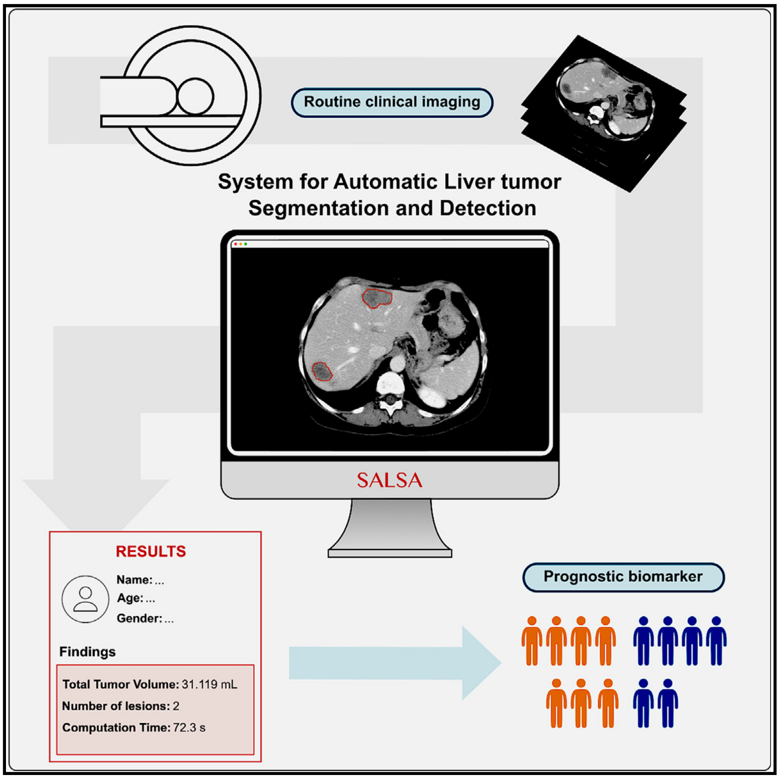
Cell Reports
Medicine

Article

A CT-based deep learning-driven tool for automatic liver tumor detection and delineation in patients with cancer

Graphical abstract



Name: ...
Age: ...
Gender: ...

Findings

Total Tumor Volume: 31.119 mL
Number of lesions: 2
Computation Time: 72.3 s

Highlights

- Quantifying liver tumors is crucial for cancer diagnosis and treatment planning
- SALSA is a fully automated tool for precise liver tumor detection and delineation
- SALSA surpasses state-of-the-art models and the radiologists' inter-reader agreement
- SALSA can arguably enhance cancer detection, treatment planning, and response evaluation

Authors

Maria Balaguer-Montero,
Adrià Marcos Morales, Marta Ligero, ...,
Rodrigo Dienstmann, Elena Garralda,
Raquel Perez-Lopez

Correspondence

rperez@vhio.net

In brief

Balaguer-Montero et al. present SALSA (system for automatic liver tumor segmentation and detection), a fully automated tool for liver tumor delineation. SALSA outperforms state-of-the-art models and expert radiologist agreement, enabling efficient and reliable liver tumor assessment.

Balaguer-Montero et al., 2025, Cell Reports Medicine 6, 102032
April 15, 2025 © 2025 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrm.2025.102032

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

A CT-based deep learning-driven tool for automatic liver tumor detection and delineation in patients with cancer

Maria Balaguer-Montero,[1,8] Adrià Marcos Morales,[1,8] Marta Ligero,[1,2,8] Christina Zatse,[1] David Leiva,[3] Luz M. Atlagich,[1,4] Nikolaos Staikoglou,[1] Cristina Viaplana,[5] Camilo Monreal,[1] Joaquin Mateo,[6] Jorge Hernando,[6] Alejandro García-Álvarez,[6] Francesc Salvà,[6] Jaume Capdevila,[6] Elena Elez,[6] Rodrigo Dienstmann,[5,7] Elena Garralda,[6] and Raquel Perez-Lopez[1,9,*]

[1]Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), 08035 Barcelona, Spain
[2]Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, TUD Dresden University of Technology, 01307 Dresden, Germany
[3]Bellvitge University Hospital, 08907 Barcelona, Spain
[4]Oncocentro Apys, Viña Del Mar 2520598, Chile
[5]Oncology Data Science (ODysSey) Group, Vall d'Hebron Institute of Oncology (VHIO), 08035 Barcelona, Spain
[6]Department of Medical Oncology, Vall d'Hebron University Hospital and Institute of Oncology (VHIO), 08035 Barcelona, Spain
[7]University of Vic – Central University of Catalonia, 08500 Vic, Spain
[8]These authors contributed equally
[9]Lead contact
*Correspondence: rperez@vhio.net

https://doi.org/10.1016/j.xcrm.2025.102032

SUMMARY

Liver tumors, whether primary or metastatic, significantly impact the outcomes of patients with cancer. Accurate identification and quantification are crucial for effective patient management, including precise diagnosis, prognosis, and therapy evaluation. We present SALSA (system for automatic liver tumor segmentation and detection), a fully automated tool for liver tumor detection and delineation. Developed on 1,598 computed tomography (CT) scans and 4,908 liver tumors, SALSA demonstrates superior accuracy in tumor identification and volume quantification, outperforming state-of-the-art models and inter-reader agreement among expert radiologists. SALSA achieves a patient-wise detection precision of 99.65%, and 81.72% at lesion level, in the external validation cohorts. Additionally, it exhibits good overlap, achieving a dice similarity coefficient (DSC) of 0.760, outperforming both state-of-the-art and the inter-radiologist assessment. SALSA's automatic quantification of tumor volume proves to have prognostic value across various solid tumors (p = 0.028). SALSA's robust capabilities position it as a potential medical device for automatic cancer detection, staging, and response evaluation.

INTRODUCTION

Liver cancer poses a significant health challenge. Primary liver cancers such as hepatocellular carcinoma and cholangiocarcinoma are often diagnosed in advanced stages, with limited treatment options available and unfavorable prognosis. Additionally, the liver is a common site for metastases originating from other primary cancers, significantly impacting patient prognosis.[1,2] Early and precise detection is crucial as it opens the possibility for localized and potentially curative treatments, thereby improving the overall outcomes for patients.

The evaluation of liver tumor burden is crucial at different stages of cancer treatment. For therapy decisions, especially when planning surgical interventions, accurately assessing the number and volume of liver tumors is critical.[3] This evaluation, typically performed on medical images, such as computed tomography (CT), is essential for planning curative-intent surgeries where balancing tumor removal with preserving healthy liver tissue is crucial. Currently, this task is carried out either subjectively or in a more quantitative way manually, which is not only time-consuming but also prone to variability between different observers and within the same observer.[4–6] Furthermore, the need for a more comprehensive assessment of liver tumors extends to treatment monitoring in patients with cancer. The evaluation of cancer volume changes throughout treatment on CT images, as opposed to relying solely on the maximum diameter of a few tumors (as defined by the standard Response Evaluation Criteria In Solid Tumors [RECIST][7]), offers a more accurate response assessment and prediction of clinical outcomes.[8–10]

Yet, accurately delineating tumors for volume analysis (i.e., drawing tumor contours) poses practical challenges and often acts as a bottleneck in numerous research projects and clinical applications that involve volumetric disease assessment. Here, a fully automated delineation tool can be transformative, enhancing accuracy in tumor detection and volume assessment. Such a tool will reduce manual workload and variability in measurements, streamline the treatment planning process, and support more accurate and robust response evaluation.

Previous studies in the field of medical imaging have aimed to advance the automated detection and segmentation of liver tumors.[11,12] While these efforts have yielded valuable insights and tools, they often confront challenges related to limited inclusion of tumor types and a primary focus on segmenting individual lesions, which limits their capacity to enable an integral assessment of liver tumor burden. Seeking to address these clinical challenges, our study introduces system for automatic liver tumor segmentation and detection (SALSA), a tool that offers a comprehensive solution for the automatic detection and delineation of all liver tumors on CT images. Our approach has been developed in an extensive and heterogeneous dataset encompassing primary tumors, liver metastases, and various CT protocols, enabling comprehensive performance evaluation. The tool is generalizable across a multicentric test cohort and four external independent datasets. It also incorporates qualitative feedback from expert radiologists, allowing comparison between their preferred segmentations (ground truth) and those produced by our tool. Furthermore, we benchmarked our tool's efficacy against the most accurate liver tumor segmentation tool to date, derived from the Liver Tumor Segmentation Challenge (LiTS),[12] compared our model's segmentations against evaluations from three expert radiologists, and explored the prognostic potential of liver cancer burden quantified automatically by this tool.

Our tool, SALSA, which can be accessed and tested through our site https://radiomics.vhio.net/salsa/, surpasses the accuracy of the top available LiTS[12] models and expert radiologists. It offers precise and automated liver cancer detection and volume quantification on CT images, with implications for clinical outcomes and with no need for user prompts such as region of interest delineation of the lesions to segment.

RESULTS

Population demographics

In this multi-center, retrospective study, we analyzed CT data from 1,598 contrast-enhanced CT scans of the entire liver in 1,306 patients with cancer. This analysis included 4,908 liver tumors identified in 1,041 of these patients, while 265 patients had no cancer in the liver. The detection and delineation model was developed and tested using 885 CT scans from 593 patients. Additionally, four independent cohorts were collected from open-access repositories for external validation[12–15] (Figure 1). The dataset comprises a wide range of images, including both primary and secondary tumors with differing sizes and visual characteristics. The predominant tumor types in the development and test cohorts are liver metastases from colorectal cancer (151 patients), lung cancer (88 patients), and neuroendocrine tumors (80 patients). Also, 19 patients had primary liver tumors (either hepatocarcinoma or cholangiocarcinoma).

Two of the four independent external validation datasets, the Liver Tumor Segmentation (LiTS) open-source dataset[12] and the Medical Segmentation Decathlon (MSD)-Hepatic Vessels dataset,[13] did not include specific information about the origin of the liver tumors, indicating only that both primary and metastatic cancers were present in the dataset, without specifying the type for each scan. Patients in the third external validation dataset, from The Cancer Imaging Archive (TCIA) Colorectal Liver Metastases (CRLM),[14] all had liver metastases from colorectal cancer, whereas the fourth external validation dataset, also from TCIA, Hepatocellular Carcinoma Transarterial Chemoembolization (HCC-TACE-Seg),[15] only included patients with primary liver cancer (hepatocarcinoma).

Additionally, the tumors in the development and test cohorts exhibit varying levels of contrast compared to the healthy liver tissue, spanning from hyperdense to hypodense tumors. Both datasets exhibit similar distributions in terms of gender, number of liver tumors, percentage of patients with primary liver cancer, and disease burden. The patient demographics and clinical characteristics per group are shown in Table 1.

CT scan characteristics

The dataset comprised a diverse collection of abdominal CT scans acquired using different scanners and acquisition protocols following intravenous contrast injection, adhering to standard clinical procedures. This dataset mirrors real-world clinical conditions, encompassing common imaging artifacts such as metal and motion artifacts and displaying variations in resolution and image quality. For a concise summary of data statistics, see Table S1.

SALSA shows high accuracy for cancer detection and precise tumor delineation

Three state-of-the-art neural network architectures for image segmentation, including the latest transformer architectures, were employed in the development of a tool designed for the automatic detection and delineation of liver tumors. The 3D U-Net cascade model implementation from nnU-Net exhibited the top performance of all the explored models (Figures 2A–2C and S1; Tables S2–S4). Therefore, the 3D U-Net cascade model is designated as the SALSA tool and has been tested to evaluate its robustness and prognostic value, with the results shown in the following sections.

The lesion detection accuracy was assessed both in a patient-wise and tumor-wise manner. SALSA showed high accuracy for detecting liver tumors, with a patient-wise precision of 80.59% and a recall of 99.08% on the test set. In the external validation cohort, which included the four independent datasets, it obtained similar results: a precision of 99.65% and a recall of 94.17%. When considering each lesion individually, SALSA obtained a lesion-by-lesion detection precision of 58.07% and 81.72% in the test and external validation cohort and a recall of 70.38% and 57.92%, respectively (see Table 2). These results are consistent with the differences among the four datasets, where different segmentation criteria, in terms of minimum lesion size, were applied, reflecting the robustness of the proposed tool.

In parallel, segmentation evaluation was performed binarily by computing patient-wise and tumor-wise metrics. Both approaches reported a good overlap among the tumor masks

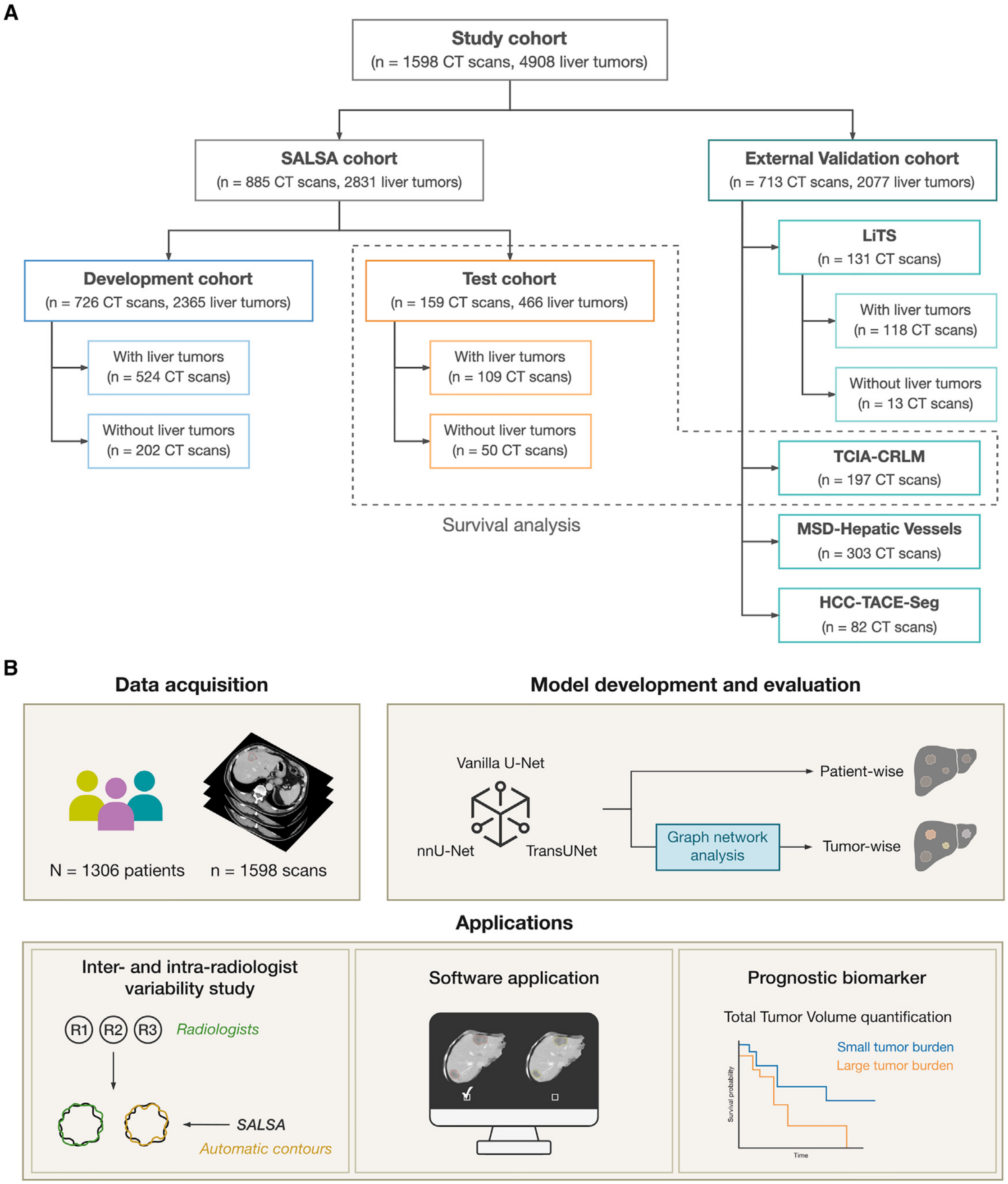


Figure 1. Overview of the study population and design

(A) Distribution of the data included in the study. It details the number of CT scans and liver tumors involved in the development and test cohorts as well as the collected external validation cohorts.

(B) Overview of the study workflow and methodological framework, including the tested architectures and evaluation metrics for per-patient and tumor-wise assessments. It features a schematic representation of two side studies: (1) benchmarking against professional radiologists through intra- and inter-reader variability studies and (2) expert preference analysis comparing manual (ground truth) segmentations with those generated by SALSA. Additionally, it explores the use of automated liver tumor quantification as a prognostic biomarker in patients with cancer.

generated by SALSA and the ground truth in both the test (patient-wise DSC of 0.737 and tumor-wise DSC of 0.761) and external validation (patient-wise DSC of 0.738 and tumor-wise DSC of 0.760) cohorts (see Table 2).

With a precision for liver tumor detection of 28.06%, 82.61% recall, and 0.714 DSC in our test set and 54.03%, 85.47%, and 0.690, respectively, in the external validation cohort, SALSA has been proved to benchmark the LiTS top-performing model,

Table 1. Description of the patient characteristics in the development and test datasets used in this study

	Development	Test	LiTS	TCIA-CRLM	TCIA-HCC	MSD	*p* values
Number of patients	452	141	131	197	82	303	–
Number of CT scans	726	159	131	197	82	303	–
Number of CT scans with liver tumors	524	109	118	197	82	303	–
Number of CT slices with tumor (%)	35,692 (19.53)	9,006 (22.44)	7,296 (23.56)	8,834 (19.31)	2,411 (39.76)	16,693 (23.38)	–
Number of CT scans with primary liver cancer (%)	20 (2.76)	8 (5.03)	–	0 (0)	82 (100)	–	0.595
Tumor volume (mL)	2.97 (0.91–10.45)	4.52 (1.42–15.64)	0.54 (0.15–2.80)	1.56 (0.49–7.24)	55.46 (17.73–185.49)	5.84 (0.94–37.90)	6.82e−06
Total tumor volume (mL)[a]	27.99 (6.88–97.88)	53.28 (18.56–151.27)	16.64 (3.54–108.53)	9.49 (3.68–32.31)	79.65 (24.79–317.53)	31.34 (8.46–104.48)	0.119
Age	69 (59–77)	63 (55–71)	–	61 (52–69)	67 (58–76)	–	1.65e−05
Gender (% female)	42.04	50.35	–	40.61	35.2	–	0.101
Number of tumors	2,365	466	893	524	118	542	–
Number of tumors per patient	2 (0–5)	2 (0–5)	4 (1–10)	2 (1–3)	1 (1–7)	1 (1–2)	0.489

Data are represented as median and interquartile range (IQR).

LiTS, Liver Tumor Segmentation Challenge; TCIA, The Cancer Imaging Archive; CRLM, Colorectal Liver Metastases; HCC, Hepatocellular Carcinoma; TACE, Transarterial Chemoembolization; MSD, Medical Segmentation Decathlon.

[a] Not including patients without liver tumors.

benefitting from having a larger, more heterogeneous, and real-world set of cohorts of liver tumors (Figures 3 and 4A; Table S5).

Impact of tumor size and density on SALSA's detection and segmentation performance

We assessed whether factors such as tumor type, size, imaging characteristics, including tumor intensity, and CT scan acquisition parameters, like slice thickness, had an impact on the accuracy of SALSA in detecting and delineating liver cancer. Remarkably, hypodense tumors compared to the liver density and larger tumors were proven to be more readily detected and delineated by the SALSA tool (Figures 2D and 2E; Table S6). Furthermore, our analysis revealed that the distribution of DSC remained consistent across different tumor types and regardless of the CT slice thickness (Figure S2; Table S7).

The SALSA tool showed a lower performance when detecting and delineating small liver tumors (Figure 2E; Table S8). Considering that tumors under 1 cm in the largest axial plane are often deemed indeterminate and non-measurable by the RECIST guidelines,[7] which were applied during the development cohort segmentation, SALSA inherently prioritizes clinically significant cancers. Moreover, its consistent performance across tumor types and CT slice thicknesses underscores SALSA's broad applicability in diagnosing prevalent solid tumors with standard CT imaging protocols.

Automatically generated contours are comparable to manual segmentations by expert radiologists

To explore the variability among radiologists in detecting and delineating liver tumors, we randomly selected a group of 25 patients from our test cohort. Three radiologists, blinded to the ground truth, delineated all liver tumors in each case. All outlines created by both the radiologists and the models were measured against a gold standard, specifically, masks segmented manually by the reference expert radiologist (radiologist 1).

We calculated how closely the contours matched the manual segmentation using the tumor-wise DSC metric for quantification. Additionally, we examined intra-radiologist variability by assessing the agreement in delineation when the same reference radiologist performed the liver tumor segmentation twice.

Our findings revealed that SALSA's performance in outlining liver tumors (F1-score of 75.89 and DSC of 0.800 in the 25-patient test cohort subset) was comparable to, or even better than, the level of agreement observed by each of the two blinded radiologists (F1-scores of 68.72 and 47.81, DSCs of 0.778 and 0.736, respectively), used for inter-radiologist variability assessment. Such performance was found to be even close to that of the reference expert radiologist (F1-score of 79.50 and DSC of 0.820), used for intra-radiologist variability assessment, indicating a high level of precision in the model's detections and delineations (Figure 4B; Table S9).

Automatic tumor burden quantification by SALSA is prognostic in patients with cancer

The prognostic value of liver cancer burden was assessed using both manual evaluations by expert radiologists and automated delineations by SALSA. Analysis was conducted on data including the test set, with 141 patients (Table S10), and all the cases from the external validation cohort for which clinical outcome was available (TCIA-CRLM), including 197 patients. The association between total tumor volume and clinical outcome was studied. The results revealed that a higher liver cancer burden is associated with poorer prognosis ($p = 0.028$, hazard ratio; 95% confidence interval = 1.692; 1.055, 2.715), regardless

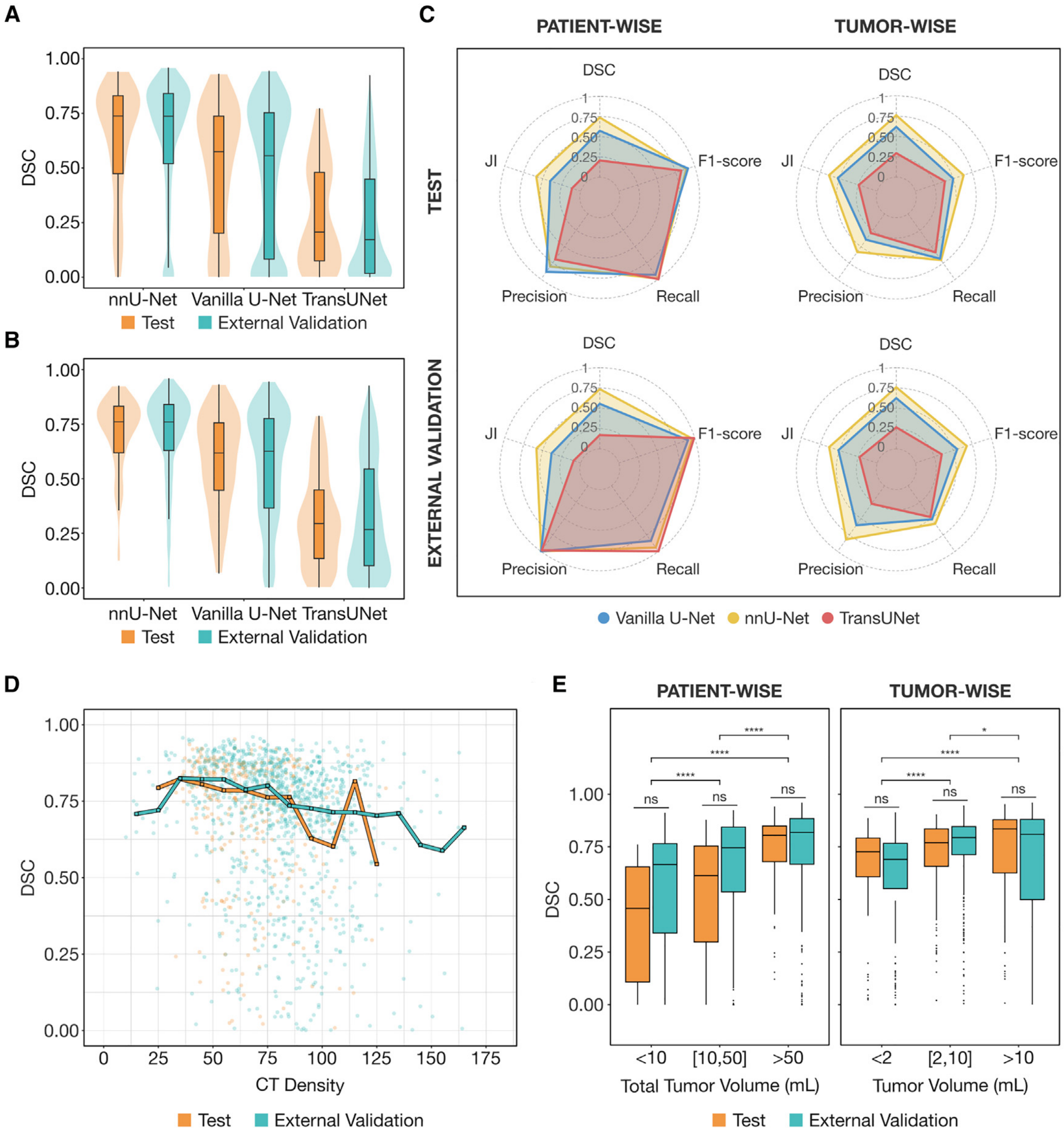


Figure 2. SALSA's performance evaluation

(A and B) Delineation performance evaluated by the intersection of ground truth masks with those automatically generated by SALSA, calculated using the dice similarity coefficient (DSC). Results from the three tested architectures in the test set (orange) and external validation set (blue) are shown for both patient-wise (A) and tumor-wise (B) levels.

(C) Evaluation of various metrics for detection (precision, recall, and F1-score) and delineation (DSC and Jaccard Index [JI]) performance by SALSA.

(D and E) Analysis of the impact of tumor density (D) and volume (E) on delineation performance, highlighting SALSA's reduced efficacy in hyperdense and small liver tumors.

Significance was calculated using independent two-sample t tests with Bonferroni correction and is represented as: *$p < 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$; **$p \leq 0.0001$; NS, not significant. Data are represented as median and interquartile range (IQR) in (A), (B), and (E). See also Tables S3, S4, and S6.

of whether the assessment was manual or automated. Notably, the automated quantification by SALSA confirmed the prognostic importance of tumor burden and demonstrated comparable risk stratification of patients to manual assessments. This finding suggests that automated tools like SALSA could offer fast and precise tumor burden quantification and associated prognostic differentiation in liver cancer (Figure 5).

Radiologists show equal preference for manual and SALSA liver tumor segmentations

For expert validation purposes, a user-friendly web application was developed to allow direct comparison of radiologist preferences between manual segmentations and those generated by the SALSA tool. The application, available at https://radiomics.vhio.net/salsa/, featured the entire liver volume as a scrollable

Table 2. Liver tumor detection and delineation performance of SALSA at both the patient and tumor-wise levels

	Tumor detection performance					
	Patient-wise			Tumor-wise		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Test (*n* = 159)	80.59	99.08	88.88	58.07	70.39	63.64
External validation (*n* = 713)	99.65	94.17	96.83	81.72	57.92	67.79

	Tumor detection performance			
	Patient-wise		Tumor-wise	
	DSC	JI	DSC	JI
Test (*n* = 159)	0.738 (0.474–0.831)	0.585 (0.310–0.710)	0.761 (0.619–0.832)	0.635 (0.534–0.721)
External validation (*n* = 713)	0.737 (0.520–0.841)	0.583 (0.351–0.725)	0.760 (0.629–0.840)	0.633 (0.502–0.733)
External validation LITS (*n* = 131)	0.769 (0.543–0.841)	0.624 (0.373–0.725)	0.773 (0.677–0.833)	0.670 (0.554–0.717)
External validation TCIA (*n* = 197)	0.724 (0.518–0.828)	0.568 (0.349–0.707)	0.760 (0.651–0.826)	0.617 (0.501–0.726)
External validation MSD (*n* = 303)	0.743 (0.533–0.841)	0.591 (0.363–0.726)	0.766 (0.602–0.846)	0.638 (0.501–0.739)
External validation HCC (*n* = 82)	0.726 (0.435–0.868)	0.569 (0.278–0.767)	0.715 (0.385–0.861)	0.590 (0.459–0.767)

Data are presented as percentages (%) for precision, recall, and F1-score and as median and interquartile range (IQR) for DSC and JI.
LiTS, Liver Tumor Segmentation Challenge; TCIA, The Cancer Imaging Archive; CRLM, Colorectal Liver Metastases; HCC, Hepatocellular Carcinoma; TACE, Transarterial Chemoembolization; MSD, Medical Segmentation Decathlon; DSC, dice similarity coefficient; JI, Jaccard's Index.

element and allowed for window adjustment and navigation to aid radiologists in accurately evaluating the quality of the contours, depicted over the scan using random colors in order to avoid biasing the choice. A subset of 200 randomly selected cases from both the test and external validation cohorts was reviewed by three experienced radiologists, who were asked to submit their preference for either manual or automated segmentation, or to express no preference.

In the test dataset, radiologists clearly preferred model segmentations over manual ones, choosing them 53.33% of the time compared to 15.00% for manual segmentations. They claimed no specific preference in the remaining 31.66% of cases. Conversely, in the TCIA-CRLM dataset used for external validation, preferences for manual and automatic segmentations were equally distributed, along with a similar rate of no preference responses (*p* value 0.855), indicating no significant bias toward either method. These observations are depicted in Figure 4C, which illustrates the frequency of each type of segmentation chosen by each radiologist across different evaluation datasets.

Within the LiTS cohort, a distinct preference for manual segmentations was observed, primarily due to the cohort's characteristic of smaller lesions, as highlighted in Table 1. This pattern reflects SALSA's slightly decreased performance in segmenting smaller lesion sizes, as shown in Figure 2. It should be noted that SALSA is designed to prioritize the detection of malignant liver tumors larger than 1 cm in diameter, intentionally excluding indeterminate lesions, also defined as those non-measurable according to standard response criteria such as RECIST.[7]

This validation by expert radiologists revealed a balanced preference for SALSA's segmentations, highlighting its strategy to exclude non-specific or non-measurable lesions.

DISCUSSION

The liver, affected by both primary hepatobiliary cancers and metastases, plays a pivotal role in cancer staging, prognosis, and treatment planning. Precision in tumor detection and delineation is pivotal for consistently and precisely measuring tumor burden on imaging at diagnosis and over multiple time points through the course of the disease. However, the detailed quantification of liver cancer currently demands the skill of experienced radiologists and manual tumor segmentation, posing significant barriers to clinical application. Our development of an automatic segmentation tool for liver tumors, both primary and metastatic, aims to overcome these challenges.

We compared three advanced deep learning methods, including a transformer-based U-Net and nnU-Net,[16–18] to develop a robust tool for automated tumor detection and delineation. Our models were trained on a diverse dataset, encompassing primary tumors, liver metastases, and various CT protocols, enabling comprehensive performance evaluation. Among the explored models, nnU-Net was found to be the best performing option, surpassing state-of-the-art transformer architectures, which exhibited poor performance due to their requirement for massive training cohorts to achieve good results.[19]

We demonstrated the tool's ability to generalize across a multicentric test cohort and four external datasets.[12] We computed standard evaluation metrics to assess its performance quantitatively and, additionally, requested qualitative feedback by asking expert radiologists to compare their preferred segmentations, ground truth with those produced by our tool, SALSA. Furthermore, we benchmarked our tool's efficacy against the most accurate liver tumor automatic segmentation tool to date, derived from the LiTS.[12] We also compared our model's segmentations against those from three expert radiologists, offering an in-depth evaluation of our tool's performance. Additionally, with the aim of exploring a possible clinical application, we explored the potential of liver cancer burden quantified automatically by this tool as a prognostic biomarker and proven that it is equal to the one obtained from manual segmentations.

SALSA enables fully automated detection and segmentation of liver tumors, providing precise quantification of tumor

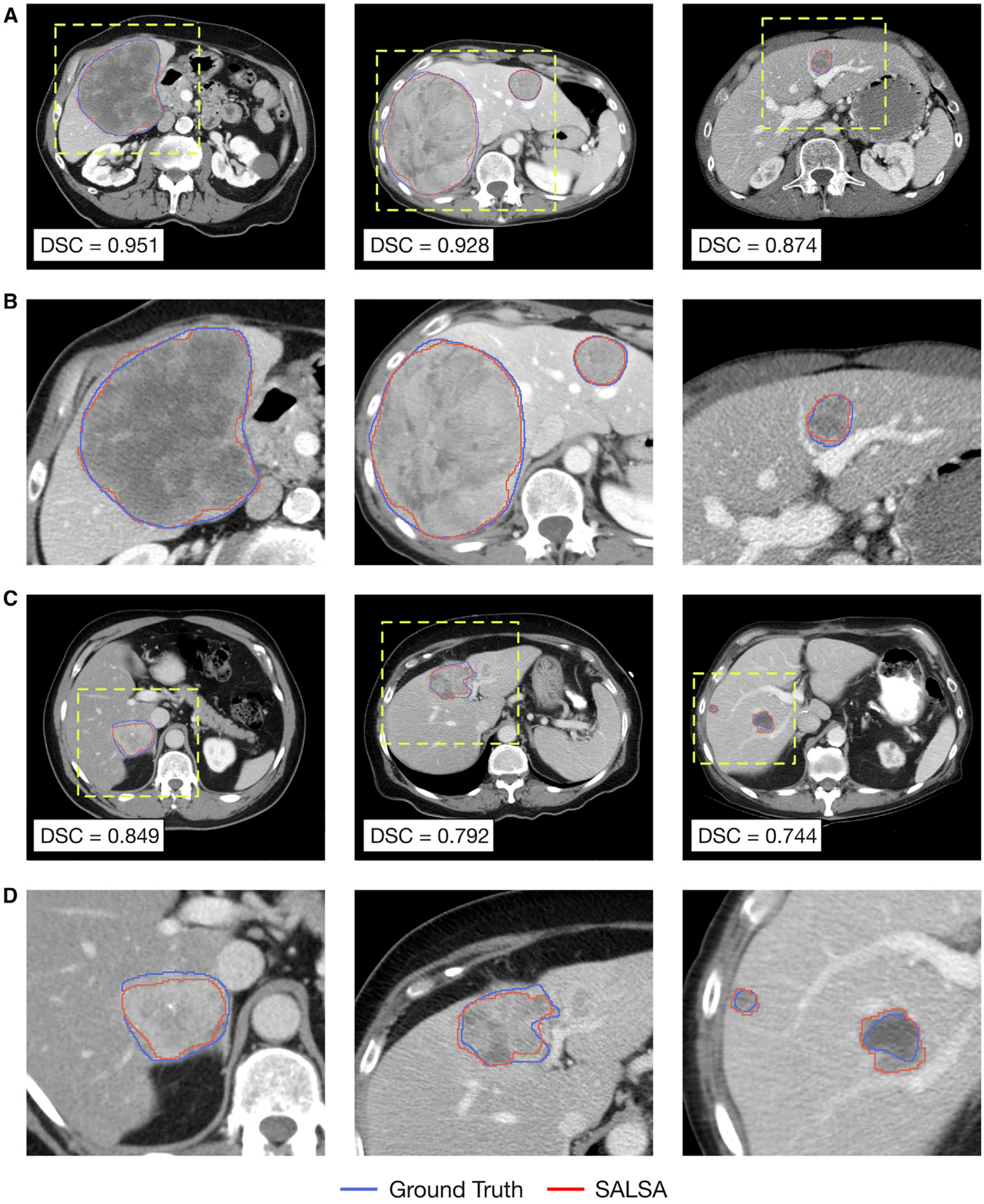


Figure 3. Visual comparison of the automatically delineated contours with radiologist-generated ground truth

Representative cases of liver tumors delineated by SALSA (red lines) compared to the ground truth (blue lines) segmented masks. Yellow dashed boxes in (A) and (C) indicate the regions magnified in (B) and (D) for improved visualization. The delineations are displayed as colored masks, highlighting areas of agreement and discrepancies between the assessments.

number and volume. It matches expert radiologists in detecting and outlining liver tumors, a capability not previously available in open-access tools. Surpassing the highest-performing model from the LiTS challenge at both patient and lesion levels, SALSA sets a new benchmark in automated medical imaging segmentation. Crucially, unlike the LiTS dataset models, SALSA's robustness and generalizability have been rigorously validated across an independent test set and four external datasets, showcasing its unparalleled effectiveness in tumor detection and delineation in real-world clinical

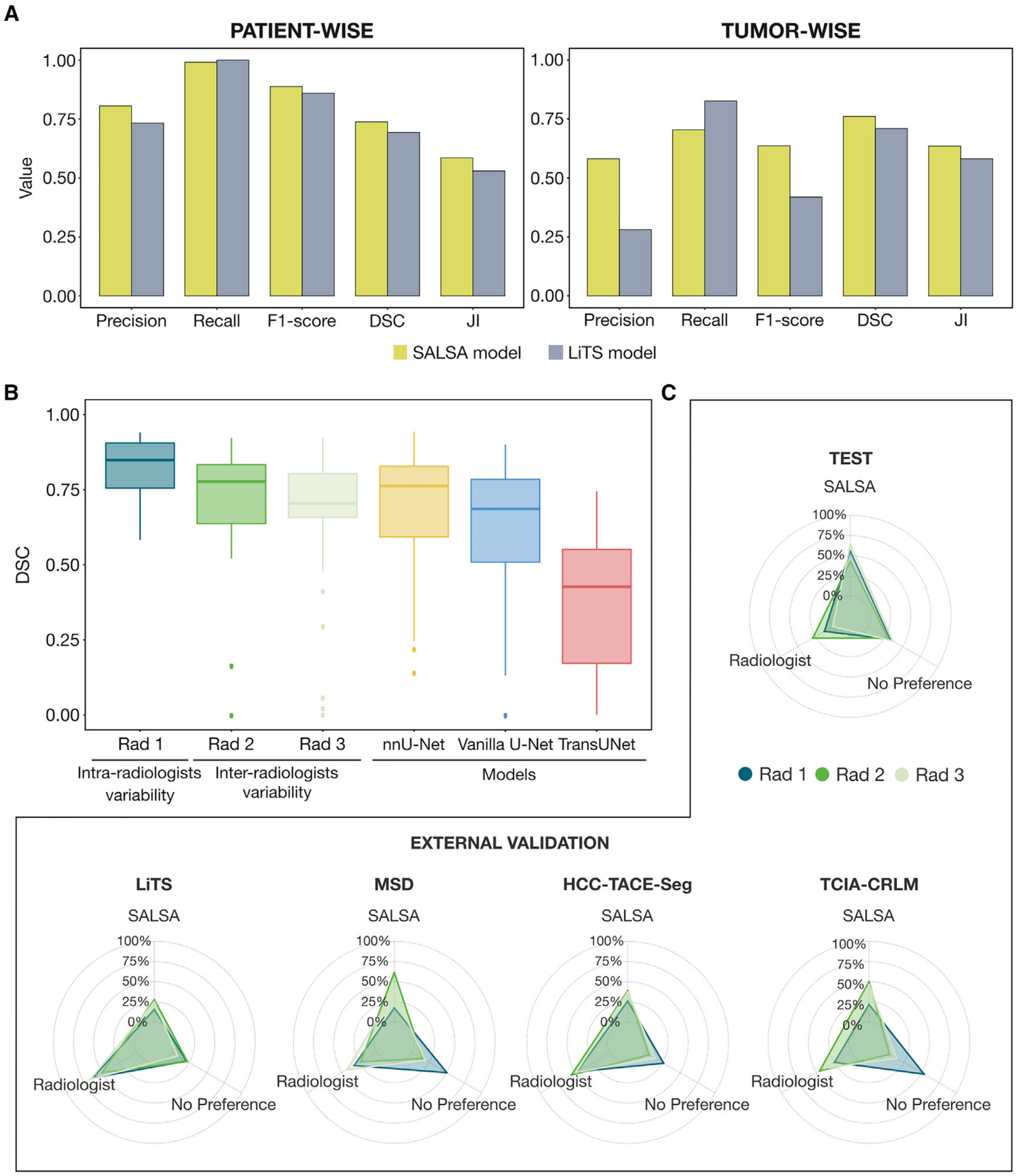


Figure 4. Benchmarking SALSA against state-of-the-art models and inter-radiologist metrics

(A) SALSA benchmarks against the top-performing LiTS model for both detecting and delineating liver tumors. Metrics include detection (precision, recall, and F1-score) and delineation (dice similarity coefficient [DSC] and Jaccard Index [JI]) across the test set using patient-wise and tumor-wise approaches.

(B) Comparison of tumor delineation overlaps. Overlap with ground truth delineations is compared for segmentations by the same radiologist on two occasions (Rad 1), two independent radiologists (Rad 2 and Rad 3), and the SALSA models (nnU-Net, Vanilla U-Net, and TransUNet), providing a benchmark of model performance against human experts.

(C) Radiologist preferences for manual versus automated liver tumor delineation. Three expert radiologists (Rad 1, Rad 2, and Rad 3) assessed their preferences between manual segmentations (performed by an expert radiologist) and automated segmentations (SALSA), with the option to express no preference.

Data are represented as median and interquartile range (IQR) in (A) and (B). See also Tables S5 and S9.

settings. We have also developed an open-source software application, accessible for testing at https://radiomics.vhio.net/salsa/. This tool may facilitate the integration of SALSA into clinical oncology imaging workflows, significantly reducing the time required for manual tumor detection and precise delineation.

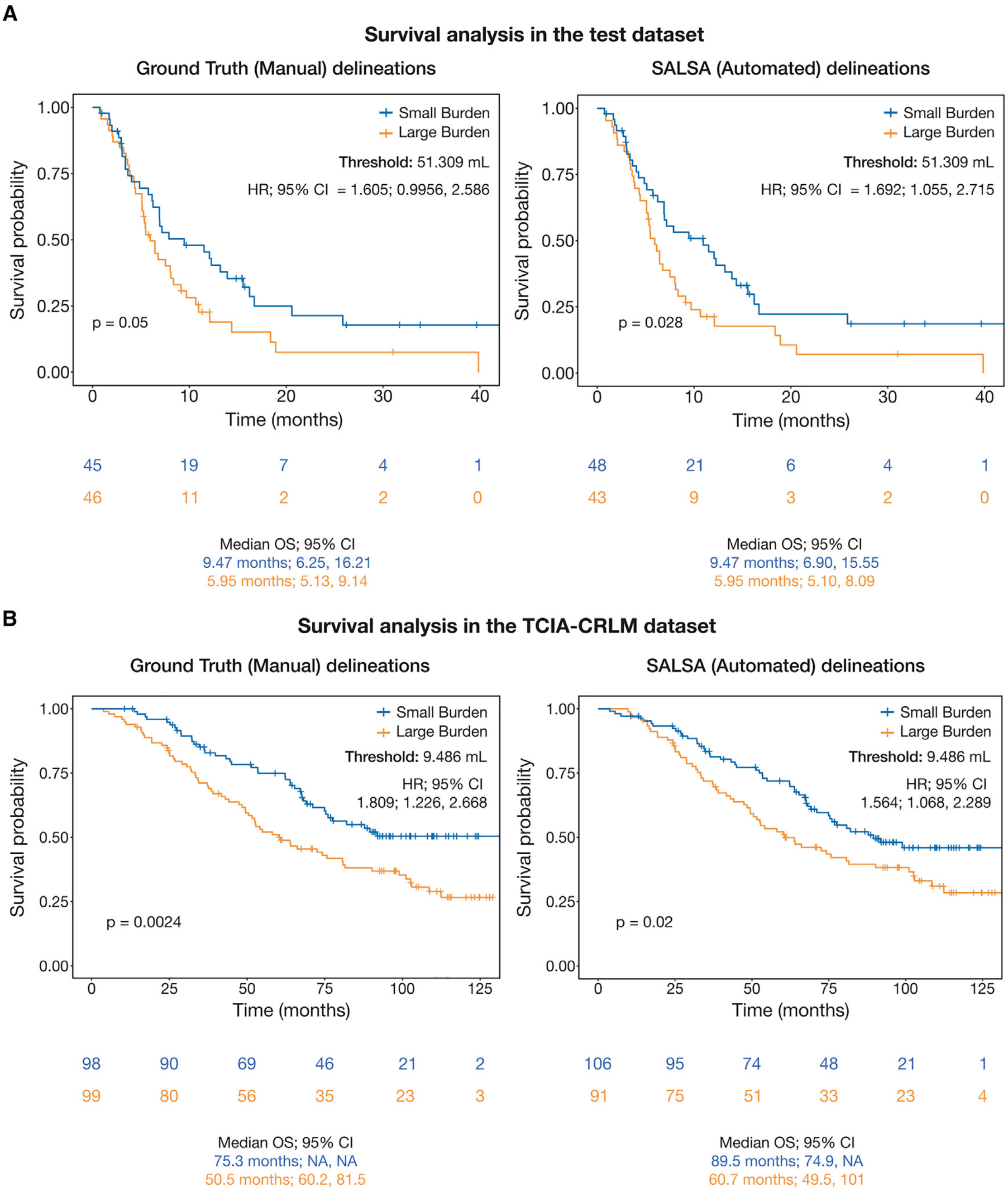


Figure 5. Total tumor volume quantification as a prognostic biomarker

Kaplan-Meier curves and log rank test results for overall survival are shown for 141 patients in the test set (A) and 197 patients from the TCIA-CRLM dataset in the external validation cohort (B). Patients were grouped by thresholding total liver tumor volume, using both ground truth (left) and SALSA-generated volumes (right), at the median value for each cohort, demonstrating SALSA's potential for biomarker research.

In conclusion, SALSA achieves precise and automated identification and delineation of liver cancer on CT images, facilitating more accurate quantification of tumor burden, a critical factor in cancer prognosis and management, with no prior manual prompt requirements. Our validation across several test and external cohorts highlights SALSA's effectiveness and reliability, matching, and often surpassing, the accuracy of expert radiologists. The inclusion of an open-source application further underscores its potential for widespread adoption in clinical routines. SALSA holds the promise of improving cancer care, enhancing patient outcomes, and increasing the efficiency of healthcare systems by offering more rapid, consistent, and reliable data for clinical trials and decision-making in clinical practice.

Limitations of the study

We acknowledge certain limitations in our study. The datasets considered as ground truth were annotated by expert radiologists, but with only one rater per scan, potentially introducing label bias. In our inter- and intra-radiologist variability study, we demonstrate that our radiologists' segmentations are subject to intrinsic manual variability, which is inevitably inherited by our model. Although this variability represents a limitation that could be mitigated by using consensus-based annotations, we chose to retain the original annotations as they were produced in the original studies and clinical routine to maintain a real-world context.

This potential limitation may be more pronounced in the segmentation of small lesions, where inter- and intra-observer variability tends to be greater. Furthermore, tumor segmentation adhered to RECIST guidelines, which stipulate that lesions under 1 cm in diameter are not segmented as they are considered non-measurable. This limitation becomes apparent when evaluating SALSA on the LiTS dataset,[12] which includes smaller liver lesions (median lesion volume of 0.54 mL, corresponding to a radius smaller than 0.5 cm, under sphericity assumption). In this cohort, SALSA's accuracy in detecting the so-called ground truth liver lesions decreases. However, considering that such small liver lesions are often regarded as indeterminate and non-measurable by RECIST, this raises questions about the true malignant nature and clinical relevance of all these small tumors.

Moreover, the vast majority of the lesions in the development cohort were metastatic and hypodense (mean lesion density interquartile range ranging 55.04–84.21 HU) compared to the background of the non-pathological liver tissue, which markedly benefited SALSA's performance in identifying hypodense lesions. Conversely, while hyperdense lesions, present in both the test and external validation datasets, led to poorer performance, no similar bias was observed for primary liver cancers. While acknowledging the limitations related to the manual annotations and variability in tumor types, densities and size, and the potential for improving SALSA's performance in small and hyperdense liver tumors, the development of SALSA offers a robust solution for liver tumor detection and volume quantification.

RESOURCE AVAILABILITY

Lead contact

All inquiries for further information regarding this work should be directed to and will be fulfilled by the lead contact, Raquel Perez-Lopez, Vall d'Hebron Institute of Oncology (VHIO), 11–13 Saturnino Calleja, 08035 Barcelona, Spain (rperez@vhio.net).

Materials availability

No materials such as reagents or other products were generated in this study.

Data and code availability

- The datasets used in this study from the Vall d'Hebron University Hospital and other collaborating centers are not publicly available due to patient privacy concerns. To request access to the data, please contact the lead contact, who will connect you with the responsible researcher at the corresponding center. Data will be accessible only if the Ethics Committee of each center where the data were collected grants permission. Therefore, the requester must describe the project for which data access is requested, detailing the objectives and data management plan. Data access will be considered for research purposes and non-commercial use only. To ensure patient privacy, access to personally identifiable information or sensitive clinical information (including medical histories) will not be provided, and requests for data access must rigorously adhere to the consent agreements established with study participants. Additional terms and conditions for accessing data by collaborating institutions may apply, as defined by the institutional Ethics Committee.
- The datasets extracted from the liver (LiTS dataset) and hepatic vessels (MSD dataset) tasks from the Medical Segmentation Decathlon (https://doi.org/10.48550/arXiv.1902.09063), can all be accessed from the Medical Segmentation Decathlon website (http://medicaldecathlon.com/).
- The TCIA-CRLM dataset, including the CT scans and the annotated liver tumors, is publicly available and can be accessed through (https://doi.org/10.7937/QXK2-QG03).
- The TCIA-HCC-TACE-Seg dataset, including the CT scans and the annotated liver tumors, is publicly available and can be accessed through (https://doi.org/10.7937/TCIA.5FNA-0924). The references and links for the datasets are listed in the key resources table.
- All the original codes and the trained models generated in this study are available under the folder "models" both in a repository at Zenodo (https://doi.org/10.5281/zenodo.14644657) and in a mirroring repository at GitHub (https://github.com/radiomicsgroup/liver-SALSA), publicly accessible as of the date of publication. The DOI and link are listed in the key resources table. Instructions for use can be found on the GitHub readme file.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

R.P.-L. is supported by "la Caixa" Foundation, a CRIS Cancer Foundation Talent Award (TALENT19-05), the Fero Foundation, the Instituto de Salud Carlos III-Investigación en Salud (PI21/01019), the Asociación Española Contra el Cancer (AECC) (PRYCO211023SERR), and the Agency for Management of University and Research Grants of Catalonia (AGAUR) (2023PROD00178).

AUTHOR CONTRIBUTIONS

Conceptualization, R.P.-L.; methodology, M.B.-M., A.M.M., M.L., and R.P.-L.; software, C.M. and A.M.M.; validation, D.L., L.M.A., N.S., C.V., J.M., J.H., A.G.-A., F.S., J.C., E.E., R.D., and E.G.; formal analysis, M.B.-M., A.M.M., and M.L.; investigation, M.B.-M., A.M.M., M.L., and R.P.-L.; resources, D.L., L.M.A., N.S., C.V., J.M., J.H., A.G.-A., F.S., J.C., E.E., R.D., and E.G.; data curation, C.Z. and C.V.; writing – original draft, M.B.-M., A.M.M., and M.L.; writing – review and editing, D.L., L.M.A., N.S., J.M., J.H., A.G.-A., F.S., J.C., E.E., R.D., E.G., and R.P.-L.; funding acquisition, R.P.-L.; supervision, R.P.-L.

DECLARATION OF INTERESTS

R.P.-L. declares research funding by AstraZeneca and Roche; she participates in the steering committee of a clinical trial sponsored by Roche, not related to this work.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author(s) used ChatGPT 4 in order to assist in reviewing and improving the grammar and writing fluency of this manuscript. The content and scientific integrity of the paper were solely developed by the authors. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Study population
- METHOD DETAILS
  - ○ Image processing
  - ○ Non-liver tissue masking and CT scan cropping
  - ○ Modeling
  - ○ Evaluation metrics
  - ○ Lesion association
  - ○ Detection evaluation
  - ○ Delineation evaluation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Impact of tumor size and slice thickness on SALSA's segmentation performance
  - ○ Inter- and intra-radiologist variability study
  - ○ Prognostic value of tumor burden study
  - ○ Radiologist feedback on the preferred delineation mask: SALSA vs. ground truth

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2025.102032.

Received: July 5, 2024
Revised: December 7, 2024
Accepted: February 24, 2025
Published: March 20, 2025

REFERENCES

1. Tumeh, P.C., Hellmann, M.D., Hamid, O., Tsai, K.K., Loo, K.L., Gubens, M.A., Rosenblum, M., Harview, C.L., Taube, J.M., Handley, N., et al. (2017). Liver Metastasis and Treatment Outcome with Anti-PD-1 Monoclonal Antibody in Patients with Melanoma and NSCLC. Cancer Immunol. Res. *5*, 417–424.

2. Tsilimigras, D.I., Brodt, P., Clavien, P.-A., Muschel, R.J., D'Angelica, M.I., Endo, I., Parks, R.W., Doyle, M., de Santibañes, E., and Pawlik, T.M. (2021). Liver metastases. Nat. Rev. Dis. Primers *7*, 27.

3. Siriwardena, A.K., Mason, J.M., Mullamitha, S., Hancock, H.C., and Jegatheeswaran, S. (2014). Management of colorectal cancer presenting with synchronous liver metastases. Nat. Rev. Clin. Oncol. *11*, 446–459.

4. Yoon, S.H., Kim, K.W., Goo, J.M., Kim, D.-W., and Hahn, S. (2016). Observer variability in RECIST-based tumour burden measurements: a meta-analysis. Eur. J. Cancer *53*, 5–15.

5. Krasovitsky, M., Lee, Y.C., Sim, H.-W., Chawla, T., Moore, H., Moses, D., Baker, L., Mandel, C., Kielar, A., Hartery, A., et al. (2022). Interobserver and intraobserver variability of RECIST assessment in ovarian cancer. Int. J. Gynecol. Cancer *32*, 656–661.

6. Iannessi, A., and Beaumont, H. (2023). Breaking down the RECIST 1.1 double read variability in lung trials: What do baseline assessments tell us? Front. Oncol. *13*, 988784.

7. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer *45*, 228–247.

8. Schwartz, L.H., Curran, S., Trocola, R., Randazzo, J., Ilson, D., Kelsen, D., and Shah, M. (2007). Volumetric 3D CT analysis - an early predictor of response to therapy. J. Clin. Orthod. *25*, 4576.

9. Hayes, S.A., Pietanza, M.C., O'Driscoll, D., Zheng, J., Moskowitz, C.S., Kris, M.G., and Ginsberg, M.S. (2016). Comparison of CT volumetric measurement with RECIST response in patients with lung cancer. Eur. J. Radiol. *85*, 524–533.

10. Sohaib, S.A., Turner, B., Hanson, J.A., Farquharson, M., Oliver, R.T., and Reznek, R.H. (2000). CT assessment of tumour response to treatment: comparison of linear, cross-sectional and volumetric measures of tumour size. Br. J. Radiol. *73*, 1178–1184.

11. Wesdorp, N.J., Zeeuw, J.M., Postma, S.C.J., Roor, J., van Waesberghe, J.H.T.M., van den Bergh, J.E., Nota, I.M., Moos, S., Kemna, R., Vadakkumpadan, F., et al. (2023). Deep learning models for automatic tumor segmentation and total tumor volume assessment in patients with colorectal liver metastases. Eur. Radiol. Exp. *7*, 75.

12. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al. (2023). The Liver Tumor Segmentation Benchmark (LiTS). Med. Image Anal. *84*, 102680.

13. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al. (2022). The Medical Segmentation Decathlon. Nat. Commun. *13*, 4128.

14. Simpson, A.L., Peoples, J., Creasy, J.M., Fichtinger, G., Gangai, N., Lasso, A., Keshava Murthy, K.N., Shia, J., D'Angelica, M.I., and Do, R.K.G. (2023). Preoperative CT and survival data for patients undergoing resection of Colorectal Liver Metastases (Colorectal-Liver-Metastases). Sci. Data *11*, 172. https://doi.org/10.7937/QXK2-QG03.

15. Moawad, A.W., Fuentes, D., Morshid, A., Khalaf, A.M., Elmohr, M.M., Abusaif, A., Hazle, J.D., Kaseb, A.O., Hassan, M., Mahvash, A., et al. (2021). Multimodality annotated HCC cases with and without advanced imaging segmentation. Sci. Data *10*, 33. https://doi.org/10.7937/TCIA.5FNA-0924.

16. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Lecture Notes in Computer Science Lecture notes in computer science (Springer International Publishing), pp. 234–241.

17. Ma, J. (2021). Cutting-edge 3D Medical Image Segmentation Methods in 2020: Are Happy Families All Alike?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2101.00232.

18. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., and Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.04306.

19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.

20. Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al. (2023). TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol. Artif. Intell. *5*, e230024.

21. C. Brett, M. Markiewicz, C.J. Hanke, M. Côté, M.-A. Cipollini, B. McCarthy, Paul J., D. Cheng, C.P. Halchenko, Y.O. Cottaar, et al., India nipy/nibabel: 5.0.0 https://doi.org/10.5281/zenodo.7516526.

22. Bell, D., and Greenway, K. (2015). Hounsfield unit. Radiopaedia.org. https://doi.org/10.53347/rid-38181.

23. Creators MONAI Consortium MONAI: Medical Open Network for AI .(2023) https://doi.org/10.5281/zenodo.8436376.

24. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

25. Creators Silversmith, W. (2021). cc3d: Connected Components on Multilabel 3D & 2D Images (Zenodo). https://doi.org/10.5281/zenodo.5719536.

26. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1505.04597.

27. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods *18*, 203–211.

28. Ye, J., Wang, H., Huang, Z., Deng, Z., Su, Y., Tu, C., Wu, Q., Yang, Y., Wei, M., Niu, J., et al. (2022). Exploring Vanilla U-Net for Lesion Segmentation from Whole-body FDG-PET/CT Scans. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.07490.

29. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

30. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support. *2017*, 240–248.

31. Dantzig, G.B., and Fulkerson, D.R. (1957). On the max-flow min-cut theorem of networks. In Linear Inequalities and Related Systems (Princeton University Press), pp. 215–222. (AM-38).

32. Müller, D., Soto-Rey, I., and Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. BMC Res. Notes *15*, 210.

33. Chair, D.S., University of Würzburg, IPA Group, Berlin, H.-U. zu, KDE Group, University of Kassel, L3S Research Center, and (Germany), H (1979). Information Retrieval (Butterworth-Heinemann). https://www.bibsonomy.org/bibtex/0edccdac9af024f458911b82f61686ab.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
LiTS Challenge training dataset	Bilic et al.[12]	https://doi.org/10.48550/arXiv.1901.04056
TCIA-CRLM dataset	Simpson et al.[14]	https://doi.org/10.7937/QXK2-QG03
MSD Hepatic Vessels dataset	Antonelli et al.[13]	https://doi.org/10.1038/s41467-022-30695-9
TCIA-HCC-TACE-Seg	Moawad et al.[15]	https://doi.org/10.7937/TCIA.5FNA-0924
Trained model for SALSA (‘‘models’’ folder)	This paper; Zenodo; GitHub	https://doi.org/10.5281/zenodo.14644657, https://github.com/radiomicsgroup/liver-SALSA
Software and algorithms		
Code for SALSA	This paper; Zenodo; GitHub	https://doi.org/10.5281/zenodo.14644657, https://github.com/radiomicsgroup/liver-SALSA
Python v3	Python Software Foundation	https://www.python.org/
TotalSegmentator	Wasserthal et al.[20]	https://github.com/wasserth/TotalSegmentator

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study population

This retrospective study included data from cancer patients treated at Vall d’Hebron University Hospital and other centers collaborating with the Vall d’Hebron Institute of Oncology (VHIO). For external validation purposes, data from four open-access, public datasets were also collected: the Liver Tumor Segmentation (LiTS) challenge training dataset,[12] the Medical Segmentation Decathlon (MSD)-Hepatic Vessels dataset,[13] The Cancer Imaging Archive (TCIA) Colorectal Liver Metastases (CRLM) dataset.[14] The total sample size of the study consisted of 1598 CT scans, accounting for 4908 tumors from 1306 patients.

Anonymized CT scans were obtained retrospectively from digital clinical records at the Vall d’Hebron Institute of Oncology following approval by VHIO’s Ethics Committee (PR(AG)70/2018, PR(AG)261/2019) and the waiver of informed consent. An anonymization process, which involved removing all metadata linking the CT scans to the corresponding patients from the DICOM files, was applied to ensure patient privacy.

The set of scans from the VHIO cohorts was divided into a training cohort, consisting of 726 scans, and an in-domain test cohort, comprising 25% of the cases and totaling 159 scans. The split was defined by identifying those cohorts coming from clinical trials that were still ongoing at the beginning of the study. As recruitment was not completed or not all the recruited patients’ scans had been segmented, these cohorts could not be used for training purposes and were defined as the test ones. For this VHIO cohort, the median overall survival (OS) was 12.16 months, and 60% of the patients had an ECOG grade of 0 or 1. Regarding previous treatments, there was considerable heterogeneity in the specific drugs administered, and information was missing for 26% of the patients (155 individuals). However, when considering the predominant drug mechanisms, immune checkpoint inhibitors were the most common treatment: a combination of PD-1/anti–PD-1 therapy was administered to 261 patients (44%), while PD-L1 immunotherapy was used in 125 patients (22.6%), together covering the vast majority of the cohort.

The four open-source datasets were reserved for external validation.

METHOD DETAILS

Image processing

From the development cohort, CT images were evaluated by experienced radiologists using 3D Slicer software, and all liver tumors larger than 1 cm in diameter were delineated. Tumor segmentations were already available for all open-access validation datasets, forming a comprehensive and independent dataset of scans and ground truth segmentations. Both the scans and the masks were converted to NIfTI format to enhance computational performance and ensure format standardization.

Before applying additional transformations to the CT scans, the TotalSegmentator tool[20] was used to extract a binary mask labeling the liver volume. This mask was preserved for subsequent masking and cropping of the scans.

The Python NiBabel library[21] was used to extract a 3D array of voxel intensities in Hounsfield Units (HU) from each CT scan,[22] along with the corresponding voxel sizes, from the NIfTI files. This data was then used to standardize the voxel size of the scans to 1 × 1 × 1 mm^3 using the Spacingd function available in the MONAI Transforms package.[23] This transformation was applied in parallel to the scans, with bilinear interpolation, and to the binary masks, using nearest neighbor interpolation.

Non-liver tissue masking and CT scan cropping

To reduce the variability of features that the models need to learn, all non-liver tissue was masked to a constant value. To optimize computational efficiency, the scans were then cropped around the remaining unmasked signal.

For this purpose, the TotalSegmentator v2 tool[20] was applied to the scans before resizing, generating a binary mask labeling the liver. This mask, along with the lesion segmentation mask and the scan, was resized using nearest neighbor interpolation.

Before using them for masking, the liver masks were dilated using the SciPy Image binary_dilation function.[24] The dilation process was iteratively applied for 15 cycles to ensure that the liver borders were not inadvertently excluded from the data. Subsequently, the dilated mask underwent pre-processing to ensure it consisted of a single connected component without any internal holes. This step was implemented using the connected-components-3d Python package,[25] retaining only the largest component from both the original and inverted masks.

Once pre-processed, the liver mask was used to set the values of all non-liver voxels to −1024 Hounsfield Units (HU), the typical minimum value in the CT scan range. The resulting array was cropped to remove all regions devoid of liver, retaining only the bounding box surrounding the liver mask with an additional 10-voxel margin. To ensure compatibility with all applied models and reduce computational costs during training, the cropping process concluded with a padding operation to ensure that each axis had a length that was a multiple of 64.

Modeling

Neural network architectures

For our study, we selected three frameworks, resulting in a total of seven different architectures. The first choice was a vanilla 3D U-Net[26] obtained from the nnU-Net package,[27] serving as our benchmark architecture.[27] Additionally, we opted for a TransUNet 2D-transformer[18] architecture featuring a field of view of 256x256 pixels, 4 heads, 4 output channels, 8 blocks, a patch dimension of 16 and an MLP dimension of 512. Finally, our third model was an nnU-Net ensemble,[27] which incorporated five distinct, independently-trained architectures for comprehensive comparison.

The vanilla U-Net was initialized with weights from a reference project focused on automatic segmentation[28] combining FDG-PET and CT images. To ensure compatibility with our CT-based study, the FDG-PET channel was deleted, retaining only the CT input channel. TransUNet was initialized using random weights as no available pretrained models were found.

Model training

Both the vanilla U-Net and TransUNet models were trained for over 800 epochs until convergence was observed. The AdamW optimizer[29] was used with an initial learning rate of 1e−4 and a weight decay of 1e−4. To adapt the learning rate dynamically, a Step Learning Rate (LR) scheduler was used, reducing the learning rate by a factor of 1/sqrt(10) every 140 epochs. A combination of Binary Cross Entropy (BCE) loss and Dice loss[30] was used for backpropagation.

The nnU-Net was trained following the default configuration of the authors' implementation, using BCE and Dice loss function, an initial learning rate of 1e−2 with a Poly LR scheduler, 100 epochs per fold and SGD optimizer with Nesterov momentum.

Evaluation metrics

The model performance was evaluated for both liver tumor detection and delineation. Moreover, these evaluations were conducted at both the patient and tumor levels. For per-tumor metrics, it was necessary to establish associations between predicted and ground truth tumors, accounting for scenarios where each predicted tumor could be linked to none, one, or multiple ground truth tumors, and vice versa. This tumor association challenge was addressed by treating it as a graph partitioning problem.

Lesion association

Lesion association was accomplished through the utilization of the max-flow/min-cut graph theory algorithm.[31] To compare segmentations, we transformed them into a graph for each pair of predicted and ground truth masks, where nodes represented segmented lesions and edges represented their overlap.

Initially, all lesions were identified using a 3D connected components algorithm.[25] Subsequently, all components from each segmentation were paired with those from the other. For each overlapping pair, the Jaccard Index[32] was computed to measure their similarity and degree of overlap, constituting the edges of a weighted graph.

Each graph was first analyzed to identify all sub-graphs that constituted isolated components. Components consisting of a single isolated vertex from a ground truth lesion were categorized as undetected lesions. Conversely, components with a single vertex from the segmentation were classified as incorrect detections. Components containing one lesion from each segmentation were classified as true positives. Similarly, those containing only a single edge from either of the two segmentations, with several from the other, were classified as instances where a single lesion was detected as multiple or vice versa. In such cases, the ground truth lesions involved were considered true positives.

Finally, components containing more than one lesion from each segmentation were regarded as complex sub-graphs that needed to be divided into simpler components as described earlier. To achieve this, the well-known max-flow/min-cut graph theory algorithm[31] was used, designating two random vertices, chosen from the segmentation with fewer lesions involved in the component, as the source and sink vertices. For those components where both segmentations had more than two lesions involved, the described process was applied iteratively until the component was properly divided.

Once all the components had been processed, the sets of lesions involved from each segmentation were combined into a single lesion per segmentation. Hence, considering only 0-to-1, 1-to-0 and 1-to-1 correspondences thereafter.

Detection evaluation

For the evaluation of tumor detection performance confusion matrices, precision, recall and F1 Score[33] metrics were employed. In the per-scan approach, the ground truth data was compared with the model predictions, classifying each case as true positive, false positive, true negative, or false negative based on the presence or absence of lesions in both the ground truth and the model prediction, regardless of the lesion location and individual detection.

To perform a per-lesion evaluation, we utilized the graph-based previously described lesion association method. This approach allowed us to associate each lesion from the predicted segmentations with either nothing, a tumor, or a set of tumors from the ground truth segmentation and vice-versa. Each ground truth tumor was counted as a true positive or false negative depending on whether it was associated with a predicted lesion or a set of predicted lesions. Conversely, all model-predicted lesions that were not associated with any ground truth lesion were considered false positives.

Delineation evaluation

The evaluation of liver tumor delineation performance was conducted using multiple, well-established metrics, both on a per-patient and per-lesion basis.

On a per-scan basis, the Dice Similarity Coefficient (DSC) and Jaccard Index (JI) were applied to compare the predicted and ground truth masks. These metrics were calculated by applying the corresponding functions to the binary delineated masks, allowing for the assessment of the overall tumor delineation quality for each scan.

To evaluate the delineation quality at the per-lesion level, the graph-based previously described lesion association method. Subsequently, DSC and JI were computed for each pair of associated ground truth and predicted lesions or sets of lesions. This approach enabled the accurate gauging of delineation precision for individual lesions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests were performed to compare: (i) patient age, number of tumors, individual tumor volume and total tumor volume distribution between development and test sets of each tumor type (unpaired two-sample Wilcoxon test, after a Shapiro-Wilk normality test); (ii) patient gender distribution and liver cancer origin, either primary or metastatic, between development and test sets (Chi-squared test). Relations are shown in Table 1. Statistical significance was considered for $p < 0.05$. Variable names and abbreviations are explained in each figure or table legend, numbers in tables are explained in the heading of each column.

Impact of tumor size and slice thickness on SALSA's segmentation performance

To assess the impact of the tumor volume and the slice thickness, the ground truth volumes were divided in groups and significance for distribution equality was calculated by performing independent two-sample t-tests, both among the produced groups and the test and external validation cohorts. Later, the *p*-values were adjusted using the Bonferroni correction. The results of these statistical tests are summarized in Figures 2E and S2B, and Tables S6 and S7.

Inter- and intra-radiologist variability study

An additional sub-study was designed to evaluate intra- and inter-radiologist tumor delineation variability, using 25 randomly selected cases from the development cohort assessed by three experienced radiologists. Radiologist 1, who has over 15 years of experience in oncological imaging, was responsible for delineating the ground truth segmentations. This radiologist evaluated the CT scans twice, with an interval of more than three months between assessments, to facilitate intra-radiologist variability studies. Radiologists 2 and 3, with 15 and 5 years of experience respectively, who were blinded to the ground truth delineations, also evaluated the CT scans to allow for inter-radiologist evaluation.

The radiologists independently segmented the cases without access to the ground truth segmentations. We then computed the previously described evaluation metrics by comparing the resulting segmentations with the ground truth masks from our dataset. The results of this evaluation were compared with those of the selected SALSA model to assess both human and machine performance.

Prognostic value of tumor burden study

Clinical outcome data from the development, test, and TCIA datasets were collected in the form of overall survival (OS). This data allowed us to explore the potential of tumor burden as a prognostic biomarker in cancer patients.

Tumor volume was categorized into two classes using the median value obtained from all the scans in the training cohort as a threshold. These two groups were then used to explore differences in patient survival through Kaplan-Meier charts. Log rank tests were conducted to assess differences in mean OS. This procedure was applied to both the tumor burden values extracted from the model-predicted segmentations and the ground truth segmentations.

To evaluate the effectiveness of both the radiologist and model-powered criteria, the Kaplan-Meier charts were compared by examining the statistical test *p*-values associated with each.

Radiologist feedback on the preferred delineation mask: SALSA vs. ground truth

To collect feedback from expert radiologists on the SALSA segmentations, an easy-to-use web interface app was developed using the Streamlit Python library for blinded delineation comparison. For each scan, both the ground truth and the model-predicted delineation masks were displayed on the liver scan, using random colors for distinction.

Three expert radiologists were asked to evaluate a subset of cases from both the test and external validation sets. They reported which delineation masks they considered more accurate or indicated that both masks were equally good if no specific preference was evident. The frequency of each of the three choices, including opting for no preference, was then recorded and a descriptive analysis of the preferences was performed (Figure 4C).