Biomedical Signal Processing and Control 106 (2025) 107663

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Multimodal learning for non-small cell lung cancer prognosis ☆

Yujiao Wu [a], Yaxiong Wang [b], Xiaoshui Huang [c], Haofei Wang [d], Fan Yang [e], Wenwen Sun [f], Sai Ho Ling [g], Steven W. Su [h,*]

[a] Commonwealth Scientific and Industrial Research Organization, Hobart, 7004, TAS, Australia
[b] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[c] School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[d] Peng Cheng Laboratory, Shenzhen, China
[e] Shenzhen Peini Medical Technology Co., Ltd., Shenzhen, China
[f] The Department of Dermatology, Affiliated Shenzhen Maternity & Child Healthcare Hospital, Southern Medical University, Shenzhen, China
[g] School of Electrical and Data Engineering, University of Technology Sydney, NSW, Sydney, Australia
[h] The College of Medical Information and Artificial Intelligence, Shandong First Medical University & Shandong Academy of Medical Sciences, Shandong, China

ARTICLE INFO

Keywords:
Multimodal learning
NSCLC
Survival analysis
Transformer

ABSTRACT

This paper focuses on the task of survival time analysis for lung cancer. Despite significant progress in recent years, the performance of existing methods is still far from satisfactory. Traditional and some deep learning-based approaches for lung cancer survival time analysis primarily rely on textual clinical information such as staging, age, and histology, etc. Unlike these existing methods that predicting on the single modality, we observe that human clinicians usually consider multimodal data, such as textual clinical parameters and visual scans when estimating survival time. Motivated by this observation, we propose Lite-ProSENet, a smart cross-modality network for survival analysis that simulates human decision-making. Specifically, Lite-ProSENet adopts a two-tower architecture that takes the clinical parameters and the CT scans as inputs to produce survival prediction. The textural tower is responsible for modeling the clinical parameters. We build a light transformer using multi-head self-attention as our textural tower. The visual tower, ProSENet, is designed to extract features from CT scans. The backbone of ProSENet is a 3D ResNet that works together with several repeatable building blocks named 3D-SE Resblocks for compact feature extraction. Our 3D-SE Resblock is composed of a 3D channel "Squeeze-and-Excitation" (SE) block and a temporal SE block. The purpose of 3D-SE Resblock is to adaptively select valuable features from CT scans. Besides, to further filter out the redundant information in the CT scans, we developed a simple yet effective frame difference mechanism, which boost the performance of our model to achieve new state- of-the-art results. Extensive experiments were conducted using data from 422 NSCLC patients from The Cancer Imaging Archive (TCIA). The results show that our Lite-ProSENet outperforms favorably against all comparison methods and achieves a new state-of-the-art concordance score of 89.3%. Our code is available at: https://github.com/wangyxxjtu/Lite_ProTrans.

1. Introduction

Lung cancer is one of the most malicious diseases, the overall five-year survival rate for lung cancer (LC) is even less than 20%. Most lung cancers can be divided into two broad histological subtypes: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Compared to SCLC, NSCLC accounts for the majority of diagnoses and is less aggressive. NSCLC spreads and grows more slowly than SCLC and causes few or no symptoms until it is advanced. As a result, patients are usually not detected until it is at a later stage. And it has caused millions of deaths in both women and men [1–7]. Lung cancer survival analysis, or prognostication, of lung cancer attempts to model the time range for a given event of interest (biological death), from the beginning of follow-up until the occurrence of the event. The survival model is an estimate of how lung cancer will develop, and it can reveal the relationship between prognostic factors and the disease. Using the accurate prognostic models, doctors can determine the most likely development(s) of the patient's cancer [8,9]. To improve predictive accuracy and automate the NSCLC survival analysis process,

☆ This paper was completed by H. Wang in collaboration with Y.Wu during his doctoral studies.

* Corresponding author.
E-mail addresses: yujiao.wu@csiro.au (Y. Wu), wangyx@hfut.edu.cn (Y. Wang), huangxiaoshui@sjtu.edu.cn (X. Huang), wanghf@pcl.ac.cn (H. Wang), yangfan@wpeony.com (F. Yang), 412416074@qq.com (W. Sun), Steve.Ling@uts.edu.au (S.H. Ling), suweidong@sdfmu.edu.cn, Steven.Su@uts.edu.au (S.W. Su).

https://doi.org/10.1016/j.bspc.2025.107663
Received 18 April 2024; Received in revised form 10 January 2025; Accepted 31 January 2025
Available online 24 February 2025
1746-8094/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

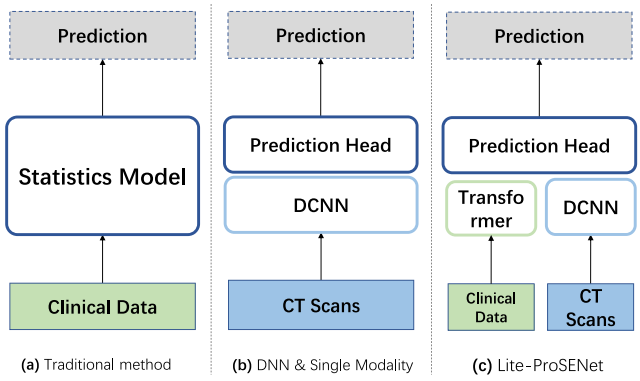


Fig. 1. The architecture comparison of existing methods and our proposed Lite-ProSENet.

as well as to assist medical experts develop precise treatment solutions, we aim to explore a novel method for NSCLC survival analysis.

Traditional statistical methods for survival analysis leverage structured data from comprehensive medical examinations. These methods primarily include time-to-event modeling tools such as the Cox proportional hazards (Cox- PH) model [10], the accelerated failure time (AFT) model [11], the Kaplan–Meier [12], etc. Besides, machine learning-based approaches for survival analysis have gained popularity, including survival trees [13,14], Bayesian methods [15,16] and Support Vector Regression [17,18] etc. These models assume a constant hazard ratio between two subjects over time [19], estimating either a risk score or the time-to-event distribution. However, when implemented in clinical practice, the interaction between covariates can be complex, and all these methods focus solely on structured data, overlooking the enormous information within the image data such as CT scans. Moreover, medical experts normally need to spend significant effort developing hand-crafted features for these models. Recently, some works have utilized pathological images to demonstrate the effectiveness of image-based features. However, obtaining pathological images requires a lung biopsy, an invasive procedure associated with potential health risks, including pneumothorax, bleeding, infection, systemic air embolism, and other side effects.

Artificial intelligence has made rapid strides over the past decade. With advancements in deep learning techniques, AI has accomplished remarkable success in various fields of research, including natural language processing, computer vision, etc. As a cutting-edge technology, deep learning holds significant promise for medical diagnostics. Some of the most innovative and novel deep learning methods have already been successfully applied to lung cancer diagnosis using CT images, with performance levels that surpass even those of human experts [20–23]. However, existing deep learning-based methods only consider the structured or visual cues. In contrast, medical professionals often consider both clinical parameters and visual information, such as CT images together to make comprehensive decisions. Consequently, the predictions generated by current methods lack sufficient reliability and credibility. To address this limitation, we propose a novel multimodal paradigm for lung cancer survival analysis inspired by the success of deep learning, as shown in Fig. 1.

To successfully build such a multimodal network, the first challenge is to encode the task-friendly features from different modalities. Although the format of the clinical parameters looks like some discrete symbols in a diagram, they are fundamentally highly correlated. To uncover correlations between different factors, we propose a Light transformer network to process the textual clinical parameters. The core building block of our model is the multi-head self-attention [24]. Moreover, self-attention mechanism is capable of correlating various disease factors, enabling the capture of more comprehensive information.

CT slices contain rich spatial and temporal information. In our previous work [25], we adopted 3D ResNet as the backbone for feature extraction. However, we found that excessive redundant information in both the spatial and temporal dimensions significantly hindered the model's ability to focus on the most important components of the visual data. To alleviate this problem, we for the first time propose a 3D Channel SE block and a 3D Temporal SE block. Both blocks are integrated into the original residual module, forming an architecture specifically designed for NSCLC prognosis, which we term ProgSE-Net. Additionally, we observe that the pixels in adjacent CT slices are similar or identical in most cases. To further enhance ProgSE-Net, we propose a frame difference mechanism. This mechanism generates two additional CT slices by subtracting adjacent pixels in two directions, a strategy that has proven effective in our practice.

In conclusion, considering the above, we have developed the first multimodal network for NSCLC survival analysis, which takes Deep Learning-based NSCLC survival analysis one step forward by simultaneously considering the textual clinical parameters and the visual CT clues. As shown in Fig. 1, our network adopts a two-tower paradigm: a clinical tower and a visual tower. The clinical tower is responsible for encoding the clinical parameters, while the visual tower aims to extract the visual representation from the CT images. Finally, the prediction head fuses the cross-modal features to provide a time prediction.

In summary, the key contributions of Lite-ProSENet are as follows:

- The first application of a two-tower DNN for NSCLC survival time analysis, utilizing both structured data and CT images simultaneously.
- The first application of transformer and 3DSE-Net block to multimodal clinical parameters for disease prognosis.
- Results from benchmark and real-world clinical datasets demonstrate that Lite-ProSENet outperforms the state of the art (*SOTA*[1]) methods by a substantial margin.

The remainder of this paper is organized as follows: Section 2 presents related work on NSCLC survival analysis, covering both traditional methods and deep learning-based approaches. Section 3 details the proposed Lite-ProSENet. In Section 4, we discuss the experiments and ablation studies. Section 5 explores various choices made when building the network, including hyper-parameter tuning. Finally, Section 6 concludes the paper and outlines future work. The details of each section are provided below.

2. Related work

In this section, we give an overview of the traditional statistical methods and deep convolutional neural networks, then highlight the correlation to our contributions.

2.1. Statistical methods

Conventional statistical methods for NSCLC survival analysis only use the textual modality and involve modeling time to an event. They can be divided into three types: non-parametric, semi-parametric and parametric methods. Kaplan–Meier analysis (KM) [26] is a typical non-parametric approach to survival outcomes. KM Analysis is suitable for small data sets with a more accurate analysis cannot include multiple variables. Life table [27] is a simple statistical method that appropriate for large data sets and has been successfully applied to European lung cancer patients [28]. The Nelson-Aalen estimator (NA) [29] is a non-parametric estimator of the cumulative hazard function (CHF) for censored data. NA estimator directly estimates the hazard probability. As for semi-parametric method, the distribution of survival is not required. For example, the Cox regression model is used in [30], which

[1] SOTA refers to the best results on benchmark datasets.

discovered the critical factor that has a greater impact on survival analysis in lung cancer. The Cox proportional hazards model [27] is the most commonly used model in survival analysis and the baseline hazard function is not specified. Coxboost can be applied to high-dimensional data to fit the sparse survival models. Better than the regular gradient boosting approach (RGBA), coxboost can update each step with a flexible set of candidate variables [31]. The parametric method is easy to interpret and can provide a more efficient and accurate result when the distribution of survival time follows a certain distribution. But it leads to inconsistencies and can provide sub-optimal results if the distribution is violated. The Tobit model [32], for example, is one of the earliest attempts to extend linear regression with the Gaussian distribution for data analysis with censored observations. Buckley-James (BJ) regression [33,34] uses least squares as an empirical loss function and can be applied to high-dimensional survival data. BJ regression is an accelerated failure time model. Bayesian survival analysis [16,35,36] encodes the assumption via prior distribution.

2.2. DNN based methods

Image-based techniques for survival analysis of lung cancer normally adopt histopathological images. The work of [37] was the first to use a deep learning approach to classify cell subtypes. Using machine learning methods, H. Wang et al., proposed a framework [38] and found a set of diagnostic image markers highly correlated with NSCLC subtype classification. The work of Kun-Hsing Yu et al. [39], extracts 9879 quantitative image features and uses regularized machine learning methods to distinguish short-term survivors from long-term survivors. In the work of Xinliang Zhu et al. [40], a deep convolutional neural network for survival analysis (DeepConvSurv) with pathological images was proposed for the first time. The mentioned methods cannot learn discriminative patterns directly from Whole Slide Histopathological Images (WSIs) and some of them predict the survival status of patients using hand-crafted features extracted from manually labeled small discriminative patches. [41] introduced a novel multimodal learning framework that integrates WSIs and CT images for survival prediction, demonstrating superior performance compared to unimodal approaches. Despite the accurate patient information from WSIs like pathology images, they often include invasive steps. With this consideration, this paper solely focuses on the CT scans and clinical parameters from patient to conduct survival analysis, thereby remedying the physical and psychological stress of invasive models. Considering the scare and precious annotations for medical data, many researchers explored to build the computer-aided diagnosis system with imperfect labels [42–44]. In the work of [42], an annotation-free method for survival prediction based on whole slide histopathology images was proposed for the first time. Liao et al. proposed a novel multi-view "divide-and-rule" model to predict the lunge Nodule Malignancy with noisy labels [44]. In [43], Xie et al. utilized both the labeled and unlabeled data to learn a lung nodule classification model. In contrast, this paper solely focuses on survival analysis on a fully supervised setting, analyzing with imperfect labels is a study-worthy topic and we will explore it in our future works.

In summary, traditional statistical methods tend to use textual data with limited information. In recent years, with the development of deep learning, more work has begun to explore methods that use histopathology images. However, it is invasive to obtain the images. There is a work that uses CT images but with hand-crafted features that require instructions from medical experts. Moreover, all these methods only use single modality and ignore the complementary information that comes from multimodality. Therefore, to capture the underlying complex relationships between multimodality medical testing results and NSCLC survival time, we propose a non-invasive, fully automated DNN method to improve the prediction accuracy of NSCLC prognosis.

3. Methodology

The proposed method is a two-tower architectural model. In this section, we describe details within the model for NSCLC prognosis.

3.1. The architecture of Lite-ProSENet

Clinical parameters and visual CT images both contain rich information but lie in different spaces, as a result, the information from different modalities cannot be integrated directly to give a comprehensive representation. To perform an effective feature fusion and alignment, we devise our model as a two-tower architecture, whose effectiveness has been well validated in the cross-modality learning field [45–48] . Fig. 2 gives the overall illustration of our framework, the proposed Lite-ProSENet contains two towers, *i.e.,* Lite-Transformer and ProSENet. Given a piece of data d, which is composed by the clinical parameters c, the CT images I and survival time t, *i.e.,* $d = \{(c, I), t\}$. The clinical parameters c is first fed into an embedding layer to obtain the dense representation, and then pass through the light transformer to get the effective features.

CT images I are first fed into the ProSENet for the feature extraction. The following prediction module fuses the features from different modalities and give the survival prediction $\hat{t}$ based on the multimodality feature. Finally, the parameters of two towers are jointly optimized by minimizing the distance between the survival time prediction $\hat{t}$ and annotated one t. In the following, we will illustrate the details of each component of our network.

3.2. Light transformer

Light transformer in our Lite-ProSENet is a simple "lightweight transformer" comprising limited number of attention layers, aiming to efficiently process the clinical parameters.[2] As shown in Fig. 3, the raw items in clinical parameters are first fed into an embedding layer to get a dense representation, then, the dense representations are fed into the multi-head self-attention layers to get the clinical features.

Clinical Embedding. A piece of clinical parameters usually contains several items, $c_i (i = 1, 2, \dots, m.)$, where c_i is a kind of clinical item. To better represent the raw clinical embedding, we assign each clinical item a dense feature using the popular embedding technique. We first give the initial item representation by the one-hot encoding, and then a matrix is introduced to project the initial representation to a dense feature:

$$c_i = \text{one\_hot}(c_i) \times W \tag{1}$$

where one_hot(·) is the function that project the item to a one hot vector, $W \in R^{V \times d}$ is the learnable map weight, V is the item vocabulary size, and the d is the dimension of dense representation. For the sake of symbol simplicity, we still use c_i to denote the dense vector of item c_i.

Multi-Head Self Attention. We adopt multi-head self attention in our model, which allows the model to jointly attend to information from different representation subspaces at different positions. Multi-head attention is an extension of self-attention, but repeat the attention mechanism several times.

Each time, the transformer uses three different representations: the Queries, Keys and Values generate from the fully-connected layers. Fig. 4 illustrates the whole process of self-attention mechanism. Let $C \in R^{m \times d}$ be the matrix formed by the item embeddings of clinical parameters c, mathematically, the outputs S by the computation of self-attention can be expressed as: where $W_q, W_k, W_v \in R^{d \times h}$ are the

[2] Our light transformer is not a special efficient variants of transformer like Linformer [49], Performer [50], and Reformer [51], developing an efficient transformer is not the focus of this paper.

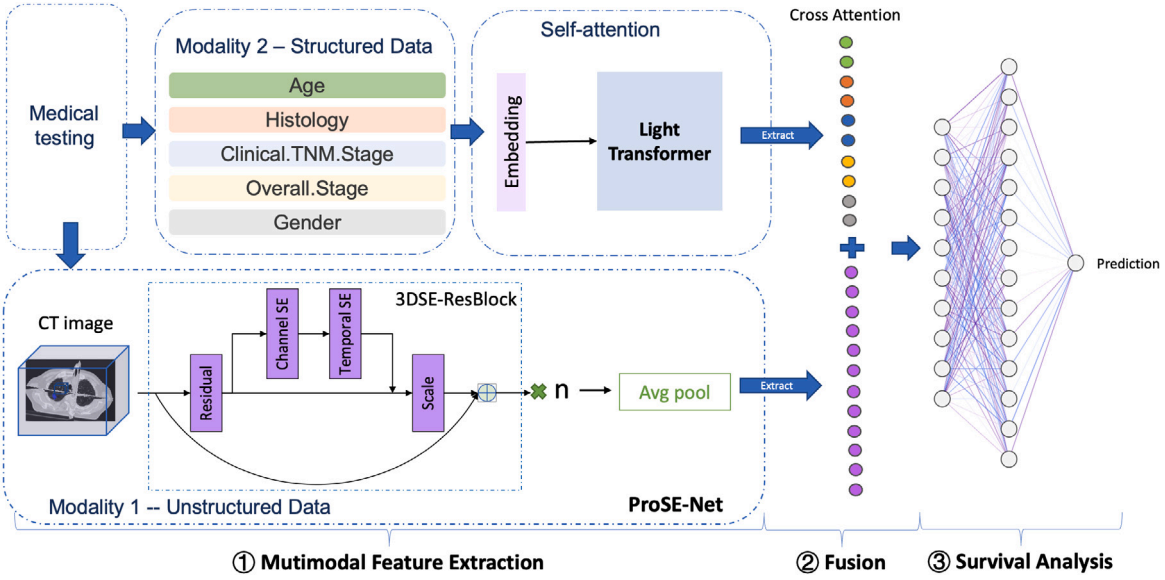


Fig. 2. A two-tower DNN model for learning similarity between textual clinical parameters and CT image representations.

learnable parameters, h is the embedding dimension. Taking the self-attention (SA) as the basic block, the multi-head self-attention (MSA) is given by repeating the SA several times, and the outputs from different heads are concatenated together. Finally, the architecture of our lite-transformer is given as follows:

$$A_k = MSA(LN(A_{k-1})) + A_{k-1}, \qquad k = 1, 2, ..K \tag{2}$$

$$A'_k = MLP(LN(A_k)) + A_k, \qquad k = 1, 2, \dots, K \tag{3}$$

$$T = LN(A'_K) \tag{4}$$

where $A_0 = C$, $LN(\cdot)$ is the layer normalization, T indicates the final clinical features, K is the total MHA layers.

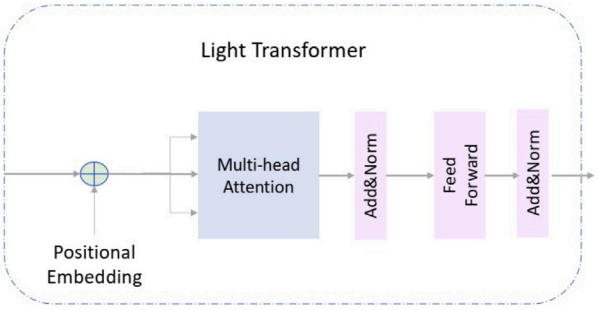


Fig. 3. The Light Transformer - model architecture.

3.3. ProSE-Net

ProSE-Net is the model that learns unstructured data representation through a 3DResnet based network with several repeatable 3DSE-ResBlocks. 3DSE-ResBlocks composes a residual block, followed by a Channel SE-block and a Temporal SE-block. Such a design can effectively improve the representational power of ProSENet. The key contributions of our ProSENet lie in the 3D channel SE block and temporal SE block, hereinafter, we will elaborate the details of these two modules.

Channel SE-block. Channel SE-block targets to produce a compact feature via a squeeze-and-excitation operation along the channel dimension. Let $F \in R^{f\times c\times h\times w}$ be an arbitrary feature map, channel SE-block first performs "Squeeze" operation:

$$p = \frac{1}{fhw}\sum_{i=1}^{f}\sum_{j=1}^{h}\sum_{k=1}^{w} F_{i,:,j,k} \tag{5}$$

where $p \in R^c$.

Excitation operation first introduces two full-connected layers to perform a interaction between different channels, and a sigmoid is introduced to produce a information filter:

$$g = \text{sigmoid}(W_1 \text{ReLU}(W_2 \times p)), \tag{6}$$

where $W_1 \in \mathbb{R}^{c\times\frac{c}{r}}, W_2 \in \mathbb{R}^{\frac{c}{r}\times c}$. $g \in R^c$ would serve as the gate to perform information selection and the channel feature in F would be updated as follow:

$$F^c = [F_{:,1,:,:} \times g_1, \dots, F_{:,c,:,:} \times g_c] \tag{7}$$

The common SE block only maintain the channel information and squeeze other dimensions, consequently, the important temporal information of 3D slices is also missed. To address this weakness, we augment the naïve SE block with a temporal excitation. First, the temporal dimension is remained when pooling the feature:

$$p^t = \frac{1}{hw}\sum_{j=1}^{h}\sum_{k=1}^{w} F_{j,k}, \tag{8}$$

$p^t \in R^{f\times c}$, we next produce a channel gate for each frame $g^t \in R^{f\times c}$, where the channel gate for ith frame is computed by Eq. (6):

$$g^t_{i,:} = \text{sigmoid}(W_1 \text{ReLU}(W_2 p^t_{i,:})). \tag{9}$$

we share the weights W_1 and W_2 when producing gates in each channel SE block. The goals for the weight sharing stems from two aspects, the first is to propagate the information inside different views, building a lighter network with fewer paramters is the second reason. In our practice, sharing parameters can also promote the performance.

Finally, we fuse two types of gates and develop our full channel SE block as follows:

$$F^c = [F_{1,1,:,:} \times G_{11}, F_{1,2,:,:} \times G_{12}, \dots, F_{f,c,:,:} \times G_{fc}] \tag{10}$$

$$G_{ij} = g^t_{ij} \times g_j \tag{11}$$

where G is the final gate filter, which is a combination of local view g^t and global g to perform a more reliable information filtering.

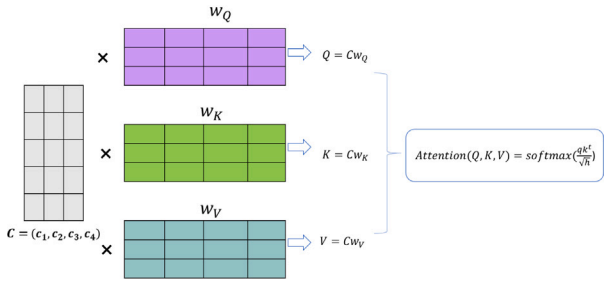
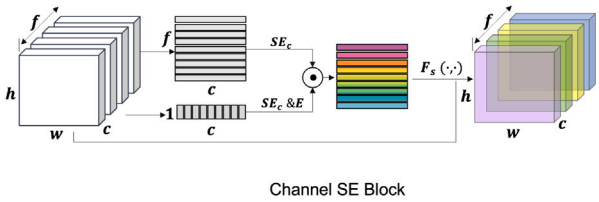
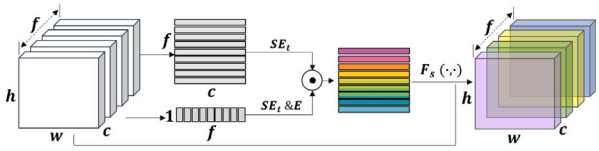


Fig. 4. The process of applying self-attention layer to the Query, Value and Key matrices.



(a) Channel Squeeze-and-Excitation block



(b) Temporal Squeeze-and-Excitation block

Fig. 5. Two types of Squeeze-and-Excitation blocks in our ProSENet.

$$q = CW_q, \quad k = CW_k, \quad v = CW_v,$$
$$S = \text{softmax}(\frac{qk^T}{\sqrt{h}}), \qquad A = Sv. \tag{12}$$

Temporal SE-block Similar with Channel SE-block, temporal SE-block also targets to filter out the redundant information but focuses on the temporal dimension. As shown in Fig. 5(b), the computation procedure is analogous with the channel SE-block. The pooling is down along the channel-spatial dimension and spatial dimension, which provide the global and local frame information, respectively. Then, two types of gates are similarly produced following Eqs. (6) and (9), which are then fused via Eq. (11) and give the joint gate G'. Finally, the temporal SE-block is formulated as:

$$F^t = [F^c_{1,1,:,:} \times G'_{11}, F^c_{1,2,:,:} \times G'_{12}, \dots, F^t_{f,c,:,:} \times G'_{fc}]. \tag{13}$$

In our 3D SE block, the channel SE and temporal SE are stacked to achieve the information filtering along the channel and temporal dimensions, which forms our entire 3D SE-Resblock with the well-known 3D Resblock. The CT slices are first fed into the ProSENet to extract multi-dimension features and then pass through a 3D global average pooling to get the final feature F_I.

3.4. Multimodal feature fusion and prediction

Given the clinical features T from lite transformer and CT image feature F_I from ProSENet, the next task is to fuse the multi-modality features and give the prediction of survival time $\hat{t}$. Thanks to the powerful features from our Lite transformer and ProSENet, we can simply concatenate the cross-modal features and predict the survival time using a MLP, and a encouraging performance can be harvested in our practice:

$$\hat{t} = MLP(concat([T, F_I])), \tag{14}$$

where the $MLP(\cdot)$ is a two-layer full-connected layers, $concate(\cdot,\cdot)$ performs concatenation for the input two vectors.

Enhance Prediction via Frame Difference. We observe that although the CT images contain rich information, there are so many duplicated pixels between the CT slices, hindering the ProSENet to perceive the key information among the CT slices. To remedy this issue, we propose a simple yet effective mechanism, *i.e.,* frame difference. The proposed frame difference performs a subtraction between two consecutive slices, such that the duplicated pixel could be ignored in the resulted slice. Following this idea, we perform the frame difference along two directions: forward and backward, the produced CT images are marked as I^f, I^b, respectively. Given this, our visual information are contain three types, *i.e.,* the raw data I, frame difference along forward and backward direction I^f and I^b. We then feed each of the visual information and the clinical parameters into our Lite-ProSENet, consequently, three time prediction could be given. Finally, we integrate the three predictions to produce the final result:

$$\bar{t} = \omega\hat{t} + (1-\omega)\frac{\hat{t}_f + \hat{t}_b}{2} \tag{15}$$

where $\hat{t}_f$ and $\hat{t}_b$ are the survival prediction from the (I^f, c) and (I^b, c), respectively, ω is the trade-off weight, and $\bar{t}$ is the final prediction of survival time.

3.5. Network optimization

With the final prediction $\bar{t}$ and the annotated survival time t, the network parameters are learned by minimizing the distance between the prediction and the annotation:

$$\mathcal{L} = \frac{1}{b}\sum_{i=1}^{b}(\bar{t} - t)^2 + \lambda\|\mathcal{W}\|_2, \tag{16}$$

where b is the batch size during training, $\|\cdot\|_2$ is the l_2 normalization, the second penalty is the parameter normalization, which is introduced to avoid overfitting, W is all of the network parameters, and λ is the trade-off hyper-parameter.

4. Experiments

We conduct extensive experiments based on NSCLC patients from TCIA to validate the performance of our proposed method with several state-of-the-art methods in terms of the prediction accuracy for the survival time for each patient. Besides, we also evaluate the prediction result by concordance. Afterward, we perform several ablation experiments regarding different network structures to determine the best structure.

4.1. Dataset

In this work, we considered 422 NSCLC patients from TCIA to assess the proposed framework. For these patients pretreatment CT scans, manual delineation by a radiation oncologist of the 3D volume of the gross tumor volume and clinical outcome data are available [52]. The corresponding clinical parameters are also available in the same collection. The patients who had neither survival time nor event status were excluded from this work.

4.2. Data preprocessing

For CT images, we resize the raw data which is the 3D volume of the primary gross tumor volume into $96 \times 96 \times 8$. After that, we transform the range linearity into [0,1]. Then, to prevent overfitting, we perform data argumentation which includes three methods: rotate, swap, and flip. Then we get $422 \times 8 = 3376$ samples, among which there are $373 \times 8 = 2984$ uncensored samples and $49 \times 8 = 392$ censored samples.

Clinical parameters contain categorical data and non-categorical data. The detailed distribution and description of the data used is

Table 1
Clinical parameters description.

	Histology	TNM stage grouping	Clinical T stage	Clinical M stage	Clinical N stage
Categories	Squamous cell carcinoma Large cell Not otherwise specified(Nos) Adenocarcinoma	Stage I Stage II Stage IIIa Stage IIIb	T1 T2 T3 T4 T5	M0 M1 M3	N0 N1 N2 N3 N4
Details	Including the major histological subtypes of NSCLC. Histology in the context of NSCLC refers to the microscopic examination of tissue to determine the specific subtype of the cancer. This information is crucial for guiding treatment decisions and prognosis.	The TNM staging system is a method used to classify the extent of cancer spread in an individual's body, based on three key components	T(Tumor): Refers to the size and extent of the primary tumor. For example: T0 means no evidence of primary tumor, while T4 indicates a very large tumor or one that has invaded certain critical structures.	M(Metastasis): Specifies whether the cancer has metastasized to distant sites in the body. For example: N0 means no regional lymph node involvement, whereas N3 indicates extensive lymph node involvement.	N(Node): Indicates whether the cancer has spread to nearby lymph nodes. For example: M0 means no distant metastases, and M1 indicates distant metastasis.

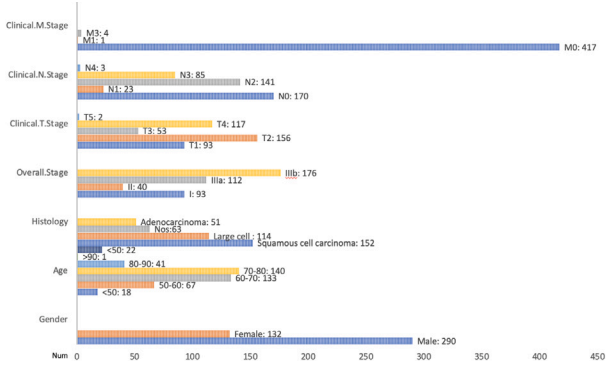


Fig. 6. The clinical parameters distribution.

shown in Fig. 6 and Table 1. Firstly, missing value is a common problem in medical data and may pose difficulties for data analyzing and modeling. Specifically, in our dataset, the 'age' category contains a few missing values. After observing the data, we find that the age of patients in the dataset is close to each other. Thus, we impute the mean value and fill it into the missing value. Afterward, in order to fit into our model, we use the one-hot encoder to encode categorical data into numbers, which allows the representation of categorical data to be more expressive.

Then, we use the min–max feature scaling method and standard score method to perform data normalization, such as age and survival time. For input x, the min–max feature scaling method's output is:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{17}$$

and the standard score method's output is:

$$x' = \frac{x - mean(x)}{std(x)} \tag{18}$$

where std is the standard deviation.

For a single patient with multiple tumors, we select the primary gross tumor volume ('GTV-1') to be processed in our work, while other tumors such as secondary tumor volumes denoted as 'GTV2', 'GTV3' to name just a few, which were occasionally present, were not considered in our work.

4.3. Experiment setup

We train and evaluate the framework on the NSCLC-Radiomic dataset following 5-fold cross-validation with the patient-level split. We divide the dataset into training, validation, and testing data into 6:2:2 respectively. Specifically, the overall dataset was split into five parts, each part contains 20% of the dataset, and 3 of 5 splits are for training, validation and test both use 1 split.

In Lite-transformer, the number of head in MHA is set as 3, and the total layers K is configured as 5, more layers and heads bring limited performance gain but large parameters in our practice. In ProSENet, the ratio of channel and temporal SE are both set as 2, *i.e.*, $r = 2$. For hyperparameters tuning such as the penalty coefficient, we use the validation dataset to fine-tune and get the optimized hyperparameters. The configuration of ω in Eq. (15) and the λ in Eq. (16) is set to 0.4 and 0.001 respectively, to balance the contributions of different terms. In the training process, we use 800 epochs in total with Adam as the optimizer. The batch size parameter is set as 64. The initial learning rate is set as 0.001, then the learning rate is decayed by 0.5 in every 40 epochs.

Since we use survival time as the label, not cumulative hazard. In the training and validation process, we only use the uncensored data for precise survival time and objective function calculation, and in the testing process, we use all data for concordance evaluation and uncensored data for MAE evaluation.

We include several SOTA survival analysis methods as baselines to compare with our work, including Cox-time [53], DeepHit [54], CoxCC [53], PC-Hazard [19] and the regular cox regression.

4.4. Quantitative results

In this subsection, we make a thorough comparison with both traditional and recent deep learning-based methods. The quantitative results of C-index and MAE are compared in Table 2. Noted the results reported here are directly derived from the original paper respectively.

As shown in Table 2, all the comparison methods except our previous work DeepMMSA only use the clinical parameters or the CT slices for prediction. For example, by building a survival function, Cox-regression can provide the probability that a certain event (e.g. death) occurs at a certain time t, the C-index of Cox- regression is only 0.601. In contrast, many experiments based on Deep Learning only use the visual information from CT scans. Although deep convolutional neural networks (DCNNs) are very powerful in feature extraction, the visual information alone is not reliable enough to accurately predict survival

Table 2
C-index and MAE comparison between Lite-ProSENet and comparison methods.

Methods	Invasive	DNN-based	Modality		Performance	
			Textual	Visual	C-index↑	MAE↓
Cox-time [53]	–	–	✓	–	0.6152	0.183
Cox-regression [27]	–	–	✓	–	0.6009	0.204
CoxCC [53]	–	–	✓	–	0.6120	0.183
PC-Hazard [19]	–	–	✓	–	0.191	0.6094
DeepHit [54]	–	–	✓	–	0.6133	0.183
DeepMMSA [25]	–	✓	✓	✓	0.6580	0.162
LASSO-Cox [55]	–	–	✓	–	0.6698	NA
Cox + SuperPC [56]	✓	–	–	✓	0.556	NA
Log–logistic [57]	–	–	✓	–	0.5924	NA
BJ-EN [58]	–	–	✓	–	0.6646	NA
RSF [59]	–	–	✓	–	0.595	NA
MTLSA.V2 [60]	–	–	✓	–	0.680	NA
BoostCI [61]	–	–	✓	–	0.6497	NA
WSISA [42]	✓	✓	–	✓	0.703	NA
DeepSurv [62]	–	–	✓	–	0.602	NA
DeepConvSurv [63]	✓	✓	–	✓	0.629	NA
DFS [64]	–	✓	–	✓	0.673	0.166
FMCIB [65]	–	✓	–	✓	0.673	0.166
Lite-ProSENet		✓	✓	✓	0.893	0.043

time. For example, the best C-index of deep learning-based methods only use visual CT is 0.703 [42]. Our previous work, DeepMMSA [25] makes the first attempts to fuse the multimodal data using a two-tower framework. Although we found that multimodal inputs could boost the performance, the final results do not surpass the deep learning based methods using only visual information such as WSISA [42]. This observation indicates that the straightforward network cannot work well for multimodal fusion. Consequently, we developed our Lite-ProSENet to build an effective multimodal network for survival analysis. Our Lite-ProSENet was able to achieve a C-index of 0.893, outperforming all comparative methods, which well validate the superiority of our method.

4.5. Ablation study

To build an effective cross-modal survival model, we design our Lite transformer for clinical parameters and propose the 3D- SE Resblock to effectively model the visual CT slices. Furthermore, we propose a frame difference mechanism to promote our performance to the new state-of-the-art. In this subsection, we will verify the effectiveness of the above modules to support our claims through extensive experiments.

The results are reported in Table 3, where we systematically examine the contribution of each component, including the Lite-Transformer, the 3D- SE Resblock in ProSENet, and the mechanism of frame difference. In the baseline method (no modules are equipped), the Lite-Transformer is replaced by several MLP layers to form a similar parameters. As is shown in Table 3, the C-index of the baseline method is only 0.796, and the C-index improves when each module is equipped. For example, the baseline with Lite-Transformer could achieve a C-index of 0.824, and the 3D-SE Resblock helps the baseline to improve the C-index from 0.796 to 0.841. Applying any two modules simultaneously could improve the performance even further. If we apply 3D-SE Resblock and frame difference, we could attain the C-index of 0.873, which is a significant improvement. When all of three modules are configured, we harvest the best performance, whose C-index could reach the new state-of-the-art 0.893. The observation on MAE shows a consistent tendency.

As one of our main motivations for the Lite-ProSENet design, verifying the effectiveness of multi-modality modeling is also a critical aspect. We also investigate the benefits of multi-modality learning from this aspect. The results are also reported in Table 3, where Lite-ProSENet$_V$ and Lite-ProSENet$_T$ refer to the Lite-ProSENet with visual tower and textural tower, respectively. We can observe that the network with any tower alone could not achieve satisfactory performance, the visual tower only achieves a C-index of 0.712. Although the 3D- SE block boosts the performance to 0.739, it is still not satisfactory. The observations of Lite-ProSENet$_T$ are also conclusive. The model with multi-modality learning could achieve a C-index of 0.796, which well demonstrates the importance of fusing the clinical parameters and the visual CT images for the survival time analysis.

5. Discussion

In this section, we will give several discussions about the many choices when building our network, including the effect of the joint gate in our 3D SEResblock, the order of two SE blocks, the impact of the bi-directional frame difference. Besides the choices of several mechanisms, the hyper-parameters, ω in Eq. (15) and λ in Eq. (16), are also presented in this section.

5.1. Validate the joint gate in 3D SEResblock.

In the 3D SEReslock, we augment the channel SE and the temporal SE with the joint gate to perform the information filtering, more details can be found in Section 3.3. In this subsection, we would validate the effectiveness of our proposed joint gate.

The results are reported in Table 4, we set the baseline as the network where visual tower is the naïve 3D Resnet, 'global SE' refers to the gate is only built by the naïve SE block. For channel SE, the output of global SE is produced by the Eq. (7).[3] 'local SE' indicates the gate is only build by the channel-wise or frame-wise information, for channel SE, the output of global SE is produced by replacing the g in Eq. (7) with g^t defined in Eq. (9). Joint gate uses both the global and local SE block, *i.e.,* our 3D SEResblock. As we can observed from Table 4, SE block is an effective module, the system benefits from both types of SE block. For channel SE block, when equipping the global SE block, the C-index is improved from 0.826 to 0.842, and the local SE block also boosts the perform from 0.826 to 0.867. When the joint gate is applied, the performance gets a significantly improvement, from 0.826 to 0.893. The observation of temporal SE block is also conclusive, which well validates the effectiveness of 3D-SE Resblock.

[3] When studying the channel SE (temporal SE), we equip the full temporal SE block (channel SE).

Table 3
Discuss the effectiveness of slices from frame difference, '✓' and '–' means applying and not applying the corresponding modules.

	Lite-transformer	3D-SE resblock	Frame difference	C-index↑	MAE↓
Lite-ProSENet	–	–	–	0.796	0.121
	✓	–	–	0.824	0.108
	–	✓	–	0.841	0.092
	–	–	✓	0.837	0.103
	✓	✓	–	0.859	0.086
	–	✓	✓	0.873	0.063
	✓	–	✓	0.862	0.071
	✓	✓	✓	0.893	0.043
Lite-ProSENet$_V$	–	–	–	0.712	0.223
	–	✓	–	0.743	0.187
Lite-ProSENet$_T$	–	–	–	0.739	0.181
	✓	–	–	0.761	0.149

Table 4
Discuss the importance of the global and local SE module in our ProSENet, '✓' and '–' means applying and not applying the corresponding modules.

	SE blocks	Global SE	Local SE	C-index↑	MAE↓
baseline	Channel SE	–	–	0.826	0.058
		✓	–	0.842	0.051
		–	✓	0.867	0.047
		✓	✓	0.893	0.043
	Temporal SE	–	–	0.819	0.061
		✓	–	0.839	0.053
		–	✓	0.862	0.049
		✓	✓	0.893	0.043

Table 5
The performance comparison of different stacking orders of channel SE and temporal SE is presented, where I and II indicate ranking in the first and second places, respectively. The symbols '✓' and '–' carry the same meanings as in Table 4.

	Channel SE	Temporal SE	C-index↑	MAE↓
Lite-ProSENet	✓	–	0.871	0.058
	–	✓	0.879	0.055
	II	I	0.881	0.049
	I	II	0.893	0.043

Table 6
Discuss the effectiveness of slices from frame difference, '✓' and '–' means applying and not applying the corresponding modules.

	Forward	Backward	C-index↑	MAE↓
Lite-ProSENet	–	–	0.854	0.058
	✓	–	0.881	0.046
	–	✓	0.879	0.045
	✓	✓	0.893	0.043

5.2. Study the stacking order of two SE blocks.

In our 3D-SE Resblock, the channel SE is applied first, and the temporal acts on the output of the channel SE block, as shown in Eqs. (11) and (13). In this subsection, we study the difference in performance between two SE blocks in different stacking order.

The performance comparison is given in Table 5, we study two types of stacking order, *i.e.,* channel SE first and then temporal SE second, and temporal SE first and then channel SE second. As shown in Table 5, the strategy of the channel SE first and the temporal SE second performs better. The C-index of channel SE first could reach 0.893, while the temporal SE first is worse, whose c-index is 0.881. Consequently, we first apply the channel SE in our network to achieve a better C-index. In addition to stacking order, in this subsection, we also investigate the importance of two SE blocks. As shown in Table 5, using the channel SE or the temporal SE performs alone performs worse than using two SE blocks simultaneously with arbitrary stacking order, which validating the effectiveness of our channel and the temporal SE blocks.

5.3. The effectiveness of the bi-directional frame difference.

When predicting the final survival time, we introduce frame difference to filter out the redundant information between different CT slices. To further boost the performance, we perform bidirectional frame difference among CT images. In this subsection, we discuss the effectiveness of our bidirectional frame difference.

To thoroughly validate the effectiveness of the proposed frame difference, we study three cases, *i.e.*, only the frame difference along forward direction and backward direction, and the bi-directional frame difference. The results can be found in Table 6, where the 'forward' and 'backward' mean the normal direction and the reverse direction, respectively. From Table 6, we can observe that both the 'forward' and 'backward' frame difference can promote the performance. When introducing the forward frame difference, the C-index gets improved from 0.854 to 0.881, and the backward frame difference can boost the C-index from 0.854 to 0.879. When we integrate the frame difference simultaneously in the forward and backward directions, we get the best C-index of 0.893. These observations reveal that our proposed frame difference is an effective mechanism.

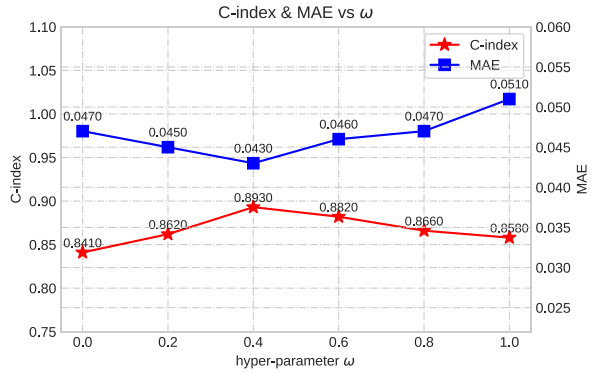
5.4. Discussion about hyper-parameter ω

We introduce a trade-off parameter when integrating the prediction of normal CT slices and the bi-directional frame difference, as shown in Eq. (15). The ω is set as 0.4 by default. In this subsection, we would study the performance change when the hyper-parameter ω varied.

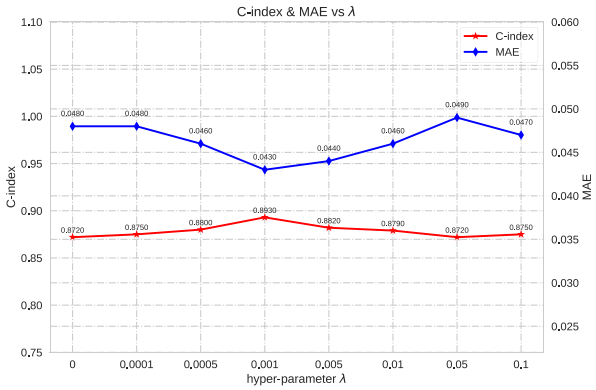
Fig. 7(a) shows the performance of our network when ω varied from 0 to 1 with step 0.2. We can observe from Fig. 7(a) that the c-index and MSE loss show consistent trendency. When $\omega = 0$, this means we only predict by the slices of frame difference and ignore the normal CT slice, this case does not achieve a satisfactory performance whose c-index is only 0.841, the reason for this observation may stem from too much information missed when completely abandoning the original CT slice. This guess is validated by the cases of $\omega \neq 0$. The case of $\omega = 1$ means we does not apply the frame difference, this case also fails to achieve the best performance, revealing that the slices of frame difference are necessary for our network. The case of $\omega = 0.4$ achieves the best performance, this means that slices from frame difference play an important role in the survival time prediction. Further enlarging the weight of frame difference does not promote the performance. Therefore, we fix ω as 0.4 in our network.

5.5. Discussion about hyper-parameter λ.

When training the network, we employ the popular parameter normalization strategy to avoid overfitting, *i.e.,* the second term in Eq. (16), and introduce a hyperparameter λ to balance the main loss and the parameter normalization. By default, we set λ to 0.001. In this section, we will study the impact of the hyper-parameter λ on the



(a) The performance comparison when ω varies from 0 to 1.



(b) Performance comparison when λ varies from 0 to 0.1.

Fig. 7. Discussion about the hyper-parameters ω in Eq. (15) and λ in Eq. (16).

performance.

The changes of c-index and MSE loss are shown in Fig. 7(b), where the y-axis represents performance and the x-axis represents λ. We study the performance under $\lambda = \{0, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$. From Fig. 7(b), we can observe a clear trend. When $\lambda = 0$, which means that we dispense with parameter normalization, the network does not achieve good performance. Then, when we increase λ, the performance starts to increase. When $\lambda = 0.001$, we were able to achieve the best c-index 0.893. If we increase λ further, we cannot obtain more performance gains. In the case of $\lambda = 0.4$, a good trade-off between the main loss and the parameter normalization is achieved.

6. Conclusion and future work

This work contributes a powerful multimodal network for more accurate prediction of NSCLC survival, aimed at assisting clinicians in developing timely treatment plans and improving patients' quality of life. Our method achieves a new state-of-the-art result with an 89.3% on the C-index. To effectively model cross-modal data, we develop a two-tower network: the textual tower processes clinical parameters, and the visual tower handles CT slices. Inspired by the success of the transformer in the NLP field, we propose a lightweight transformer leveraging the core of self-attention mechanisms. For the visual tower, we design a ProSENet based on the 3D-SE Resblock, where channel Squeeze-and-excitation and temporal Squeeze-and-excitation are proposed to suppress the redundant information among the CT slices. Besides, we further introduce a frame difference mechanism to help promote our network up to the new state-of-the-art in terms of both C-index and MAE. In experiments, we conduct comprehensive comparisons, ablation studies, and discussions, all of which verify the superiority of our Lite-ProSENet. The practice of this work bolster our confidence in deep learning-based survival analysis. We believe that the deep learning-based method holds significant potential for survival time analysis. In the future, we will further investigate this problem from the following two aspects:

- **Effective fusion of cross-modal features.** In this work, the fusion of multimodal features is straightforward; we simply concatenate the features from Lite-transformer and ProSENet. In the future, we aim to explore more advanced and effective fusion strategies.
- **Leveraging information from large-scale pretrained models.** Large-scale pretrained cross-modal models have shown great potential in various tasks, such as Visual question answering, images captioning, and cross-media retrieval, et al. After training with millions of data, the large-scale models contain powerful knowledge, how to adapt these knowledge to survival time analysis is a promising direction. We plan to explore this direction in the future.
- **Survival analysis with noisy labels.** In the future, we also plan to explore survival analysis with noisy labels [43,44]. This direction is motivated by the challenge of acquiring accurate labels, which are often difficult and costly to obtain. Developing robust methods that can effectively handle label noise will enhance the applicability and efficiency of survival analysis in real-world scenarios where label acquisition is limited.

CRediT authorship contribution statement

Yujiao Wu: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yaxiong Wang:** Writing – review & editing, Methodology, Conceptualization. **Xiaoshui Huang:** Software, Methodology. **Haofei Wang:** Software, Resources. **Fan Yang:** Supervision, Conceptualization. **Wenwen Sun:** Validation. **Sai Ho Ling:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Steven W. Su:** Writing – review & editing, Resources, Project administration, Conceptualization.

Declaration of competing interest

None Declared.

Acknowledgment

The authors would like to thank all the co-authors who contributed their invaluable expertise and support to the successful completion of this study.

Data availability

Data will be made available on request.

References

[1] William D. Travis, Pathology of lung cancer, Clin. Chest Med. 23 (1) (2002) 65–81.
[2] Stephen G. Spiro, Gerard A. Silvestri, One hundred years of lung cancer, Am. J. Respir. Crit. Care Med. 172 (5) (2005) 523–529.
[3] E.J. Feuer, N. Howlader, A.M. Noone, M. Krapcho, J. Garshell, D. Miller, et al., National cancer institute SEER cancer statistics review, Natl. Cancer Inst. 103 (7) (2015) 1975–2012.
[4] John D. Minna, Jack A. Roth, Adi F. Gazdar, Focus on lung cancer, Cancer Cell 1 (1) (2002) 49–52.

[5] Cesare Gridelli, Antonio Rossi, David P. Carbone, Juliana Guarize, Niki Karachaliou, Tony Mok, Francesco Petrella, Lorenzo Spaggiari, Rafael Rosell, Non-small-cell lung cancer, Nat. Rev. Dis. Prim. 1 (1) (2015) 1–16.

[6] Roy S. Herbst, Daniel Morgensztern, Chris Boshoff, The biology and management of non-small cell lung cancer, Nature 553 (7689) (2018) 446–454.

[7] Yen-Tsung Huang, Rebecca S. Heist, Lucian R. Chirieac, Xihong Lin, Vidar Skaug, Shanbeh Zienolddiny, Aage Haugen, Michael C. Wu, Zhaoxi Wang, Li Su, et al., Genome-wide analysis of survival in early-stage non–small-cell lung cancer, J. Clin. Oncol. 27 (16) (2009) 2660–2667.

[8] Antonio L. Visbal, Brent A Williams, Francis C. Nichols III, Randolph S. Marks, James R. Jett, Marie-Christine Aubry, Eric S. Edell, Jason A. Wampfler, Julian R. Molina, Ping Yang, Gender differences in non–small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002, Ann. Thorac. Surg. 78 (1) (2004) 209–215.

[9] Se-Jun Park, Chong-Suh Lee, Sung-Soo Chung, Surgical results of metastatic spinal cord compression (MSCC) from non–small cell lung cancer (NSCLC): analysis of functional outcome, survival time, and complication, Spine J. 16 (3) (2016) 322–328.

[10] David R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B Stat. Methodol. 34 (2) (1972) 187–202.

[11] C. Silambarasanand, R. Elangovan, Accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data, 2020.

[12] Edward L. Kaplan, Paul Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (282) (1958) 457–481.

[13] T.M. Hoàng, V.L. Parsons, Bagging survival trees for prognosis based on gene profiles, Compstat — Proc. Comput. Stat. (2004) 1201–1208.

[14] H. Ishwaran, U.B. Kogalur, RandomSurvivalForest: Random survival forests, 2013.

[15] Rber Ziegel, Bayesian survival analysisby Joseph G. Ibrahim; Ming-Hui Chen; Debajyoti Sinha, Technometrics 44 (2) (2002) 201.

[16] K.H. Lee, Bayesian variable selection in parametric and semiparametric high dimensional survival analysis, 2011.

[17] P.K. Shivaswamy, W. Chu, M. Jansche, A support vector approach to censored targets, in: Icdm, 2008.

[18] F.M. Khan, V.B. Zubek, Support vector regression for censored data (SVRc): A novel tool for survival analysis, in: IEEE, 2008.

[19] Håvard Kvamme, Ørnulf Borgan, Continuous and discrete-time survival prediction with neural networks, 2019, arXiv preprint arXiv:1910.06724.

[20] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[21] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, Clara I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[22] Fuyong Xing, Yuanpu Xie, Hai Su, Fujun Liu, Lin Yang, Deep learning in microscopy image analysis: A survey, IEEE Trans. Neural Networks Learn. Syst. 29 (10) (2017) 4550–4568.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. (IJCV) 115 (3) (2015) 211–252.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[25] Yujiao Wu, Jie Ma, Xiaoshui Huang, Sai Ho Ling, Steven Weidong Su, Deep-MMSA: 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 2021, Melbourne, Australia, October 17-20, 2021, IEEE, 2021, pp. 1468–1472.

[26] W.N. Dudley, Ra Rita Wickham, M.N. Coombs, An introduction to survival statistics: Kaplan-Meier analysis, J. Adv. Pr. Oncol. 7 (1) (2016) 91–100.

[27] By D.R. Cox, Regression models and life-tables, J. R. Statal Soc. 34 (2) (1972).

[28] M.L.G. Janssen-Heijnen, G. Gatta, D. Forman, R. Capocaccia, J.W.W. Coebergh, EUROCARE Working Group, et al., Variation in survival of patients with lung cancer in Europe, 1985–1989, Eur. J. Cancer 34 (14) (1998) 2191–2196.

[29] R.S. Singh, D.P. Totawattage, The statistical analysis of interval-censored failure time data with applications, Open J. Stat. 03 (2) (2013) 155–166.

[30] Jeffrey L. Port, Michael S. Kent, Robert J. Korst, Daniel Libby, Mark Pasmantier, Nasser K. Altorki, Tumor size predicts survival within stage IA non-small cell lung cancer, Chest 124 (5) (2003) 1828–1833.

[31] H. Binder, M. Schumacher, Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models, Bmc Bioinform. 9 (1) (2008) 14.

[32] J. Tobin, Estimation of relationships for limited dependent variables, 1956.

[33] J. Buckley, I. James, Linear regression with censored data, Biometrika 66 (3) (1979) 429–436.

[34] S. Wang, B. Nan, J. Zhu, D.G. Beer, Blackwell Publishing Inc., in: Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates, (No. 1) 2008.

[35] Joseph G. Ibrahim, Chen Debajyoti Sinha, Bayesian semiparametric models for survival data with a cure fraction, Biometrics 57 (2) (2015) 383–388.

[36] K.H. Lee, S. Chakraborty, J. Sun, Survival prediction and variable selection with simultaneous shrinkage and grouping priors, Stat. Anal. Data Min. 8 (2) (2015) 114–127.

[37] Jiawen Yao, Sheng Wang, Xinliang Zhu, Junzhou Huang, Imaging biomarker discovery for lung cancer survival prediction, in: Sébastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gözde B. Ünal, William M. Wells III (Eds.), MICCAI, in: Lecture Notes in Computer Science, vol. 9901, 2016, pp. 649–657.

[38] Hongyuan Wang, Fuyong Xing, Hai Su, Arnold Stromberg, X. Arnold, Lin Yang, Novel image markers for non-small cell lung cancer classification and survival prediction, BMC Bioinformatics (2014).

[39] Kun-Hsing Yu, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, Daniel L. Rubin, Michael Snyder, Predicting non-small cell lung cancer diagnosis and prognosis by fully automated microscopic pathology image features, in: AMIA, AMIA, 2017.

[40] Xinliang Zhu, Jiawen Yao, Junzhou Huang, Deep convolutional neural network for survival analysis with pathological images, in: Tianhai Tian (Ed.), IEEE BIBM, IEEE Computer Society, 2016, pp. 544–547.

[41] Zhe Li, Yuming Jiang, Mengkang Lu, Ruijiang Li, Yong Xia, Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution, IEEE Trans. Med. Imaging 42 (9) (2023) 2678–2689.

[42] X. Zhu, J. Yao, F. Zhu, J. Huang, WSISA: Making survival prediction from whole slide histopathological images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[43] Yutong Xie, Jianpeng Zhang, Yong Xia, Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT, Med. Image Anal. 57 (2019) 237–248.

[44] Zehui Liao, Yutong Xie, Shishuai Hu, Yong Xia, Learning from ambiguous labels for lung nodule malignancy prediction, IEEE Trans. Med. Imaging 41 (7) (2022) 1874–1884.

[45] X. Yi, J. Yang, L. Hong, D.Z. Cheng, E. Chi, Sampling-bias-corrected neural modeling for large corpus item recommendations, in: The 13th ACM Conference, 2019.

[46] P. Neculoiu, M. Versteegh, M. Rotaru, Learning text similarity with siamese recurrent networks, in: Repl4NLP Workshop At ACL2016, 2016.

[47] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, Smart reply: Automated response suggestion for Email, ACM (2016).

[48] Y. Yang, S. Yuan, D. Cer, S.Y. Kong, R. Kurzweil, Learning semantic textual similarity from conversations, in: Proceedings of the Third Workshop on Representation Learning for NLP, 2018.

[49] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, Hao Ma, Linformer: Self-attention with linear complexity, 2020, arXiv preprint arXiv:2006.04768.

[50] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al., Rethinking attention with performers, 2020, arXiv preprint arXiv:2009.14794.

[51] Nikita Kitaev, Łukasz Kaiser, Anselm Levskaya, Reformer: The efficient transformer, 2020, arXiv preprint arXiv:2001.04451.

[52] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al., The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (6) (2013) 1045–1057.

[53] Håvard Kvamme, Ørnulf Borgan, Ida Scheel, Time-to-event prediction with neural networks and Cox regression, J. Mach. Learn. Res. 20 (129) (2019) 1–30.

[54] Changhee Lee, William Zame, Jinsung Yoon, Mihaela van der Schaar, Deephit: A deep learning approach to survival analysis with competing risks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (no. 1) 2018.

[55] Robert Tibshirani, The Lasso method for variable selection in the Cox model, Stat. Med. (1997) 385–395.

[56] Eric Bair, Trevor Hastie, Debashis Paul, Robert Tibshirani, Prediction by supervised principal components, J. Amer. Statist. Assoc. 101 (473) (2006) 119–137.

[57] E.T. Lee, J. Wang., Statistical Methods for Survival Data Analysis, vol. 476, Wiley. Com, 2003.

[58] Zhu Wang, C.Y. Wang, Buckley-james boosting for survival analysis with high-dimensional biomarker data, 2010, boosting, accelerated failure time model, Buckley-James estimator, censored survival data, LASSO, variable selection.

[59] Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, Andy J. Minn, Random survival forests for high-dimensional data, Stat. Anal. Data Min. 4 (1) (2011) 115–132.

[60] Yan Li, Jie Wang, Jieping Ye, Chandan K. Reddy, A multi-task learning formulation for survival analysis, in: Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, Rajeev Rastogi (Eds.), ACM SIGKDD, ACM, 2016, pp. 1715–1724.

[61] Andreas Mayr, Matthias Schmid, Boosting the concordance index for survival data – A unified framework to derive and evaluate biomarker combinations, PLoS One 9 (1) (2014) e84483.

[62] Jared Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, Yuval Kluger, Deep survival: A deep Cox proportional hazards network, 2016, CoRR abs/1606.00931.

[63] Xinliang Zhu, Jiawen Yao, Junzhou Huang, Deep convolutional neural network for survival analysis with pathological images, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2016, pp. 544–547.

[64] Benedetta Gottardelli, Varsha Gouthamchand, Carlotta Masciocchi, Luca Boldrini, Antonella Martino, Ciro Mazzarella, Mariangela Massaccesi, René Monshouwer, Jeroen Findhammer, Leonard Wee, et al., A distributed feature selection pipeline for survival analysis using radiomics in non-small cell lung cancer patients, Sci. Rep. 14 (1) (2024) 7814.

[65] Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, et al., Foundation model for cancer imaging biomarkers, Nat. Mach. Intell. 6 (3) (2024) 354–367.