Check for updates

Received: 28 December 2023 | Revised: 19 March 2024 | Accepted: 2 April 2024

DOI: 10.1002/mp.17076

RESEARCH ARTICLE

MEDICAL PHYSICS

Fully automated explainable abdominal CT contrast media phase classification using organ segmentation and machine learning

Yazdan Salimi[1] | **Zahra Mansouri**[1] | **Ghasem Hajianfar**[1] | **Amirhossein Sanaat**[1] | **Isaac Shiri**[1,2] | **Habib Zaidi**[1,3,4,5]

[1]Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland

[2]Department of Cardiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[3]Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

[4]Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

[5]University Research and Innovation Center, Óbuda University, Budapest, Hungary

Correspondence
Habib Zaidi, Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva, Switzerland.
Email: habib.zaidi@hcuge.ch

Funding information
Euratom research and training programme 2019–2020 Sinfonia project, Grant/Award Number: 945196

Abstract

Background: Contrast-enhanced computed tomography (CECT) provides much more information compared to non-enhanced CT images, especially for the differentiation of malignancies, such as liver carcinomas. Contrast media injection phase information is usually missing on public datasets and not standardized in the clinic even in the same region and language. This is a barrier to effective use of available CECT images in clinical research.

Purpose: The aim of this study is to detect contrast media injection phase from CT images by means of organ segmentation and machine learning algorithms.

Methods: A total number of 2509 CT images split into four subsets of non-contrast (class #0), arterial (class #1), venous (class #2), and delayed (class #3) after contrast media injection were collected from two CT scanners. Seven organs including the liver, spleen, heart, kidneys, lungs, urinary bladder, and aorta along with body contour masks were generated by pre-trained deep learning algorithms. Subsequently, five first-order statistical features including average, standard deviation, 10, 50, and 90 percentiles extracted from the above-mentioned masks were fed to machine learning models after feature selection and reduction to classify the CT images in one of four above mentioned classes. A 10-fold data split strategy was followed. The performance of our methodology was evaluated in terms of classification accuracy metrics.

Results: The best performance was achieved by Boruta feature selection and RF model with average area under the curve of more than 0.999 and accuracy of 0.9936 averaged over four classes and 10 folds. Boruta feature selection selected all predictor features. The lowest classification was observed for class #2 (0.9888), which is already an excellent result. In the 10-fold strategy, only 33 cases from 2509 cases (∼1.4%) were misclassified. The performance over all folds was consistent.

Conclusions: We developed a fast, accurate, reliable, and explainable methodology to classify contrast media phases which may be useful in data curation and annotation in big online datasets or local datasets with non-standard or no series description. Our model containing two steps of deep learning and machine learning may help to exploit available datasets more effectively.

KEYWORDS
contrast-enhanced CT, data curation, deep learning, machine learning, segmentation

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2024 The Authors. *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

1 | INTRODUCTION

Computed tomography (CT), as a versatile non-invasive imaging modality, plays a significant role in clinical diagnosis and follow-up of patients presenting with a wide range of pathologies. The administration of contrast media (mostly positive contrast media) improves the diagnostic value and accuracy of this imaging procedure by enhancing the contrast in tissues.[1] This enhanced contrast is useful in the differentiation of multiple pathologic situations reflecting improved discrimination between different tissue types.[2] The most common method of contrast media administration for contrast enhanced CT (CE-CT) is intravenous injection using radiopaque materials mostly containing iodine. For instance, the management of patients treated with selective internal radiation therapy (SIRT) for the treatment of liver malignancies requires CE-CT for the delineation of the tumoral tissues, perfused liver lobe and organs at risk towards personalized dosimetry.[3]

Artificial intelligence (AI) and particularly deep learning has shown very promising performance in multiple tasks, including image segmentation,[4–6] image generation,[7,8] dosimetry,[9–11] and classification.[12,13] However, the number of clean and reliable data available is still the bottle neck for generalizability and robustness of deep learning models. In this regard, data and model sharing between imaging centers may help to collect diverse training datasets. However, even for federated learning strategies, the data should be in the same format to be able to train a model among different centers.[14] Although the DICOM standard is capable of storing information about the injection phase, privacy issues require that the image data has to be anonymized by removing private tags, which might include removing the injection phase information. We evaluated the number of online available datasets and came to the conclusion that most datasets in DICOM format did not have any straightforward information about the contrast media injection phase, while the image only data formats, such as NIFTI, do not include any related information; Even within the same hospital when using different scanners, the naming of study/series/contrast injection phase are commonly not consistent, even in the same department.[15] For example, in our dataset, terms such as "tardif, Tard, Tardivo, Tardive, Delayed" were used for the delayed phase whereas "SPC, Plain, NoContrast, SANS IV, NATIF, native" were used for unenhanced images. The formal written language between different cities and imaging centers is another barrier in effective data sharing between centers even in a single country. These limitations and lack of trustable information about the contrast media administrative phases underscore the need for developing an easily accessible solution based on minimum shared data, which is the image matrix only.

Philbrick et al.[16] trained deep neural network classifiers to separate contrast enhanced phases in abdominal CTs for kidney pathologies indications. They attempted to explain what the network sees during training and inference by generating GradCam and saliency maps after each decision. They reported F1-scores between 0.781 and 0.999 for different contrast media phases. Zhou et al.[17] trained a deep neural network to separate four phases of liver CE-CT including non-contrast, arterial, venous, and delay. They used natural language processing to extract the ground truth data from DICOM tags and reported an average F1-score of 0.977 on a very large dataset. Tang et al.[18] used a generative adversarial network to classify the contrast phases in abdominal CE-CTs and achieved an accuracy of 0.93. They used 2D axial slices to train and evaluate the algorithm. Dao et al.[19] used random sampling and deep neural networks for automated classification of CE-CT images. They reported F1 score of 0.9209 on the internal test consisting of 358 scans using 2D slices as input to the network. The evaluation of their algorithm on external datasets reported F1 scores of 0.7679 and 0.8694 on two manually annotated online databases. Muhamedrahimov et al.[20] used CE-CT images and the time from injection as the ground truth to train a regression machine learning model to predict the time from injection. Then they used these times to classify images and reported an overall accuracy of 0.933 in classification. Rocha et al.[15] used machine learning to classify axial 2D slices from liver CE-CTs in four phases and reported an accuracy of more than 0.98. Reis et al.[21] independently adopted a strategy similar to the one used in our study and developed a machine learning method to classify contrast media phase based on few organs segmentation. They used online available datasets to train their network and reported F1 scores of 0.966, 0.789, 0.922, and 0.95 for four contrast phases consisting of non-contrast, arterial, venal, and delayed, respectively.

This study is an extension of previous preliminary work aimed to develop explainable, fast, and accurate methodology to classify the contrast media phase using imaging only data. Inspired by the logic human follows to understand the injection phase, which includes checking the enhancement ratio and the hounsfield unit (HU) in few organs, such as aorta and kidney, we attempted to use the very simple features extracted from the above-mentioned regions and machine learning algorithms to classify CT images.

2 | MATERIALS AND METHODS

This study included 2509 Thorax/Abdomen CT images acquired between 2011 and 2022 on scanners commercialized by two different vendors (Siemens Healthcare

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and Philips Medical Systems) in different imaging centers for evaluation of different liver pathologies, mainly for initial diagnosis and follow-up of liver malignancies, such as hepatocellular carcinoma (HCC). All images were acquired in Geneva University Hospital. CT images were acquired in four phases, including non-enhanced (W/O) (class 0, 683 images), arterial (class 1, 714 images), venous (class 2, 419 images), and delayed (class 3, 693 images) acquisition with the same scan range covering the abdominal area or more. The total number of patients included in this study is 581 patients. The pipeline of this study is showed in Figure 1. First, the technologists classified the images in one of the phases according to the DICOM series information header series description according to their local definitions. These classes for each 3D image were used as the ground truth for training and evaluation of our methodology. Meanwhile, seven organs including the liver, spleen, heart, kidneys, lungs, urinary bladder, and aorta plus body contour were segmented from CT images automatically using previously developed algorithm.[22] Second, five first-order features, including the average, standard deviation (SD), 10 percentile, median, and 90 percentiles were extracted using basic Python software-based image processing using SimpleITK (version 2.2) and Numpy (version 1.2) libraries. Overall, five features from seven organs (35 features) were extracted. These five features for seven organs (in total 35 features) were used to train a machine learning algorithm to perform classification task in a 10-fold data split strategy. Finally, the performance of the whole algorithm was evaluated in terms of F1 score, accuracy, specificity, sensitivity, precision, and area under the curve (AUC). Receiver operating characteristic (ROC) curves and confusion matrices were then generated for further evaluation.

2.1 | Feature selection and machine learning

Three feature selection methods, including Boruta,[23] Maximum Relevance Minimum Redundancy (MRMR),[24] and Recursive Feature Elimination (RFE)[25] were used to reduce the number of features. Boruta and RFE are wrapper-based whereas MRMR is filter-based algorithm. In Boruta and RFE, the number of selected features is determined by specific criteria, while in MRMR, it is based on the top 15 features. Subsequently, six different classifiers, including K-nearest neighbors (KNN),[26] eXtreme Gradient Boosting (XGB),[27] Decision Tree (DT),[28] Support Vector Machine (SVM),[29] Naive Bayes (NB),[30] and Random Forest (RF)[31] were used to classify images. In this study, we used nested cross-validation with outer and inner validation. We used 10-fold cross-validation for outer validation. In each repetition of outer validation, feature selection methods were performed on nine portions of data followed by training the classifier model. Hyperparameter optimization for each classifier was performed using inner 10-fold cross-validation scheme; the detailed parameters modified during optimization are summarized in Table S1. The trained models with the best hyperparameter were tested on one portion of dataset. This procedure was repeated 10 times. Finally, among these 3×6 feature selection/models, the best model based on performance was selected. For model evaluation, we used confusion matrix, AUC, F1_score, sensitivity, specificity, precision, and accuracy.

2.2 | Segmentation model

A previously developed and validated ResUNET-based model was used for CT image segmentation.[22] First, the body contour was generated on CT images via image processing algorithms and then this contour was used to crop the images to foreground. Then, the trained models were used to generate segmentation masks through inference on every 2D axial CT slice. Organ-specific post-processing was performed to refine the network output by implementing prior knowledge about the position of organs in the human body. Information from the liver and lung segmentations was used to remove false positive segmented voxels from the network output. For example, if the deep learning (DL) model assigns some voxels in the neck region to the liver, we used the lung segmentation boundaries to remove the false positive segmentations. Through this strategy, we implemented prior knowledge about the anatomical location of organs to improve the segmentation accuracy. The Dice coefficients (in %) for this segmentation model reported on external dataset were 96.98 ± 1.62, 94.68 ± 9.28, 91.95 ± 6.13, 94.13 ± 5.05, 97.63 ± 1.18, 83.78 ± 17.97, 93.55 ± 2.22, for the liver, spleen, heart, kidneys, lungs, urinary bladder, aorta, respectively. It should be noted that we expected an error in the segmentations as the Dice values deviated from the ideal value of 1. However, if the segmentation model missed parts of the aorta, liver, heart, or other organs, the average value and the simple features selected and used were not significantly affected. We used an average value in the aorta and other organs for partly segmented aorta and the average difference was less than 5%.

The five mentioned features including average, SD, 10 percentile, median, and 90 percentiles were calculated inside the eight segmentation masks including the liver, spleen, heart, kidneys, lungs, urinary bladder, aorta, and body contour and recorded for the next step.

3 | RESULTS

Table 1 summarizes the demographic information of the patients included in this study plus the

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

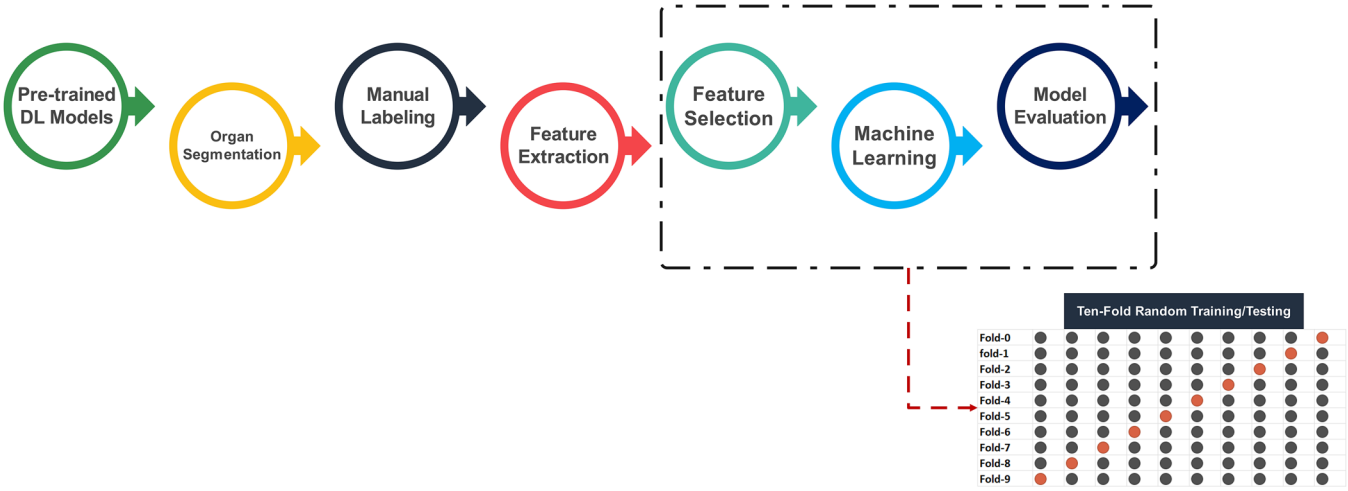

FIGURE 1 Flowchart of the steps followed in this study. The manual image labels were extracted according to the DICOM metadata.

TABLE 1 Demographic and acquisition parameters of patients included in the study protocol, kVp (Tube voltage).

Demographics		**Acquisition/Reconstruction parameters**	
Age (years)	**65.23 ± 11.36**	**Manufacturer**	**Philips, Siemens**
Date	2011 to 2022	kVp	70, 80, 90, 110, 130
Weight (Kg)	79.25 ± 16.64	Pitch factor	0.90 ± 0.25
Class #0	683	$CTDI_{vol}$ (mGy)	9.98 ± 10.14
Class #1	714	Slice thickness (mm)	2.35 ± 1.68
Class #2	419	Acquisition diameter (mm)	498 ± 43
Class #3	693	Reconstruction diameter (mm)	414 ± 84
total	2509	Average tube current (mA)	288.89 ± 137.93

acquisition/reconstruction CT parameters. Patients' age was 65.23 ± 11.36 years, and the average tube current was 288.8989 ± 137.9301 mA. This was a balanced distribution of data between classes where there were 683, 714, 419, and 693 images in classes 0 to 3, respectively.

3.1 | Segmentation

Figures 2 and 3 depict examples of visualization of segmented organs on axial images. The segmentation shows almost the same region in four phases. Please note the enhancement and differences in the segmented regions on Figures 2 and 3. The urinary bladder was not included in all images. Hence, we excluded this organ as a predictor in our model that is, six organs plus body contour (seven segmentation masks) were used to train the machine learning model. Figure 4 presents a 3D visualization of the segmented organs in different views. 3D rotating views showing the segmentation outcome are provided in supplementary material section.

3.2 | Selected model and feature selection

The best performance was achieved by Boruta feature selection and RF model. This feature selection method found all the expected important features, because the features were selected according to the specific task. Figure S1 shows the feature importances indicated by this FS method. The highest importance in Boruta feature selection was for aorta, cardiac, and spleen organs.

3.3 | Classification results average for the best model/feature selection

Table 2 summarizes the results of the classification model calculated in 10-fold data split strategy. F1-score was 0.9842 in average for the four classes, while the highest score was for non-enhanced (0.9978) and the lowest was for venous phase (0.9667). An AUC higher than 0.99 was achieved for all four classes.

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

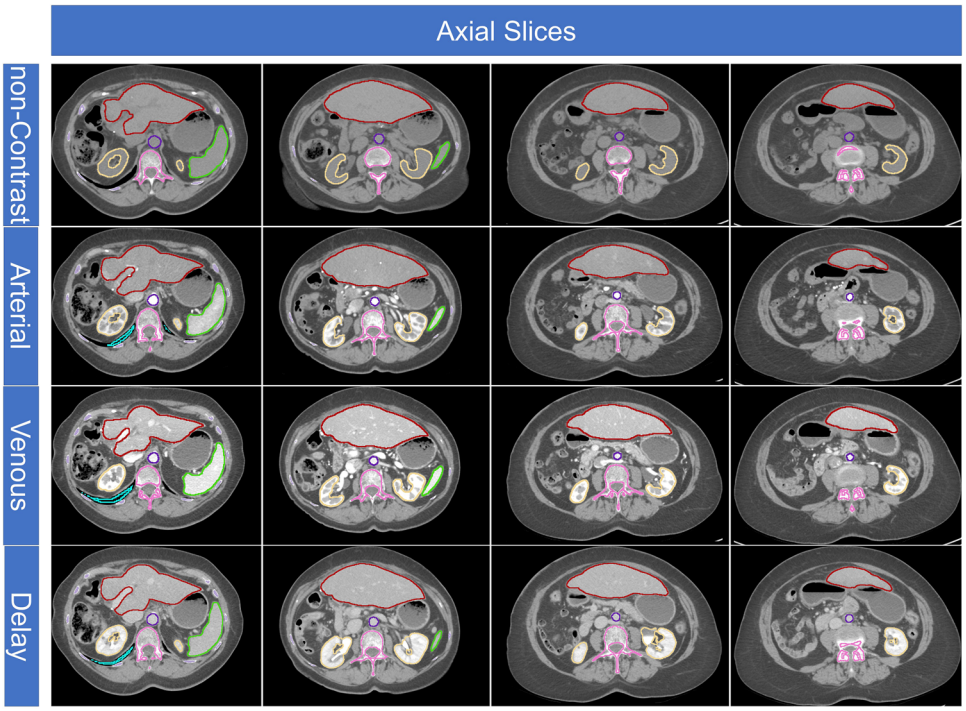

FIGURE 2 Axial CT slices of the same patient presenting with HCC in four phases including non-contrast (#0, top row), arterial (#1, second row), Venous (#2, third row) and delayed (#3, fourth row). The same line colors show the segmented organs. Please note the enhancement in the segmented regions. A window/level of 500/0 HU was used for all images to be comparable. Green: spleen, yellow: kidneys, dark red: liver, light blue: lungs, purple: aorta, pink: vertebrae (was not used in the model). CT, computed tomography; HCC, hepatocellular carcinoma.

TABLE 2 The classification performance of the best model for each of the four classes.

Class	F1_score	Sensitivity	Specificity	Precision	Accuracy	AUC
0 (W/O)	0.9978	0.9956	1.0000	1.0000	0.9988	0.9989
1 (Arterial)	0.9951	0.9916	0.9994	0.9986	0.9972	1.0000
2 (Vein)	0.9667	0.9690	0.9928	0.9644	0.9888	0.9983
3 (Delay)	0.9813	0.9856	0.9912	0.9771	0.9896	0.9991
Total classes	0.9852	0.9854	0.9959	0.9850	0.9936	0.9991

Abbreviation: AUC, area under the curve.

Figure 5 represents the normalized confusion matrix for all data included in 10-fold strategy. The lowest performance was achieved for the venous phase once again. Figure 6 shows the ROC curves for all classes demonstrating excellent performance.

3.4 | Classification results—10-folds

Table 3 shows the performance of our proposed model for each fold, averaged over all four classes. The results are consistent among folds which shows the robustness and reproducibility of our model. Figure S2 shows the normalized confusion matrix for each fold, demonstrating consistent performance.

Table 4 summarizes the performance metrics for different combinations of feature selection and models, indicating Boruta/RF selected in this work as the best model.

4 | DISCUSSION

Identifying the contrast injection phase in contrast-enhanced CT images is a crucial step in using online or local databases for various machine learning or modeling tasks. Unfortunately, the DICOM headers, and naming in metadata among imaging centers is not standardized, which is a hurdle in developing models based on large datasets. We proposed a novel method based on DL-based segmentation developed in our group and a RF machine learning model to classify CT images. Our results were promising, achieving an average AUC of 0.9991 and accuracy of 0.9936.

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

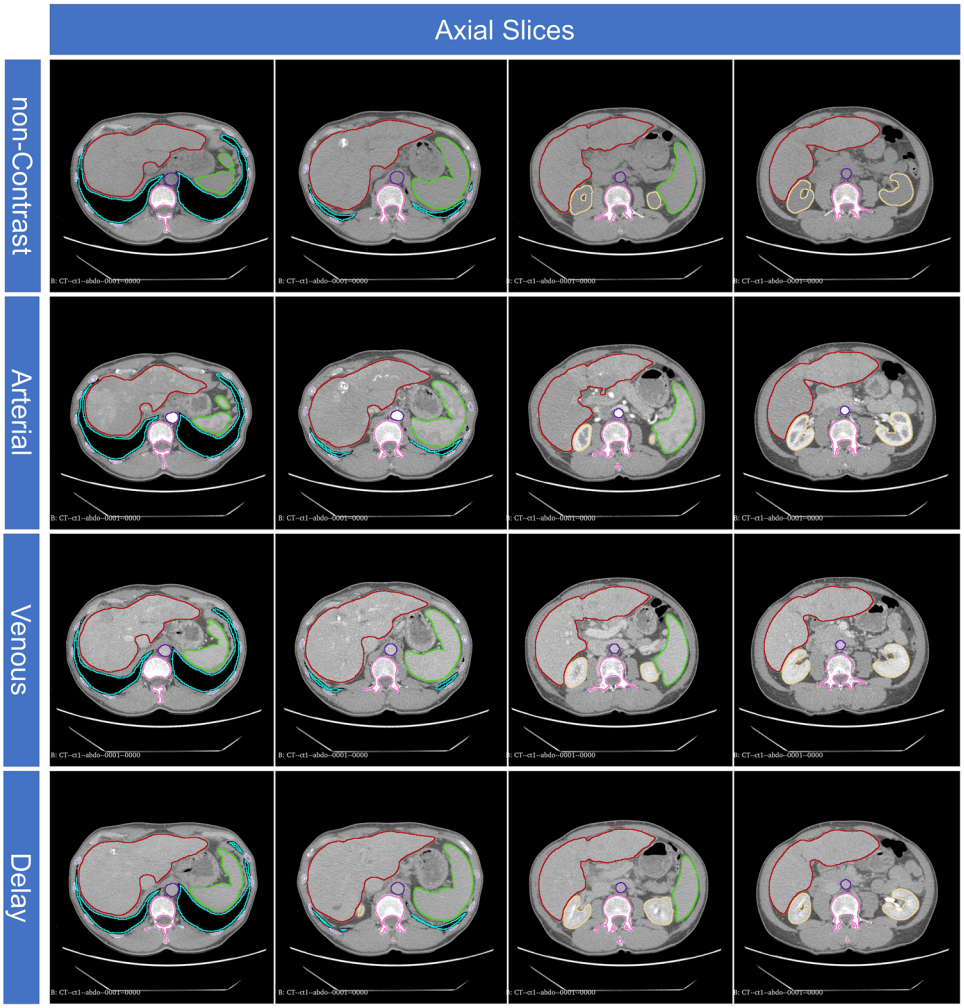

FIGURE 3 Axial CT slices for a patient referred for metastatic liver malignancies indication (same caption as Figure 2). The window/level is 500/0 HU and the difference in lung is not visible. Green: spleen, yellow: kidneys, dark red: liver, light blue: lungs, purple: aorta, pink: vertebrae (was not used in the model). CT, computed tomography.

TABLE 3 The classification performance separated by folds.

Fold	F1_score	sensitivity	specificity	precision	accuracy	AUC
0	0.9820	0.9831	0.9949	0.9810	0.9920	0.9988
1	0.9767	0.9821	0.9940	0.9734	0.9901	0.9998
2	0.9777	0.9799	0.9937	0.9759	0.9901	0.9973
3	0.9952	0.9940	0.9986	0.9964	0.9980	0.9999
4	0.9856	0.9846	0.9960	0.9869	0.9940	0.9998
5	0.9880	0.9892	0.9960	0.9870	0.9940	0.9995
6	0.9810	0.9812	0.9949	0.9811	0.9920	0.9975
7	0.9821	0.9808	0.9946	0.9835	0.9920	0.9997
8	0.9903	0.9881	0.9972	0.9930	0.9960	0.9999
9	0.9904	0.9904	0.9974	0.9904	0.9960	0.9998

Abbreviation: AUC, area under the curve.

10-fold data split with consistent performance over all folds demonstrated the robustness of our methodology on a fairly large database consisting of 2509 3D CT images. The overall error in our study is an accumulation of errors arising from the segmentation and machine learning classification tasks. However the whole pipeline is automated from scratch and showed excellent results. Our light-weight segmentation model with an average

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

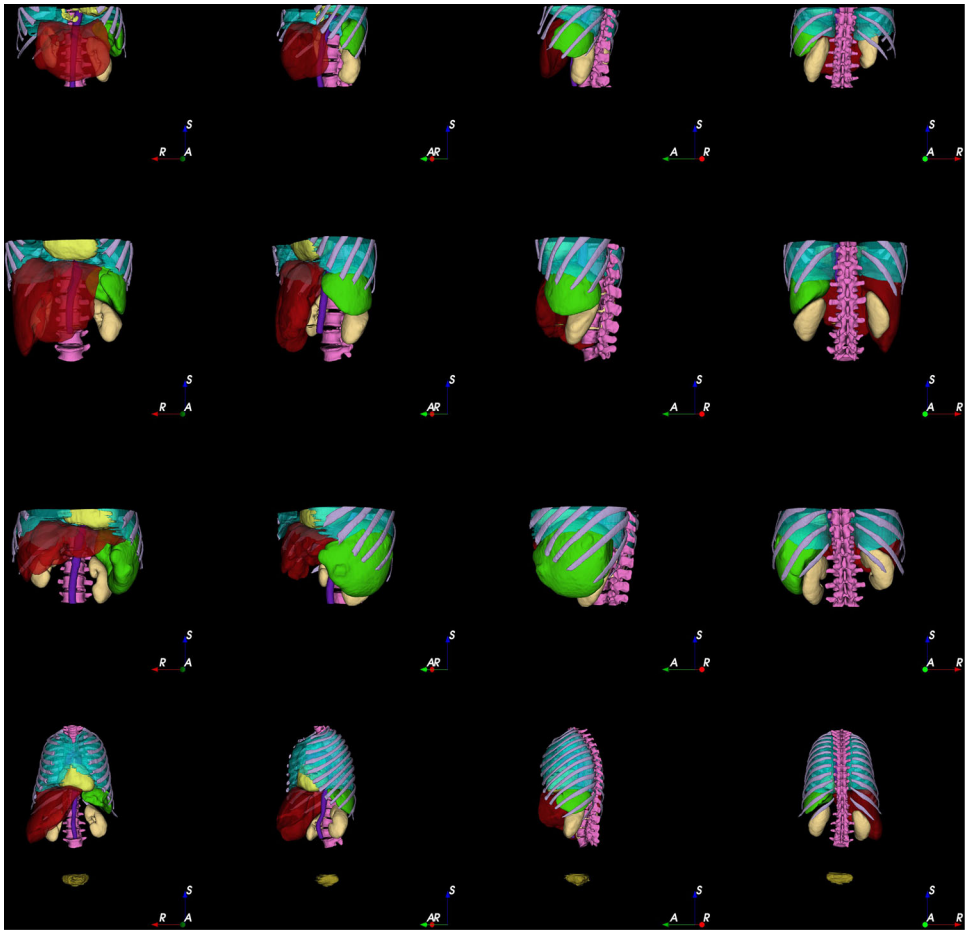

FIGURE 4 Examples of segmented organs displayed in 3D views. Yellow: heart, purple: aorta, brown: urinary bladder. green: spleen, light yellow: kidney, dark red: liver, light blue: lungs. The vertebrae and ribs were shown for better visualization. Each row belongs to a single patient and the columns represent 3D views from four different views from anterior to posterior.

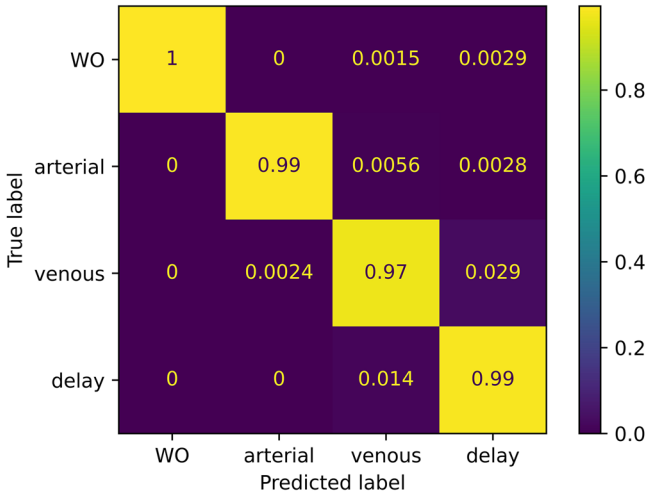

FIGURE 5 Normalized confusion matrix showing the excellent performance of the model in 10-fold strategy.

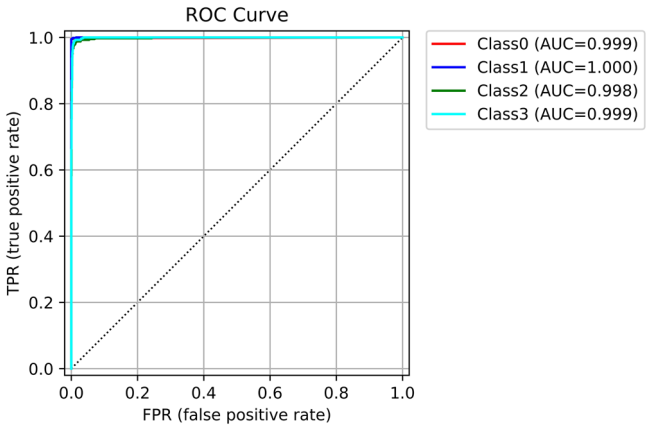

FIGURE 6 ROC curve showing excellent classifications. The legend at the top-right shows the calculated AUC for each class. AUC, area under the curve; ROC, receiver operating characteristic.

inference time of ~8 s per image for all seven organs and less than 1 s for machine learning part offers an affordable and accurate methodology for handling large datasets. We attempted to reproduce what human logic follows to indicate the contrast media phase in a typical CT image, which makes our methodology completely explainable. After intra-venous injection of the

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Performance comparison of different FS/Model combinations averaged over 10 folds.

FS–Model	F1_score	Sensitivity	Specificity	Accuracy	AUC
Boruta–DT	0.9573	0.9536	0.9870	0.9809	0.9817
Boruta–KNN	0.9786	0.9778	0.9940	0.9908	0.9967
Boruta–NB	0.9614	0.9631	0.9894	0.9835	0.9931
Boruta–RF	0.9849	0.9853	0.9957	0.9934	0.9991
Boruta–SVM	0.9807	0.9805	0.9943	0.9914	0.9991
Boruta–XGB	0.9634	0.9588	0.9891	0.9698	0.9841
MRMR–DT	0.9493	0.9457	0.9854	0.9783	0.9831
MRMR–KNN	0.9783	0.9777	0.9938	0.9906	0.9956
MRMR–NB	0.9494	0.9512	0.9862	0.9785	0.9928
MRMR–RF	0.9796	0.9804	0.9944	0.9912	0.9990
MRMR–SVM	0.9829	0.9830	0.9950	0.9924	0.9986
MRMR–XGB	0.9810	0.9815	0.9947	0.9918	0.9980
RFE–DT	0.9523	0.9485	0.9862	0.9795	0.9807
RFE–KNN	0.9681	0.9662	0.9909	0.9862	0.9939
RFE–NB	0.9212	0.9248	0.9780	0.9655	0.9889
RFE–RF	0.9705	0.9691	0.9917	0.9874	0.9969
RFE–SVM	0.9723	0.9701	0.9920	0.9880	0.9953
RFE–XGB	0.9716	0.9703	0.9920	0.9878	0.9971

Abbreviations: AUC, area under the curve; DT, decision tree; KNN, K-nearest neighbors; MRMR, maximum relevance minimum redundancy; NB, naive Bayes; RF, random forest; SVM, support vector machine; XGB, eXtreme gradient boosting.

contrast media, depending on the time passed, different combinations of enhancement in different organs are observed. First, enhancement in the descending aorta, then the rest of organs as shown in Figures 2 and 3. We used the basic information from these important organs to predict the injection phase. To do so, the input to machine learning classifier is a very basic statistical measurement of the average, SD, and 10, 50, and 90 percentiles. Boruta feature selection (FS) selected all 35 features which was expected as we only calculated meaningful first-order features. Our aim was to make the procedure as simple and explainable as possible without using deep learning classifiers and without using higher order radiomics/texture features. By breaking down the decision process into segmentation and classification steps, we tried to avoid using AI as a black box and follow human logic to perform the task. This two-step approach may be safer and more resilient to errors and our results demonstrated superior performance compared to direct deep learning classification studies. It should be clarified that we did not manipulate the decision-making process by machines (e.g., Random Forest machine) during training or testing phase. We fed the model with the features that humans consider for identifying the contrast injection phase. We used 90 percentile, average and 10 percentile features to avoid misclassification resulting from situations, such as aorta calcification or kidney stone presenting with high HUs, which might affect the model's decision-making process. High-level texture features, such as Gray Level Co-Occurrence Matrix (GLCM) extracted from the volumes of interest night provide relevant additional information. However, the aim of this study was to generate a reliable and explainable model using the most simplistic possible features. Considering high level texture features raises the question about their reproducibility when using different scanners, reconstruction kernels, kVps and other acquisition parameters, as reported in the literature.[32,33] Conversely, the features used, such as average, are more reproducible as they regularly checked by QA protocols. Besides the overall accuracy of our purposed pipeline was 0.9936, which is already very good.

In terms of accuracy, our proposed model outperformed or is comparable with results reported in the published literature, with prediction of only 33/2509 cases ($\sim < 1.4\%$) in the wrong class. The misclassified cases were reviewed, and it appears that the main reason was an error in the segmentation module. The accuracy was excellent for all four classes without a drop in performance in a single class. Our F1-score (0.9852) was better than the one reported by Zhou et al.[17] (0.977) and Dao et al.[19] (0.9209). The achieved accuracy (0.9936) was also higher than the one reported by Tang et al.[18] (0.93) and Muhamedrahimov et al.[20] (0.933).

Our study has a lot in common with what Reis et al. study.[21] The authors have almost followed the same steps, except that they used 48 radiomics features from each organ to predict the contrast media phase and used TotalSegmentator[34] trained models to generate organs masks. The number of features was lower in our study (five simple features), which makes our methodology more explainable. We also used our own light-weight segmentation model,[22] which is much faster than TotalSegmentor[34] (10 s vs 3 min). In terms of accuracy, our model outperformed their reported results. The F1-scores (our study vs Ries et al. study) were 0.9978 versus 0.966, 0.9951 versus 0.789, 0.9667 versus 0.922, and 0.9813 versus 0.950 for the four classes consisting of non-contrast, arterial, venous, and delayed, respectively. However, we did not have access to datasets used in previous studies to provide a fair comparison with our model in a realistic way, and as such, we compared our results with the classification accuracy metrics reported in the literature. As an external dataset, we used our model to predict the contrast injection phase on LiTs dataset[35] and provided the indicated label in Table S2.

Because of using multiple organs information, our model may be robust for cases with pathological conditions, such as renal failure, which can affect deep learning-based image classification methodologies.[15] It should be mentioned that our segmentation model was inferred on each CT image at each phase and performed well on images from different contrast media

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

injection phases. Due to the fact that our model inference is very fast (1 s per image/organ), generating the training dataset was not time consuming. Besides, our dataset contains patients referred to radiology department for severe pathologies in the liver and the segmentation module showed robust performance in those cases (one of these cases is shown in Figure 2).

Among the limitations of this study is that the adopted methodology depends on the accuracy of the segmentation model.

5 | CONCLUSION

We developed a fast, accurate, reliable, and explainable methodology to classify the contrast media phases in CT images, which may be useful in data curation and annotation of large online datasets or local datasets with non-standard or no series description. This algorithm might help in better exploitation of shared datasets more effectively and help to gain better solutions for a variety of AI tasks by providing more reliable data tags and information about contrast media phases.

ACKNOWLEDGMENTS

This work was supported by the Euratom research and training programme 2019–2020 Sinfonia project under grant agreement No945196.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Miles KA, Lee TY, Goh V, et al. Current status and guidelines for the assessment of tumour vascular support with dynamic contrast-enhanced computed tomography. *Eur Radiol*. 2012;22(7):1430-1441.
2. Ruhling S, Navarro F, Sekuboyina A, et al. Automated detection of the contrast phase in MDCT by an artificial neural network improves the accuracy of opportunistic bone mineral density measurements. *Eur Radiol*. 2022;32(3):1465-1474.
3. Riveira-Martin M, Akhavanallaf A, Mansouri Z, et al. Predictive value of (99m)Tc-MAA-based dosimetry in personalized (90)Y-SIRT planning for liver malignancies. *EJNMMI Res*. 2023;13(1):63.
4. Salimi Y, Shiri I, Akhavanallaf A, et al. Deep learning-based fully automated Z-axis coverage range definition from scout scans to eliminate overscanning in chest CT imaging. *Insights Imaging*. 2021;12(1):162.
5. 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.
6. Vrtovec T, Mocnik D, Strojan P, Pernus F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. *Med Phys*. 2020;47(9):e929-e950.
7. Salimi Y, Shiri I, Akhavanallaf A, et al. Deep learning-based calculation of patient size and attenuation surrogates from localizer image: toward personalized chest CT protocol optimization. *Eur J Radiol*. 2022;157:110602.
8. Salimi Y, Shiri I, Akavanallaf A, Mansouri Z, Arabi H, Zaidi H. Fully automated accurate patient positioning in computed tomography using anterior-posterior localizer images and a deep neural network: a dual-center study. *Eur Radiol*. 2023;33(5):3243-3252.
9. Salimi Y, Akhavanallaf A, Mansouri Z, Shiri I, Zaidi H. Real-time, acquisition parameter-free voxel-wise patient-specific Monte Carlo dose reconstruction in whole-body CT scanning using deep neural networks. *Eur Radiol*. 2023;33(12):9411-9424.
10. Myronakis M, Stratakis J, Damilakis J. Rapid estimation of patient-specific organ doses using a deep learning network. *Med Phys*. 2023;50(11):7236-7244.
11. Maier J, Klein L, Eulig E, Sawall S, Kachelriess M. Real-time estimation of patient-specific dose distributions for medical CT using the deep dose estimation. *Med Phys*. 2022;49(4):2259-2269.
12. Shiri I, Salimi Y, Pakbin M, et al. COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: analysis of a multi-institutional dataset of 14,339 patients. *Comput Biol Med*. 2022;145:105467.
13. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med*. 2020;26(8):1224-1228.
14. Shiri I, Vafaei Sadr A, Akhavan A, et al. Decentralized collaborative multi-institutional PET attenuation and scatter correction using federated deep learning. *Eur J Nucl Med Mol Imaging*. 2023;50(4):1034-1050.
15. Rocha BA, Ferreira LC, Vianna LGR, et al. Contrast phase recognition in liver computer tomography using deep learning. *Sci Rep*. 2022;12(1):20315.
16. Philbrick KA, Yoshida K, Inoue D, et al. What does deep learning see? insights from a classifier trained to predict contrast enhancement phase from CT images. *AJR Am J Roentgenol*. 2018;211(6):1184-1193.
17. Zhou B, Harrison AP, Yao J, et al. CT Data Curation for Liver Patients: Phase recognition in dynamic contrast-enhanced CT. *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer International Publishing; 2019:139-147. doi:10.1007/978-3-030-33391-1_16
18. Tang Y, Gao R, Lee HH, et al. Phase identification for dynamic CT enhancements with generative adversarial network. *Med Phys*. 2021;48(3):1276-1285.
19. Dao BT, Nguyen TV, Pham HH, Nguyen HQ. Phase recognition in contrast-enhanced CT scans based on deep learning and random sampling. *Med Phys*. 2022;49(7):4518-4528.
20. Muhamedrahimov R, Bar A, Laserson J, Akselrod-Ballin A, Elnekave E. Using machine learning to identify intravenous contrast phases on computed tomography. *Comput Methods Programs Biomed*. 2022;215:106603.
21. Reis EP, Blankemeier L, Chaves JMZ, et al. Automatic contrast phase detection on abdominal computed tomography using clinically-inspired techniques. paper presented at: medical imaging with deep learning, short paper track2023.
22. Salimi Y, Shiri I, Mansouri Z, Zaidi H. Deep learning-assisted multiple organ segmentation from whole-body CT images. *Medrxiv*. 2023. doi:10.1101/2023.10.20.23297331:2023.2010.2020.23297331
23. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36(11):1-13.
24. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G. Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*. 2013;29(18):2365-2368. PMID:23825369
25. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom Intell Lab Syst*. 2006;83(2):83-90.
26. Peterson LE. K-nearest neighbor. *Scholarpedia*. 2009;4(2):1883.
27. Chen T, He T. Xgboost: extreme gradient boosting. R packege version 0.4-1 1.4 (2015): 1-4.

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

28. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27(2):130-135.
29. Kecman V. Support vector machines—An introduction. In: Wang L, ed. *Support Vector Machines: Theory and Applications*. Springer Berlin Heidelberg; 2005:1-47. doi:10.1007/10984697_1
30. Murphy KP. Naive bayes classifiers University of British Columbia 18.60 (2006): 1-8.
31. Rigatti SJ. Random forest. *J Insur Med*. 2017;47(1):31-39.
32. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology*. 2018;288(2):407-415.
33. Reiazi R, Abbas E, Famiyeh P, et al. The impact of the variation of imaging parameters on the robustness of computed tomography radiomic features: a review. *Comput Biol Med*. 2021;133:104400.
34. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell*. 2023;5(5):e230024.
35. Bilic P, Vorontsov E, Chlebusr G, et al. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis* 2023;84:102680. doi: 10.48550/arxiv.1901.04056

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Salimi Y, Mansouri Z, Hajianfar G, Sanaat A, Shiri I, Zaidi H. Fully automated explainable abdominal CT contrast media phase classification using organ segmentation and machine learning. *Med Phys*. 2024;51:4095–4104. https://doi.org/10.1002/mp.17076

24734209, 2024, 6, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17076 by Bibliotheque de l'Universite de Geneve, Division de l'information, Wiley Online Library on [03/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License