PLOS ONE

RESEARCH ARTICLE

Improving lung cancer diagnosis and survival prediction with deep learning and CT imaging

Xiawei Wang[1], **James Sharpnack**[2], **Thomas C.M. Lee**[1,2]*

1 Graduate Group in Biostatistics, University of California - Davis, Davis, California, United States of America, **2** Department of Statistics, University of California - Davis, Davis, California, United States of America

* tcmlee@ucdavis.edu

Citation: Wang X, Sharpnack J, Lee TCM (2025) Improving lung cancer diagnosis and survival prediction with deep learning and CT imaging. PLoS One 20(6):e0323174. https://doi.org/10.1371/journal.pone.0323174

Editor: Xu Yanwu, South China University of Technology, CHINA

Received: August 22, 2024

Accepted: April 3, 2025

Published: June 11, 2025

Data availability statement: The dataset used in this study is publicly available from the National Lung Screening Trial (NLST) via the Cancer Data Access System (CDAS). All data necessary to replicate the findings reported in this manuscript, including relevant metadata and documentation, can be accessed without restriction at: https://cdas.cancer.gov/nlst/. Reference: National Lung Screening Trial Research Team. "The national lung screening trial: overview and study design." Radiology 258.1 (2011): 243–253.

Abstract

Lung cancer is a major cause of cancer-related deaths, and early diagnosis and treatment are crucial for improving patients' survival outcomes. In this paper, we propose to employ convolutional neural networks to model the non-linear relationship between the risk of lung cancer and the lungs' morphology revealed in the CT images. We apply a mini-batched loss that extends the Cox proportional hazards model to handle the non-convexity induced by neural networks, which also enables the training of large data sets. Additionally, we propose to combine mini-batched loss and binary cross-entropy to predict both lung cancer occurrence and the risk of mortality. Simulation results demonstrate the effectiveness of both the mini-batched loss with and without the censoring mechanism, as well as its combination with binary cross-entropy. We evaluate our approach on the National Lung Screening Trial data set with several 3D convolutional neural network architectures, achieving high AUC and C-index scores for lung cancer classification and survival prediction. These results, obtained from simulations and real data experiments, highlight the potential of our approach to improving the diagnosis and treatment of lung cancer.

1 Introduction

Lung cancer is one of the most common causes of cancer-related deaths worldwide. Early diagnosis and treatment are crucial for improving patients' survival rates [1,2]. Survival analysis, a branch of statistics that has been widely used in public health research, provides valuable insights into the impact of different conditions on the survival time of patients; e.g., [3,4]. In the context of lung cancer, early detection through screening methods helps identify the tumor in its early stage and applying survival analysis to lung cancer patients can aid in early detection and ultimately improve patients' survival outcomes. Meanwhile, in recent years, computer-aided diagnosis has gained significant attention, particularly in medical image data analysis [5–9]. Deep learning techniques have been increasingly applied to analyze various kinds of medical images due to their effectiveness, for example, [10–16].

Funding: This research was partially supported by the National Science Foundation under grants DMS-2113605 and DMS-2210388 in 2021 and 2022, respectively. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Despite the promising results obtained by using these techniques, the accessibility of high-quality medical images poses a challenge in applying these techniques. For example, Hou et al. [10] required whole slide tissue images obtained from invasive procedures, Gao et al. [11] required multiple longitudinal CT images captured over time, and Wang et al. [12] required both demographic information and chest CT images.

In addition, most of these studies focused on patients already diagnosed, neglecting those who may be prospective candidates undergoing regular CT screening for early detection. Furthermore, there are few works that have utilized survival analysis, which limits the statistical efficiency of these methods. Considering the significant impact of early detection on patients' survival chances [17], there is an urgent need to develop a new approach that can enhance both the early detection and survival prediction for individuals currently diagnosed and those potentially at risk of lung cancer, while considering the accessibility of the medical image data.

This paper introduces a novel deep learning approach that fundamentally advances existing survival analysis methods for lung cancer. While previous approaches like DeepSurv[18] were limited to demographic information and DeepConvSurv[19] only utilized 2D pathological images, our method uniquely leverages 3D convolutional neural networks (CNNs) to capture the complex three-dimensional morphology of lungs from CT images. This represents a significant methodological advancement as it allows direct modeling of the non-linear relationship between survival hazards and complete volumetric lung structure. We develop a specialized mini-batched loss function that not only handles the non-convexity inherent in neural networks but also efficiently processes large-scale 3D imaging datasets. Furthermore, we innovate beyond existing frameworks by introducing a dual-objective approach that simultaneously predicts both cancer presence and survival risk through a novel combination of binary cross-entropy and mini-batched loss functions. This unified framework represents a substantial improvement over previous methods that typically address these tasks separately. The promising empirical properties of the proposed method are illustrated by simulation experiments and the application to the National Lung Screening Trial (NLST) dataset [20].

Our approach advances the state-of-the-art through several key innovations: (i) it pioneers the integration of 3D medical image classification with survival analysis, moving beyond the 2D image analysis of previous methods; (ii) it uniquely addresses both existing and potential patients through a dual-task framework, enabling earlier detection than traditional single-task approaches; and (iii) it achieves superior accessibility by requiring only a single raw CT scan, eliminating the dependence on longitudinal data or expert annotations that limit existing methods. These innovations collectively enable more robust and widely applicable survival prediction than previous approaches like DeepSurv and DeepConvSurv.

The rest of this paper is organized as follows: Section 2 introduces related works in computer-aided diagnosis and basic knowledge about survival data and the Cox proportional hazards model. Section 3 derives the mini-batched loss function of the extended Cox model and introduces the idea of the two-task method and corresponding metrics. Section 4 presents the simulation study of the mini-batched loss based on the MNIST dataset and the simulation of the two-task method based on the Nodule-CIFAR dataset. Section 5 presents the real data experiment with the two-task method, which includes CT images from potential lung cancer patients.

2 Background

2.1 Related work

The Cox proportional hazards model [21] was first proposed to explore the relationship between the survival chance of a patient and a group of explanatory variables through the concept of hazard rate, see Eq 1. Later, Breslow [22] and Cox [23] discussed the estimation of model parameters, particularly for the baseline hazard function. Despite it being proposed more than 50 years ago, the Cox model continues to be one of the most widely used models in medical research for investigating patients' survival chances.

The use of medical images to aid the diagnosis and treatment of diseases has become increasingly popular. Much research has been conducted on the use of deep learning techniques to analyze medical images as a computer-aided diagnosis. For example, Hou et al.[10] studied the feature of whole slide tissue image patches with a CNN. Wang et al. [12] detected lung cancer with CT images and clinical demographics. Ardila et al. [13] proposed a CNN-based method to predict lung cancer risk. Gao et al. [11] performed research in detecting lung cancer with long short-term models. Liu et al. [14] studied detecting nodules from CT images for lung cancer with adversarial attacks. However, some of these images or data may not be readily available or collected. These methods required whole slide tissue images from an invasive procedure [10], or longitudinal medical images captured over time [11, 13], or demographic information in addition to medical images [12]. For more details, refer to [24,25] for a comprehensive review of deep learning techniques applied to medical images.

While these imaging methods have produced excellent results for the tasks that they were designed for, they did not establish a correlation with patients' survival. Katzman et al. [18], for the first time, developed the DeepSurv model to study the non-linear relationship between survival hazards and clinical features. It replaced the linear part $\beta^{\mathsf{T}}\boldsymbol{x}$ in the Cox proportional hazards model (1) with multi-layer perceptrons $f(\boldsymbol{x})$. However, this model has a limitation in that it can only process clinical information. To address this limitation, DeepConvSurv was then proposed by Zhu et al. to predict patients' survival directly from the 2D region of interests (ROI) of pathological images, using CNNs for $f(\boldsymbol{x})$.

In this paper, we aim to expand previous research by developing a model that classifies lung cancer occurrence from potential lung cancer patients with only one 3D CT scan and further predicts the patient's relative hazards of dying from lung cancer. Our approach integrates 3D CNNs, binary classification, and the Cox proportional hazards model. By combining these techniques, we aim to establish a direct correlation between potential patients' 3D medical images and patients' survival, which could have significant implications for early lung cancer diagnosis.

2.2 Survival data

Survival analysis typically considers time-to-event data. Let $T^* = \min(T, C)$ be the observed time, where T denotes the event time and C denotes the censored time. Here, T is the time from the beginning of the observation to an event, usually death, disease occurrence, or other experience of interest, which can be unobserved if censoring occurs first. The censored time C is the time after which nothing is observed about the object. In addition to observing T^*, we also have the event indicator: $\delta_i = 1_{\{T_i \leq C_i\}}$ that tells us if the i–th observation T_i is censored or not. In our study, T^* is the observed time from the beginning of the study to either observed death or censoring. If death is observed, $T^* = T$ and $\delta = 1$, if censoring is observed, $T^* = C$ and

$\delta = 0$. The objective is to model the event distribution of T,

$$F(t) = P(T \leq t) = \int_0^t f(u)du,$$

where the density function $f(t)$ is

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t}.$$

In survival analysis, it is common to alternatively study the survival function $S(t)$, or the hazard function $\lambda(t)$, or the cumulative hazard function $\Lambda(t)$, defined respectively as

$$S(t) = P(T > t) = \int_t^\infty f(u)du,$$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta_t},$$

and

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

Their relationships can be expressed as

$$\lambda(t) = \frac{f(t)}{S(t)},$$

and

$$S(t) = \exp(-\Lambda(t)),$$

so it's equivalent to studying either of them. In this paper, we focus on the density function $f(t)$ and the corresponding likelihood function.

Given a set of right-censored samples $\{T_i^*, \delta_i\}_{i=1}^n$, the likelihood function L is:

$$\begin{aligned} L &= \prod_{i=1}^n f(T_i^*)^{\delta_i} S(T_i^*)^{1-\delta_i} \\ &= \prod_{i=1}^n \lambda(T_i^*)^{\delta_i} S(T_i^*), \end{aligned}$$

which can be further used for parameter estimation.

2.3 Cox proportional hazards model and DeepSurv

The Cox proportional hazards model is one of the most used models for exploring the relationship between the hazards $\lambda(t|\boldsymbol{x})$ and the explanatory covariates $\boldsymbol{x}$. In particular, it assumes proportional hazards and linear contribution of the covariates to the log relative hazards function:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp(\beta^\intercal \boldsymbol{x}), \tag{1}$$

where t represents time, $\lambda_0(t)$ is the baseline hazard function (an infinite dimensional parameter), $\boldsymbol{x}$ is a set of covariates, and β is the corresponding coefficient that measures the effect

of the covariates. Cox [21,23] proposed to use the partial likelihood for estimating β with the advantage of circumventing $\lambda_0(t)$. Let $R(t) = \{i : T_i^* > t\}$ be the risk set at time t; i.e., the set of all individuals who are "at risk" for failure at time t. The partial likelihood is the product of the conditional probabilities of the observed individuals being chosen from the risk set to fail:

$$L(\beta)_{\text{partial}} = \prod_{i=1}^{n} \left[\frac{\exp(\beta^\intercal \boldsymbol{x_i})}{\sum_{j \in R(T_i^*)} \exp(\beta^\intercal \boldsymbol{x_j})} \right]^{\delta_i},$$

where $R(T_i^*)$ denotes the set of individuals that are "at risk" for failure at time T_i^* in the sample.

The estimate $\hat{\beta}$ for β can be obtained by minimizing the averaged negative partial log–likelihood $\mathcal{L}(\beta)$, which is convex:

$$\mathcal{L}(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \delta_i \left[\beta^\intercal \boldsymbol{x_i} - \log \sum_{j \in R(T_i^*)} \exp(\beta^\intercal \boldsymbol{x_j}) \right].$$

The cumulative baseline hazard function can be estimated with the Breslow estimator:

$$\begin{aligned} \hat{\Lambda}_0(t;\beta) &= \sum_{j \notin R(t)} \Delta \hat{\Lambda}_0(T_j^*) \\ &= \sum_{j \notin R(t)} \frac{\delta_j}{\sum_{k \in \mathcal{R}(T_j^*)} \exp(\beta^\intercal \boldsymbol{x_k})}. \end{aligned}$$

The DeepSurv method can be seen as a non-linear version of the Cox model. It replaces the linear log relative hazards term $\beta^\intercal \boldsymbol{x}$ in the Cox model with a non–linear multi–layer perceptron (MLP)$f(\boldsymbol{x};\theta)$:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\big(f(\boldsymbol{x};\theta)\big),$$

where $f(\boldsymbol{x};\theta)$ is a fully–connected MLP parameterized by θ.

3 Methodology

3.1 Extended Cox model with convolution neural network

In this study, we modeled patients' hazard function of a certain disease based on 3D medical images. We cannot directly apply the DeepSurv or DeepConvSurv model because MLP or 2D CNN is deficient for 3D image data. Therefore, we extended the DeepSurv model by replacing MLP with a 3D convolution neural network $f(\boldsymbol{x};\Theta)$, which predicted the effects of a patient's morphological features $\boldsymbol{x}$ on their hazard rate and parameterized by the weights of the network Θ:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\big(f(\boldsymbol{x};\Theta)\big).$$

3.2 Loss function derivation

Let

$$\Lambda(t) = \Lambda_0(t) \exp\big(f(\boldsymbol{x};\Theta)\big)$$

and

$$S(t) = \exp\Big(-\Lambda_0(t)\exp\big(f(\boldsymbol{x};\Theta)\big)\Big),$$

so the full likelihood function is

$$L(\Lambda_0,\Theta) = \prod_{i=1}^{n}\Big\{\Big[\lambda_0(T_i^*)\exp\big(f(\boldsymbol{x_i};\Theta)\big)\Big]^{\delta_i}\times \exp\Big(-\Lambda_0(T_i^*)\exp\big(f(\boldsymbol{x_i};\Theta)\big)\Big)\Big\}.$$

Moreover, the negative log-likelihood becomes

$$\mathcal{L}(\Lambda_0,\Theta) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\delta_i\Big[f(\boldsymbol{x_i};\Theta) + \log\lambda_0(T_i^*)\Big] - \Lambda_0(T_i^*)\exp\big(f(\boldsymbol{x_i};\Theta)\big)\Big\}, \tag{2}$$

$$\ell(\Lambda_0,\Theta) = -\sum_i\left[\delta_i\left(f(x_i;\Theta) + \log\lambda_0(T_i^*)\right) - e^{f(x_i;\Theta)}\Lambda_0(T_i^*)\right],$$

which depends on both Λ_0 and parameters Θ in f.

In practice, the prior knowledge of Λ_0 is not available. To overcome this issue, we adopted the non-parametric Breslow estimator, which treated the baseline as a piece-wise constant between event failure times:

$$\hat{\Lambda}_0(t;\Theta) = \sum_{j\notin R(t)}\Delta\hat{\Lambda}_0(T_j^*) = \sum_{j\notin R(t)}\frac{\delta_j}{\sum_{k\in\mathcal{R}(T_j^*)}\exp\big(f(\boldsymbol{x_k};\Theta)\big)}.$$

By plugging it into the negative log-likelihood Eq 2, we derived the partial likelihood without $\lambda_0(t)$:

$$\mathcal{L}_{\text{fb}}(\Theta) = -\frac{1}{n}\sum_i\delta_i\left[f(\boldsymbol{x_i};\Theta) - \log\sum_{j\in\mathcal{R}(T_i^*)}\exp\big(f(\boldsymbol{x_j};\Theta)\big)\right]. \tag{3}$$

We refer to this as the *full-batched loss* in this paper. In fact, the procedure of getting partial likelihood of the Cox proportional model can lead us to the equivalent loss function. Given the model $\lambda(t) = \lambda_0(t)\exp\big(f(\boldsymbol{x};\Theta)\big)$, the partial likelihood now becomes

$$L(\Theta)_{\text{partial}} = \prod_{i=1}^{n}\left[\frac{\exp\big(f(\boldsymbol{x_i};\Theta)\big)}{\sum_{j\in R(T_i^*)}\exp\big(f(\boldsymbol{x_j};\Theta)\big)}\right]^{\delta_i}, \tag{4}$$

The full-batched loss function can be obtained by taking the average of the negative log of the partial likelihood.

Even though the full-batched loss is convex in f, due to the non-convexity of the neural network, the full-batched loss is non-convex. Also, the full-batched loss involves complicated

sums over the risk set, which can be as large as the full data set, making it computationally expensive.

To deal with the non-convexity and make it scalable to large datasets, we modified the full-batched loss by first subsampling the data and collecting them to a batch Ω, and then restricting the risk set $R(T_i^*)$ only to contain the subsampled data in the current batch:

$$\tilde{\mathcal{L}}_{\mathrm{mb}}(\Theta) = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\delta_i\left[f(\boldsymbol{x_i};\Theta) - \log\sum_j \exp\left(f(\boldsymbol{x_j};\Theta)\right)\right] \tag{5}$$

with $j \in \mathcal{R}(T_i^*) \cap \Omega$. We refer to this expression as the *mini-batched loss* in the paper. If we set the batch as the full data set, then the mini-batched loss is equivalent to the full-batched loss. The batch size can be as small as 2. By restricting data to a randomly sampled batch, we avoided massive calculations. The mini-batched loss is unlike the minibatch gradient descent with i.i.d. (independent and identically distributed) data with respect to the full-batched loss since taking the expectation over random minibatch samples does not give the averaged negative log-likelihood.

As an aside, we can see that the partial likelihood in (4) is the likelihood of observing the given order of events, which in this case is the order of individuals' deaths. By evaluating the partial likelihood, we are in effect ignoring any information of the timing of the events beyond just their ordering. This objective and the mini-batch gradient descent described above appear in recommendation system applications where user preferences are expressed via the relative ordering of click-through events. The resulting method is called listwise ranking in the recommendation system literature [26,27].

3.3 Two-task method for disease diagnosis and survival hazard prediction

Lung cancer is one of the most common cancers. Computed Tomography (CT) images, which include a series of axial image slices that visualize the tissues and nodules within the lung area, can be extremely useful for diagnostic purposes. When given a patient's pulmonary CT images, one objective is to diagnose whether the patient has lung cancer or not, i.e., lung cancer classification. In addition, we hope to predict the severity of cancer by estimating the patient's risk of dying from lung cancer, i.e., survival hazard prediction. Traditionally, to fulfill the two tasks, one option is to train separate models with different losses, respectively: binary cross entropy for lung cancer classification and mini-batched loss for survival hazard prediction. However, it raises concerns about divergent predictions, which may result in predicting a case without lung cancer but with a high risk of mortality of dying from lung cancer.

The link between lung cancer diagnosis and survival prediction is established through the comprehensive analysis of imaging studies. Extracted information from CT images, such as the presence of lung nodules and detailed characteristics (including size, shape, location, and tumor spread), is not only instrumental in confirming the presence of cancer, but also provides critical details that inform prognosis, guide treatment decisions, and influence survival predictions for individual patients. The higher the probability of having lung cancer inferred from CT images, the more likely it is that the cancer exhibits features associated with an advanced or aggressive nature. These features contribute to an increased risk of mortality, forming the basis for the correlation between the probability of having lung cancer and survival prediction. The integration of imaging data into a holistic approach enhances the precision and personalized nature of lung cancer care.

Recognizing the clinical need to integrate these tasks, we present a novel method capable of simultaneously performing lung cancer classification and survival hazard prediction

using the same input––a two-task neural net framework, as illustrated in Fig 1. The output layer, which predicted the log relative hazards $f(\boldsymbol{x};\Theta)$, was also used for lung cancer classification with sigmoid activation. This choice is intuitive as the function f represents hazard, implying that a higher hazard is indicative of a higher probability of having lung cancer. Instead of having separate losses, we defined the loss as the sum of binary cross entropy and the batched loss. Let y_i be the indicator of having lung cancer, $\boldsymbol{x_i}$ be the image input to the deep neural network, and $f(\boldsymbol{x_i};\Theta)$ be the neural network output for log relative hazards, $P(\boldsymbol{x_i};\Theta) = sigmoid\big(f(\boldsymbol{x_i};\Theta)\big)$ is predicted cancer probability:

$$L(\Theta) = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\Bigg\{\delta_i\Big[f(\boldsymbol{x_i};\Theta) - \log\sum_j \exp\big(f(\boldsymbol{x_j};\Theta)\big)\Big] + \Big[y_i \log P(\boldsymbol{x_i};\Theta) + (1-y_i)\log\big(1-P(\boldsymbol{x_i};\Theta)\big)\Big]\Bigg\}, \tag{6}$$

with $j \in \mathcal{R}(T_i^*) \cap \Omega$.

One advantage of this approach is consolidating the goals of cancer classification and survival hazard prediction into a singular model, motivated by the clinical reality that the CT image shows information that is critical for both cancer diagnosis and survival prediction. Training a unified model concurrently for both objectives with shared neural net parameters promises a more comprehensive understanding and superior predictive performance, while conventional approaches of training separate models with binary cross entropy for cancer classification and mini-batched loss for hazard prediction focus exclusively on one aspect. This two-task method provides a holistic view, bridging the diagnostic and prognostic aspects of lung cancer, and offers a more clinically relevant perspective for personalized patient care decisions. Another advantage lies in the dual losses, which enable more comprehensive supervision of the neural net's fit, thereby preventing overfitting during training.

3.4 Evaluation metrics

For the cancer classification task, we used AUC (area under the ROC curve) to evaluate the model performance. In the hazard prediction task, we employed the concordance index (C-index) for evaluation. C-index, introduced by Harrell et al. [28], is a goodness of fit measure for models that produce risk scores for censored data. In our context, it estimates the probability that, for any random pair of individuals, the predicted survival times would exhibit the

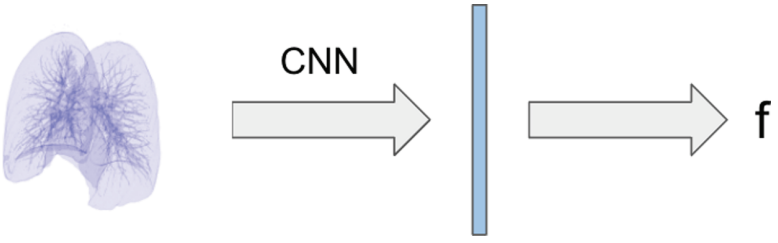

Fig 1. Two-task convolution neural network illustration. The network processes a 3D lung CT mesh through a shared CNN backbone to extract imaging features. These learned representations are then used to simultaneously predict both lung cancer classification probability (via sigmoid activation) and survival hazard estimation. The unified architecture enables joint optimization of both tasks through a combined loss function (Eq 6). This integrated approach leverages the complementary nature of cancer detection and survival prediction tasks, potentially improving the performance of both predictions compared to separate single-task models.

https://doi.org/10.1371/journal.pone.0323174.g001

same ordering as their actual survival times. This is equivalent to determining whether, for any random pair of patients, the predicted hazard has the reverse order in comparison to their actual survival times, as patients with higher predicted survival times correspond to lower predicted hazards. The C-index in our context is defined by the following formula:

$$\begin{aligned} C &= \frac{\#\text{ concordant pairs}}{\#\text{ concordant pairs} + \#\text{ disconcordant pairs}} \\ &= P\{\hat{T}_i > \hat{T}_j |\ T_i > T_j, \delta_j = 1\} \\ &\approx P\{\hat{f}_i < \hat{f}_j |\ T_i > T_j, \delta_j = 1\} \qquad (7) \\ &= \frac{\sum_{i \neq j} 1\{\hat{f}_i < \hat{f}_j\} 1\{T_i > T_j\} \delta_j}{\sum_{i \neq j} 1\{T_i > T_j\} \delta_j}, \end{aligned}$$

where approximation (7) follows from the argument that a patient with a higher hazard score should have a shorter survival time.

When C-index = 1, it indicates that the order of the predictions matches exactly with the order of the true survival times. On the other hand, C-index = 0.5 suggests that the predictions are random. Generally, a C-index above 0.7 is considered indicative of a good model. However, for cancer survival analysis, studies such as [29–32] suggest that a C-index between 0.6 and 0.7 is sufficient to provide valuable insights.

4 Simulation studies

This section reports results from three simulation experiments. Both Simulations A and B focused on the extended Cox model and its prediction of the log relative hazards function f. Simulation A was under the setting where there were event cases only, while Simulation B involved both censored and event cases. Both simulations used the same images from the MNIST dataset and the same generated survival time, but different censoring statuses. We compared the performance of the oracle loss, full-batched loss, and mini-batched loss under the settings of Simulations A and B. Simulation C was designed for the two-task framework, involving both the disease occurrence classification and the survival hazard prediction with the log relative hazards function. We generated a new dataset from the CIFAR-10 dataset, called Nodule-CIFAR. We compared the loss function performance of the combination of binary cross-entropy and full-batched/mini-batched in terms of AUC and C-index.

4.1 Simulations A and B

4.1.1 MNIST dataset and time-to-event data. We used the MNIST image dataset and generated artificial survival times for digits in our simulations. The MNIST dataset is an image dataset of handwritten digits from 0 to 9; see [33]. We selected 2 digits from the MNIST dataset as input images of the neural network with different patterns, w.l.o.g., we selected zeros and ones. We generated the survival time for each digit using an exponential distribution with different constant hazards $\lambda_j = 1 \times \exp(\phi_j)$, $j = 0, 1$, where the baseline hazard $\lambda_0(t)$ was set to 1, and the true log relative hazards was ϕ_j. In Simulation A, all cases were labeled as events. In Simulation B, we randomly labeled half of the individuals who lived beyond the median as censored cases within each digit. The distribution of the test set is shown in Fig 2.

4.1.2 Architecture. Simulations A and B were trained under the same feed-forward convolution neural network, which consisted of a stack of convolution and dense layers. The net structure is listed in Table 1.

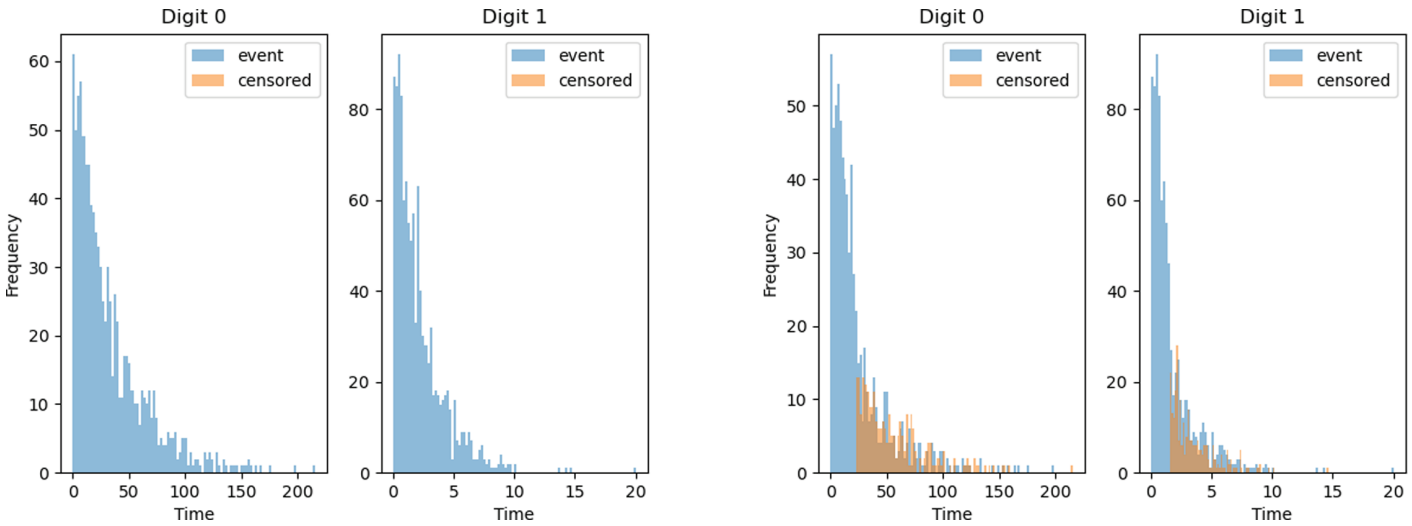

Fig 2. Simulated survival time distributions. (Left) Survival time distributions for the two digits in Simulation A, without the censoring mechanism; (Right) Survival time distributions for the two digits in Simulation B, with the censoring mechanism. The censored cases are labeled in orange, which overlaps the upper half of the event cases.

https://doi.org/10.1371/journal.pone.0323174.g002

Table 1. Convolution neural net architecture for simulations A and B.

Layer Type	Number of Kernels	Kernel Size	Output Size
Convolution	32	5×5	$28 \times 28 \times 32$
Max Pooling		2×2, stride = 2	$14 \times 14 \times 32$
Convolution	64	5×5	$14 \times 14 \times 64$
Max Pooling		2×2, stride = 2	$7 \times 7 \times 64$
Flatten			3136
Fully Connected			1024
Fully Connected			128
Fully Connected			1

https://doi.org/10.1371/journal.pone.0323174.t001

4.1.3 Results of simulations A and B. We introduced the oracle loss in this section. It leverages the prior knowledge of the baseline hazard $\lambda_0(t)$ when compared with the full-batched loss (Eq 3) and mini-batched loss (Eq 5). In our simulations, w.l.o.g., we set $\lambda_0(t) = 1$ when generating survival time, so that $\Lambda_0(t) = t$. Plugging the baseline hazard into the averaged negative full log-likelihood (Eq 2) provided us the oracle loss, for which f can be trained:

$$\mathcal{L}_{\text{orc}}(\Theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i f(\boldsymbol{x_i};\Theta) - \exp\big(f(\boldsymbol{x_i};\Theta)\big)T_i^*\right]. \tag{8}$$

Due to the non-convexity of neural network f, we used the stochastic gradient descent (SGD) method to minimize the non-convex loss function. Correspondingly, the batched version is provided below.

$$\tilde{\mathcal{L}}_{\text{orc}}(\Theta) = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\left[\delta_i f(\boldsymbol{x_i};\Theta) - \exp\big(f(\boldsymbol{x_i};\Theta)\big)T_i^*\right], \tag{9}$$

where Ω is the selected batch for a training iteration. We will later refer to this as the *oracle loss*.

We also calculated the true loss as the baseline for benchmark comparisons. When both the baseline hazard $\lambda_0(t)$ and the log relative hazards ϕ_j were available, we could directly plug them into the averaged negative full log-likelihood (Eq 2), which gave the true loss.

Results of Simulations A and B are reported in Fig 3 and Table 2. In both simulations, the oracle loss settled to the true loss, the oracle loss was less than the batched losses, both batched losses settled to the same value, and the mini-batched loss settled faster than the full-batched loss. This met our expectations since the oracle loss had access to the base rate. In addition, due to the extra information, the C-index trained by the oracle loss is expected to be larger, which was validated in both Simulations A and B, see Table 2. In Simulation A, though the C-index curve fluctuated after loss converges, it achieved a high value for both full batched loss and mini-batched loss, showing good rank prediction on the hazards when there is no censoring. In Simulation B, two C-indexes were calculated: C_1 involved both censored and event case, while C_2 involved event cases only. Here, C_1 exceeds 0.7, which means good rank predictions for pairs across censored and event groups and pairs within the event group. Moreover, the faster convergence and small difference between C_{orc} and C_{mb} indicated the feasibility of mini–batched loss for training parameters without prior information of $\lambda_0(t)$.

4.2 Simulation C: nodule-CIFAR simulation with classification and hazard prediction

4.2.1 Nodule-CIFAR dataset We introduced a new dataset, called Nodule-CIFAR, which was generated from the CIFAR-10 dataset [34]. Nodule-CIFAR was inspired by Tumor-CIFAR from Gao et al. [11] and simulated benign and malignant nodules on the CIFAR-10 images.

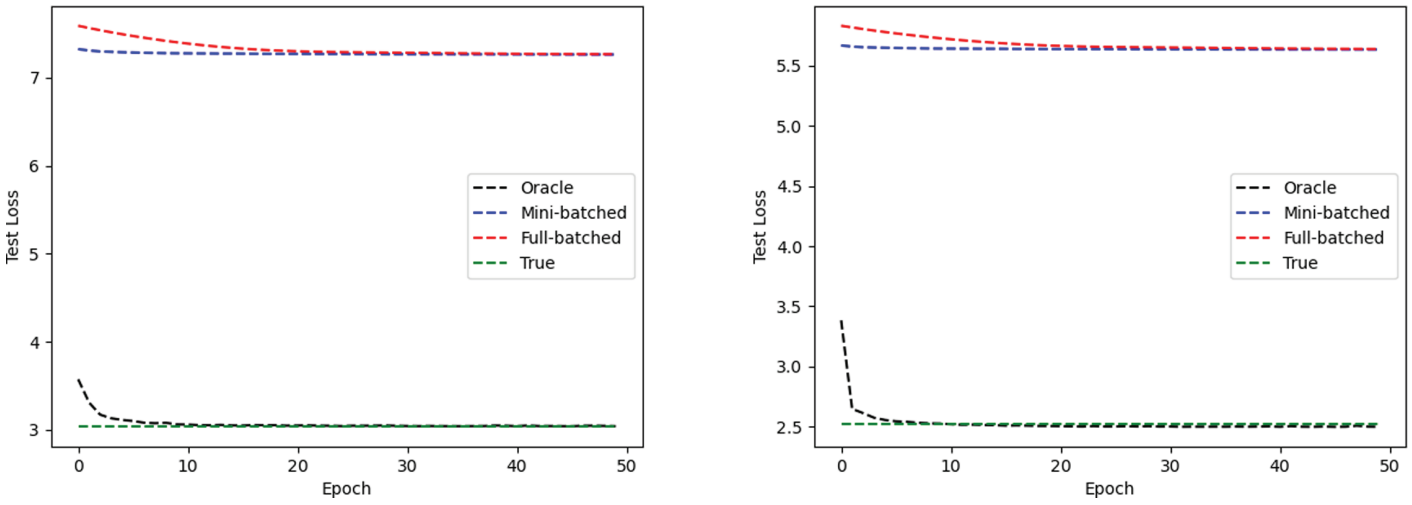

Fig 3. Simulation losses by epoch. (Left) Simulation A. (Right) Simulation B.

https://doi.org/10.1371/journal.pone.0323174.g003

Table 2. Simulations A and B: C-indexes under three losses.

	Oracle	Full-batched	Mini-batched
A	0.7268	0.7165	0.7189
B w/ censored (C1)	0.7184	0.7146	0.7166
B w/o censored (C2)	0.6845	0.6770	0.6790

https://doi.org/10.1371/journal.pone.0323174.t002

In reality, benign nodules typically exhibit smaller sizes with regular round shapes and are non-cancerous, while malignant nodules tend to be larger in size and exhibit irregular shapes. Healthy individuals possess benign nodules, but patients may have both benign and malignant nodules. To simulate this, we introduced black and white dots onto CIFAR-10 images to simulate benign nodules, while dummy nodules were represented as white blobs to simulate malignant nodules.

The dataset inherently followed a 10:1 train/test split of CIFAR-10, consisting of 10,000 training samples and 1,000 testing samples, with no additional custom splitting applied. We randomly assigned images to non-cancerous and cancerous groups with equal probability, so that cancer prevalence was 50% in both training and test sets. Among the cancerous cases, we randomly labeled 50% as censored, and the remaining were labeled as events, the events of failure of dying from cancer. For the non-cancerous cases, they would not die of cancer, so all of them were labeled as censored. Next, we incorporated simulated nodules, either benign or malignant, onto CIFAR-10 images based on their assigned group. The non-cancer images yet featuring benign nodules, displayed numerous small black and white dots distributed across the image to simulate benign nodules. In contrast, the images in the cancer groups had two additional big white patches randomly located in the images, mimicking malignant nodules. Within the cancer group, the censored had relatively smaller white patches compared to the event, because the censored group had not yet reached a deadly stage. The original image categories from the CIFAR-10 dataset were irrelevant in this context; the distinctions between cancer and non-cancer were determined by the presence of simulated white patches. Moreover, within the cancer group, the censoring status was solely associated with the sizes of the simulated white patches. Fig 4 is an example of images in the Nodule-CIFAR dataset.

Time-to-event data corresponding to Nodule-CIFAR images were generated based on the largest size of simulated nodules in each image. The recorded time followed an exponential distribution with a parameter of $\lambda = 1 \times \exp(\phi)$, where $\phi \propto size$, the largest size of simulated nodules in each image. This was consistent with our expectation that the larger the nodule size, the larger the hazards, and the smaller the survival time.

Fig 5 shows the distribution of nodule size and survival time for each group. The non-cancer group had smaller nodules on average compared to the cancer group. Within the cancer group, those event cases (eventually died of cancer in simulation) had larger malignant nodules. The time-to-event for the non-cancer group was larger than the cancer group. Within the cancer group, the time-to-event of censored cases was larger than the event cases.

4.2.2 Architecture. Like Simulations A and B, Simulation C was trained under a feed-forward convolution neural network, which consisted of a stack of convolution and dense layers. The output was used for both disease occurrence classification and hazard prediction evaluation. See Table 3 for the structure of the neural network.

4.2.3 Results of simulation C. The loss function for the two-task network was the sum of the binary cross entropy and the full-batched/mini-batched loss. To compare the model performance trained with different losses under the same network architecture, see Fig 6 for the epoch-wise losses, AUC, and C-index, and Table 4 for their stabilized values after the losses converge. As shown in Fig 6 (top-left panel), the one with mini-batched loss (blue) converged much faster than the one with full-batched loss (red); it reached a minimum after a few epochs and stabilized. Fig 6 (top-right) showed that both losses outperformed the baseline AUC 50% significantly, which was achieved by predicting all cases as non-cancer, and the model trained with mini-batched loss achieved a slightly higher AUC. As for the hazard prediction evaluation, we calculated two C-indexes $C1$ and $C2$, where $C1$ was for all cases (cancer

Fig 4. Example images from nodule-CIFAR dataset showing the addition of simulated nodules. Top row (Benign): Original CIFAR-10 image (left) and the same image with added small black and white dots simulating benign nodules (right). Bottom row (Malignant): Original image (left) and the same image with both small dots and larger white patches (right), where the small dots represent benign nodules and the larger white patches simulate malignant nodules. Both malignant and benign cases contain the small dots to reflect the real-world scenario where both types of nodules can coexist, with malignant cases distinguished by the additional larger white patches. Note that these artificial features are distinct from natural white areas in the original images.

https://doi.org/10.1371/journal.pone.0323174.g004

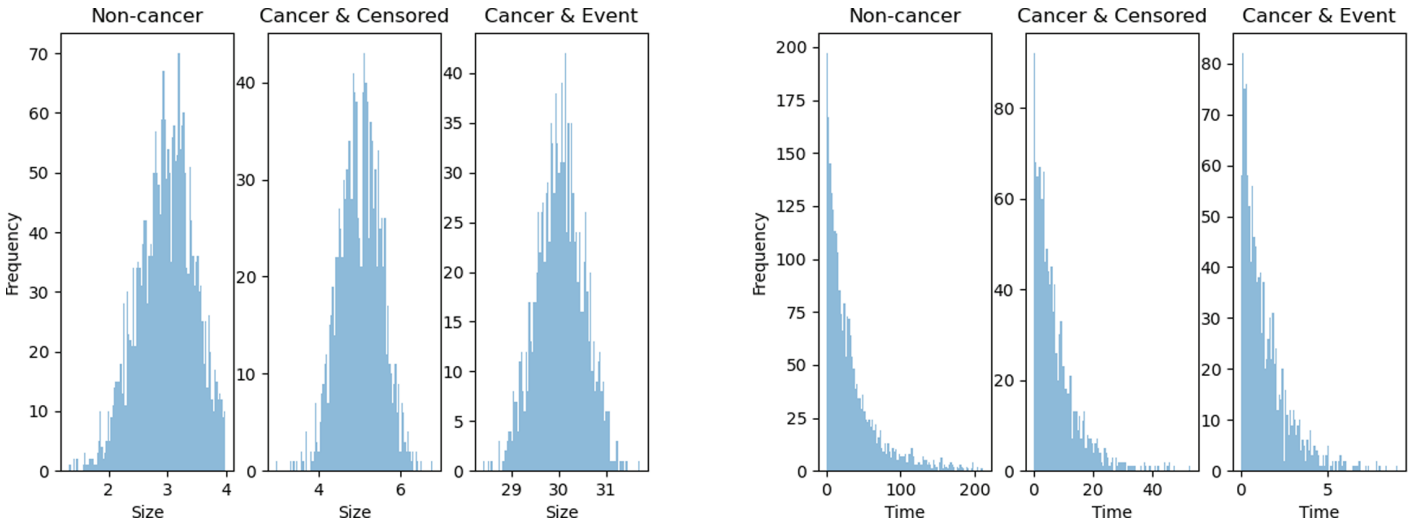

Fig 5. Nodule size and survival time distribution by group. (Left) Nodule size distribution by group. The non-cancer group has smaller nodules on average when compared with the cancer group. Within the cancer group, event cases (those who eventually die of cancer in simulation) have larger malignant nodules. (Right) Survival time distribution by group in Nodule-CIFAR. The time-to-event for the non-cancer group is larger than the cancer group. Within the cancer group, the time-to-event of censored is larger than that of the event cases.

https://doi.org/10.1371/journal.pone.0323174.g005

and non-cancer, Fig 6 (bottom-left panel)) and *C*2 was for the cancer group (Fig 6 (bottom-right panel)). Both losses achieved competitive *C*1 and *C*2 values, especially within the cancer group, where *C*1 exceeded 0.75 for both losses. Comparing Fig 6 (bottom-left panel) and

Table 3. Convolution neural net architecture for simulation C.

Layer Type	Number of Kernels	Kernel Size	Output Size
Convolution	32	5 × 5	28 × 28 × 32
Max Pooling		2 × 2, stride = 2	14 × 14 × 32
Convolution	64	5 × 5	14 × 14 × 64
Max Pooling		2 × 2, stride = 2	7 × 7 × 64
Flatten			3136
Fully Connected			100
Fully Connected			10
Fully Connected			1

https://doi.org/10.1371/journal.pone.0323174.t003

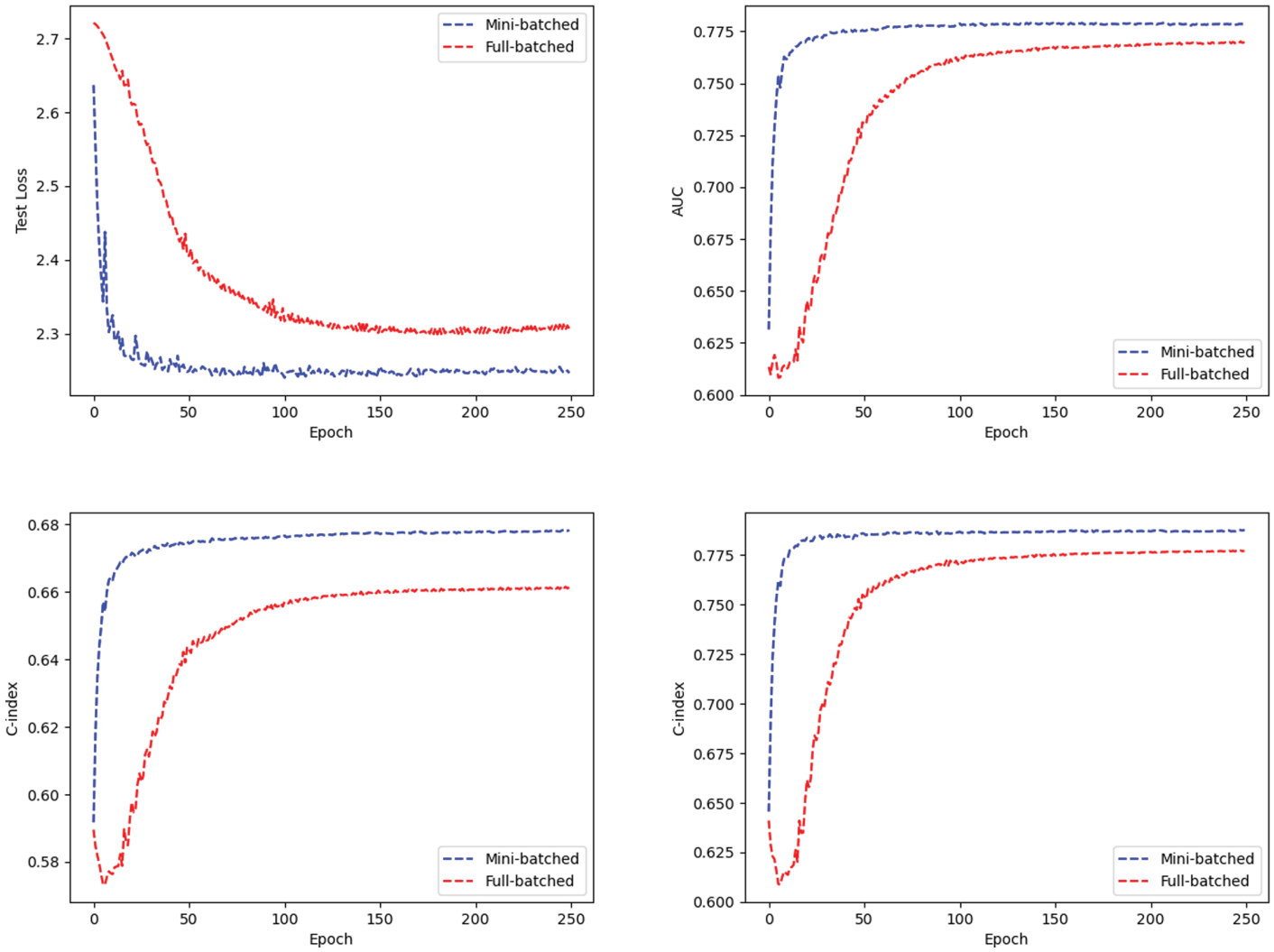

Fig 6. Result of simulation C. (Top-left): test loss by epoch; (Top-right): test AUC by epoch; (Bottom-left): test c-index C1 of all cases by epoch; (Bottom-right): test c-index C2 of the cancer group by epoch. The sum of binary cross entropy and mini-batched loss performed better in both classification and hazard prediction by achieving higher stabilized AUC, C1, and C2 within fewer epochs.

https://doi.org/10.1371/journal.pone.0323174.g006

Table 4. Simulation C: AUC and C-index under two losses.

	Full-batched	Mini-batched
AUC	0.770	0.783
C1	0.661	0.677
C2	0.779	0.785

https://doi.org/10.1371/journal.pone.0323174.t004

Fig 6 (bottom-right panel), we noticed the C-index decreased to around 0.65 when it involved the non-cancer group, which was caused by the trade-off between the classification and hazard prediction tasks. Overall, the sum of binary cross entropy and the mini-batched loss performed better in both classification and hazard prediction by achieving higher stabilized AUC and C-index values within fewer epochs.

5 Real data experiment

5.1 NLST dataset

The National Lung Screening Trial (NLST) collected medical images and survival information from potential lung cancer patients during 2002–2009, see [20]. It was a randomized controlled trial to determine whether screening for lung cancer with low-dose helical computed tomography (CT) reduced mortality from lung cancer in high-risk individuals relative to screening with chest radiography (X-ray). Participants were randomly assigned to two study arms in equal proportions. One arm received low-dose helical CT, while the other received single-view chest radiography.

CT images are a set of axial slice images of the human body. They can reveal both normal and abnormal tissues inside the organs. The abnormal tissues of the lungs are called nodules. Nodules usually are spherical but may have other shapes. Each sub-type of nodules has a different cancer probability. Hence, doctors take into consideration all nodules when diagnosing lung diseases with CT images.

In our experiment, we chose 991 patients who developed cancer during the trial period from a pool of 15,000 patients who received CT imaging. Subsequently, we collected the most recent CT images from these 991 patients confirmed to have lung cancer, among whom 427 passed away due to lung cancer. For the classification task, we similarly gathered the most recent CT images from an equal number of potential patients who did not have lung cancer. Among the total of 1882 patients, those with confirmed lung cancer cases were assigned a label of $y_i = 1$, while all others were labeled as $y_i = 0$. In addition, those who experienced lung cancer-related mortality were categorized as events of failure (non-censored) with $\delta_i = 1$, whereas the rest were considered censored with $\delta_i = 0$. Each patient's most recent CT examination was utilized as the input image denoted as X. Furthermore, we collected patients' survival time T^* by subtracting their latest exam date from the date they were last known alive.

5.2 Preprocessing

In terms of preprocessing the CT images from NLST datasets, we utilized the open-source code [35] to segment the lungs from the CT images and applied the nodule detection method described in [36] to obtain the top 5 suspicious nodule crops as input. For completeness, we provide a brief summary of their method below.

5.2.1 Lung segmentation. The CT images are a set of cross-sectional images of the body. Preprocessing for lung segmentation was required before they were ready for the CNN. First,

the CT scans should be resampled to $1 \times 1 \times 1\,mm^3$ isotropic resolution, then the resampled CT scans were preprocessed with the following main steps:

i. Mask extraction: The first step was to extract the lungs' mask by converting the image to Hounsfield unit (HU) and binarizing the image with the lungs' HU values. HU is a standard quantitative scale for describing radiodensity. Each organ has a specific HU range, and the range remains the same for different people. Here, we used a –320 HU value as the threshold for the lungs. The largest connected component located in the center of the image was extracted as the lungs' mask.
ii. Convex hull computation: The second step was to compute the convex hull of the lungs' mask. Because some nodules might be connected to the outer lung wall and might not be covered by the mask obtained in the previous step, a preferred approach was to obtain the convex hull of the mask. However, it could include other unrelated tissues if one directly computes the convex hull of the mask. To overcome this issue, we first divided the mask into left and right lung masks, then computed their respective convex hulls, and lastly merged them to form the final, whole lungs' convex hull.
iii. Lung segmentation: We obtained a segmentation of the lungs by first multiplying the CT image with the mask and then filling the masked region with tissue luminance.

After completing these three steps, 3D segmented lungs can be extracted. An example is shown in Fig 7.

5.2.2 Nodule detection. The sizes of the segmented lung images varied for each patient, which went against the requirement for identical image sizes in CNNs to work properly. To resolve this, the segmented images were resampled to the same resolution and fixed slice distance. Although the size of each cropped image might differ due to varying lung sizes among patients, zero padding was used if the image size is less than $224 \times 224 \times 224 \times 1$; otherwise, the central 224-width cubes were extracted. An attempt was made to directly input this preprocessed 224-cube into a 3D network for lung cancer classification and hazard prediction. Still, it was computationally time-consuming, and the results were unsatisfactory due to the large size of 3D images and potential memory issues. To address the issue, we followed Liao et al.'s nodule detection process [36]. The nodule detector took in the 3D segmented lung CT image and output predicted nodule proposals with their center coordinates, radius, and confidence. The five most suspicious lung proposals were selected as input X for our network, as Liao et al. determined that $k = 5$ was sufficient for recall when different top k proposals with the highest confidence were selected for inference [36]. For each selected proposal, a $96 \times 96 \times 96 \times 1$ patch centered on the proposed nodule was cropped, resulting in an input size of $5 \times 96 \times 96 \times 96 \times 1$, where one channel represented the number of channels.

5.3 Network structure

The top five regions with the highest nodule confidence were considered for cancer occurrence classification and hazard prediction tasks for each patient. The network had two phases: feature extraction from each lung crop using convolutional layers, and feature combination through integration, as shown in Fig 8. The final output f was evaluated with AUC and C-index metrics.

5.3.1 Convolution phase. We had three different convolution structures to extract features from the top five nodule crops: Alex3D, VGG163D, and Res-net18. Each took a nodule proposal as input and output a 128-D feature. We also adopted the pre-trained cancer classifier from Liao et al. [36] as a performance benchmark.

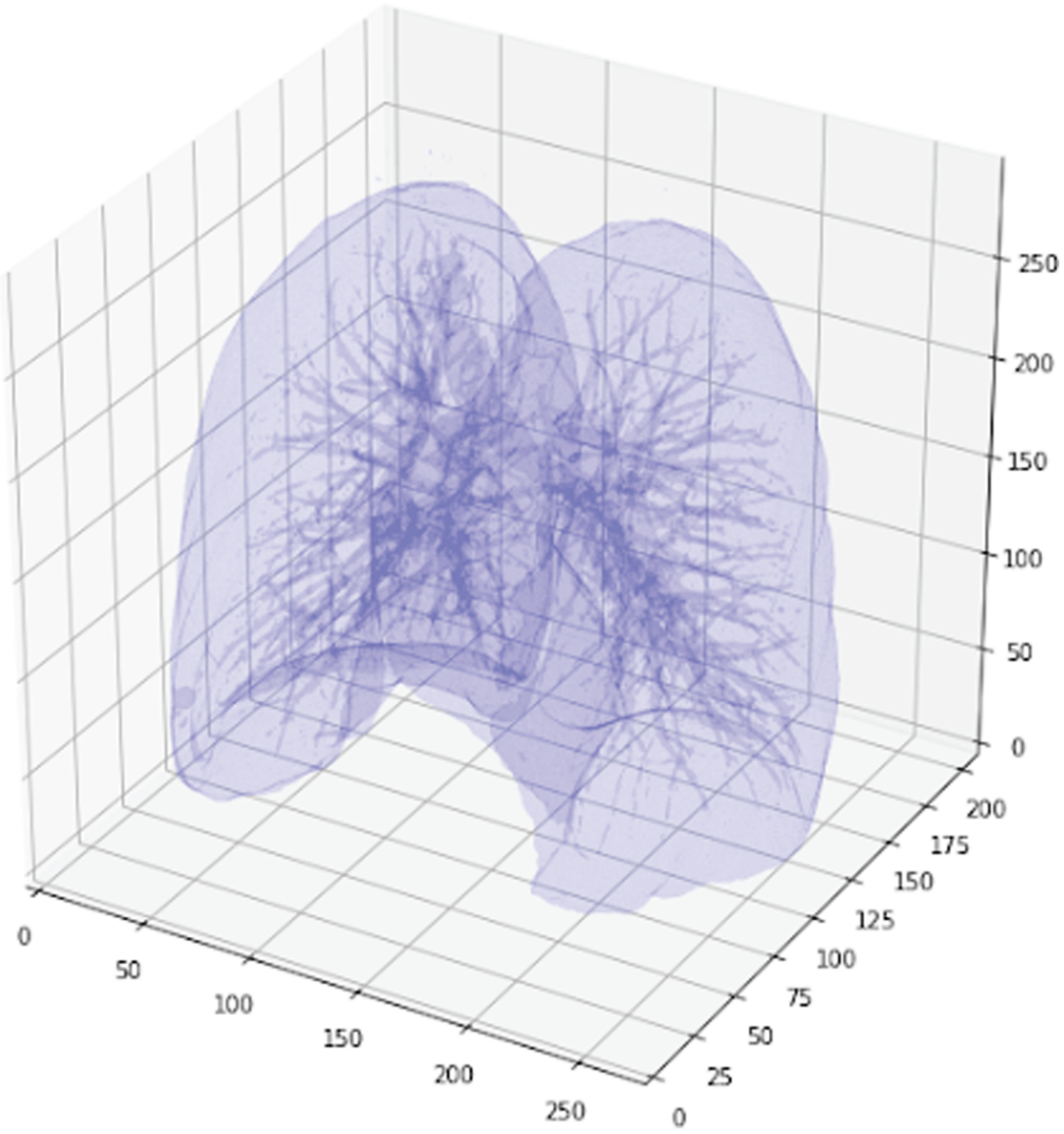

Fig 7. An example for 3D segmented lungs from CT images.

https://doi.org/10.1371/journal.pone.0323174.g007

5.3.2 3D Alex Net. Table 5 lists layers in Alex 3D. The network was based on the classic 2D Alex Net architecture with modifications specifically tailored for the NLST dataset.

5.3.3 3D VGG16. Table 6 lists the layers in 3D VGG16 developed from 2D VGG16 [37], with modifications specifically tailored for the NLST dataset.

5.3.4 3D ResNet-18. Table 7 lists the layers in 3D ResNet-18 developed from a 2D residual network [38]. Downsampling was performed by Res-block2_1, Res-block3_1, and Res-block4_1 with a stride of 2.

5.3.5 Pretrained cancer classifier. We adopted the pre-trained cancer classifier from Liao et al. [36] as a performance benchmark. Liao et al. [36] propose a 3D deep neural network based on U-net for cancer probability reference, which has 2 modules: a nodule detection

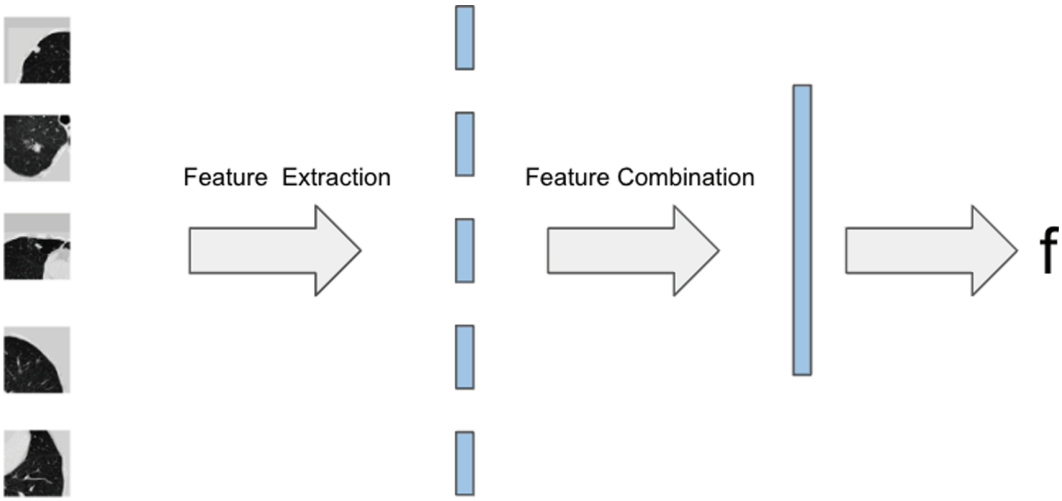

Fig 8. Network structure with 2 phases. Convolution and integration phases.

https://doi.org/10.1371/journal.pone.0323174.g008

Table 5. 3D Alex Net architecture for lung CT images.

Layer Type	Number of Kernels	Kernel Size	Output Size
Convolution	96	$3 \times 3 \times 3$	$48 \times 48 \times 48 \times 96$
Max Pooling		$3 \times 3 \times 3$	$23 \times 23 \times 23 \times 96$
Convolution	256	$5 \times 5 \times 5$	$23 \times 23 \times 23 \times 256$
Max Pooling		$3 \times 3 \times 3$	$11 \times 11 \times 11 \times 256$
Convolution	384	$3 \times 3 \times 3$	$9 \times 9 \times 9 \times 384$
Convolution	256	$3 \times 3 \times 3$	$9 \times 9 \times 9 \times 256$
Max Pooling		$3 \times 3 \times 3$	$4 \times 4 \times 4 \times 256$
Flatten			16384
Fully Connected			4096
Fully Connected			128

https://doi.org/10.1371/journal.pone.0323174.t005

module and a cancer classification module. Because of the limited data size, the classification module (called N-net) integrates the pre-trained detection module as part of the classifier. We followed Liao et al. 's process to obtain the features from image patches: For each selected crop, we fed it to the N-net and obtained the last convolutional layer of the nodule classifier, whose size is $24 \times 24 \times 24 \times 128$. The central $2 \times 2 \times 2$ voxels of each proposal feature were extracted and max-pooled, resulting in a 128-D feature, as shown in Fig 9.

5.3.6 Integration phase. After the convolution phase, the network had five 128D features for each patient. To obtain a single output from these multiple nodule features, three integration methods were explored. The best-performing integration method is shown in Table 8, and its graphical representation can be found in Fig 10. The features from the top five nodules were individually input into a fully connected layer with 32 hidden units. The maximum value of each feature was considered for the final result after concatenating into a single 5D feature, and following a fully connected layer generated the final output *f*.

5.4 Results

The performance of our proposed deep learning frameworks was evaluated using AUC for lung cancer occurrence classification and C-index for survival hazard prediction, as shown in Table 9. We compared four different models: our baseline N-net [36] and three advanced

Table 6. 3D VGG Net architecture for lung CT images.

Layer Type	Number of Kernels	Kernel Size	Output Size
Convolution	64	3 × 3 × 3	96 × 96 × 96 × 64
Convolution	64	3 × 3 × 3	96 × 96 × 96 × 64
Max Pooling		3 × 3 × 3	48 × 48 × 48 × 64
Convolution	128	3 × 3 × 3	48 × 48 × 48 × 128
Convolution	128	3 × 3 × 3	48 × 48 × 48 × 128
Max Pooling		3 × 3 × 3	24 × 24 × 24 × 128
Convolution	256	3 × 3 × 3	24 × 24 × 24 × 256
Convolution	256	3 × 3 × 3	24 × 24 × 24 × 256
Convolution	256	3 × 3 × 3	24 × 24 × 24 × 256
Max Pooling		3 × 3 × 3	12 × 12 × 12 × 256
Convolution	512	3 × 3 × 3	12 × 12 × 12 × 512
Convolution	512	3 × 3 × 3	12 × 12 × 12 × 512
Convolution	512	3 × 3 × 3	12 × 12 × 12 × 512
Max Pooling		3 × 3 × 3	6 × 6 × 6 × 512
Convolution	512	3 × 3 × 3	6 × 6 × 6 × 512
Convolution	512	3 × 3 × 3	6 × 6 × 6 × 512
Convolution	512	3 × 3 × 3	6 × 6 × 6 × 512
Max Pooling		3 × 3 × 3	3 × 3 × 3 × 512
Flatten			13824
Fully Connected			4096
Fully Connected			4096
Fully Connected			128

https://doi.org/10.1371/journal.pone.0323174.t006

Table 7. 3D ResNet-18 architecture for lung CT images.

Layer Name	3D Resnet-18	Output Size
Conv1	7 × 7 × 7,64,stride 2	48 × 48 × 48 × 64
Max pooling	3 × 3 × 3, stride 2	24 × 24 × 24 × 64
Res-block1	$\begin{bmatrix} 3\times3\times3, 64 \\ 3\times3\times3, 64 \end{bmatrix}\times2$	24 × 24 × 24 × 64
Res-block2	$\begin{bmatrix} 3\times3\times3, 128 \\ 3\times3\times3, 128 \end{bmatrix}\times2$	12 × 12 × 12 × 128
Res-block3	$\begin{bmatrix} 3\times3\times3, 256 \\ 3\times3\times3, 256 \end{bmatrix}\times2$	6 × 6 × 6 × 256
Res-block4	$\begin{bmatrix} 3\times3\times3, 512 \\ 3\times3\times3, 512 \end{bmatrix}\times2$	3 × 3 × 3 × 512
Average-pool		512
Fully Connected		128

https://doi.org/10.1371/journal.pone.0323174.t007

architectures (AlexNet, ResNet, and VGG). The C-index calculations incorporated both cancer and non-cancer groups, as the non-cancer cohort in the NLST dataset consisted of individuals with potential risk of developing cancer.

All three advanced architectures demonstrated statistically significant improvements over the baseline N-net model (all p-values <0.001). For cancer classification, ResNet achieved the highest AUC of 0.6885, representing a 26.8% improvement over the baseline's 0.5428. Similarly, for survival prediction, VGG yielded the best C-index of 0.6036, an 18.6% improvement compared to the baseline's 0.5088. Despite falling below the conventional 0.7 threshold, these results should be interpreted within the challenging context of survival prediction for lung cancer. As noted by prior studies in cancer prognostication [29–32] , C-index values between 0.6 and 0.7 can still provide valuable clinical insights, particularly when dealing with complex

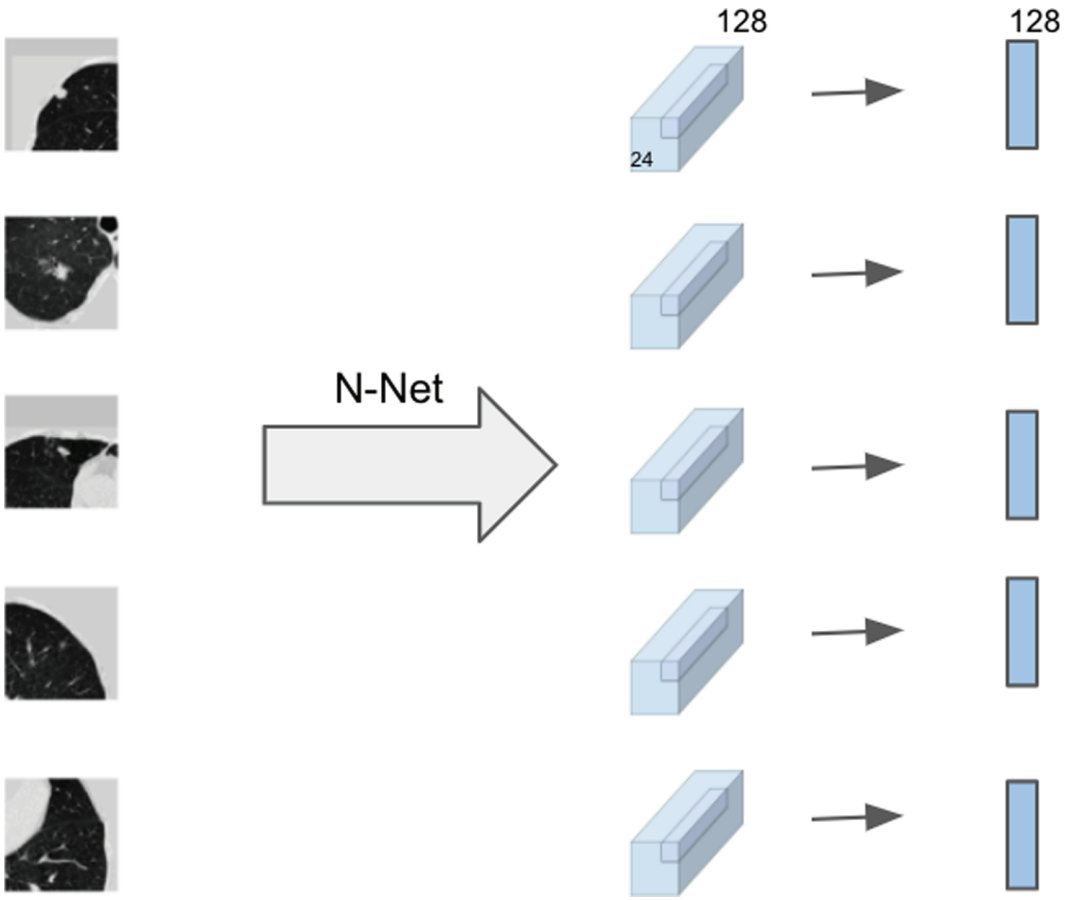

Fig 9. Feature extraction. Using a pre-trained classifier to get features from the top five suspicious crops.

https://doi.org/10.1371/journal.pone.0323174.g009

Table 8. Integration phase structure.

Layer Type	Output Size
Convolutional Phase Output	128 × 5
Fully Connected	32 × 5
Max Pool	1 × 5
Fully Connected	1

https://doi.org/10.1371/journal.pone.0323174.t008

diseases like lung cancer where numerous factors influence survival outcomes. The statistical significance of our improvements ($p < 0.001$) compared to the baseline indicates that our approach captures meaningful prognostic information from CT images.

The narrow confidence intervals observed for most models indicate high stability in both cancer classification and survival prediction performance. Interestingly, while ResNet demonstrated superior classification performance (highest AUC), VGG achieved the best survival prediction (highest C-index), suggesting that different architectural characteristics may be optimal for different predictive tasks in lung cancer analysis.

These results highlight both the promise of our approach and the inherent challenges in lung cancer survival prediction from imaging data alone, pointing to opportunities for further refinement by potentially incorporating additional clinical variables or more sophisticated modeling techniques.

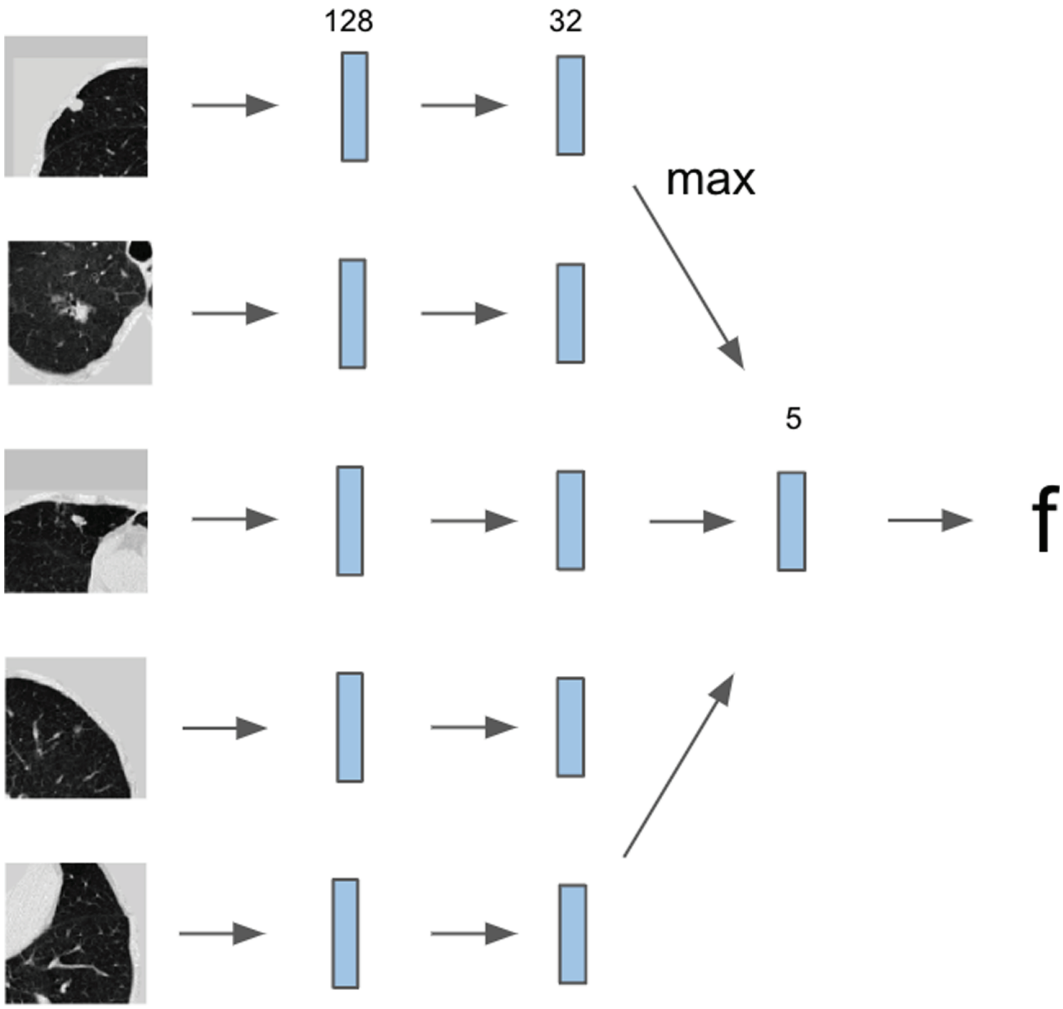

Fig 10. Feature integration.. Graphical representation of feature integration process.

https://doi.org/10.1371/journal.pone.0323174.g010

Table 9. Model performance and statistical significance compared to baseline N-net.

Model	AUC		C-index	
	Score (95% CI)	p-value	Score (95% CI)	p-value
N-net [36]	0.5428 (0.5423-0.5438)	ref	0.5088 (0.5082-0.5094)	ref
AlexNet	0.6441 (0.6329-0.6553)	<0.001	0.5823 (0.5754-0.5891)	<0.001
ResNet	**0.6885** (0.6881-0.6890)	<0.001	0.5987 (0.5981-0.5993)	<0.001
VGG	0.6790 (0.6788-0.6793)	<0.001	**0.6036** (0.6033-0.6039)	<0.001

ref: reference baseline

https://doi.org/10.1371/journal.pone.0323174.t009

6 Discussion

The results of our study demonstrate that combining binary cross-entropy and mini-batched loss, obtained by extending the Cox proportional hazards model with 3D CNNs, can significantly improve lung cancer diagnosis and survival prediction. Our approach achieves an AUC of 0.6885 for lung cancer classification and a C-index of 0.6036 for survival prediction on the NLST dataset, substantially outperforming the baseline method [36] which obtained an AUC of 0.5428 and C-index of 0.5088. These improvements were found to be statistically significant ($p<0.001$), indicating robust performance enhancements.

Our approach features three key strengths. First, the use of mini-batched loss effectively handles the non-convexity induced by neural networks, enabling efficient training on large datasets. The combination with binary cross-entropy allows joint optimization of both classification and hazard prediction tasks, providing a more comprehensive understanding of lung cancer compared to training separate models for diagnosis and prognosis.

Second, our framework offers exceptional generalizability, providing a versatile blueprint that can be adapted to various diseases and imaging modalities. A significant advantage is its applicability to any medical image—whether 3D or 2D—requiring survival prediction. The architectural flexibility allows the model to be tailored based on unique image characteristics, disease pathology, and data availability. This versatility positions our method as a broadly applicable tool across diverse medical imaging contexts.

Third, our unified framework provides a flexible architecture for integrating multiple data modalities beyond imaging. While our current implementation focuses on CT image features, the model's structure inherently supports the incorporation of additional clinical variables such as patient demographics, smoking history, genetic markers, and other relevant clinical metadata. We can integrate these linear features alongside medical images within the mini-batched loss framework. This multi-modal approach directly addresses potential confounding factors that could lead to misclassification or overdiagnosis. By allowing seamless integration of diverse data types, our method can provide a more nuanced and comprehensive risk assessment that transcends the limitations of single-modality analysis. The ability to incorporate contextual clinical information represents a significant advancement in personalized medical risk prediction, enabling a more holistic approach to understanding disease progression and individual patient outcomes.

Despite these advances, our approach has some limitations that suggest directions for future work. First, while 3D CNNs enable capturing complex patterns in CT scans, their black-box nature makes it challenging to interpret the features driving predictions. Developing techniques to improve CNN interpretability, such as visual explanations of informative regions [39], would help build clinician trust. Second, potential biases could arise if the NLST data does not fully represent the target screening population. Evaluating on additional diverse datasets and employing bias mitigation methods are important future steps.

Taken together, our work demonstrates the significant potential of deep learning-based survival models for improving medical image analysis across multiple clinical applications. While our implementation focused on lung cancer, the fundamental integration of CNN architectures with survival analysis provides a versatile framework applicable to numerous diseases and imaging modalities. The statistical improvements over baseline methods underscore the value of our unified approach to diagnosis (disease occurrence classification) and prognosis (survival hazard prediction). As we continue to refine these methods through clinical validation across different medical contexts, we envision this technology becoming an increasingly valuable complement to expert assessment, ultimately contributing to more precise and personalized healthcare.

7 Conclusion

In this paper, we presented a novel deep learning-based method that integrates lung cancer classification and survival prediction into a unified framework. Our approach extends the Cox proportional hazards model with 3D CNNs and leverages a combination of binary cross-entropy and mini-batched loss to efficiently train on large-scale CT datasets. Empirical results on the NLST dataset demonstrate statistically significant improvements over prior methods, with our best models achieving an AUC of 0.6885 (ResNet) and C-index of 0.6036 (VGG).

Our framework makes important strides towards harnessing the power of deep learning and medical imaging for advancing healthcare. The approach is particularly valuable for its generalizability across different medical imaging contexts—capable of handling both 2D and 3D images while providing robust survival predictions. This versatility positions our method as a broadly applicable tool in medical image analysis beyond just lung cancer.

The significant statistical improvements over baseline methods underscore the value of our unified approach to classification and prognosis. As we continue to refine these methods through clinical validation across different medical contexts, we envision this technology becoming an increasingly valuable complement to expert assessment, ultimately contributing to more precise and personalized healthcare across multiple diseases and imaging modalities.

Acknowledgments

The authors are most grateful to the reviewers for their most constructive and helpful comments.

Author contributions

Conceptualization: Xiawei Wang, James Sharpnack, Thomas C.M. Lee.

Data curation: Xiawei Wang.

Formal analysis: Xiawei Wang, James Sharpnack, Thomas C.M. Lee.

Funding acquisition: Thomas C.M. Lee.

Investigation: Xiawei Wang, James Sharpnack.

Methodology: Xiawei Wang, James Sharpnack, Thomas C.M. Lee.

Project administration: Xiawei Wang, James Sharpnack, Thomas C.M. Lee.

Resources: Thomas C.M. Lee.

Software: Xiawei Wang, James Sharpnack.

Supervision: James Sharpnack, Thomas C.M. Lee.

Validation: Xiawei Wang, James Sharpnack.

Visualization: Xiawei Wang, James Sharpnack.

Writing – original draft: Xiawei Wang.

Writing – review & editing: Xiawei Wang, James Sharpnack, Thomas C.M. Lee.

References

1. Alberg AJ, Samet JM. Epidemiology of lung cancer. Chest. 2003;123(1 Suppl):21S–49S. https://doi.org/10.1378/chest.123.1_suppl.21s PMID: 12527563

2. Spiro SG, Silvestri GA. One hundred years of lung cancer. Am J Respir Crit Care Med. 2005;172(5):523–9. https://doi.org/10.1164/rccm.200504-531OE PMID: 15961694
3. Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access. 2021;9:39707–16.
4. Lee C, Yoon J, Schaar M van der. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng. 2020;67(1):122–33. https://doi.org/10.1109/TBME.2019.2909027 PMID: 30951460
5. Chen Q, Xie W, Zhou P, Zheng C, Wu D. Multi-crop convolutional neural networks for fast lung nodule segmentation. IEEE Trans Emerg Top Comput Intell. 2022;6(5):1190–200. https://doi.org/10.1109/tetci.2021.3051910
6. Du J, Guan K, Zhou Y, Li Y, Wang T. Parameter-free similarity-aware attention module for medical image classification and segmentation. IEEE Trans Emerging Topics Comput Intell. 2023;7(3):845–57. https://doi.org/10.1109/TETCI.2022.3199733
7. Li T, Zhou F, Zhu Z, Shu H, Zhu H. A label-fusion-aided convolutional neural network for isointense infant brain tissue segmentation. Proc IEEE Int Symp Biomed Imaging. 2018;2018:692–5. https://doi.org/10.1109/ISBI.2018.8363668 PMID: 30555624
8. Mielke MM, Kozauer NA, Chan KCG, George M, Toroney J, Zerrate M, et al. Regionally-specific diffusion tensor imaging in mild cognitive impairment and Alzheimer's disease. Neuroimage. 2009;46(1):47–55. https://doi.org/10.1016/j.neuroimage.2009.01.054 PMID: 19457371
9. Du W, Hannig J, Lee T, Su Y, Zhang C. AutoGFI: streamlined generalized fiducial inference for modern inference problems. arXiv, preprint, 2024. https://doi.org/10.48550/arXiv.2404.08169
10. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016:2424–33. https://doi.org/10.1109/CVPR.2016.266 PMID: 27795661
11. Gao R, Huo Y, Bao S, Tang Y, Antic SL, Epstein ES, et al. Distanced LSTM: time-distanced gates in long short-term memory models for lung cancer detection. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2019, pp. 310–8.
12. Wang J, Gao R, Huo Y, Bao S, Xiong Y, Antic SL, et al. Lung cancer detection using co-learning from chest CT images and clinical demographics. In: Medical Imaging 2019: Image Processing. vol. 10949. International Society for Optics and Photonics; 2019, p. 109491G. https://doi.org/10.1117/12.2512965 PMID: 31602088
13. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954–61. https://doi.org/10.1038/s41591-019-0447-x PMID: 31110349
14. Liu S, Setio AAA, Ghesu FC, Gibson E, Grbic S, Georgescu B, et al. No surprises: training robust lung nodule detection for low-dose CT scans by augmenting with adversarial attacks. IEEE Trans Med Imaging. 2021;40(1):335–45. https://doi.org/10.1109/TMI.2020.3026261 PMID: 32966215
15. Zhang D, Li L, Sripada C, Kang J. Image-on-scalar regression via deep neural networks. arXiv, preprint, 2020. https://arxiv.org/abs/2006.09911
16. Han Y, Lee TCM. Structural break detection in non-stationary network vector autoregression models. IEEE Trans Netw Sci Eng. 2024;11(5):4134–45. https://doi.org/10.1109/tnse.2024.3398002
17. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. Open Biol. 2017;7(9):170070. https://doi.org/10.1098/rsob.170070 PMID: 28878044
18. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24. https://doi.org/10.1186/s12874-018-0482-1 PMID: 29482517
19. Zhu X, Yao J, Huang J. Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2016, pp. 544–7.
20. National Lung Screening Trial Research Team, Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, et al. The National Lung Screening Trial: overview and study design. Radiology. 2011;258(1):243–53. https://doi.org/10.1148/radiol.10091808 PMID: 21045183
21. Cox DR. Regression models and life-tables. J R Stat Soc B Stat Methodol. 1972;34(2):187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
22. Breslow NE. Discussion of professor Cox's paper. J Royal Stat Soc B. 1972;34:216–7.
23. Cox DR. Partial likelihood. Biometrika. 1975;62(2):269–76. https://doi.org/10.1093/biomet/62.2.269

24. Cao W, Wu R, Cao G, He Z. A comprehensive review of computer-aided diagnosis of pulmonary nodules based on computed tomography scans. IEEE Access. 2020;8:154007–23. https://doi.org/10.1109/access.2020.3018666
25. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D deep learning on medical images: a review. Sensors (Basel). 2020;20(18):5097. https://doi.org/10.3390/s20185097 PMID: 32906819
26. Cao Z, Qin T, Liu TY, Tsai MF, Li H. Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 129–36.
27. Wu L, Hsieh CJ, Sharpnack J. Sql-rank: A listwise approach to collaborative ranking. In: International Conference on Machine Learning. PMLR; 2018, pp. 5315–5324.
28. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–6. https://doi.org/10.1001/jama.1982.03320430047030 PMID: 7069920
29. Sun W, Jiang Y-Z, Liu Y-R, Ma D, Shao Z-M. Nomograms to estimate long-term overall survival and breast cancer-specific survival of patients with luminal breast cancer. Oncotarget. 2016;7(15):20496–506. https://doi.org/10.18632/oncotarget.7975 PMID: 26967253
30. Huang L, Chen J, Hu W, Xu X, Liu D, Wen J, et al. Assessment of a radiomic signature developed in a general NSCLC cohort for predicting overall survival of ALK-positive patients with different treatment types. Clin Lung Cancer. 2019;20(6):e638–51. https://doi.org/10.1016/j.cllc.2019.05.005 PMID: 31375452
31. de Boer JD, Putter H, Blok JJ, Alwayn IPJ, van Hoek B, Braat AE. Predictive capacity of risk models in liver transplantation. Transplant Direct. 2019;5(6):e457. https://doi.org/10.1097/TXD.0000000000000896 PMID: 31321293
32. Chai H, Xia L, Zhang L, Yang J, Zhang Z, Qian X, et al. An adaptive transfer-learning-based deep Cox neural network for hepatocellular carcinoma prognosis prediction. Front Oncol. 2021;11:692774. https://doi.org/10.3389/fonc.2021.692774 PMID: 34646759
33. Deng L. The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag. 2012;29(6):141–2. https://doi.org/10.1109/msp.2012.2211477
34. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009. Available from: https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf
35. Zuidhof G. Full preprocessing tutorial. 2017. Available from: https://www.kaggle.com/code/gzuidhof/full-preprocessing-tutorial.
36. Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network. IEEE Trans Neural Netw Learn Syst. 2019;30(11):3484–95. https://doi.org/10.1109/TNNLS.2019.2892409 PMID: 30794190
37. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv, preprint, 2014. https://doi.org/arXiv:1409.1556
38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 770–8.
39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 618–26.