Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260119

Predicting 2-Year Overall Survival in NSCLC from CT Scans Using 2D CNNs and Soft Attention

Domenico PAOLO[a,1] and Carlo GRECO[e,f] and Edy IPPOLITO[e,f] and Michele FIORE[e,f] and Sara RAMELLA[e,f] and Paolo SODA[a,c] and Matteo TORTORA[*,b] and Alessandro BRIA[*,d] and Rosa SICILIA[*,g]

[a] *Unit of Artificial Intelligence and Computer Systems, Department of Engineering, University Campus Bio-Medico of Rome, Italy.*
[b] *Department of Naval, Electrical, Electronics and Telecommunications Engineering, University of Genoa, Italy.*
[c] *Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umea University, Sweden.*
[d] *Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Italy.*
[e] *Department of Medicine and Surgery, Research Unit of Radiation Oncology, Università Campus Bio-Medico di Roma, Italy.*
[f] *Operative Research Unit of Radiation Oncology, Università Campus Bio-Medico di Roma, Italy.*
[g] *UniCamillus-Saint Camillus International University of Health Sciences, Rome, Italy.*

ORCiD ID: Domenico Paolo https://orcid.org/0009-0001-8997-2839

Abstract. Accurate overall survival (OS) prediction in non-small cell lung cancer (NSCLC) is crucial but challenging due to high-dimensional 3D computed tomography (CT) data, limited annotations, and time-to-event outcomes. Traditional 3D CNNs are computationally expensive and prone to overfitting on small datasets. We propose a lightweight framework that aggregates 2D CT slice embeddings via soft attention to form a 3D patient representation. In our approach, features are extracted with EfficientNetB0, and DeepHit models time-to-event survival. Validated on LUNG1 (415 patients), our method outperforms 3D ResNet (+0.077) and alternative aggregation strategies (+0.005) in time-dependent concordance index (Ctd-index). Furthermore, transfer learning from LUNG1 improves performance on the small private CLARO dataset (0.579 vs 0.503). This shows that 2D CNNs with soft attention provide a computationally efficient yet effective alternative to 3D CNN architectures for NSCLC OS prediction, with a substantially lower computational cost (54.3 GFLOPs vs. 2924.6 GFLOPs for ResNet3D-18).

Keywords. Non-small cell lung cancer, survival analysis, prognosis, soft attention

[1] Corresponding Author: Domenico Paolo, email: domenico.paolo@unicampus.it.
[*] These authors share last authorship.

1. Introduction

Non-small cell lung cancer (NSCLC) is the leading cause of cancer deaths worldwide [1]. Accurate survival prediction from computed tomography (CT) scans is critical yet remains challenging due to tumor heterogeneity and the high dimensionality of imaging data. Traditional radiomic approaches rely on manually engineered features, limiting reproducibility and scalability across datasets [2]. To overcome these constraints, deep learning methods have emerged, enabling the automatic extraction of high-level image representations. Among these, 3D CNNs capture volumetric information but are computationally demanding and prone to overfitting, whereas 2D CNNs are more efficient but process slices independently, missing inter-slice context. Recent work addresses this by aggregating slice-level features with attention mechanisms [3]. However, these approaches typically produce fixed-time binary predictions that ignore the temporal nature of survival outcomes [4–8]. In contrast, we propose a framework that aggregates 2D CNN slice features via soft attention and models time-to-event outcomes with DeepHit [9]. This slice-based aggregation follows the Multiple Instance Learning paradigm [10], treating each patient volume as a bag of slice-level instances, where soft attention implicitly highlights the most prognostically informative slices. Our main contributions include: (i) an automated CT-based survival prediction pipeline, (ii) a lightweight attention-based 2D fusion strategy outperforming 3D CNNs, and (iii) improved generalization via transfer learning.

2. Methods

Our method, depicted in **Figure 1**, processes CT scans in three stages: (i) slice-level feature extraction using 2D CNN (EfficientNetB0 [11]), (ii) volume representation via soft attention-based aggregation, and (iii) risk prediction using DeepHit survival network.

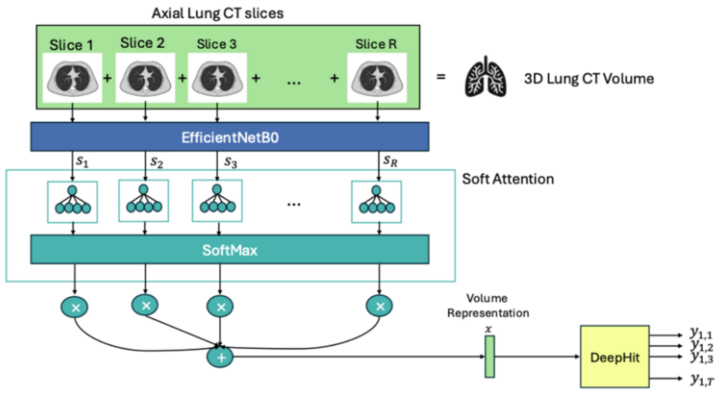

Figure 1. Overview of the proposed method.

We use EfficientNetB0, pretrained on ImageNet, finetuning only the final block to extract 256-dimensional embeddings s_j for each CT slice. For a volume with R slices, embeddings $\{s_1, s_2, \ldots, s_R\}$ are aggregated via soft (gated) attention mechanism that

learns slice-specific weights w_j to emphasize the most prognostically relevant slices, forming a patient-level representation $x = \sum_{j=1}^{R} w_j s_j$. This focuses on prognostically relevant slices, reducing noise and computational cost compared to 3D CNNs. The aggregated vector x is fed into DeepHit [12], a neural network modeling time-to-event outcomes with censoring, i.e., patients whose follow-up ended before the time horizon (2-year/24 months) and who are known to survive up to their last visit. DeepHit was selected for its ability to model short-term OS without proportional hazards assumptions. It predicts monthly survival probabilities up to 24 months using a composite loss $L = \alpha L1 + \beta L2$, which combines a log-likelihood term L1 and a ranking term L2 to handle observed deaths and censored case (α and β denote their relative weights).

3. Dataset and Experiments

We evaluate our approach on two datasets: LUNG1 (415 patients, stage I–III NSCLC), a publicly available dataset [12], and CLARO (119 patients, stage III NSCLC), a private cohort collected under ethical approvals (Ethical Committee approvals: 30 Oct 2012 and 16 Apr 2019; ClinicalTrials.gov Identifier: NCT03583723) [13]. We use pretreatment CT scans and 2-year survival endpoints. Due to its small size, CLARO is not indeed as a benchmark dataset but is used primarily to assess transfer learning effectiveness. Preprocessing includes resampling to 1×1×3 mm, lung segmentation via U-Net, keeping slices where lung area > 2% of slice area, cropping to lung bounding boxes, and resizing slices to 224×224. We compare our method against three types of baselines: (i) 3D CNNs; (ii) slice aggregation strategies; (iii) backbone architectures. First, 3D CNNs are represented by ResNet3D-18 pretrained on Kinetics [14], which directly processes 3D volumes to produce a volume-level representation. Second, we evaluate two different slice aggregation strategies, including average pooling of 2D slice embeddings and self-attention with a class token, which generate volume-level representations from slice-level features. Third, we test four backbones with soft attention, namely ResNet50, VGG16, AlexNet, and ViT [15], all producing volume-level representations for the survival network. All baselines feed the resulting representations into the same DeepHit survival network to ensure a fair comparison. Training is performed in PyTorch using AdamW (learning rate = 1e−4), batch size 4, 100 epochs without early stopping, as preliminary experiments showed convergence after 70–80 epochs, and 10-fold cross-validation on the LUNG1 dataset. The loss weights and the DeepHit hyperparameters were kept fixed throughout training ($\alpha = \beta = 0.5$, ReLU activations, a dropout rate of 0.2, 10 layers, and 100 hidden neurons per layer). Performance is evaluated using the time-dependent concordance index (Ctd-index) [16] by accounting for changes in risk ranking over time in survival analysis. A Ctd-index of 0.5 corresponds to random risk ranking, while higher values indicate better discrimination over time.

4. Results and Discussion

This section compares our approach with 3D CNN baselines, analyzes the impact of slice aggregation strategies, and evaluates transfer learning on the CLARO dataset.

Comparative Analysis. **Table 1** reports the comparison results. Our 2D backbone with soft attention outperforms the 3D ResNet model while requiring significantly fewer

computations: 54.3 GFLOPs for our approach compared to 2924.6 GFLOPs for ResNet3D-18. In both case only the final layer is trained, demonstrating that attention-based fusion of 2D slice features produces more discriminative volume-level representations at a much lower computational cost. When analyzing slice aggregation strategies using EfficientNet-B0 as backbone, soft attention yields the best performance. Its advantage over Average Pooling derives from its ability to assign higher importance to the most informative slices, allowing the survival network to focus on the clinically relevant regions of the volume. Compared to self-attention, soft attention performs better likely due to its stronger generalization ability and reduced number of parameters, which is particularly beneficial in limited-data scenarios. Among the evaluated backbone architectures, EfficientNet-B0 consistently achieves the highest performance, likely thanks to its compound scaling strategy and parameter efficiency. Overall, the proposed fusion-based approach outperforms 3D CNN across different backbones, demonstrating that constructing volume representations by aggregating 2D slice features is a more effective strategy for OS prediction, particularly when data availability is limited.

Table 1. Comparison of our approach with 3D CNNs, different aggregation mechanisms (Average Pooling and Self-Attention), and 4 backbones. Results are reported using the Ctd-index over 10-fold cross-validation.

Category	Model	Ct-index (mean $\pm$ std)
3D CNNs	ResNet3d-18	0.507 ± 0.074
Slice Aggregation Strategies	Average Pooling	0.570 ± 0.033
	Self-Attention (class token)	0.579 ± 0.079
Backbone Architectures	ResNet50 + Soft Attention	0.564 ± 0.067
	Vgg16 + Soft Attention	0.524 ± 0.063
	AlexNet50 + Soft Attention	0.550 ± 0.051
	ViT + Soft Attention	0.546 ± 0.043
Our proposal	**EfficientNetB0 + Soft Attention**	$\mathbf{0.584 \pm 0.054}$

Impact of transfer learning. We evaluate transfer learning from LUNG1 to CLARO. When transferring to CLARO, models pretrained on LUNG1 were further fine-tuned on CLARO, while 'training from scratch' refers to initializing weights from ImageNet only, without any domain-specific pretraining. As reported in Table 4, transfer learning results in a +0.076 improvement, demonstrating that pretraining on the domain-specific LUNG1 dataset provides superior performance on CLARO compared to training from scratch. This improvement can be attributed to the limited size of the CLARO dataset, which makes it challenging to train a model effectively. By leveraging knowledge learned from LUNG1, the pre-trained model can exploit previously acquired features and patterns, leading to enhanced performance on the smaller CLARO dataset. These results highlight the effectiveness of domain-specific transfer learning in scenarios with constrained data availability.

Table 2. Results of Ctd-index on CLARO. Standard deviation is not reported as the Ctd-index was computed on aggregated out-of-fold predictions across all folds.

Model	Fine-tuned on Lung1?	Ct-index
EfficientNetB0+Soft Attention	No	0.503
EfficientNetB0+Soft Attention	Yes	0.579

5. Conclusions

We propose an efficient framework for 2-year NSCLC survival prediction that aggregates 2D CT slice features through soft attention and models time-to-event outcomes with DeepHit. The method outperforms 3D ResNet baselines on LUNG1 with substantially lower computational cost, showing that attention-based aggregation of informative 2D slices can capture clinically relevant volumetric information more effectively than full 3D convolutions. Transfer learning from LUNG1 to the smaller CLARO dataset further improves performance, highlighting the benefit of leveraging related clinical data in limited-data settings. The soft attention mechanism also provides inherent interpretability by emphasizing prognostically relevant slices, and the lightweight EfficientNet-B0 backbone demonstrates the effectiveness of efficient architectures with attention-based fusion. The framework is generalizable to other clinical prediction tasks and can be extended to multimodal systems integrating imaging and EHR data, as explored in [17].

References

[1] Shong, Lynn Y-W., and David C-L. Lam. "Emerging Trends in Global Lung Cancer Burden." *Seminars in respiratory and critical care medicine*. Thieme Medical Publishers, Inc., 2025.

[2] Chen, Mitchell, et al. "Radiomics and artificial intelligence for precision medicine in lung cancer treatment." *Seminars in cancer biology*. Vol. 93. Academic Press, 2023.

[3] Morvan, Ludivine, et al. "Learned deep radiomics for survival analysis with attention." *International Workshop on PRedictive Intelligence In MEdicine*. Cham: Springer International Publishing, 2020.

[4] Braghetto, Anna, et al. "Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset." *Scientific Reports* 12.1 (2022): 14132.

[5] Zheng, Sunyi, et al. "Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning." *Radiotherapy and oncology* 180 (2023): 109483.

[6] Pai, Suraj, et al. "Foundation model for cancer imaging biomarkers." *Nature machine intelligence* 6.3 (2024): 354-367.

[7] Torres, Felipe Soares, et al. "End-to-end non–small-cell lung cancer prognostication using deep learning applied to pretreatment computed tomography." *JCO Clinical Cancer Informatics* 5 (2021): 1141-1150.

[8] Haarburger, Christoph, et al. "Image-based survival prediction for lung cancer patients using CNNS." *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019.

[9] Lee, Changhee, et al. "Deephit: A deep learning approach to survival analysis with competing risks." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

[10] Ilse, Maximilian, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning." *International conference on machine learning*. PMLR, 2018.

[11] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.

[12] Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5.1 (2014): 4006.

[13] Caruso, Camillo Maria, et al. "A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer." *Journal of Imaging* 8.11 (2022): 298.

[14] Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.

[15] Suganyadevi, S., V. Seethalakshmi, and Krishnasamy Balasamy. "A review on deep learning in medical image analysis." *International Journal of Multimedia Information Retrieval* 11.1 (2022): 19-38.

[16] Antolini, Laura, Patrizia Boracchi, and Elia Biganzoli. "A time-dependent discrimination index for survival data." *Statistics in medicine* 24.24 (2005): 3927-3944.

[17] Tortora, Matteo, et al. "RadioPathomics: multimodal learning in non-small cell lung cancer for adaptive radiotherapy." *IEEe Access* 11 (2023): 47563-47578.