METASTATIC LUNG CANCER PROGNOSIS VIA DEEP IMAGE-BASED LESION PRIORITIZATION

Forest Yang, Skander Jemaa†, Thomas Bengtsson°, Laurent El Ghaoui*+*

*UC Berkeley, Berkeley, CA, USA †Genentech, South San Francisco, CA, USA
°N-Power Medicine, Redwood City, CA, USA +VinUniversity, Hanoi, Vietnam

ABSTRACT

We propose to perform cancer prognosis via lesion prioritization, which proceeds in two steps. First, a deep neural network is run on CT image patches containing lesions to predict lesion-level risks. Then, lesion risks are aggregated to form patient risk. We show our approach outperforms other deep learning-based approaches in a low data regime and utilize lesion risks for interpretability, showing presence in bone, growth likelihood, and lung containment margin to be potential survival-relevant lesion features detected by the model.

1. INTRODUCTION

Cancer is a leading cause of death across the world. Within cancer, lung cancer is the leading cause of death due to its high incidence rate coupled with high mortality rate.

Here, we focus on deep learning prognosis models, which estimate patient survival from computed tomography (CT) scans. Accurate prognoses can better inform treatment decisions, quicken the treatment development process, and an interpretable prognosis model could aid human doctors in understanding the disease. Towards these goals, we propose a deep learning model that performs cancer prognosis by lesion prioritization.

Lesion prioritization assigns a risk to each individual lesion, where risk signifies the impact of a lesion on patient survival. A simple, intuitive, and clinically utilized measure of the risk of a lesion is its size, hence the use of size-based Response Evaluation Criteria in Solid Tumors (RECIST) and TNM staging clinical criteria. We hypothesize that a deep learning model could produce more informed risk estimates by accounting for factors other than lesion size.

Lesion-level risk predictions, as opposed to previous approaches which only predict patient-level risk, lend interpretability. A doctor could peruse the lesions identified as high-risk and, in case these contradict medical intuitions, know to trust the model less, or in case these conform to medical intuitions, have more confidence in the model. In more ambiguous cases, high risk lesions selected by the model may offer new insight as to what constitutes a dangerous lesion. For treatment development, changes in high risk lesions at followup may be more informative about treatment efficacy than RECIST, where the choice of which lesions to measure is subjective.

Our approach uses a convolutional neural network (CNN) to predict lesion risks and formulates patient risk as an aggregation of lesion risks. Thus, in the process of learning to correlate patient risk with survival, the model learns to correlate lesion risk with the impact of the lesion on survival. Our contributions are as follows:

- We propose and implement lesion prioritization, which predicts lesion-level risks and then aggregates them to predict patient survival, as a framework for deep metastatic lung cancer prognosis.
- We show that our model outperforms alternative deep learning approaches in a low data regime (around 100-200 patients) on metastatic lung cancer.
- By using the predicted lesion risks for model interpretation, we find that the model predicts higher risks for lesions outside the lung, particularly in bone. Furthermore, we found that predicted lesion risk is predictive of whether a lesion will grow, suggesting sensitivity of the model to features of a proliferation phenotype.

2. RELATED WORK

Machine learning and, more recently, deep learning, have been used in radiology for a broad set of tasks [1]. In this work, we focus on prognosis, the task of predicting patient outcome, in patients with metastatic non-small cell lung cancer.

Previous work on image-based cancer prognosis with deep learning often address lung cancer [2, 3, 4], but colorectal, pancreatic, and brain cancer have also been addressed [5, 6, 7]. Most approaches feed a single 2D or 3D box-shaped patch containing the primary tumor as input to the network [3, 2, 6, 7, 4]. Lu et al. [5], similarly to us, deal with metastatic cancer and feed in multiple lesions to the network. However, they output a diameter-weighted average of the lesion representations, whereas we predict a risk score for each lesion, allowing lesion-level comparisons. Furthermore, they ignore small lesions of less than $1\,\mathrm{cm}$ in diameter. Table 1 summarizes relevant previous work.

Most attempts at interpreting prior deep cancer prognosis models [2, 4, 5] consist of applying Grad-CAM [8], and show these models "pay attention" to tumor border in potentially meaningful ways. Hosny et al. [2] also conduct a gene-set enrichment analysis (GSEA), correlating gene expression measurements with model predictions to find enriched pathways. However, these works do not consider interpretation based on lesion-level risk scores, as they only predict patient-level risk scores.

3. DATASET

Our dataset consists of 258 patients with non-small cell lung cancer from a phase III clinical trial. 198 (77%) of cases are stage IV, and 246 (95%) are adenocarcinoma. The median survival is 1 year and 6.5 months, with observed survival for 186 patients and censoring times for the other 72. The lesions of each patient are segmented by one to four radiologists (average: 2.92), each radiologist producing a distinct segmentation mask. The median of the per-patient lesion count, averaged over radiologists, is 10.7, with interquartile range [5, 19.4], reflecting a high number of lesions per patient due to advanced stages of cancer.

Lesions vary widely in size, shape, and location. Axial slices from the chest of two example CT images are shown in Figure 1. Lesions occur frequently in the lung, mediastinum, bone, liver, and unlabeled regions, with rare occurrences in the spleen, kidney, and stomach. An ideal lesion scoring model must model risk for a wide range of lesion characteristics and surrounding tissues.

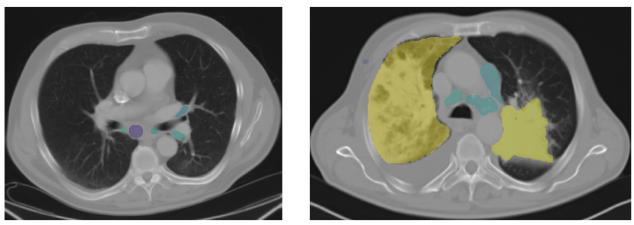

Fig. 1. Chest axial slices with lesions highlighted. Note the small, round nature of the lesions of the patient on the left and the large, irregular nature of the lesions of the patient on the right. The patient on the left had a survival time of 283 days and the patient on the right had a survival time of 855 days.

4. METHOD

Conceptually, our approach maps each lesion of a patient to a lesion-level risk score, and aggregates the lesion risk scores to form a patient-level risk score, trained using the negative proportional log likelihood (NPLL) loss [10, 11]. We extract a $2d \times 2d$ patch of each lesion of a patient where d is the lesion bounding box diameter and resize it to 72×72. Then we feed the lesion patch to a ResNet18 lesion scoring network to obtain risk score r_{l_i} and the lesion volume to a fully connected size scaler network to obtain scaling factor s_{l_i}. The patient risk score is the sum of the scaled lesion risks:

$$r_p = \sum_{i=1}^{nl_p} r_{l_i} s_{l_i}$$

where nl_p is the number of lesions in the annotation. In practice, we limit the number of lesions considered per patient annotation to 40. During training, we consider each annotation as a distinct example, but in evaluation, we average scores produced by different annotations for the same patient. A schematic of the approach is shown in Figure 2.

During training, we perform standard data augmentations to each lesion patch, namely: random shifting, scaling, flipping, and rotation.

For evaluation purposes, we ran 4-fold cross validation, partitioning the dataset into 4 equally sized parts with matched joint survival $\times$ treatment distributions, using two parts for training and one for validation and test each, cycling the parts to obtain 4 splits.

5. RESULTS

Overall, we found our model components to be performant, and the approach of aggregating lesion risks across all or near-all lesions to outperform other deep CT-based prognosis approaches on metastatic cancer.

5.1. Comparison to pretraining

We assessed the efficacy of a ResNet18 trained from scratch for the lesion scoring network by comparing with pretrained models finetuned on our task – a ResNet50 pretrained on ImageNet, and a ResNet50 pretrained on radiological image tasks [12]. Overall, while pretraining obtained higher training performance, the test and validation performance were not significantly higher. See the first three rows of Table 2.

5.2. Size scaler ablations

Next, we assessed the effect of the size scaler on performance. In the "no FC network" ablation, we replaced the fully connected size scaler network with multiplication by a learned constant, changing the scaling factor from a nonlinear to a linear function of volume. In another ablation, we removed the size scaler entirely and simply summed the lesion risks to generate patient risk. The ablation results are in columns 4 and 5 of Table 2.

A linear function of volume performed similarly to a nonlinear one, but removing size-based scaling altogether performed slightly worse. This emphasizes the benefit of including lesion size information for prognosis, though there may be room to optimize the way it is incorporated.

Ref	Cancer	Treatment	N_P	N_L	N_T	Model	Loss	AUC / C-index
[2]	Lung, Stage I-III	Radiotherapy	771	1	1	Custom 3D CNN	Binary (2 year)	0.70 AUC
[3]	Lung, Stage III	Radio– & chemotherapy	179	1	2–4	ResNet +GRU	Binary (2 year)	0.74 AUC
[4]	Lung, Stage I-II	Surgery	800	1	1	Custom 3D CNN	Nnet-survival[9]	0.74 C
[5]	Colorectal, Stage IV	Chemo– & targeted therapy	1028	1–10	2–4	Inception-v3 +BiLSTM	Binary (1 -year)	0.649 C
[6]	Pancreatic*	Unknown	205	1	1	3D-ResNet18 +LSTM	NPLL	0.683 C

Table 1. Previous approaches on deep cancer prognosis from images. N_P, N_L, N_T stand for the number of patients, number of lesions considered per patient, and number of scans taken at different time points considered respectively. If the model has "3D" in its name, inputs are 3D lesion-centered CT patches. Otherwise, inputs are 2D patches. *: used contrast-enhanced CT.

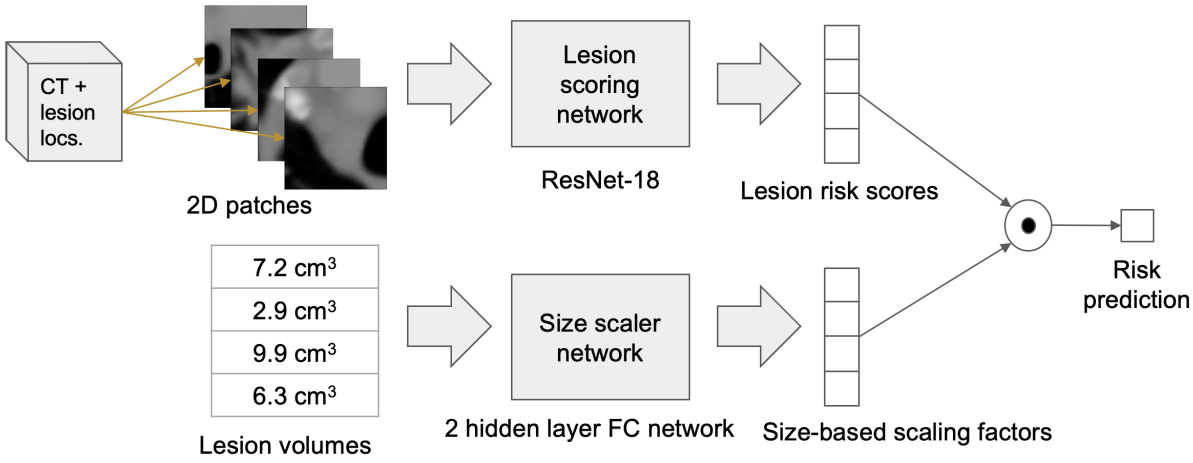

Fig. 2. A schematic of the proposed approach. The patient risk scores are trained using the NPLL loss.

5.3. Comparison to alternative models

Though the primary benefit of lesion-level risks is interpretability, we compared the survival prediction performance of our approach with that of other viable deep learning approaches to ensure a reasonable level of performance.

In fact, our approach outperformed other deep learning approaches. The alternatives were, predicting survival from a size-weighted average embedding of the 5 largest lesions, similarly to [5], and using whole CT volumes as input. When using whole CT volumes, we resampled all volumes to a spacing of $(0.78125, 0.78125, 5)$, used a $(480, 480, 120)$-shaped crop, and performed analogous data augmentations to the ones done on 2D lesion patches.

Our method outperformed both approaches (Table 2). Outperforming lesion averaging suggests that in metastatic cancer, accumulating risk over all lesions is more accurate than basing risk on a summary derived from large lesions. Finally, the whole volume approach likely performed the worst due to data scarcity. Our dataset has 258 patients, while a successful whole volume-based approach in lung cancer detection used over 10,000 patients [13].

6. MODEL INTERPRETATION USING LESION RISKS

By assigning risk at the lesion level, one can better understand the medical reasoning of the model and what kinds of lesions predict a better or worse prognosis. The following interpretability analyses all use lesion risks predicted by a trained model from an arbitrarily chosen split on its test set.

6.1. The lesion risks in each organ

The distributions of lesion risks in each organ were obtained using radiologist-annotated organ masks and plotted in Fig-

	Train	Valid	Test
ImageNet ResNet50*	0.782±0.06	0.662±0.03	0.631±0.04
RadImgNet ResNet50*	0.660±0.03	0.640±0.01	0.645±0.03
Proposed	0.658±0.02	0.656±0.04	0.662±0.02
No FC network	0.632±0.02	0.643±0.01	0.658±0.02
No size scaler	0.621±0.03	0.577±0.06	0.574±0.04
Mean of largest 5	0.643±0.03	0.565±0.03	0.567±0.03
Whole volume	0.713±0.01	0.572±0.03	0.535±0.02

Table 2. Model C-indices from 4-fold cross validation. For the starred models, the best validation checkpoint was chosen for evaluation, while for the other models, the last model checkpoint (after 30 epochs) was chosen.

ure 3. Interestingly, the risk distribution in the lung is lower than in other organs, recapitulating the increased risk associated with metastasis. Furthermore, the model highlights bone lesions as particularly high risk, agreeing with the observed association of bone metastases with morbidity [14].

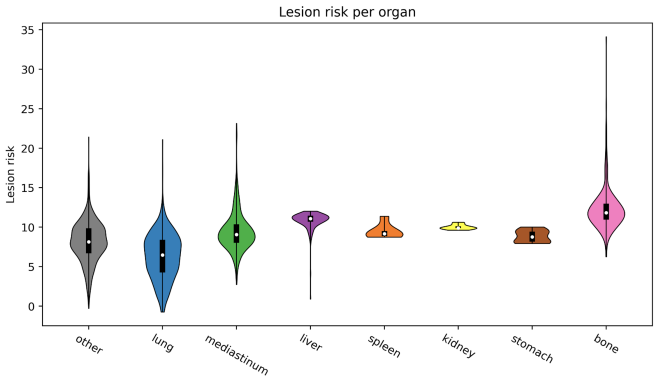

Fig. 3. Violin plots showing the distribution of predicted lesion risks in each organ.

6.2. Lesion risk is correlated with lesion growth

Intuitively, risky lesions are more likely to grow. To see whether this notion is borne out in the risk predictions of our model, we tested the hypothesis that the mean risk of growing lesions exceeds the mean risk of shrinking lesions. We labeled a lesion as growing or shrinking if it came from an annotation with a second timepoint available (89.6%) based on whether its volume increased in the second timepoint.

Because lesion risks from the same patient are likely correlated, instead of using a t-test, we used a mixed linear model including patient identity coefficients to test our hypothesis.

The p-value on the test set was 0.006, indicating significantly higher risk among growing lesions. The risk distributions are shown in Figure 4. Thus, we extracted from the model that growing lesions negatively predict survival, and are partially identifiable from CT image features.

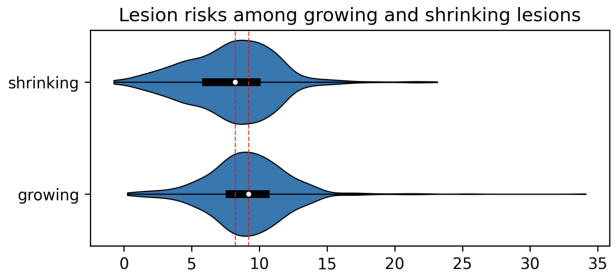

Fig. 4. Risk distributions for growing and shrinking lesions. The white dots represent medians.

6.3. Visualization of high and low risk lesions

We visualized the top 2 highest predicted risk lesions and the top 2 lowest predicted risk lesions on the test set in Figure 5. We observe that the highest risk lesions are bone lesions, agreeing with our earlier observation that predicted risks in bone are the highest. The lowest risk lesions are well contained within the lung as opposed to near the edges, suggesting that the model picked up on high-margin containment in the original organ as a predictor of low risk cancer.

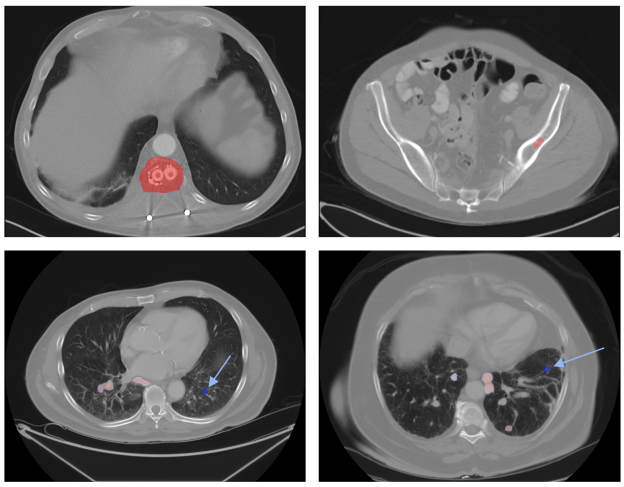

Fig. 5. The top row shows the 2 highest risk lesions on the test set. The bottom row shows the 2 lowest risk lesions on the test set. Lesions, including other ones in the same image, are colored according to their risks, with blue denoting low risk and red denoting high risk.

7. COMPLIANCE WITH ETHICAL STANDARDS

We complied with ethical standards as experiments only used de-identified patient data from a past clinical trial.

8. ACKNOWLEDGEMENTS

Forest Yang acknowledges Genentech for funding, computing resources, and provision of the dataset.

9. REFERENCES

[1] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, Aug 2018.

[2] Ahmed Hosny, Chintan Parmar, Thibaud P. Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J. Gillies, Raymond H. Mak, and Hugo J. W. L. Aerts, "Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study," *PLoS medicine*, vol. 15, no. 11, pp. e1002711, Nov. 2018.

[3] Yiwen Xu, Ahmed Hosny, Roman Zeleznik, Chintan Parmar, Thibaud Coroller, Idalid Franco, Raymond H. Mak, and Hugo J. W. L. Aerts, "Deep learning predicts lung cancer treatment response from serial medical imaging," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, vol. 25, no. 11, pp. 3266–3275, June 2019.

[4] Hyungjin Kim, Jin Mo Goo, Kyung Hee Lee, Young Tae Kim, and Chang Min Park, "Preoperative ct-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas," *Radiology*, vol. 296, no. 1, pp. 216–224, July 2020.

[5] Lin Lu, Laurent Dercle, Binsheng Zhao, and Lawrence H. Schwartz, "Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging," *Nature Communications*, vol. 12, no. 1, pp. 6654, Nov. 2021.

[6] Jiawen Yao, Yu Shi, Le Lu, Jing Xiao, and Ling Zhang, "Deepprognosis: preoperative prediction of pancreatic cancer survival and surgical margin via contrast-enhanced ct imaging," Aug. 2020, arXiv:2008.11853 [cs, eess].

[7] Zhenyu Tang, Yuyun Xu, Lei Jin, Abudumijiti Aibaidula, Junfeng Lu, Zhicheng Jiao, Jinsong Wu, Han Zhang, and Dinggang Shen, "Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 2100–2109, Jun 2020.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb 2020, arXiv:1610.02391 [cs].

[9] Michael F. Gensheimer and Balasubramanian Narasimhan, "A scalable discrete-time survival model for neural networks," *PeerJ*, vol. 7, pp. e6257, Jan 2019, arXiv:1805.00917 [cs, stat].

[10] Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 24, Dec. 2018, arXiv:1606.00931 [cs, stat].

[11] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel, "Time-to-event prediction with neural networks and cox regression," July 2019.

[12] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang, "Radimagenet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, pp. e210315, Sept. 2022.

[13] Peter G. Mikhael, Jeremy Wohlwend, Adam Yala, Ludvig Karstens, Justin Xiang, Angelo K. Takigami, Patrick P. Bourgouin, PuiYee Chan, Sofiane Mrah, Wael Amayri, Yu-Hsiang Juan, Cheng-Ta Yang, Yung-Liang Wan, Gigin Lin, Lecia V. Sequist, Florian J. Fintelmann, and Regina Barzilay, "Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography," *Journal of Clinical Oncology*, vol. 41, no. 12, pp. 2191–2200, Apr. 2023.

[14] Chiara D'Antonio, Antonio Passaro, Bruno Gori, Ester Del Signore, Maria Rita Migliorino, Serena Ricciardi, Alberto Fulvi, and Filippo de Marinis, "Bone and brain metastasis in lung cancer: recent advances in therapeutic strategies," *Therapeutic Advances in Medical Oncology*, vol. 6, no. 3, pp. 101–114, May 2014.