Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253310

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright© Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Automated detection and characterization of small cell lung cancer liver metastasis on computed tomography

Sophia Ty[1]
Fahmida Haque[1]
Parth Desai[2]
Nobuyuki Takahashi[3,4]
Usamah Chaudhary[1,5]
Peter L. Choyke[1]
Anish Thomas[4]
Barış Türkbey[1]
Stephanie A. Harmon[1]

[1]National Cancer Institute, Artificial Intelligence Resource, Maryland, USA

[2] Fox Chase Cancer Center at Temple University Hospital, Philadelphia, USA

[3]National Cancer Center Hospital East, Department of Medical Oncology, Kashiwa, Japan

[4]National Cancer Institute, Developmental Therapeutics Branch, Maryland, USA

[5]University of Texas Southwestern Medical Center, Texas, USA

Corresponding author: Barış Türkbey

E-mail: turkbeyi@mail.nih.gov

Received 28 February 2025; revision requested 24 March 2025; last revision received 14 July 2025; accepted 11 September 2025.

PURPOSE

Small cell lung cancer (SCLC) is an aggressive disease with diverse phenotypes that reflect the heterogeneous expression of tumor-related genes. Recent studies have shown that neuroendocrine (NE) transcription factors may be used to classify SCLC tumors with distinct therapeutic responses. The liver is a common site of metastatic disease in SCLC and can drive a poor prognosis. Here, we present a computational approach to detect and characterize metastatic SCLC (mSCLC) liver lesions and their associated NE-related phenotype as a method to improve patient management.

METHODS

This study utilized computed tomography scans of patients with hepatic lesions from two data sources for segmentation and classification of liver disease: (1) a public dataset from patients of various cancer types (segmentation; n = 131) and (2) an institutional cohort of patients with SCLC (segmentation and classification; n = 86). We developed deep learning segmentation algorithms and compared their performance for automatically detecting liver lesions, evaluating the results with and without the inclusion of the SCLC cohort. Following segmentation in the SCLC cohort, radiomic features were extracted from the detected lesions, and least absolute shrinkage and selection operator regression was utilized to select features from a training cohort (80/20 split). Subsequently, we trained radiomics-based machine learning classifiers to stratify patients based on their NE tumor profile, defined as expression levels of a preselected gene set derived from bulk RNA sequencing or circulating free DNA chromatin immunoprecipitation sequencing.

RESULTS

Our liver lesion detection tool achieved lesion-based sensitivities of 66%–83% for the two datasets. In patients with mSCLC, the radiomics-based NE phenotype classifier distinguished patients as positive or negative for harboring NE-like liver metastasis phenotype with an area under the receiver operating characteristic curve of 0.73 and an F1 score of 0.88 in the testing cohort.

CONCLUSION

We demonstrate the potential of utilizing artificial intelligence (AI)-based platforms as clinical decision support systems, which could help clinicians determine treatment options for patients with SCLC based on their associated molecular tumor profile.

CLINICAL SIGNIFICANCE

Targeted therapy requires accurate molecular characterization of disease, which imaging and AI may aid in determining.

KEYWORDS

Computer vision, segmentation, neuroendocrine gene expression, radiomics, tumor classification, transcriptomics, molecular tumor profile

Epub: 06.10.2025

Publication date: xx.xx.2025

DOI: 10.4274/dir.2025.253310

You may cite this article as: Ty S, Haque F, Desai P, et al. Automated detection and characterization of small cell lung cancer liver metastasis on computed tomography. *Diagn Interv Radiol.* 06 October 2025 DOI: 10.4274/dir.2025.253310 [Epub Ahead of Print].

Small cell lung cancer (SCLC) is an aggressive form of lung cancer strongly associated with smoking and accounts for 13%–15% of all lung cancer cases.[1,2] Patients often present with advanced disease, resulting in a poor prognosis with a 5-year survival rate of 7%.[3] These outcomes reflect the challenges in clinical management of a recalcitrant cancer marked by the ubiquitous presence of TP53, RB1 loss-of-function events, and high chromosomal instability,[4,5] which drive rapid progression, widespread metastasis,[5,6] and treatment resistance following initial response to therapy.[7] Among many complications associated with cancer progression, SCLC typically leads to hepatic metastasis, which is seen in 21%–27% of patients at presentation and 69% at autopsy.[8] This makes the liver the most prevalent metastatic site after mediastinal lymph nodes—an important characteristic since liver metastasis is also an independent marker of poor prognosis.[9,10]

SCLC also demonstrates a high degree of heterogeneity, manifesting under various transcriptional subtypes. The classification of SCLC subtypes is defined by the expression levels of four transcription regulators, namely neuronal differentiation 1 (NEUROD1), achaete-scute family basic helix-loop-helix transcription factor 1 (ASCL1), POU class 2 homeobox3, and yes-associated protein (YAP1).[11] The relative expression of these regulators leads to heterogeneous neuroendocrine (NE) gene expression, which has therapeutic implications.[12,13] The SCLC tumors associated with relatively high NEUROD1 and ASCL1 expression are considered NE positive and demonstrate greater susceptibility to DNA-damaging agents;[14,15] non-NE SCLC tumors have greater POU2F2 and YAP1 expression and have been shown to possess better response to immunotherapy.[13,16-18]

Despite emerging insights into its molecular subtypes, SCLC is still currently treated as a homogenous disease. As we gain more insights into the molecular underpinnings of SCLC that drive tumor response to treatments, there is a need for clinical workflows that can stratify patients based on their tumor profile. Methods that identify subpopulations of patients with SCLC who are likely to benefit from specific targeted treatments without requiring additional invasive testing can offer physicians actionable insights, especially when treatment response status can be determined at the time of diagnosis. Furthermore, computational platforms that can accurately detect and characterize tumors offer practical utility in supporting physicians from diagnosis to treatment. They can automate critical tasks, integrate different types of medical data (e.g., radiology scans, biopsy findings, blood panel information), extract clinically relevant tumor characteristics, and consolidate medical information for health practitioners.

Within the past decade, artificial intelligence (AI) has been integrated into automating medical image processing tasks. More specifically, deep learning has shown promise in segmenting objects at different imaging scales, including tissue[19] and cellular[20] levels, for a variety of medical conditions, including lung cancer[21,22] and hepatic diseases.[23-25] Preceding the popularity of deep learning for medical image segmentation is the use of radiomics, a common research approach to describe tumors quantitatively, including their intensity, shape, and texture, which may be used as image-based biomarkers for downstream analysis and association studies. Radiomics has been adopted to address various clinical decision tasks, such as lesion classification[26,27] and treatment response prediction,[28,29] including applications in liver-associated malignancies.[30-32]

Given the implications of transcriptional subtypes to treatment response in SCLC and the association between SCLC hepatic metastasis and prognosis, we investigated whether NE status, as determined by tumor gene expression analysis, can be determined from computed tomography (CT) scans of confirmed SCLC metastasis in the liver. In this study, we present a two-step machine learning framework for automated detection and characterization of SCLC liver metastasis. We employ deep learning for three-dimensional (3D) segmentation of hepatic lesions followed by radiomics-based analysis to characterize and classify image scans by NE status, defined as high expression of a pre-selected gene set.

Methods

A graphical summary of the study objective is shown in Figure 1.

Main points

- The liver is a frequent site of metastases in small cell lung cancer, and artificial intelligence helps identify and segment tumors.
- The computed tomography (CT) imaging characteristics of liver lesions have a moderate correlation with neuroendocrine transcription factors.
- An end-to-end machine learning pipeline may help characterize the molecular profile of liver lesions in CT.

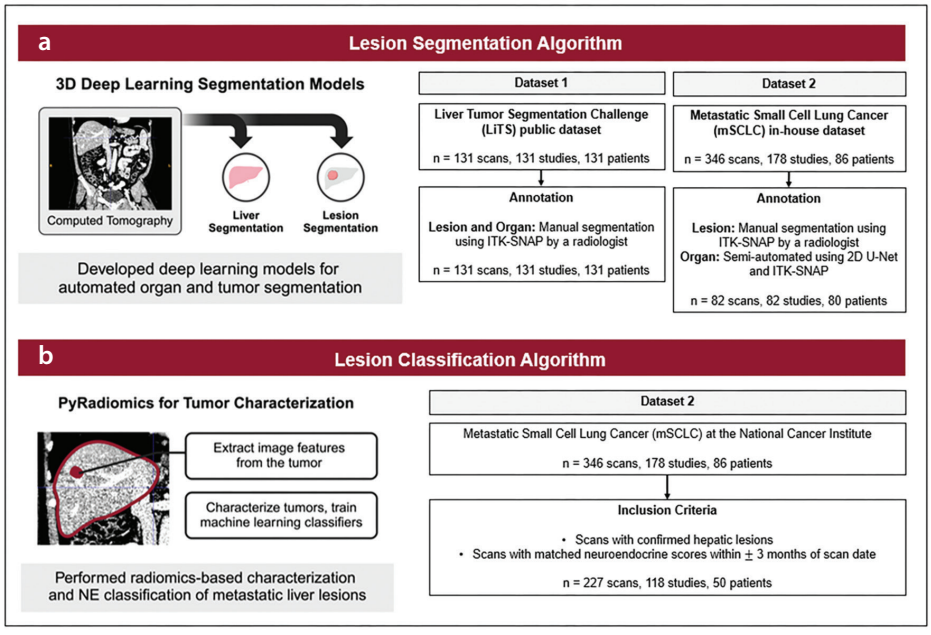


Figure 1. Graphical summary of the machine learning framework for automated detection and characterization of small cell lung cancer (SCLC) liver metastases on computed tomography. **(a)** Three-dimensional deep learning models were developed to simultaneously segment the liver and tumors within the region. **(b)** Radiomics-based characterization of metastatic SCLC liver tumors was performed, and subsequently supervised machine learning models were trained to classify patients' neuroendocrine (NE) status as NE positive or NE negative.

Study population and data description

Two datasets were utilized in this study: (1) the liver tumor segmentation (LiTS) dataset, a publicly available dataset containing 131 CT scans; and (2) a retrospective cohort of patients with metastatic SCLC (mSCLC) underwent CT at the National Cancer Institute (Bethesda, MD, USA).

The LiTS dataset consists of multi-center scans of primary and secondary hepatic tumors. All scans were manually annotated by a radiologist (>3 years' experience) using the ITK-SNAP open source software platform to obtain liver and lesion labels. Annotations were confirmed by three additional radiologists; the most senior reader's findings were used in any labeling conflicts. This research study was conducted retrospectively using human participant data made available as open-access materials by Bilic et al.[25] This open-source cohort was used for training and development of a segmentation algorithm for the detection and segmentation of focal liver lesions on CT. This cohort was used exclusively for the segmentation task (Figure 1).

An initial query for the mSCLC dataset identified 88 patients diagnosed with SCLC and undergoing disease monitoring or treatment at the institution under one or more clinical protocols, including the following ClinicalTrials.gov identifiers: NCT02769962 (IRB 16-C-0107; 2016-05-09), NCT03554473 (IRB 18-C-0110; 2018-09-11), NCT02487095 (IRB 15-C-0150; 2015-07-30), NCT02484404 (IRB 15-C-0145; 2015-06-29), NCT02146170 (IRB 14-C-0105, 2014-05-28). Each protocol was approved by the local institutional review board, and written informed consent was obtained from all patients. From these patients, a total of 346 abdominal CT scans obtained during 178 CT sessions were identified for possible inclusion. This cohort was used for both the segmentation and classification tasks (Figure 1). Multiple series were included from each study date: for example, thick-slice and soft tissue thin-slice reconstructions to evaluate model robustness (Supplementary Table 1). Radiology reports were manually reviewed for each CT scan to confirm the presence or absence of hepatic lesions. From all available scans, a subset of 82 scans was manually reviewed by an expert radiologist (>15 years' experience), and liver lesions were segmented using ITK-SNAP. Liver organ annotations were obtained using a previously developed two-dimensional (2D) U-Net liver segmentation model[33] and were manually adjusted using ITK-SNAP. All annotated scans were used for the tumor segmentation task (Figure 1).

All patients in the mSCLC cohort underwent either tissue or blood sampling for bulk RNA or circulating free DNA (cfDNA) chromatin immunoprecipitation sequencing at multiple timepoints, corresponding with CT study dates (± 3 months). Expression profiles from sequencing data were used to classify patients broadly into NE-positive and NE-negative phenotype groups based on previously published methods.[13,34] Briefly, single-sample gene set enrichment analysis from a 50-gene signature panel was used to classify samples as NE (score >0) or non-NE (score <0), with a lower score in the non-NE group reflecting more confidence that the sample does not exhibit NE differentiation.[35] Strong correlation observed between cfDNA-derived and RNA-derived expression scores for NE phenotyping has been previously reported;[13,34-36] therefore, either reference standard was used for ground truth assignment in this cohort. The NE phenotype expression scores (range: −1,1) and classification (NE, non-NE) were recorded for use in this study (Table 1).

Deep learning model development for tumor segmentation

Three deep learning algorithms were selected to build a hepatic lesion detection model: (1) a 3D U-Net, (2) a 3D SegResNet, and (3) a 3D nnU-Net. During initial model development and selection, each algorithm was trained solely using the LiTS dataset, and mSCLC data were used as an independent test set. For all training, data partitions were stratified at the patient level and are summarized in Table 2.

The U-Net and SegResNet models were built using the Medical Open Network for AI platform (version 1.3.0).[37] For these two network architectures, training was conducted with the following data pre-processing and augmentations: CTs were resampled to uniform spacing (0.5 mm × 0.5 mm × 1 mm), foreground cropping, variable CT windowing, and random cropping by labels with a sampling ratio of 4:1:3 for the background, liver, and lesion, with 12 samples taken per image. Each sample crop was of size 512 × 512 × 16. Both models were trained using an adaptive moment estimator to minimize lesion-level DICE loss, with a learning rate of 0.0001 for 1,500 epochs. The final model was selected based on the highest validation DICE reported during training. Inference was conducted using the sliding window technique.

The nnU-Net model, an auto-configuring semantic segmentation model, was implemented using the built-in *3dfullres* five-fold cross-validation.[38] Pre-processing configurations selected by the model included spacing (0.789 mm × 0.789 mm × 2 mm), patch size 80 × 80 × 60, and per-image z-score standardization.

For all three models, performance was evaluated in the test set using lesion-level DICE coefficients and tumor detection sensitivity on both the LiTS and mSCLC datasets.

Due to differences in the burden and imaging characteristics of the mSCLC cohort compared with the LiTS cohort, which may potentially impact generalizability, a final nnUNet model was trained from all LiTS training data along with a subset of mSCLC scans partitioned with an approximate training/test

Table 1. Summary of key characteristics of the mSCLC dataset describing the distribution of patient scans, lesions, and neuroendocrine status in the cohort

mSCLC data characteristics	Quantity
Age (years)	63 (23–82)
Sex: male, female	45, 41
Number of scans per patient	Median: 3, range: (1–14)
Number of lesions per scan	Median: 10, range: (1–148)
Hepatic lesions	
Present	252
Absent	94
Final study cohort for classification†	
Neuroendocrine classification by sequencing-based assessment†	
Positive	172
Negative	55

†, reported for the final 227 series volumes for the classification task; mSCLC, metastatic small cell lung cancer.

split of 80%/20% scans. Inference of the final finetuned segmentation model was completed for all mSCLC scans for use in the classification model.

Radiomics characterization and neuroendocrine phenotype classification

Each mSCLC scan was labeled based on matched gene expression-based NE score as NE positive (1) or NE negative (0). All CT studies with confirmed hepatic lesions that were matched to NE scores within ± 3 months of the imaging date were included in the NE phenotype classification. Scans were partitioned with an approximate training/test split of 80%/20% images using the same stratification applied during segmentation. Splits were determined at the patient level to avoid bias, resulting in 177 scans for training and 50 scans for testing (Table 2).

Liver lesion contours obtained from the final segmentation model were characterized using radiomics. Quantitative image features were extracted using PyRadiomics (v3.0.1) with a resampling pixel spacing of (1 mm, 1 mm, 1 mm) for the (x, y, z) voxel coordinates and default image standardization parameters. A total of 107 radiomic features were obtained, representing first-order statistics, shape (2D and 3D), gray level co-occurrence matrix, gray level run length matrix (glrlm), gray level size zone matrix (glszm), neighboring gray tone difference matrix, and gray level dependence matrix (gldm). The number of lesions per image was determined using connected-components-3D (v3.12.1) and served as an additional feature, resulting in a total of 108 features. From these, a subset of imaging characteristics correlated with NE phenotype was selected using least absolute shrinkage and selection operator (LASSO) regression (scikit-learn v1.2.2).

Radiomics-based NE phenotype classification was conducted using three machine learning models: (1) logistic regression (scikit-learn v1.2.2), (2) random forest (scikit-learn v1.2.2), and (3) XGBoost (v2.0.3). For all models, default parameters were used. Code and raw data for how these models were trained are available at https://github.com/NIH-MIP/mSCLC_Segmentation_Classification. All models incorporated the subset of imaging features selected using LASSO regression for binary classification of tumors as NE (1) or non-NE (0) phenotype.

Each model was trained with and without class-based weights. Five-fold cross-validation was implemented, and the best model was selected using the F1 score and area under the receiver operating characteristic curve (AUC) from cross-validation as the primary performance criteria. When applied to the test set, the ensemble of all five-folds was utilized for test set evaluation (average prediction of five-folds).

Statistical analysis

The DICE coefficient,[39] a measure that describes spatial agreement between two image sets, was calculated to quantify the performance of the model compared with ground truth annotation from radiologists. To evaluate detection performance metrics at the lesion level, connected-components-3D (v3.12.1) was used to identify unique lesions in both the ground truth segmentations and model output. Next, each lesion was classified as a true positive (i.e., ground truth lesion correctly segmented by AI), false negative (i.e., ground truth lesion was not segmented by the model), or false positive (i.e., model segmented a lesion with no ground truth correlate) per scan. Each segmentation model's sensitivity, positive predictive value (PPV), and false positive trends were calculated and reported as summary statistics.

The relationship between tumor burden and NE scores was examined using two tests: (1) Spearman correlation analysis (SciPy v.1.11.1) for continuous NE scores, and (2) Wilcoxon rank sum tests (R v4.4.1) for binarized NE scores. For this analysis, one series per patient per scan date was selected. Tumor volume estimates were calculated using AI-predicted tumor regions from the final segmentation model.

Finally, the performance of each binary classifier was evaluated. Model accuracy, sensitivity, specificity, PPV, negative predictive value, F1 scores, and AUC were calculated (scikit-learn v1.2.2) and compared. Due to potential bias in multiple scans coming from the same CT study, bootstrap sampling was performed at the study level to select one scan per study per iteration randomly. The mean and 95% confidence intervals (CIs) of each performance metric are reported in the test set.

Table 2. Data splits per task

Task	Level	Training/validation	Test¥
Segmentation†	Pts	LiTS 121, mSCLC 62	LiTS 10, mSCLC 18
	Scans	LiTS 121, mSCLC 64	LiTS 10, mSCLC 18
	Series volumes	LiTS 121, mSCLC 64	LiTS 10, mSCLC 18
Classification (mSCLC only)**	Pts	36	14
	Scans	92	26
	Series volumes	177	50

†, For the segmentation task, two training schemes were used: with and without mSCLC. For the training scheme without mSCLC, the entire cohort was reserved for testing. Labels are created at the series volume level. **, For the classification task, labels are assigned at the scan level. Patients can have multiple labels (i.e., positive or negative) depending on the time point at which the scan/sequencing was completed. ¥, The test set for mSCLC is preserved across both tasks on the patient-level (i.e., one unique set of patients was reserved for end-to-end testing of the segmentation and classification models). mSCLC, metastatic small cell lung cancer; LiTS, liver tumor segmentation.

Table 3. Comparison of lesion segmentation model performance presents the median DICE coefficient, sensitivity, PPV, and number of false positives per scan for all deep learning models implemented for hepatic lesion segmentation. Results are shown for the publicly available LiTS dataset and an in-house mSCLC dataset

	U-Net		SegResNet		nnU-Net		Finetuned nnU-Net	
Metrics	LiTS	mSCLC	LiTS	mSCLC	LiTS	mSCLC	LiTS	mSCLC
Median DICE	0.439	0.297	0.417	0.418	0.750	0.607	0.771	0.640
Sensitivity	0.395	0.244	0.662	0.447	0.813	0.539	0.826	0.667
PPV	0.971	0.975	0.962	0.987	1.0	0.998	1.0	1.0
FP/scan	0.1 (0-1)	0.1 (0-1)	0.2 (0-1)	0.1 (0-1)	0	0.02 (0-1)	0	0

PPV, positive predictive value; mSCLC, metastatic small cell lung cancer; LiTS, liver tumor segmentation.

Results

The segmentation model utilized all LiTS data and incorporated a subset of scans from the mSCLC cohort. Of the 88 patients identified for possible study inclusion, a total of 86 patients were included in the final study cohort, with two exclusions due to insufficient data records (no sampling within the required timeframe from a scan date; no segmentations). Key characteristics of the mSCLC cohort are provided in Table 1. Image acquisition characteristics are summarized in Supplementary Table 1 for each cohort and task.

Lesion detection

The cohort information for model training is shown in Table 2. In addition to the LiTS cohort, 82 annotated scans (82 unique studies) from the mSCLC dataset were utilized for the segmentation task, of which 50 scans were positive for containing hepatic lesions. First, models were trained only on the LiTS cohort and applied to mSCLC. Among the three deep learning models implemented for automated hepatic lesion detection and segmentation, the 3D *fullres* nnU-Net model provided the most accurate and robust results for the test set from both datasets. Its detection performance had a median DICE score of 0.75 and 0.607, lesion-level sensitivity of 0.813 and 0.539, and PPV of 1.0 and 0.99 for the LiTS and mSCLC test sets, respectively. The 3D *fullres* nnU-Net model also had a range of 0–1 false positive lesions per scan for both datasets. The U-Net and SegResNet models had median DICE scores of 0.439, 0.417 for the LiTS dataset, and 0.297, 0.418 for the mSCLC dataset. The U-Net model had lesion sensitivities of 0.395 and 0.244, whereas the SegResNet model achieved 0.662 and 0.447 for LiTS and mSCLC. A summary of each model's performance is provided in Table 3.

Next, an evaluation was performed to determine how finetuning of the nnU-Net for liver lesion segmentation on the mSCLC data may improve performance. Here, the 3D *fullres* nnU-Net model achieved a median DICE of 0.771 and 0.640, with sensitivities of 0.826 and 0.667 for the LiTS test set and mSCLC test set, respectively. For both datasets, the model had a PPV of 1.0 and 0 false positives per scan. Representative images of cases with high and low concordance between the AI-predicted and ground truth annotations of mSCLC liver lesions from the test set are provided in Figure 2.

Correlation analysis of tumor burden and neuroendocrine status

A statistically significant correlation was found between tumor burden and NE scores for continuous data (Spearman: 0.252, *P* value: 0.0059) and binarized data (Wilcoxon rank sum *P* value: 0.028). Further analysis revealed that cfDNA samples had greater dependence on tumor volume, with a statistically significant correlation between tumor burden and cfDNA-derived NE scores (Spearman: 0.446, *P* value: 0.00057; Wilcoxon rank sum *P* value: 0.0013). Conversely, biopsy-derived NE scores did not have a statistically significant correlation with tumor volume (Spearman: –0.0738, *P* value: 0.5680; Wilcoxon rank sum *P* value: 0.253). Correlation plots are provided in Figure 3.

Neuroendocrine phenotype classification

A total of 227 scans from 118 CT studies were used for the NE classification task after excluding scans with no hepatic lesions by radiologist read (n = 92 scans), scans with no corresponding RNA sequencing (RNAseq) or cfDNA data (n = 19), or false negatives by segmentation model (n = 6 volumes). Of the usable data, 172 scans (89 studies) were identified as NE positive and 55 scans (29 studies) were identified as NE negative (Table 1). The LASSO feature selection was performed within the training set, identifying 20/108 radiomic features correlated with NE-related tumor phenotypes for inclusion in the classification model. The distribution of selected radiomic feature categories was as follows: 30% shape, 25% glszm, 20% gldm, 15% first order statistics, and 10% glrlm. Among these, shape was the most dominant radiomic feature type found in the subset. The top five imaging feature characteristics determined during feature selection were minor axis length, maximum 2D diameter row, major axis length, gldm large dependence emphasis, and first order variance. A full list of the selected radiomic features and a distribution summary of the selected feature type are provided in Figure 4.

For all evaluated models, weighted training did not substantially boost the predictive performance of the algorithms (Supplementary Table 2), nor did they generalize better to the test set (Supplementary Table 3), regardless of the optimization strategy for determining the best weight. Based on the cross-validation performance, the random forest classifier had the highest F1 (mean 0.86; range 0.81, 0.91) and AUC (mean: 0.68; range: 0.56, 0.81) (Supplementary Table 2). All calculated metrics for the NE phenotype classification task are summarized in Table 4, where the logistic regression model generalized better than the random forest model by AUC (mean: 0.71; 95% CI: 0.59–0.82 vs. mean 0.58; CI: 0.48–0.69, respectively), though F1 and accuracy metrics were similar across all models, likely due to the imbalance in favor of the positive class (NE). The 50 scans in the test cohort represented 26 unique CT studies. Breaking accuracy down further by scan-level classification agreement across series volumes: 18/24 scans with multiple series volumes were correctly classified in

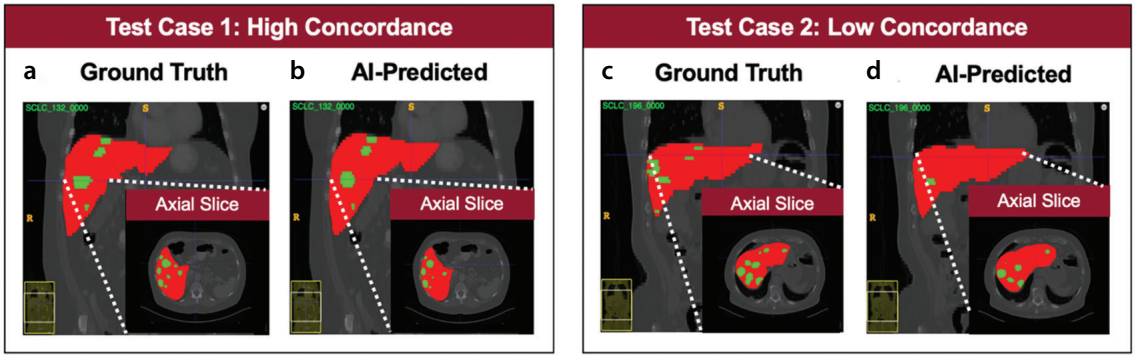


Figure 2. Representative results of the finetuned nnU-Net lesion detection model. Two metastatic small cell lung cancer (mSCLC) test cases are provided with liver (shown in red) and lesion (shown in green) segmentations from ground truth annotations and artificial intelligence (AI) predictions. A single axial slice from the three-dimensional segmentation is shown. **(a, b)** Example of an mSCLC test case with high concordance between the ground truth and AI-predicted lesions; DICE: 0.71. **(c, d)** Example of an mSCLC test case with low concordance between ground truth and AI-predicted lesions; DICE: 0.51.

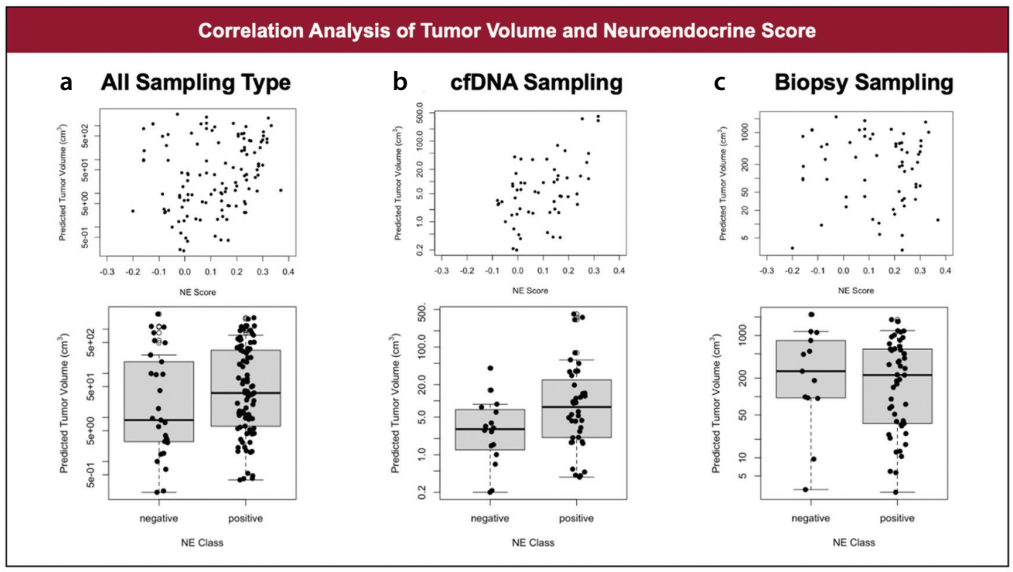
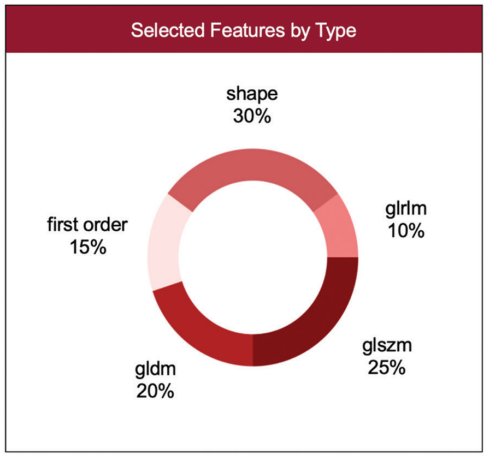


Figure 3. Correlation analysis of tumor burden to neuroendocrine scores. **(a)** There is a statistically significant correlation between tumor burden and neuroendocrine (NE) scores for both continuous data (Spearman: 0.252, *P* value: 0.0059) and binarized NE scores (Wilcoxon rank sum test statistic: 2.191, *P* value: 0.028). This analysis utilizes NE scores from both circulating free DNA (cfDNA) and biopsy samples. **(b)** Correlation plots of tumor burden and NE scores (top) and NE class (bottom) derived from cfDNA samples. There is a statistically significant (Spearman: 0.446, *P* value: 0.00057; Wilcoxon rank sum test statistic: 3.225, *P* value: 0.0013) tumor volume dependence in cfDNA-derived NE scores. **(c)** Correlation plots for biopsy-derived analysis of tumor burden and NE scores (top) or NE class (bottom) showing no statistically relevant relationship between biopsy-derived NE scores and tumor volume estimates.



Group	Feature
Shape	Major Axis Length** Maximum 2D Diameter Row** Mesh Volume* Minor Axis Length** Surface Area Voxel Volume
Gray Level Size Zone Matrix (glszm)	Gray Level Non-Uniformity* Large Area Emphasis Large Area High Gray Level Emphasis Large Area Low Gray Level Emphasis Zone Variance
Gray Level Dependence Matrix (gldm)	Dependence Non-Uniformity* Gray Level Non-Uniformity Large Dependence Emphasis** Large Dependence High Gray Level Emphasis
First Order Statistics	Energy Total Energy Variance**
Gray Level Run Length Matrix (glrlm)	Gray Level Non-Uniformity* Run Length Non-Uniformity*

** included in top five imaging features
*included in top ten imaging features

Figure 4. Distribution of selected feature types and list of selected radiomics features. The distribution of each feature type is as follows: 30% shape, 25% gray level size zone matrix, 20% gray level dependent matrix, 15% first order statistics, 10% gray level run length matrix. Least absolute shrinkage and selection operator regression identified 20 features correlated to binarized neuroendocrine scores among 108 radiomic quantities. All selected radiomic features are listed by group.

Table 4. Scan-level mSCLC NE phenotype classification results in the test set, derived from ensemble of five-fold cross-validation. Reported are mean and 95% CIs from 1000 bootstrap samples, sampling one scan (series) per CT study date per iteration

Model	Accuracy	PPV	NPV	Sensitivity	Specificity	F1	AUC
LR	0.79 (0.73, 0.85)	0.81 (0.78, 0.83)	0.64 (0.33, 1.00)	0.95 (0.90, 1.00)	0.25 (0.17, 0.33)	0.87 (0.84, 0.91)	0.71 (0.59, 0.82)
RF	0.77 (0.69, 0.85)	0.82 (0.77, 0.86)	0.50 (0.25, 0.75)	0.90 (0.85, 0.95)	0.33 (0.17, 0.50)	0.86 (0.81, 0.90)	0.58 (0.48, 0.69)
XGB	0.79 (0.73, 0.85)	0.84 (0.81, 0.86)	0.56 (0.40, 0.75)	0.90 (0.85, 0.95)	0.42 (0.33, 0.50)	0.87 (0.83, 0.90)	0.63 (0.48, 0.78)

PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; mSCLC, metastatic small cell lung cancer; CT, computed tomography; CI, confidence interval

both series volumes, 4/24 scans were incorrectly classified in both series volumes, and 2/24 scans had different classification results across the different series. The two scans with only a single series volume available were both correctly classified.

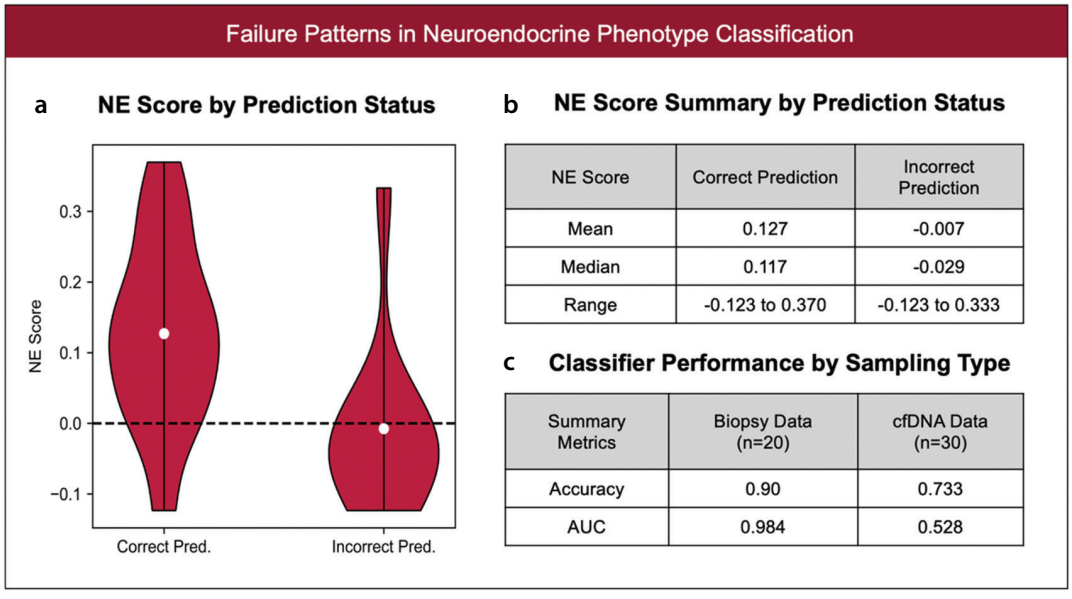
Failure analysis revealed that 8/10 of misclassifications were associated with NE scores near the boundary (NE score: 0) and with cfDNA-derived NE scores. Misclassified observations skewed to the NE-negative phenotype (Figure 5a), with a mean NE score of −0.007 as presented in Figure 5b. The NE classification results showed an AUC of 0.528 when evaluated on cfDNA-derived data alone, although they showed an AUC of 0.984 when strictly evaluated on biopsy-derived transcriptomic data, as shown in Figure 5c.

Discussion

Patients with SCLC often present with hepatic metastases.[9,40] Characterizing NE profiles of mSCLC lesions offers a pathway to stratify patients based on distinct therapeutic vulnerabilities of their molecular subtypes.[13] We demonstrate the potential of deep learning on automated liver lesion detection and the feasibility of using radiomics to describe properties of mSCLC tumors. This framework can enable the determination of patients' NE status as positive or negative for bearing an NE-like phenotype without resorting to invasive biopsy procedures and relying on scans obtained as part of routine staging studies. Our 3D nnU-Net segmentation model demonstrated that hepatic lesions can be accurately detected for patient populations with highly variable disease characteristics (e.g., a wide range of tumor size, varying number of tumors in a single scan) using a fully automated platform. We also showed the possibility of stratifying patients with SCLC by their NE status using radiomic features extracted from routinely acquired abdominal CT scans of metastatic liver lesions.

The finetuned nnU-Net segmentation model was able to identify liver lesions with a median DICE score of 0.771 for the publicly available data (LiTS) and 0.640 for our internal mSCLC dataset. It also showed adeptness in locating regions with suspected lesions in CT images with high sensitivity (LiTS: 0.826 and mSCLC: 0.667) that outperformed previously reported models (ISBI 2017: 0.458, MICCAI 2017: 0.515, MICCAI 2018: 0.554)[25] while having low false positive rates (0–1 false positive per scan) and comparable tumor-level DICE scores (ISBI 2017: 0.674, MICCAI 2017: 0.702, MICCAI 2018: 0.739)[25] for the public dataset. Further characterization of the model's behavior on mSCLC cases without ground truth reports of a lesion (n = 94) revealed that false positive predictions typically were small in volume (0.638 ± 1.73 cm^3). Overall, our detection tool not only offers valuable improvements to previous liver lesion segmentation benchmarks but also enables automating the lesion annotation process for radiomics analysis.

Correlation analysis of AI-predicted lesions showed that tumor burden has a statistically significant association with NE status. Our results revealed that NE scores exhibit dependence on tumor volume, especially when expression profiles were derived from cfDNA samples. This finding is consistent with previous studies showing a correlation between tumor burden and cfDNA in NE-related[34,41] and lung[42] neoplasms, although their exact relationship remains elusive.[43] This is also supported by our radiomics analysis, shows that shape—which includes measurements of tumor size—is the most frequently selected radiomic feature type during model development. These findings indicate that tumor burden may be an indirect measure of NE expression in mSCLC tumors. Recent studies have demonstrated a strong correlation between cfDNA-derived and RNA-derived expression scores for NE phenotyping.[13,34-36] However, the same volume-based relationship was not observed in biopsy-based RNAseq determination of NE phenotype. This discrepancy may be explained by tumor heterogeneity, where cfDNA-derived expression metrics are an aggregate of multiple lesions throughout



NE Score	Correct Prediction	Incorrect Prediction
Mean	0.127	-0.007
Median	0.117	-0.029
Range	-0.123 to 0.370	-0.123 to 0.333

Summary Metrics	Biopsy Data (n=20)	cfDNA Data (n=30)
Accuracy	0.90	0.733
AUC	0.984	0.528

Figure 5. Failure analysis of the neuroendocrine (NE) phenotype classifier. **(a)** The NE score distribution based on prediction status (correct vs. incorrect). The mean NE score of observations associated with incorrect predictions is close to the NE score boundary (NE score: 0) that divides between NE-positive and NE-negative phenotypes. **(b)** Summary of the NE scores by prediction status. Correctly predicted observations had an NE score range of −0.123 to 0.370, a mean NE score of 0.127, and a median NE score of 0.117. Incorrectly predicted observations had a score range of −0.123 to 0.333, a mean NE score of −0.007, and a median NE score of −0.029. **(c)** Evaluation of classifier performance based on sampling type [biopsy vs. circulating free DNA (cfDNA)] revealed better predictive performance on biopsy data than cfDNA data. The NE phenotype classifier had an accuracy of 0.90 and an area under the receiver operating characteristic curve (AUC) of 0.984 for observations associated with biopsy sampling; the classifier achieved an accuracy of 0.733 and an AUC of 0.528 for cases associated with cfDNA tissue sampling.

the patient, whereas biopsy-derived RNAseq expression is sampled directly from a single lesion. This heterogeneity component can also explain why more errors in NE prediction were observed in cfDNA-derived samples, as metastatic lesions elsewhere in the body may contribute to this expression, despite the predominance of liver lesions in these patients. However, these hypotheses cannot be evaluated within the patients evaluated in this study due to the limited sample size and lack of multiple targeted biopsy-based samples for RNAseq expression.

To our knowledge, techniques for predicting patient-level SCLC NE status have yet to be explored. Our approach in integrating the molecular phenotypic landscape of SCLC with image-extracted tumor markers offers a path towards building translational computing workflows that may help tailor SCLC treatment. In this study, we show that a logistic regression classifier can distinguish NE phenotypes using radiomics data with 80% accuracy (0.79 ± 0.04) and an AUC of 0.73 (0.70 ± 0.08). We noted that the phenotype classifier demonstrated high sensitivity for NE-positive tumors but low specificity for NE-negative tumors. This classifier was trained and tested on heterogeneous image acquisition settings, leading to consistency in performance for multiple reconstructions (series volumes) within a single study; however, further validation is warranted.

Our approach has several limitations. This is a relatively small patient cohort of metastatic patients with SCLC; the classification algorithm was trained and validated in 227 scans from 50 patients, and further research is warranted. We were underpowered to perform classification based on the Riley et al.[44] criterion, increasing the likelihood of our algorithm overfitting to the training population. Our data utilize a mixture of biopsy-sampled and cfDNA-based NE scores. The cfDNA sampling is comparatively easier but captures signals from both cancerous and non-cancerous components. This impacts our ability to describe with confidence the NE expression profiles specific to each lesion. Next, our framework utilizes a cascaded algorithm that analyzes biomedical imaging data in a stepwise fashion. This workflow inherently propagates segmentation errors and failures at the first step to the radiomics analysis, which occurs downstream. Since radiomics is volume dependent, the effectiveness of the tumor phenotype classifier relies on the performance of the lesion detection tool. Our DICE performance indicates that the model did not perform as well in mSCLC cases (DICE <0.7) compared with LiTS (DICE >0.7). This may be influenced by high disease burden and complex anatomy from extensive prior treatments, or it can affect cases with very small lesions, which are difficult to identify. Further work on how changes in predicted volumes impact radiomics and downstream classification is needed. Similarly, further investigation into the effects of contrast use and acquisition parameters is warranted. Third, our NE classification task is conducted at the image level rather than the lesion level. This primarily describes the bulk NE profile rather than tumor-specific characteristics. Any heterogeneity reflected across the patients' disease burden cannot be evaluated through the current workflow. Future validation in lesion-based assessment for samples with targeted sequencing may provide more context to individual signatures. Finally, the data distribution of NE positive to NE negative is largely unbalanced, making it difficult to further optimize the machine learning classifier to better detect NE-negative phenotypes.

In conclusion, deep learning and radiomics-based analysis enable automated detection and characterization of SCLC liver metastasis. Using AI-based platforms, routinely acquired CT scans may be used to determine the NE status of patients with mSCLC liver lesions. This could enable clinicians to tailor SCLC treatments based on a patient's NE status and its associated molecular tumor profile.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This research was supported in part by the Center for Cancer Research, National Cancer Institute, National Institutes of Health Intramural Research Program project number ZIABC012163 and ZIABC011793. The research was supported in part by the NIH Undergraduate Scholarship Program (S.T.). The contributions of the NIH author(s) were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered Works of the United States Government. However, the findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

1. van Meerbeeck JP, Fennell DA, De Ruysscher DK. Small-cell lung cancer. *Lancet*. 2011;378(9804):1741-1755. [Crossref]
2. Thomas A, Pommier Y. Small cell lung cancer: time to revisit DNA-damaging chemotherapy. *Sci Transl Med*. 2016;8(346):346fs12. [Crossref]
3. Thomas A, Mohindroo C, Giaccone G. Advancing therapeutics in small-cell lung cancer. *Nat Cancer*. 2025;6(6):938-953. [Crossref]
4. George J, Lim JS, Jang SJ, et al. Comprehensive genomic profiles of small cell lung cancer. *Nature*. 2015;524(7563):47-53. [Crossref]
5. Sher T, Dy GK, Adjei AA. Small cell lung cancer. *Mayo Clin Proc*. 2008;83(3):355-367. [Crossref]
6. Thomas A, Pattanayak P, Szabo E, Pinsky P. Characteristics and outcomes of small cell lung cancer detected by CT screening. *Chest*. 2018;154(6):1284-1290. [Crossref]
7. Herzog BH, Devarakonda S, Govindan R. Overcoming chemotherapy resistance in SCLC. *J Thorac Oncol*. 2021;16(12):2002-2015. [Crossref]
8. Jackman DM, Johnson BE. Small-cell lung cancer. *Lancet*. 2005;366(9494):1385-1396. [Crossref]
9. Cai H, Wang H, Li Z, Lin J, Yu J. The prognostic analysis of different metastatic patterns in extensive-stage small-cell lung cancer patients: a large population-based study. *Future Oncol*. 2018;14(14):1397-1407. [Crossref]
10. Kagohashi K, Satoh H, Ishikawa H, Ohtsuka M, Sekizawa K. Liver metastasis at the time of initial diagnosis of lung cancer. *Med Oncol*. 2003;20(1):25-28. [Crossref]
11. Rudin CM, Poirier JT, Byers LA, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nat Rev Cancer*. 2019;19(5):289-297. [Crossref]
12. Schroeder BA, Thomas A. SCLC subtypes and biomarkers of the transformative immunotherapy responses. *J Thorac Oncol*. 2023;18(9):1114-1117. [Crossref]
13. Lissa D, Takahashi N, Desai P, et al. Heterogeneity of neuroendocrine transcriptional states in metastatic small cell lung cancers and patient-derived models. *Nat Commun*. 2022;13(1):2023. [Crossref]
14. Thomas A, Takahashi N, Rajapakse VN, et al. Therapeutic targeting of ATR yields durable regressions in small cell lung cancers with high replication stress. *Cancer Cell*. 2021;39(4):566-579.e7. [Crossref]
15. Takahashi N, Kim S, Schultz CW, et al. Replication stress defines distinct molecular subtypes across cancers. *Cancer Res Commun*. 2022;2(6):503-517. [Crossref]
16. Roper N, Velez MJ, Chiappori A, et al. Notch signaling and efficacy of PD-1/PD-L1 blockade in relapsed small cell lung cancer. *Nat Commun*. 2021;12(1):3880. [Crossref]

17. Gay CM, Stewart CA, Park EM, et al. Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. *Cancer Cell*. 2021;39(3):346-360.e7. [Crossref]
18. Nabet BY, Hamidi H, Lee MC, et al. Immune heterogeneity in small-cell lung cancer and vulnerability to immune checkpoint blockade. *Cancer Cell*. 2024;42(3):429-443.e4. [Crossref]
19. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. *Phys Med*. 2021;85:107-122. [Crossref]
20. Jiang H, Zhou Y, Lin Y, Chan RCK, Liu J, Chen H. Deep learning for computational cytology: a survey. *Med Image Anal*. 2023;84:102691. [Crossref]
21. Alakwaa W, Nassef M, Badr A. Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *Int J Adv Comput Sci Appl*. 2017;8(8). [Crossref]
22. Yu X, Jin F, Luo H, Lei Q, Wu Y. Gross tumor volume segmentation for stage III NSCLC radiotherapy using 3D ResSE-Unet. *Technol Cancer Res Treat*. 2022;21:15330338221090847. [Crossref]
23. Christ PF, Elshaer MEA, Ettlinger F, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Vol 9901. Lecture Notes in Computer Science. Springer International Publishing; 2016:415-423. [Crossref]
24. Chlebus G, Schenk A, Moltz JH, van Ginneken B, Hahn HK, Meine H. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Sci Rep*. 2018;8(1):15497. [Crossref]
25. Bilic P, Christ P, Li HB, et al. The liver tumor segmentation benchmark (LiTS). *Med Image Anal*. 2023;84:102680. [Crossref]
26. Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. *Strahlenther Onkol*. 2020;196(10):879-887. [Crossref]
27. Gitto S, Interlenghi M, Cuocolo R, et al. MRI radiomics-based machine learning for classification of deep-seated lipoma and atypical lipomatous tumor of the extremities. *Radiol Med*. 2023;128(8):989-998. [Crossref]
28. Menon N, Guidozzi N, Chidambaram S, Markar SR. Performance of radiomics-based artificial intelligence systems in the diagnosis and prediction of treatment response and survival in esophageal cancer: a systematic review and meta-analysis of diagnostic accuracy. *Dis Esophagus*. 2023;36(6):doad034. [Crossref]
29. Abbaspour S, Barahman M, Abdollahi H, et al. Multimodality radiomics prediction of radiotherapy-induced the early proctitis and cystitis in rectal cancer patients: a machine learning study. *Biomed Phys Eng Express*. 2023;10(1). [Crossref]
30. Lubner MG, Stabo N, Lubner SJ, et al. CT textural analysis of hepatic metastatic colorectal cancer: pre-treatment tumor heterogeneity correlates with pathology and clinical outcomes. *Abdom Imaging*. 2015;40(7):2331-2337. [Crossref]
31. Kiryu S, Akai H, Nojima M, et al. Impact of hepatocellular carcinoma heterogeneity on computed tomography as a prognostic indicator. *Sci Rep*. 2017;7(1):12689. [Crossref]
32. Akai H, Yasaka K, Kunimatsu A, et al. Predicting prognosis of resected hepatocellular carcinoma by radiomics analysis with random survival forest. *Diagn Interv Imaging*. 2018;99(10):643-651. [Crossref]
33. Chaudhary U, Desai PA, Takahashi N, et al. Automated detection and segmentation of small cell lung cancer liver metastases on CT. *Journal of Clinical Oncology*. 2022;40(16_suppl):e13555. [Crossref]
34. Fialkoff G, Takahashi N, Sharkia I, et al. Subtyping of small cell lung cancer using plasma cell-free nucleosomes. *bioRxiv*. 2022. [Crossref]
35. Zhang W, Girard L, Zhang YA, et al. Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes. *Transl Lung Cancer Res*. 2018;7(1):32-49. [Crossref]
36. Hiatt JB, Doebley AL, Arnold HU, et al. Molecular phenotyping of small cell lung cancer using targeted cfDNA profiling of transcriptional regulatory regions. *Sci Adv*. 2024;10(15):eadk2082. [Crossref]
37. Medical Open Network for Artificial Intelligence (MONAI). Accessed March 11, 2024. [Crossref]
38. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. [Crossref]
39. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*. 2004;11(2):178-189. [Crossref]
40. Nakazawa K, Kurishima K, Tamura T, et al. Specific organ metastases and survival in small cell lung cancer. *Oncol Lett*. 2012;4(4):617-620. [Crossref]
41. Mettler E, Fottner C, Bakhshandeh N, Trenkler A, Kuchen R, Weber MM. Quantitative Analysis of Plasma Cell-Free DNA and Its DNA Integrity and hypomethylation status as biomarkers for tumor burden and disease progression in patients with metastatic neuroendocrine neoplasias. *Cancers (Basel)*. 2022;14(4):1025. [Crossref]
42. Hyun MH, Lee ES, Eo JS, et al. Clinical implications of circulating cell-free DNA quantification and metabolic tumor burden in advanced non-small cell lung cancer. *Lung Cancer*. 2019;134:158-166. [Crossref]
43. Nygaard AD, Holdgaard PC, Spindler KL, Pallisgaard N, Jakobsen A. The correlation between cell-free DNA and tumour burden was estimated by PET/CT in patients with advanced NSCLC. *Br J Cancer*. 2014;110(2):363-368. [Crossref]
44. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296. [Crossref]

Supplementary Table 1. Image acquisition characteristics

Cohort	Scanner	#Scans	KVP	In-plane resolution	Slice thickness	Contrast
LİTS cohort	N/A	131	N/A	0.768 (0.557,1)	<1 (51) 1 (38) 1-5 (33) 5 (9)	None (1) Venous (130)
mSCLC segmentation (n = 82 series volumes)	SIEMENS	73	120	0.820 (0.691-0.977)	1 (3)	None (5) Venous (66) Arterial (2)
	Canon Medical Systems	3	120	0.858 (0.782-0.858)	1 (1) 5 (2)	Venous
	GE Medical Systems	1	120	0.7813	1-5 (1)	Venous
	TOSHIBA	5	120	0.938 (0.743-0.938)	1 (1) 5 (4)	Arterial (1) Venous (4)
mSCLC classification (n = 227 series volumes)	SIEMENS	207	120 (205) 100 (2)	0.820 (0.677-0.977)	1 (105) 5 (102)	None (7) Venous (200)
	Canon Medical Systems	12	120	0.839 (0.782, 0.9)	1 (6) 5 (6)	Venous
	GE Medical Systems	2	120	0.7813, 0.7031	1-5 (2)	Arterial (1) Venous (1)
	TOSHIBA	6	120	0.782 (0.743, 0.782)	1 (3) 5 (3)	Arterial (1) Venous (5)

LiTS, liver tumor segmentation; mSCLC, metastatic small cell lung cancer; KVP, kilovolt peak; N/A, not available.

Supplementary Table 2. Cross-validation performance across weighted and unweighted models

Weighted	Optim	Pos class weight	Model	F1	AUC
No	-	-	Logistic regression	0.85 (0.81, 0.89)	0.55 (0.45, 0.67)
No	-	-	Random forest	0.86 (0.81, 0.91)	0.68 (0.56, 0.81)
No	-	-	XGBoost	0.84 (0.79, 0.91)	0.59 (0.48, 0.69)
Yes	AUC	0.25	Logistic regression	0.85 (0.82, 0.89)	0.56 (0.46, 0.69)
Yes	BAL	-	Logistic regression	0.70 (0.64, 0.83)	0.54 (0.43, 0.65)
Yes	F1	4.9	Logistic regression	0.85 (0.82, 0.89)	0.55 (0.46, 0.69)
Yes	AUC	0.3	Random forest	0.84 (0.77, 0.91)	0.67 (0.55, 0.80)
Yes	BAL	-	Random forest	0.84 (0.76, 0.91)	0.67 (0.56, 0.79)
Yes	F1	8.6	Random forest	0.85 (0.80, 0.91)	0.67 (0.54, 0.81)
Yes	AUC	0.05	XGBoost	0.30 (0.14, 0.50)	0.58 (0.48, 0.68)
Yes	BAL	0.321	XGBoost	0.72 (0.65, 0.79)	0.59 (0.48, 0.69)
Yes	F1	0.8	XGBoost	0.84 (0.78, 0.91)	0.59 (0.48, 0.71)

Optim: Method used to select positive class weight. AUC: Best AUC from all evaluated class weights. F1: Maximum F1 score from all evaluated class weights. BAL: Dataset distribution-based weighting. AUC, area under the receiver operating characteristic curve.

Supplementary Table 3. Test set performance for weighted training. 5-fold cross-validation bootstrap mean (95% CI) from 1000 random samplings of one series/study. Performance based on ensemble of models

Weight	Optim	Model	Accuracy	PPV	NPV	Sensitivity	Specificity	F1	AUC
Y	AUC	LR	0.81 (0.77, 0.85)	0.80 (0.77, 0.83)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	0.16 (0.00, 0.33)	0.89 (0.87, 0.91)	0.70 (0.59, 0.82)
Y	BAL	LR	0.58 (0.50, 0.65)	0.95 (0.89, 1.00)	0.34 (0.29, 0.40)	0.48 (0.40, 0.55)	0.91 (0.83, 1.00)	0.63 (0.55, 0.71)	0.72 (0.59, 0.86)
Y	F1	LR	0.81 (0.77, 0.85)	0.80 (0.77, 0.83)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	0.16 (0.00, 0.33)	0.89 (0.87, 0.91)	0.68 (0.55, 0.81)
Y	AUC	RF	0.69 (0.62, 0.77)	0.80 (0.75, 0.85)	0.33 (0.17, 0.50)	0.80 (0.70, 0.90)	0.33 (0.17, 0.50)	0.80 (0.74, 0.86)	0.61 (0.50, 0.72)
Y	BAL	RF	0.71 (0.62, 0.81)	0.81 (0.75, 0.86)	0.36 (0.17, 0.60)	0.82 (0.75, 0.90)	0.33 (0.17, 0.50)	0.81 (0.75, 0.88)	0.62 (0.52, 0.71)
Y	F1	RF	0.77 (0.73, 0.81)	0.78 (0.76, 0.80)	0.51 (0.00, 1.00)	0.97 (0.95, 1.00)	0.09 (0.00, 0.17)	0.87 (0.84, 0.89)	0.53 (0.43, 0.63)
Y	AUC	XGB	0.33 (0.31, 0.35)	1.00 (1.00, 1.00)	0.26 (0.25, 0.26)	0.13 (0.10, 0.15)	1.00 (1.00, 1.00)	0.22 (0.18, 0.26)	0.63 (0.53, 0.75)
Y	BAL	XGB	0.48 (0.42, 0.54)	0.79 (0.69, 0.90)	0.24 (0.15, 0.31)	0.45 (0.45, 0.45)	0.58 (0.33, 0.83)	0.57 (0.55, 0.60)	0.64 (0.49, 0.78)
Y	F1	XGB	0.73 (0.65, 0.81)	0.84 (0.79, 0.89)	0.43 (0.29, 0.60)	0.80 (0.70, 0.90)	0.50 (0.33, 0.67)	0.82 (0.76, 0.88)	0.64 (0.50, 0.78)

Optim: Method used to select positive class weight. AUC: Best AUC from all evaluated class weights. F1: Maximum F1 score from all evaluated class weights. BAL: Dataset distribution-based weighting. RF, random forest; LR, logistic regression; XGB, XG boost; AUC, area under the receiver operating characteristic curve; CI, confidence interval.