectiveness of *preoperative features*, the particularly *intraoperative features*, and *our causally-aware* module for OS time prediction, we implement ablation studies and summarize the results in Tab. 2. Specifically, Tab. 2(a), *i.e.*, *w/o CaDAG* replaces our causally-aware module with an end-to-end estimation network from (MRI, preoperative features) to intraoperative features; Tab. 2(b), *i.e.*, *w/o Intra* is the same to Tab. 1(c), *i.e.*, *MRI+pre* in Tab. 1 that directly predict OS time using MRI and preoperative features without estimated intraoperative features; Tab. 2(c), *i.e.*, *w/o Pre* is the same to our CAWIM except that it does not utilize preoperative features for intraoperative estimation and OS prediction.

The results in Tab. 2 show that the deletion of each module can lead to a descent of all metrics. Specifically, Tab. 2(a) shows a significant performance drop by 5.6% in F$_1$-Score, which indicates the effectiveness of our causally-aware module in estimating intraoperative features in OS time prediction. This phenomena can be explained by more accurate intraoperative estimation equipped with our causally-aware module, as shown in Fig. 4 of F_1-score, accuracy and R-squared score. We will leave detailed discussions in Section of "Further Discussion". Tab. 2(b), as the same to *MRI+pre* in Tab. 1, shows a 8.3% performance drop, which indicates the additional information provided by the MRI image and preoperative features. Finally, the 3.78% degradation of Tab. 2(c) compared to ours validates the effectiveness of preoperative features in OS prediction, which can also be explained by the estimation results of intraoperative features in Fig. 4. In a word, we can observe a significant drop of F_1-score if we remove causally-aware module and intraoperative features, which can demonstrate the effectiveness of our methods.

Besides, it is also interesting to compare Tab. 2(b) and Tab. 2(c) with we observe that the performance drop for *w/o Intra* is more significant than that for *w/o Pre*, with respectively an average of 8.4% and 3.8% drop in F_1-Score. This again, indicates more information of intraoperative indexes provided than preoperative indexes in predicting OS time. Indeed, we obtained Precision 27.68%, Recall 26.65% and F1-Score 23.85% if we only predicted OS time using preoperative features, which is comparable to the result from random guess. On the other hand, we can observe from Fig. 4 that preoperative indexes are beneficial to the estimation of intraoperative features, which indicating the necessity of using intraoperative features as a bridge between preoperative information and final OS time prediction.

4.3. Medical Interpretation of Causal Graph

To further explain the effectiveness of our causally-aware module, we in this subsection present our learned causal graph over intraoperative features in Fig. 5 and the corresponding medical interpretation of learned causal relations. In general, most of the parent-child relationships in our CaDAG are in line with common sense or can be explained by prior medical knowledge, *e.g., sum of tumor diameter* is determined by *number of tumors*, the edge from *G-score* to *S-score* also correspond to the strong correlation between fibrosis and hepatitis (please refer to appendix for detailed descriptions of intraoperative indexes).

Further, those indicators with non-visual features that access severity of pathology are often determined by other attributes of the liver; therefore, these features often come up in deeper layers in Fig. 5. For instance, *G-score* and *S-score* reflect the level of inflammation and fibrosis that are artificially defined according to *Metavir scoring system* [2]. Those indicators tend to rely on other intraoperative features besides merely MRI and preoperative information, *e.g. clinicopathologic hepatocirrhosis* is reasoned from *hepatocirrhosis* and other covariants. For this reason, the additional leverage of *hepatocirrhosis* lead to more accurate estimation of *clinicopathologic hepatocirrhosis*, as shown in Fig. 4 (a). On the other hand, we also observe some associations that cannot be explained well, *e.g.* the causal relationship between *sum of tumor diameter* and *cell type*, which may due to the existence of unobserved confounders.

Table 2. Results of Ablation Studies. (a) estimates each intraoperative features naively from (MRI, preoperative indexes), without the causally-aware module. (b) is the same to *MRI+Pre* in Tab. 1. (c) is the CAWIM without preoperative features for intraoperative indexes estimation and OS time prediction.

	Model	Precision	Recall	F_1-Score
(a)	w/o CaDAG	36.98±2.38	37.78±1.91	36.62±1.87
(b)	w/o Intra	36.99±5.30	35.62±3.50	33.83±3.64
(c)	w/o Pre	38.92±7.71	41.57±4.53	38.43±5.84
(d)	CAWIM	**45.58**±5.86	**43.70**±5.89	**42.21**±4.92

4.4. Visualization

In this section, we visualize the high-response area of our CAWIM Fig. 6(c) and the version without CaDAG Fig. 6(b) using Grad-CAM [29]. As shown in Fig. 6, the detected regions of our method can be more concentrated on the liver. This result can be explained with Fig.4 and Fig. 5 in a more complementary way. Specifically, as the learned causal graph is medical explainable and with better estimation of intraoperative features, the model is driven to locate on liver-related regions. More visualization results are left in appendix due to space limit.

4.5. Further Discussion

Intraoperative prediction. In our method, the estimation of intraoperative indicators are important for final prediction. Fig. 4 shows the metrics of some typical intraoperative prediction by **i)** *w/o CaDAG* and **ii)** *w/o Pre* with F_1-Score and accuracy for discrete variants (Fig. 4 (a, b)) and R-square for continuous ones (Fig. 4 (c)). Fig. 4 (a, b) shows that our CaDAG can generally enhance the estimation of categorical variables *hepatocirrhosis, cell type, differentiation,* etc. Nevertheless, we also find that such improvements will decrease if corresponding indicators are in deep layers of the graph in Fig. 5. For example, the accuracy of *G-Score* and *S-score* is only comparable and even worse than that of *w/o CaDAG*. A possible reason is the accumulated error along the directed path for estimation. On the other hand, as shown by the results of R-square (closer to 1, the better the prediction) does not bring about notable improvement in regression tasks, (*e.g.,* ascites and tumor diameter).

We also observe an evident performance degradation of *w/o Pre* for most of those discrete or continuous variants. This implies that although operative alone has limited contribution to OS time prediction, it still matters in our CAWIM, because of the less missing information due to shorter time interval from the preoperative stage to the

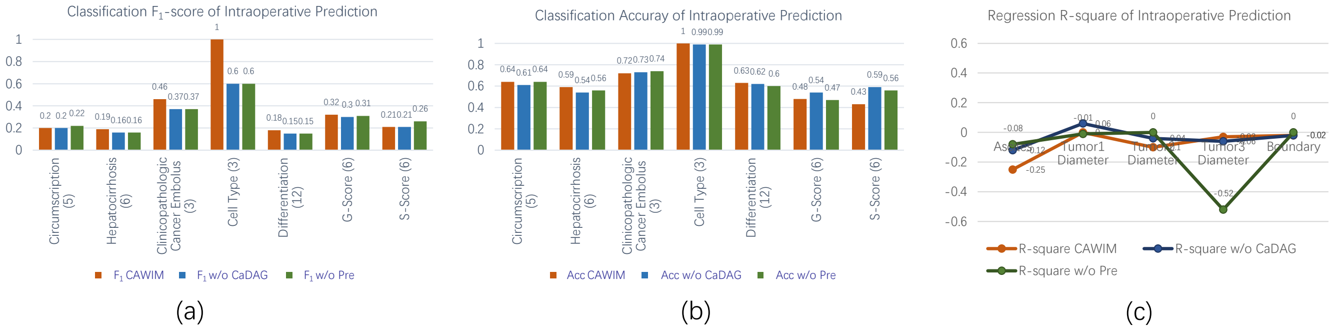

Figure 4. Quantity Results of Intraoperative Prediction. Intraoperative prediction are categorized into classification ((a) and (b)) and regression problem ((c)) according to the data form. (a) and (b) shows the prediction F$_1$-Score and accuracy rate of the discrete variants respectively. The numbers after the variant name in (a) and (b) are the number of classes, *i.e.* hepatocirrhosis is a 6-category classification task. (c) shows the R-squared score for continuous variants.

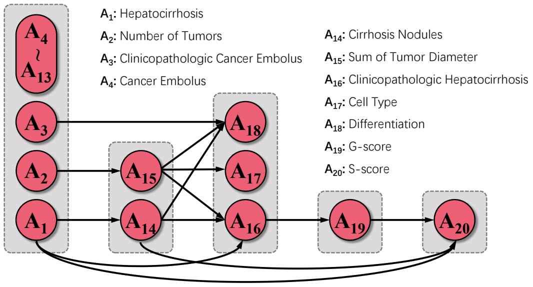

Figure 5. Learned DAG over Intraoperative Indexes **A**.

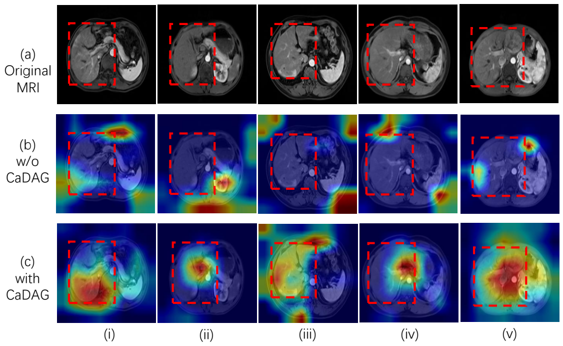

Figure 6. Heat Maps via Grad-CAM [29] on 5 Patients. (a) shows original images and the bounding box of liver. (b) and (c) shows the area of interest *w/o* and with CaDAG, respectively.

surgery stage than to the final stage. This observation validates the effect of using intraoperative features.

General Framework for OS time prediction. In general, predicting OS time from preoperative images and indexes is medically important but challenging, as there exists a lot of missing information due to a long time interval between the preoperative stage to the final OS stage. To amend this problem, it is informative to leverage some intermediate information. In this paper, we show that with this leverage and the modeling of this information via causal discovery, our CAWIM model enjoys convincing medical interpretability, more concentrated location of liver region, more accurate estimation of intraoperative features, and finally better classification results on OS time. We thus believe that our framework can be beneficial to OS time prediction in other scenarios, *e.g.*, other diseases or intermediate information in addition to intraoperative indexes.

Limitations. We find in Fig. 5 that although most of learned causal relations are consistent with medical priors, there exist relations that cannot be explained well: it is hard to determine the causal order for some pairs of variables: *e.g.*, the causal relationship between the sum of tumor diameter and cell type. This may due to unobserved confounders between these pairs of variables, which may be alleviated by learning hidden representations that can explain the associations among these pairs of variables. We will leave this exploration in our future work.

5. Conclusion

We propose a novel OS time prediction paradigm, which is the first to leverage intraoperative attributes by causal structure learning. Our method significantly outperforms baselines by a large margin. We demonstrate that this improvement is contributed to highly interpretable learned causal structure, accurate estimation of intraoperative indexes, and identification of disease-related regions. We believe our method, especially the leverage of intraoperative information equipped with causal discovery can potentially benefit other scenarios. For limitation, we shall relax the SCM assumption by allowing the existence of unobserved confounders. We leave it as the future work.

Acknowledgment. Ethics Committee of Zhongshan Hospital, Fudan University approved the protocol of this study and waived the requirement for patient-informed consent (B2021-325R). The authors are also with Shanghai Key Lab of Intelligent Information Processing, and Fudan ISTBI–ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.

References

[1] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021. 2

[2] Pierre Bedossa and Thierry Poynard. An algorithm for the grading of activity in chronic hepatitis c. *Hepatology*, 24(2):289–293, 1996. 7

[3] Nicholas A Christakis, Julia L Smith, Colin Murray Parkes, and Elizabeth B Lamont. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort studycommentary: Why do doctors overestimate? commentary: Prognoses should be based on proved indices not intuition. *Bmj*, 320(7233):469–473, 2000. 2

[4] Agostino Colli, Mirella Fraquelli, Marco Andreoletti, Barbara Marino, Enrico Zuccoli, and Dario Conte. Severe liver fibrosis or cirrhosis: accuracy of us for detection—analysis of 300 cases. *Radiology*, 227(1):89–94, 2003. 2

[5] Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021. 2, 3

[6] James J Driscoll and Oliver Rixe. Overall survival: still the gold standard: why overall survival remains the definitive end point in cancer clinical trials. *The Cancer Journal*, 15(5):401–405, 2009. 2, 3

[7] J Faivre, D Forman, J Esteve, M Obradovic, M Sant, EUROCARE Working Group, et al. Survival of patients with primary liver cancer, pancreatic cancer and biliary tract cancer in europe. *European Journal of Cancer*, 34(14):2184–2190, 1998. 2

[8] Ming Fan, Zuhui Liu, Sudan Xie, Maosheng Xu, Shiwei Wang, Xin Gao, and Lihua Li. Integration of dynamic contrast-enhanced magnetic resonance imaging and t2-weighted imaging radiomic features by a canonical correlation analysis-based feature fusion method to predict histological grade in ductal breast carcinoma. *Physics in Medicine & Biology*, 64(21):215001, 2019. 2

[9] Sheung Tat Fan, Chung Mau Lo, Ronnie TP Poon, Chun Yeung, Chi Leung Liu, Wai Key Yuen, Chi Ming Lam, Kelvin KC Ng, and See Ching Chan. Continuous improvement of survival outcomes of resection of hepatocellular carcinoma: a 20-year experience. *Annals of surgery*, 253(4):745–758, 2011. 2

[10] Xue Feng, Nicholas J Tustison, Sohil H Patel, and Craig H Meyer. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in computational neuroscience*, 14:25, 2020. 3

[11] Paul Glare, Kiran Virik, Mark Jones, Malcolm Hudson, Steffen Eychmuller, John Simes, and Nicholas Christakis. A systematic review of physicians' survival predictions in terminally ill cancer patients. *Bmj*, 327(7408):195, 2003. 2

[12] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005. 3, 4

[13] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007. 3

[14] Lu Guo, Ping Wang, Ranran Sun, Chengwen Yang, Ning Zhang, Yu Guo, and Yuanming Feng. A fuzzy feature fusion method for auto-segmentation of gliomas with multi-modality diffusion and perfusion magnetic resonance images in radiotherapy. *Scientific reports*, 8(1):1–11, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[16] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019. 2

[17] Karin W Houben and John L McCall. Liver transplantation for hepatocellular carcinoma in patients without underlying liver disease: a systematic review. *Liver Transplantation and Surgery*, 5(2):91–95, 1999. 2

[18] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020. 2, 3, 4

[19] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007. 3

[20] Christophe Laurent, Jean Frédéric Blanc, Steeve Nobili, Antonio Sa Cunha, Brigitte le Bail, Paulette Bioulac-Sage, Charles Balabaud, Maylis Capdepont, and Jean Saric. Prognostic factors and longterm survival after hepatic resection for hepatocellular carcinoma originating from noncirrhotic liver. *Journal of the American College of Surgeons*, 201(5):656–662, 2005. 2

[21] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE, 2014. 2

[22] Tao Li, Wu Li, Yehui Yang, and Wensheng Zhang. Classification of brain disease in magnetic resonance images using two-stage local feature fusion. *PloS one*, 12(2):e0171749, 2017. 2

[23] Dong Nie, Han Zhang, Ehsan Adeli, Luyan Liu, and Dinggang Shen. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *International conference on medical image computing and computer-assisted intervention*, pages 212–220. Springer, 2016. 3

[24] Byeonggwan Noh, Young Mok Park, Yujin Kwon, Chang In Choi, Byung Kwan Choi, Yo-Han Park, Kwangho Yang, Sunju Lee, Taeyoung Ha, YunKyong Hyon, et al. Machine learning-based survival rate prediction of korean hepatocellular carcinoma patients using multi-center data. *BMC gastroenterology*, 22(1):1–9, 2022. 2

[25] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017. 1

[26] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 3, 4

[27] Markus Peck-Radosavljevic. Drug therapy for advanced-stage liver cancer. *Liver cancer*, 3(2):125–131, 2014. 2

[28] Whitney B Pope, James Sayre, Alla Perlina, J Pablo Villablanca, Paul S Mischel, and Timothy F Cloughesy. Mr imaging correlates of survival in patients with high-grade gliomas. *American Journal of Neuroradiology*, 26(10):2466–2474, 2005. 3

[29] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7, 8

[30] Wei Shao, Tongxin Wang, Liang Sun, Tianhan Dong, Zhi Han, Zhi Huang, Jie Zhang, Daoqiang Zhang, and Kun Huang. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. *Medical Image Analysis*, 65:101795, 2020. 2, 3

[31] Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. Constructing bayesian network models of gene expression networks from microarray data. 2000. 3

[32] Wen Tang, Haoyue Zhang, Pengxin Yu, Han Kang, and Rongguo Zhang. Mmmna-net for overall survival time prediction of brain tumor patients. *arXiv preprint arXiv:2206.06267*, 2022. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[34] Jing-Houng Wang, Chi-Sin Changchien, Tsung-Hui Hu, Chuan-Mo Lee, Kwong-Ming Kee, Chih-Yun Lin, Chao-Long Chen, Tai-Yi Chen, Yu-Jie Huang, and Sheng-Nan Lu. The efficacy of treatment schedules according to barcelona clinic liver cancer staging for hepatocellular carcinoma–survival analysis of 3892 patients. *European journal of cancer*, 44(7):1000–1006, 2008. 2

[35] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):1–18, 2019. 2

[36] Naoki Yamanaka, Eizo Okamoto, Tsuyosi Oriyama, Jiro Fujimoto, Kazutaka Furukawa, Eisuke Kawamura, Tsuneo Tanaka, and Fumito Tomoda. A prediction scoring system to select the surgical treatment of liver cancer. further refinement based on 10 years of use. *Annals of surgery*, 219(4):342, 1994. 2

[37] Jianpeng Zhang, Yutong Xie, Qi Wu, and Yong Xia. Medical image classification using synergic deep learning. *Medical image analysis*, 54:10–19, 2019. 2

[38] Fan Zhou, Tengfei Li, Heng Li, and Hongtu Zhu. Tpcnn: two-phase patch-based convolutional neural network for automatic brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 274–286. Springer, 2017. 3

[39] Tao Zhou, H. Fu, Yu Zhang, Changqing Zhang, Xiankai Lu, Jianbing Shen, and Ling Shao. M2net: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients. *ArXiv*, abs/2006.10135, 2020. 2, 3