Time to Event Data and Machine Learning
Final Project

Itamar Tzafrir (ID: 211400007) Anat Cohen (ID: 209796663)

July 2025

Abstract

Lung cancer is known as one of the most common and deadly types of cancer. Survival analysis is vital to accurately estimate the expected time to death of lung cancer patients, enabling clinicians to make informed treatment decisions and provide appropriate patient counseling. In this study, we investigate and attempt to reproduce the DeepMMSA framework for multi-modal survival analysis in non-small cell lung cancer (NSCLC) which combines CT imaging and structured clinical data with deep learning. While reproducing the original results, we identify several methodological limitations—most notably the treatment of survival prediction as a regression task using only uncensored data. We propose an alternative approach that incorporates censored data using a Cox-based deep learning survival model and introduce architectural and preprocessing improvements.

Introduction and Motivation

Lung cancer accounts for approximately 18% of all cancer-related deaths worldwide [2]. Despite significant advances in diagnosis and treatment, it remains one of the deadliest cancers, with more than half of patients dying within a year of diagnosis and a persistently low 5-year survival rate. Most lung cancer statistics include both small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). In general, approximately 13% of all lung cancers are SCLC, and about 87% are NSCLC [14].

Survival analysis plays an important role in healthcare research by modeling the time between the start of patient follow-up and the occurrence of clinically significant events, such as death. Prognostic models derived from survival analysis not only allow exploration of complex interactions between risk factors, but also enable personalized outcome predictions for new patients by leveraging historical healthcare data. These insights can empower clinicians to make earlier and more informed treatment decisions, an essential factor in improving outcomes for patients with aggressive diseases like lung cancer.

Conventional survival models rely on structured clinical data (e.g., age, sex, and tumor-stage), however, medical imaging technology now provides rich unstructured data for diagnosis, treatment planning, and also survival analysis. A commonly used approach for incorporating such imaging data is *Radiomics*, a field focused on the manual extraction of quantitative features from medical images. In our context, these features capture information related to tumor shape (e.g., volume, sphericity), texture (e.g., GLCM, GLRLM), and intensity-based statistics [9]. Several open-source and commercial tools have been developed to support radiomic feature extraction, with Pyradiomics [6] being one of the most commonly used libraries in the field. Despite their utility, hand-crafted radiomic features may fall short in capturing complex, abstract patterns in medical images [17].

DeepMMSA [15] addresses this limitation through a multi-modal deep learning framework designed for survival analysis of non-small cell lung cancer (NSCLC) patients by integrating CT images and clinical data. It uses 3D convolutional neural networks for feature extraction from volumetric images and combines them with fully connected layers processing clinical features.

In this study, inspired by the DeepMMSA framework, we initially sought to replicate their results, but diverged from their methodology fairly early because of critical errors made in their method, particularly in relation to survival analysis. Specifically, DeepMMSA treats survival prediction as a regression task by excluding censored patients during training and optimizing a mean absolute error (MAE) loss. This approach neglects the fundamental nature of survival data, where censoring carries crucial information about the time-to-event distribution. In contrast, we adopt a principled survival modeling framework that retains censored instances and incorporates them through a partial likelihood-based loss, aligning more closely with the Cox proportional hazards model. This shift enables us to model event times more accurately and make better use of the available data.

Survival Analysis

Survival analysis differs from standard regression tasks in that it models not just whether an event occurs, but when it occurs, while accounting for incomplete observations. A critical challenge in survival data is right-censoring, where the true event time T_i may not be observed. Instead, we observe $Y_i = \min(T_i, C_i)$ and event indicator $\delta_i = I(T_i \leq C_i)$ for each patient i, where C_i is the censoring time. This happens when patients are lost to follow-up or if the study ends before the event occurs. The hazard function $h(t)$ represents the instantaneous risk of the event occurring at time t, given survival up to that point. The Cox proportional hazards (Cox PH) model [5], one of the most widely used survival models, assumes that the hazard for an individual can be expressed as

$$h(t|x_i) = h_0(t) \exp\left(\beta^T x_i\right)$$

where $h_0(t)$ is the baseline hazard and $\beta \in \mathbb{R}^p$ represents the coefficients for covariates x_i. This assumption leads to what is known as the Cox PH assumption, that the hazard ratio between any two individuals is constant over time. Formally,

$$\frac{h(t|x_i)}{h(t|x_j)} = \frac{h_0 \exp(x_i^T \beta)}{h_0 \exp(x_j^T \beta)} = \exp((x_i - x_j)^T \beta).$$

If the Cox PH assumption does not hold, the estimated parameters may be biased and the statistical inference may be invalid, resulting in poor model fit. The model is fitted by maximizing the partial likelihood:

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\beta^T x_i)}{\sum_{j\in R(t_i)} \exp(\beta^T x_j)}$$

Where $R(t_i) = \{j : Y_j \geq t_i\}$ denotes the risk set at time t_i. Modern deep learning extensions, such as DeepSurv [10], replace the linear predictor $\beta^T x_i$, with a neural network $f_\theta(x_i)$ yielding

$$h(t|x_i) = h_0(t) \exp\left(f_\theta(x_i)\right).$$

The network parameters θ are optimized by minimizing the negative log partial likelihood

$$\mathcal{L}(\theta) = -\sum_{i:\delta_i=1} \left[f_\theta(x_i) - \log \sum_{j\in R(t_i)} \exp(f_\theta(x_j)) \right].$$

This preserves the semi-parametric nature of the Cox model while allowing for complex nonlinear covariate effects.

A common metric used to evaluate the predictive accuracy of survival analysis models is the concordance index (C-index) [7], which is a generalization of the AUC for time-to-event data . The C-index is defined as

$$C = \frac{\sum_{i,j} 1_{\{T_i<T_j\}}\delta_i \cdot 1_{\{\hat{r}_i>\hat{r}_j\}}}{\sum_{i,j} 1_{\{T_i<T_j\}} \cdot \delta_i}$$

where $\hat{r_i}, \hat{r_j}$ are the predicted risk of patients i, j respectively. Intuitively, the C-index is the probability that for any two randomly selected pair of observations, the one that experiences the event first is of higher risk than the other. The C-index ranges in $[0, 1]$ where $C = 1$ means perfect prediction, $C \approx 0.5$ is not better than a random guess, and lastly, $C < 0.5$ is worse than a random guess, a systematically wrong model. The C-index is a strong evaluation metric because it handles censored data and is intuitive.

Dataset

This study utilizes the publicly available *Lung1* dataset from The Cancer Imaging Archive (TCIA) [4], which consists of clinical records and 3D CT chest scans collected from 422 NSCLC patients (35.7GB total). The clinical data is in traditional tabular format and includes patient ID, age, clincial T, N and M stages, overall stage, histology, survival time (in days) and death indicator (1 if death occurred, 0 if censored). Censored observations comprise 11.6% of the dataset. A description of the clinical data is given in Table 1. See Figure 1 for the distributions of the clinical data features and survival data.

The 3D CT scans are pre-treatment chest scans of each patient in DICOM (Digital Imaging and Communications in Medicine) format. Tumor segmentations are provided through RTSTRUCT (Radiotherapy Structure Set) files containing radiologist delineations of tumor boundaries. An example of an axial slice from the

Lung1 dataset with the tumor segmentation is provided in Figure 3.

Methodology

Whilst DeepMMSA aims to predict survival time using both clinical and imaging data through a regression approach, we adopt a Cox PH framework that appropriately models the hazard function and incorporates all patients, including those with censored follow-up. This approach predicts log-hazard ratios rather than direct survival times, enabling proper statistical inference on time-to-event data. Our methodology integrates deep learning feature extraction with classical survival analysis; CT scans and clinical data undergo separate preprocessing and feature extraction through neural networks, followed by multimodal fusion. The combined features are processed through a neural network that outputs log-hazard ratios, which are optimized using the Cox partial negative log-likelihood loss function. We validate the proportional hazards assumption using statistical tests from the 'lifelines' library (using the the Schoenfeld residuals test), confirming the appropriateness of our modeling framework. Model performance is evaluated using the C-index, which properly accounts for censored observations and handles tied survival times in its ranking calculations. [1]

Preprocessing

As mentioned, we pre-processes the 3D scans and the clinical records separately, as handled in the paper. Two patients are dropped due to failed preprocessing, or corrupted RTSTRUCT files, leaving a total of 420 patients with both clinical data and CT scans.

Clinical Data

Missing values for 'age' and 'histology' are filled in using the mean and mode values respectively. Two specific patients with missing data for 'clinical T stage' and 'overall stage' respectively, are filled in according to suggestions from [13]. Categorical features are one-hot encoded and continuous features ('age') are standardized using min-max scaling, following the steps done in the paper. This results in 24 features per patient. Note that all clinical features represent baseline measurements at diagnosis, collected concurrently with the pre-treatment CT scans. Since our model predicts survival from this fixed baseline point, we do not require time-varying covariates or adjustments for temporal changes in clinical features.

CT Data

Medical scans often have anisotropic spacing, as an example from the dataset, $0.97 \times 0.97 \times 3$ mm, where the slice thickness (z) is much larger than the in-plane resolution (x and y). 3D CNNs assume uniform voxel dimensions, and anisotropy can distort kernel operations. To standardize the data and ensure anatomical consistency across patients we retrieve the original spacing from the CT metadata and resample to $1 \times 1 \times 1$ mm through interpolation, as suggested in [9]. This standardization ensures anatomical consistency between patients and accounts for variations in imaging protocols and scanner settings.

[1] Code available at: github.com/anatcohen/SurvivalAnalysisFP

In CT imaging, intensities are measured in Hounsfield Units (HU), typically +1000 for bone, zero for water, and -1000 for air. To focus on relevant soft-tissue structures (such as lung tissue and tumors), and remove irrelevant high-density tissues (like bone), we clip the HU values to [-1000,400] improving contrast in the range relevant for lung cancer [11]. The clipped values are then normalized to [0, 1].

Unlike DeepMMSA's simplistic ”resize to $96 \times 96 \times 8$,” our method rigorously preserves physical consistency. Tumor contours are extracted from the RTSTRUCT DICOM files and used to create a 3D tumor mask. After analyzing tumor dimensions across the dataset Figure 2, we set a fixed bounding box of $120 \times 120 \times 120$ mm. This size accommodates the majority of tumors. This box is centered at the tumor's calculated center and cropped from the CT and masked according to the contours of the segmented tumor. The masked tumor data is then resized to a CNN-friendly fixed input size of $64 \times 64 \times 64$.

Feature Extraction Architecture

Clinical Feature Extraction

We employ a multilayer perceptron (MLP) to extract meaningful representations from the preprocessed clinical features. The architecture consists of a 24-dimensional input layer (after one-hot encoding and normalization) a hidden layer with 32 neurons with batch normalization and ReLU activation and a final 16-dimensional hidden layer.

3D Convolutional Network for CT Images

For volumetric feature extraction from CT scans, we utilize a 3D ResNet-18 architecture [8], it extends the classic ResNet-18 by replacing all 2D operations with 3D convolutions. This network comprises of an initial 3D convolutional layer ($7 \times 7 \times 7$ kernel, stride 2) with 64 filters, batch normalization and ReLU. Then a 3D max pooling layer ($3 \times 3 \times 3$ kernel, stride 2) followed by four residual blocks with progressively increasing channels. Finally, a fully connected layer for an output of 256 dimensions.

Unlike DeepMMSA, we initialize the network with pre-trained Med3D weights [3], which has been reported to accelerate the training convergence speed of 3D medical tasks by up to 10 times compared to training from scratch. To adapt the network for our specific task while preventing overfitting, early layers (conv1, layer1, layer2) are frozen during training, preserving low-level feature detectors. Later layers (layer3, layer4, and fc) remain trainable to learn task-specific features.

Multimodal Fusion and Survival Prediction

Following DeepMMSA's method, first feature extraction is applied on both the clinical data and the CT scans separately. Prior to fusion, features from each modality undergo batch normalization to ensure comparable scales. That is, given features $z_1, ..., z_m$ in a batch, we calculate

$$\hat{z_i} = \gamma_i \frac{z_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i$$

where μ_i, σ_i are the mean and standard deviation of z_i respectively, γ_i, β_i are learnable scale and shift parameters, and ϵ a small constant for numerical stability. The normalized image features (256-dimensional) and clinical features (16-dimensional) are concatenated to form a 272-dimensional combined representation.

The fused features are processed through a survival prediction head that outputs a single risk score compatible with the Cox PH framework. This includes a hidden layer with 32 neurons with batch normalization, ReLU, and dropout of 0.6, then a one dimensional output layer. The network is trained to minimize the negative partial log-likelihood.

Training Procedure

Data Augmentation

Given the limited dataset size and the substantial complexity of the deep learning model (containing millions of parameters from the 3D ResNet-18 alone), data augmentation is crucial to combat overfitting and improve generalization. We employ online augmentation during training, where each CT scan has a probability of 80% to undergo random 3D rotations of up to 15° in each axis. A visual example for the augmentation pipeline can be seen in Figure 5. This approach ensures the model encounters different transformations of each patient across epochs, effectively creating unlimited variations while maintaining the original training set size of 252 patients. Online augmentation provides superior regularization compared to pre-computed augmentations, as the stochastic nature prevents the model from memorizing specific transformed versions.

We apply augmentation exclusively to the training set. This contrasts with DeepMMSA's flawed approach of applying 8x ”offline” augmentation to the entire dataset, including validation and test sets. Their method introduces severe methodological issues: (1) potential data leakage if augmentation precedes train-test splitting, allowing the model to see variations of test patients during training, and (2) inflated performance metrics by evaluating on augmented test data rather than realistic, unmodified patient scans. Such practices fundamentally compromise the validity of their reported results and prevent accurate assessment of generalization to truly unseen patients.

Optimization Details

The model is trained using Adam optimizer (learning rate: 1×10^{-3}, weight decay: 1×10^{-2}) with batch size 32 for up to 100 epochs. The learning rate is reduced by factor 0.5 when the validation C-index plateaus (patience=5), and training stops early if no improvement is observed for 10 epochs.

Regularization Techniques

To combat overfitting on our small dataset, we employ multiple regularization strategies: high dropout (0.6), strong L2 weight decay (0.01), frozen early convolutional layers (preserving pre-trained features), and online data augmentation on training samples. This aggressive regularization is necessary due to the substantial imbalance between model capacity and dataset size, addressing the fundamental challenge of training millions of parameters on a limited dataset.

Experiment Setup

The dataset is split into training, validation, and testing sets using a 6:2:2 ratio. The model is fitted on the training set, early stopping is determined by the validation set, and final performance is evaluated on the test set. This process is repeated 50 times to produce 50 trained models on various train/validation/test splits evaluated separately. The data is preprocessed as described above according only to the training data. Specifically, scaling parameters and imputed values for missing data are computed from the training set alone, not from the entire dataset.

Results

A summary of the results is presented in Table 2. Regarding DeepMMSA's reported C-index of 0.658, we exclude this from our comparison due to fundamental methodological flaws in their evaluation procedure, particularly their inclusion of augmented test data and exclusion of censored patients. In our opinion, these flaws compromise the validity of their reported performance.

We first establish baseline performance using traditional survival models applied only to clinical data. Ridge-penalized Cox regression and Random Survival Forest models were trained and evaluated across 50 different train-validation-test splits, with hyperparameters optimized on the validation set. It is important to note that we also checked that indeed the PH assumption holds on the clinical data. The baseline models achieved mean C-indices of 0.540 (Cox regression) and 0.548 (Random Survival Forest). The baseline results show that the traditional models perform only slightly better than random guessing.

Our proposed multi-modal approach, combining clinical and CT imaging features, achieved a mean C-index of 0.572 across 50 runs. While this represents a modest improvement over the clinical-only baselines, the performance remains suboptimal. We would like to note that computational constraints restricted architecture exploration and hyperparameter optimization which would likely result in better performance.

For additional context, we note that a radiomics-based Cox model using hand-crafted imaging features [1] achieved a C-index of 0.609, averaged across 100 train/validation/test splits [12], outperforming our deep learning approach. This suggests that automatic feature learning from raw CT data may require larger datasets or more sophisticated architectures to surpass engineered features.

Discussion

Results Interpretation

Our study demonstrates a principled approach to multi-modal survival analysis in NSCLC, properly incorporating censored observations through deep Cox PH modeling. While our achieved C-index of 0.572 ± 0.045 represents modest predictive performance, it consistently outperforms clinical-only baselines (C-index 0.54), validating the added value of imaging data for survival prediction. The performance gap between our deep learning approach and traditional radiomics-based methods (C-index: 0.609) suggests that automatic feature learning from raw CT data requires careful consideration of dataset size and model complexity.

Our modest results reflect the relative immaturity of deep learning in survival analysis compared to other medical AI domains. This delayed adoption stems from fundamental technical challenges: standard neural network loss functions cannot accommodate censored observations, which violate the assumption of fully observed outcomes required by conventional supervised learning. Additionally, survival datasets rarely exceed thousands of patients, far smaller than the millions of examples used to train successful deep learning models in computer vision or natural language processing. These constraints explain why traditional methods like Cox regression continue to dominate clinical practice despite the deep learning revolution elsewhere. However, recent methodological advances, including the development of survival-specific neural architectures (DeepSurv, DeepHit) and increasing availability of large-scale medical imaging cohorts, suggest the field is approaching a turning point. As these technical barriers are overcome and datasets expand, we expect multimodal deep learning approaches like ours to substantially improve and eventually complement or surpass traditional survival models in clinical applications.

Methodological Concerns with DeepMMSA

While our approach builds upon DeepMMSA's multimodal framework, we identified and addressed several methodological issues that compromise the validity of their results. Most significantly, DeepMMSA treats survival prediction as a regression task that directly predicts survival times, violating core survival analysis principles by failing to model hazard functions or survival probabilities and entirely ignoring censored observations, which contain valuable partial information about survival beyond the follow-up period. By training exclusively on uncensored patients their model introduces severe selection bias toward patients with poorer outcomes and fails to represent the full patient population. This selectiveness of the uncensored patients also causes the train-test split of the data to not be inherently random. Moreover, DeepMMSA uses the lack of censored training data to justify an inappropriate mean absolute error (MAE) metric to evaluate training data but survival metrics (C-index) on the test set including censored observations.

Secondly, during the pre-processing stage, the paper resizes the 3D CT scans to $96 \times 96 \times 8$. From our understanding, this approach drastically reduces the dimensions of the z-axis, disproportionally to the other axes. This is especially problematic because the histograms in Figure 2 clearly show that 8 pixels for the z-axis is too little to capture relevant information. They also do not address the anisotropic and variable nature of spacing in CT scans. When combined with their rotation-based augmentation, this oversight transforms anatomically meaningful structures into distorted representations, as rotations applied to non-isotropic data produce artifacts and false anatomical relationships. The cumulative effect is substantial information loss and biologically implausible image representations.

Thirdly, as mentioned earlier, it is unclear whether data augmentation on the 3D scans is applied prior to the train-test split. This could cause data leakage. Additionally, from our understanding, their clinical preprocessing pipeline apparently computes normalization parameters and imputation values from all 422 patients rather than exclusively from the training set. This allows information from test patients to contaminate the preprocessing steps, violating the fundamental principle of maintaining strict train-test separation. Furthermore, DeepMMSA's proposed 8x offline augmentation which was performed on the validation and test sets, prevent accurate assessment of the model's ability to generalize to truly unseen data.

While working on this study, we came across a more recent paper [16], published by the same authors, that introduces a more complex deep learning survival analysis model. As sophisticated and complex their new approach is, it still violates the same principles of survival analysis; predicting survival times, training on uncensored data and using an MAE based loss function. This suggests a continued misunderstanding of key survival analysis principles and raises concerns about the validity of the reported results.

Study Limitations

First, our dataset of 420 patients, while carefully curated, remains small for deep learning applications, potentially explaining why simpler radiomics approaches outperformed our method. Second, computational constraints limited our exploration of more sophisticated architectures and comprehensive hyperparameter optimization. Third, the absence of cause-specific mortality data prevents modeling competing risks, non-cancer deaths may bias survival estimates, particularly in elderly patients where cardiovascular disease or other conditions may intervene. Finally, our single-institution cohort may limit generalization across different populations and imaging protocols.

Future Directions

Despite modest absolute performance, our results confirm that CT imaging contains prognostic information beyond traditional clinical staging. The consistent improvement from multimodal integration suggests potential clinical utility with further development. Future work should focus on: (1) assembling larger, multi-institutional datasets to support more complex models; (2) incorporating temporal imaging to capture tumor evolution; (3) developing interpretability methods to understand which imaging features drive predictions.

Figures and Tables

Feature	Type	Range/Values	Description
Age	Continuous	33.68–91.70	Age at diagnosis (years)
Clinical T Stage	Ordinal	1, 2, 3, 4, 5	Size and extent of primary tumor (1 = smallest, 5 = largest)
Clinical N Stage	Ordinal	0, 1, 2, 3, 4	Regional lymph node involvement (0 = no spread, 4 = extensive spread)
Clinical M Stage	Ordinal	0, 1, 3	Distant metastasis (0 = none, 3 = extensive)
Overall Stage	Categorical	I, II, IIIa, IIIb	TNM classification of cancer extent
Histology	Categorical	Adenocarcinoma, Squamous cell carcinoma, Large cell, NOS	Histological subtype of cancer
Gender	Binary	Male, Female	Patient gender
Survival Time	Continuous	10–4,454	Time to death or last follow-up (days)
Death Status	Binary	0, 1	Vital status (0 = alive, 1 = deceased)

Table 1: Summary of features in the clinical data.

Model	**Input Modality**		**C-Index**
	Clinical (Tabular)	**CT Imaging**	
Regularized Cox	✓		0.540 ± 0.027
Random Survival Forest	✓		0.548 ± 0.029
Radiomics + Cox (Aerts et al.)† [12, 1]	✓	✓	0.609 ± 0.041
Proposed Model	✓	✓	0.572 ± 0.045

† Uses radiomic features extracted from CT images, not raw CT images directly.

Table 2: Comparison of different model performances by input modality. Results are reported as mean ± standard deviation over 50 iterations of train-val-test splits.

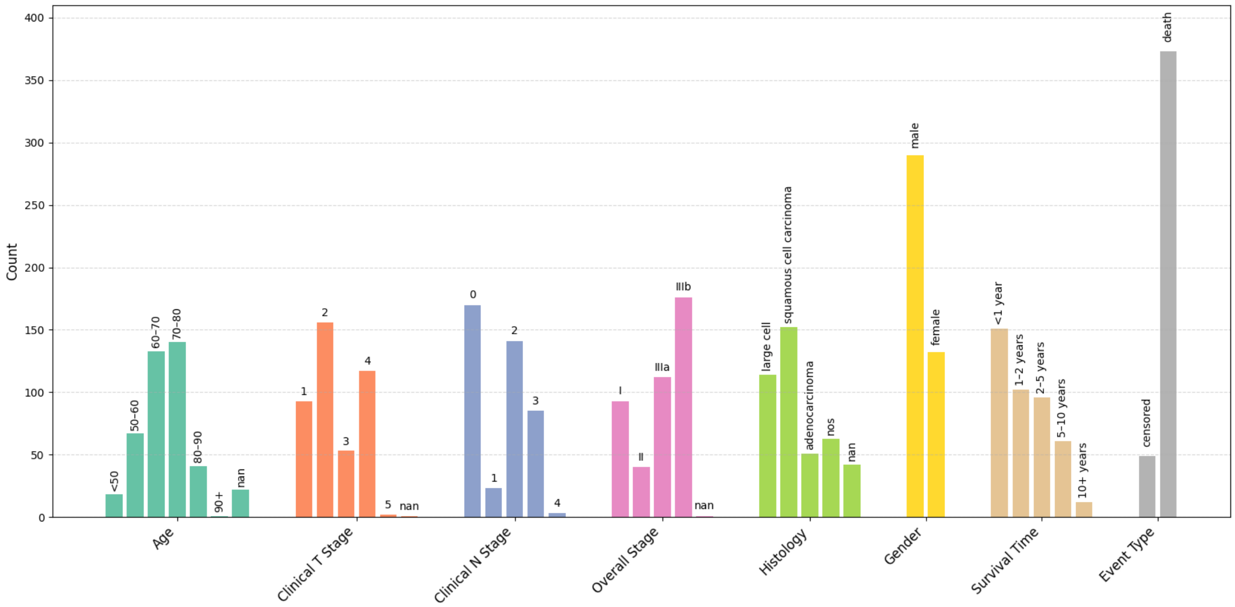

Figure 1: Distribution of clinical features and survival data prior to preprocessing.

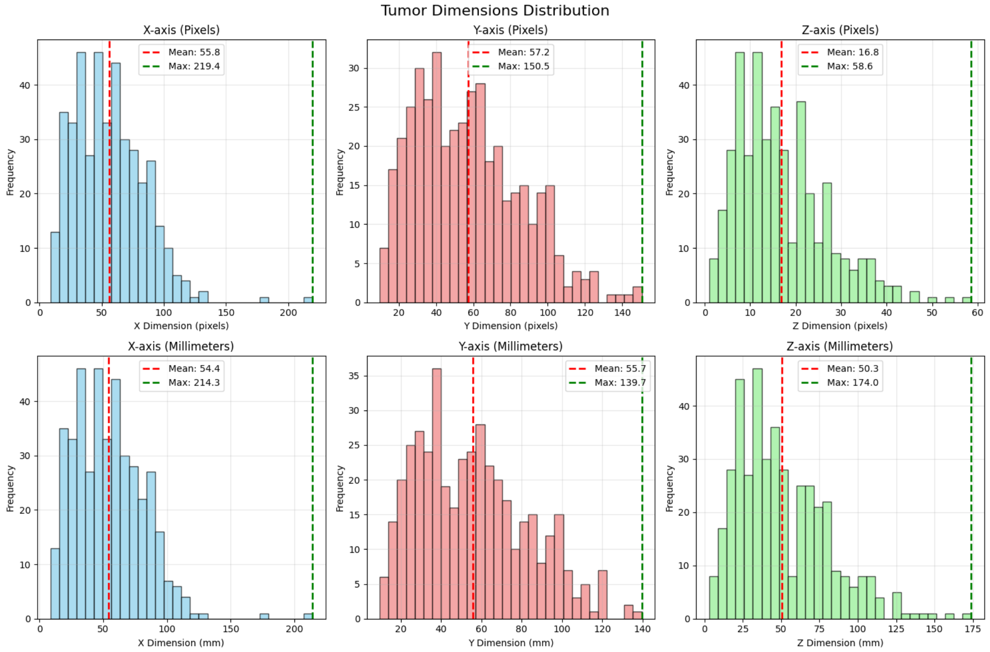

Figure 2: Distribution of tumor sizes based on contour annotations, shown both in pixels and in physical dimensions (mm). Note that in physical space, tumors are approximately of same length across dimensions. The discrepancy between pixel-based and physical measurements arises from anisotropic and variable spacing across different scans.

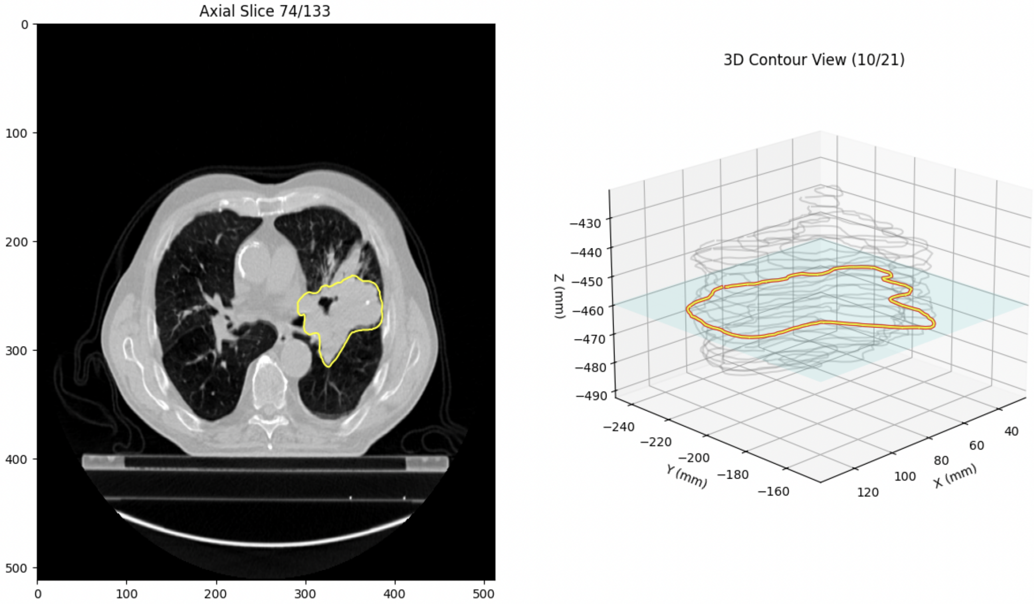

Figure 3: Tumor visualization for patient LUNG1-001 with a survival time of 2165 days. (A) Axial CT slice showing the GTV-1 (Gross Tumor Volume) contour in yellow. (B) 3D reconstruction of the tumor volume using RTSTRUCT-derived contours, highlighting the spatial extent of the tumor.

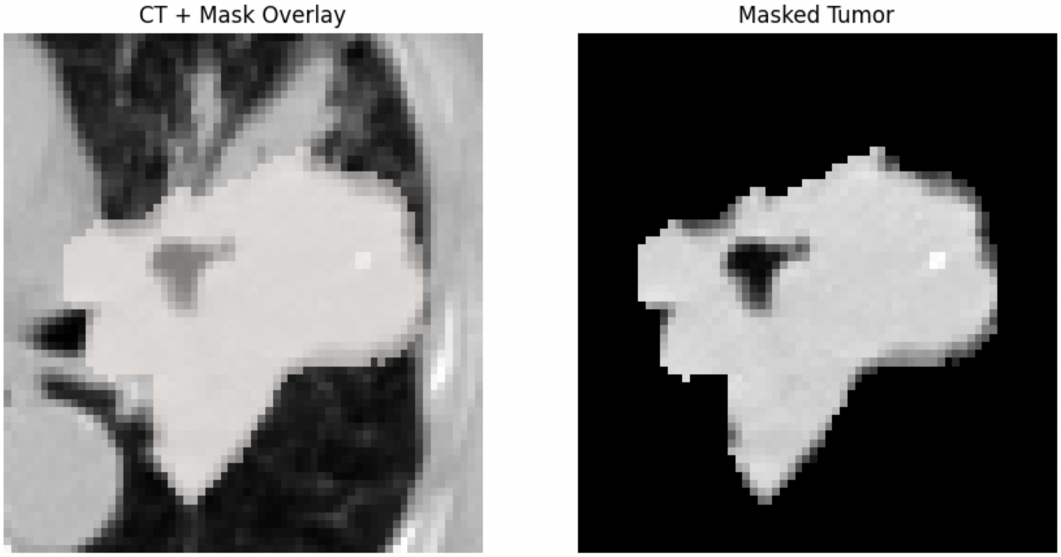

Figure 4: Preprocessed tumor volume visualization after automated extraction and masking pipeline. Axial slice from a $64 \times 64 \times 64$ preprocessed volume showing (A) the cropped CT region centered on the tumor with binary mask overlay, displaying both tumor and surrounding anatomical context, and (B) the isolated masked tumor region used as input to the 3D convolutional neural network.

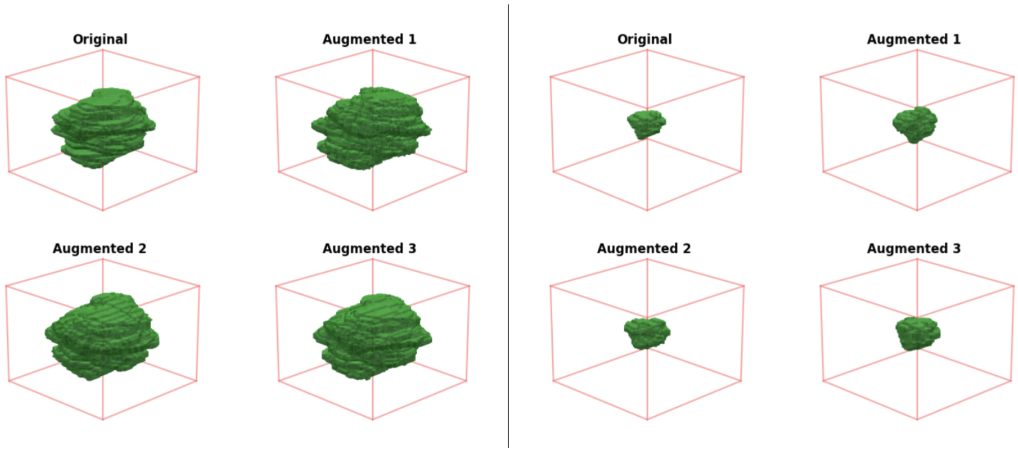

Figure 5: 3D visualization of lung tumor augmentation pipeline. Two patient examples (LUNG1-007 and LUNG1-012) showing original tumor masks and three rotational augmentations within $64 \times 64 \times 64$ voxel bounding boxes used as CNN input.

Software

All models were implemented in Python 3.12 with the following key dependencies:

- Deep Learning: PyTorch 2.7.1 for model implementation and training
- Survival Analysis: lifelines 0.30.0 (Cox models, C-index calculation, PH assumption testing), scikit-survival 0.24.1 (Random Survival Forest)
- Medical Imaging: SimpleITK 2.5.2 (DICOM I/O, resampling), pydicom 3.0.1 (RTSTRUCT parsing), scikit-image 0.25.2 (polygon drawing)
- Scientific Computing: NumPy 2.2.6, SciPy 1.16.0 (image transformations, ndimage operations), pandas 2.3.1
- Visualization: matplotlib 3.10.3

References

[1] Hugo J.W.L. Aerts, Emmanuel R. Velazquez, Ralph T.H. Leijenaar, Chintan Parmar, Patrick Grossmann, Stephanie Cavalho, Johan Bussink, Rene Monshouwer, Benjamin Haibe-Kains, Dirk Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5:4006, 2014.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[3] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[4] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Samir Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[5] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[6] J. J. M. Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Robin, S. Pieper, and H. J. W. L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.

[7] Frank E. Harrell, Kerry L. Lee, Robert M. Califf, David B. Pryor, and Robert A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 1(2):143–152, 1982.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Petros Kalendralis. *Artificial intelligence applications in radiotherapy: The role of the FAIR data principles*. Doctoral thesis, Maastricht University, 2022.

[10] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[11] Yucheng Li, Tao Liu, Zhiqiang Wang, Fang Li, Xuan Luo, Yuchen Zhou, Yanan Chen, Wenxin Zhang, and Dinggang Meng. Ct-lungnet: A deep learning framework for precise lung tissue segmentation in 3d thoracic ct scans. *arXiv preprint arXiv:2212.13971*, 2022.

[12] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. D. Cooper. Image-based survival prediction for lung cancer patients using cnns. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9284–9292. IEEE, 2018.

[13] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. Survival analysis in lung cancer: A comparative study of different approaches using nsclcradiomics (lung1) data. *Scientific Reports*, 5:11385, 2015.

[14] American Cancer Society. Key statistics for lung cancer. `https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html`.

[15] Yujiao Wu, Jie Ma, Xiaoshui Huang, Sai Ho Ling, and Steven Weidong Su. Deepmmsa: A novel multimodal deep learning method for non-small cell lung cancer survival analysis. In *Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021. arXiv:2106.06744.

[16] Yujiao Wu, Yaxiong Wang, Xiaoshui Huang, Fan Yang, Sai Ho Ling, and Steven Weidong Su. Multimodal learning for non-small cell lung cancer prognosis. *arXiv preprint arXiv:2211.03280*, 2022.

[17] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547, 2016.