Check for updates

Received: 12 September 2021 | Revised: 17 January 2022 | Accepted: 31 January 2022

DOI: 10.1002/mp.15551

RESEARCH ARTICLE

MEDICAL PHYSICS

# Phase recognition in contrast-enhanced CT scans based on deep learning and random sampling

**Binh T. Dao**[1] | **Thang V. Nguyen**[1] | **Hieu H. Pham**[1,2,3] | **Ha Q. Nguyen**[1,2]

[1]Smart Health Center, VinBigData JSC, Hanoi, Vietnam

[2]College of Engineering & Computer Science (CECS), VinUniversity, Hanoi, Vietnam

[3]VinUni-Illinois Smart Health Center, Hanoi, Vietnam

**Correspondence**
Hieu H. Pham, VinUni-IllinoisSmart Health Center & CECS, VinUniversity, Hanoi, Vietnam.
Email: hieu.ph@vinuni.edu.vn

Binh T. Dao and Thang V. Nguyen contributed equally to this work.

**Funding information**
Vingroup Big Data Institute

## Abstract

**Purpose:** A fully automated system for interpreting abdominal computed tomography (CT) scans with multiple phases of contrast enhancement requires an accurate classification of the phases. Current approaches to classify the CT phases are commonly based on three-dimensional (3D) convolutional neural network (CNN) approaches with high computational complexity and high latency. This work aims at developing and validating a precise, fast multiphase classifier to recognize three main types of contrast phases in abdominal CT scans.

**Methods:** We propose in this study a novel method that uses a random sampling mechanism on top of deep CNNs for the phase recognition of abdominal CT scans of four different phases: noncontrast, arterial, venous, and others. The CNNs work as a slicewise phase prediction, while random sampling selects input slices for the CNN models. Afterward, majority voting synthesizes the slicewise results of the CNNs to provide the final prediction at the scan level.

**Results:** Our classifier was trained on 271 426 slices from 830 phase-annotated CT scans, and when combined with majority voting on 30% of slices randomly chosen from each scan, achieved a mean F1 score of 92.09% on our internal test set of 358 scans. The proposed method was also evaluated on two external test sets: CTPAC-CCRCC ($N = 242$) and LiTS ($N = 131$), which were annotated by our experts. Although a drop in performance was observed, the model performance remained at a high level of accuracy with a mean F1 scores of 76.79% and 86.94% on CTPAC-CCRCC and LiTS datasets, respectively. Our experimental results also showed that the proposed method significantly outperformed the state-of-the-art 3D approaches while requiring less computation time for inference.

**Conclusions:** In comparison to state-of-the-art classification methods, the proposed approach shows better accuracy with significantly reduced latency. Our study demonstrates the potential of a precise, fast multiphase classifier based on a two-dimensional deep learning approach combined with a random sampling method for contrast phase recognition, providing a valuable tool for extracting multiphase abdomen studies from low veracity, real-world data.

**KEYWORDS**
CT scans, deep learning, phase recognition

## 1 | INTRODUCTION

Contrast enhancement in computed tomography (CT) scans, especially of the abdomen, is crucial for successful lesion diagnosis.[1,2] Certain types of lesions can only be observed on the CT scans taken after the injection of contrast agents into the blood veins. As illustrated in Figure 1, contrast enhancement process generally consists of three main phases[3] as follows.

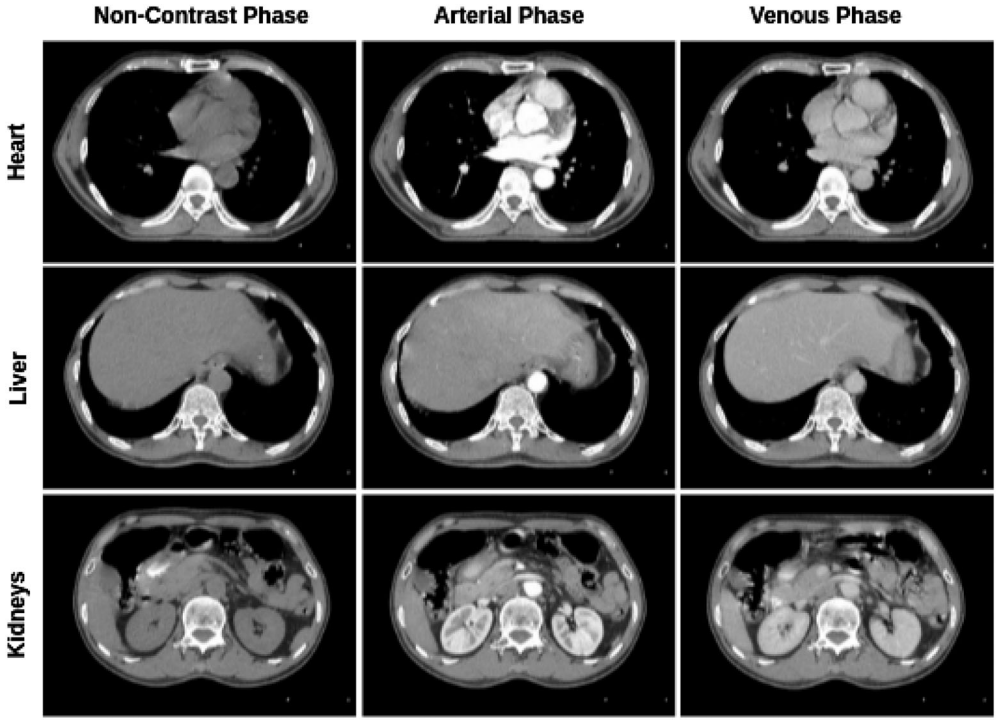


**FIGURE 1** Visual differences between the Non-Contrast (NC), Arterial (A), and Venous (V) phases in CT scans. The radiation enhancement in the different phases helps to detect different lesions in CT scans such as metastases, central tumor necrosis, and other pathologies. Radiologists usually look at arteries or veins and parenchyma to distinguish the phases

- Noncontrast: The CT scan is acquired without injection of any contrast agents.
- Arterial: The CT scan is acquired 35–40 s after the bolus injection, which can help with identifying hepatocellular carcinoma (HCC), focal nodular hyperplasia (FNH), and adenoma in the liver.
- Venous: The CT scan is acquired 70–80 s after the bolus injection, in which the liver parenchyma is enhanced through the blood supply by the portal vein, highlighting hypovascular liver lesions.

Machine learning algorithms over the past decades have achieved great success in the interpretation and diagnosis of medical imaging data,[4] including the automatic detection of liver lesions in contrast-enhanced CT images.[5–8] For instance, a multiphase analysis of abdominal CT scans[9] was performed to detect cirrhosis and HCC liver. To obtain a robust performance, however, such an algorithm often requires training from a large-scale dataset of patient's preoperative multiphase CT scans and clinical features.[8] Hence, a reliable method is needed for the collection and annotation of imaging data of abdominal structures.[10] The data mining process usually starts with accessing and collecting retrospective medical imaging data through picture archiving and communication systems (PACS). Unfortunately, the current generation of PACS systems does not support the curation of large-scale, multiphase contrast-enhanced CT datasets. The key obstacle is that DICOM tags related to a series of descriptions (e.g., noncontrast, arterial, or venous) are manually input, nonstandardized, and often incomplete.[11] These limitations lead to the impossibility of automatically categorizing medical image data based solely on their DICOM metatags, as around 15% of all studies were mislabeled due to human factors.[12] As a result, these datasets often rely on physicians for manual reannotation of CT scans, which is typically expensive and time consuming.

In addition, a typical machine learning–based computer-aided diagnosis (CAD) system for interpreting abdominal CT scans is often trained images from a specific phase, or stacks of images from different phases in some fixed order.[13] Consequently, in the deployment scenario, it is essential that the system knows precisely which phase each CT scan (series) belongs to so that the right phase scans can be fed to models. This leads to a dire need for a phase identification module for abdominal CT series.

Several approaches[14,15] have been proposed to identify multiple phases from CT scans. For instance, Zhou et al.[14] focus on volumetric characteristics of the scan, resize the entire scan to a three-dimensional (3D) block of size 32× 128×128 by interpolation, and then feed them through a 3D convolutional neural network (CNN). In the original paper, the proposed model is trained with 43 000 scans, which is a significant amount of data to be acquired, and this sparks the question of whether this method would perform well on our smaller dataset.

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

Another work by Tang et al.[15] suggests to using a generative adversarial network on each slice instead of the whole 3D scan. The authors state that CD-GAN is trained for a period of approximately 36 h on an NVIDIA 2080Ti GPU with 11G memory, showing that this method is too computationally expensive.

Different from previous approaches,[14,15] we aim to develop a fast, highly accurate deep learning system for recognizing phases from CT scans. Specifically, we propose an efficient strategy solely based on a two-dimensional (2D) representation of the slices. The proposed system consists of two main stages: (1) *random sampling* that randomly picks *R*% of slices from the input CT scan and inputs to the deep learning model. Here, *R%* denotes the percentage of slices selected from the input scan; and (2) *slice-level prediction* identifies phases of each chosen slice amd then uses majority voting to conclude the phase of the given scan. Our experimental results on internal and external (i.e., CTPAC-CCRCC,[16] LiTS[17]) datasets showed that the proposed method significantly outperforms the state-of-the-art 3D approaches while requiring less computation time for inference.

To summarize, the main contributions of this work are the following:

- We develop and evaluate a novel deep learning system for the recognition of multiphase in contrast-enhanced CT scans. The proposed system exploits a random sampler to reduce the computational time of the input examples. Majority voting is used to boost the final prediction of the system. Our extensive experiments show that the proposed approach surpasses previous state-of-the-art 3D approaches in terms of both accuracy and inference time. The proposed deep learning system can be easily reused or fine-tuned and therefore has potential benefits for several applications in clinical settings.
- The imaging dataset used in this study will be shared on our project website at https://vindr.ai/datasets/abdomen-phases , while the codes will be published at https://github.com/vinbigdata-medical/abdomen-phases. To the best of our knowledge, this is the largest annotated dataset for the recognition of multiphase in contrast-enhanced CT scans.

## 2 | PROPOSED APPROACH

### 2.1 | Overview of approach

Our main goal in this study is to develop and evaluate a fast, accurate deep learning system for the recognition of multiphase in contrast-enhanced CT scans. To this end, we first randomly sample *R*% of the slices from the whole original scan. Each of these chosen slices is then passed through a CNN model, which was trained to output the phase classification at the slice level. Finally, the scan-level prediction is predicted by a majority vote of the results obtained in the previous step. The proposed scheme for the phase recognition of abdominal CT scans is illustrated in Figure 2.

### 2.2 | Data collection and annotation

To develop deep learning algorithms for contrast phase recognition, we built an internal dataset of abdominal CT scans. The construction of this dataset was divided into three main steps: (1) data collection, (2) data deidentification, and (3) data annotation.

#### 2.2.1 | Data collection

A total of 265 abdominal studies comprising 1188 CT scans in the digital imaging and communications in medicine (DICOM) format were retrospectively randomly selected from the PACS databases of Hospital 108 and Hanoi Medical University Hospital—two major hospitals in Vietnam—within the period from 2015 until 2020. The ethical clearance of this study was approved by the Institutional Review Board of each hospital before any data-processing steps. The need for obtaining patient consent was waived because these studies did not impact clinical care.

Data characteristics, including patient demographics and the prevalence of each contrast-phase class, are summarized in Table 1. The general statistics of the slice and scan distribution for each class are featured in Figure 3. The distribution of the CT scanner models and their manufacturers are shown in Figure 4. There are six different levels of slice thickness in the entire dataset, whose distribution is captured in Table 2. Additionally, the number of slices per scan ranges from 30 to 2350 with a mean of 281. The distribution of the number of slices per scan is illustrated in Figure 5.

#### 2.2.2 | Data deidentification

To protect patient's privacy, all personally identifiable information associated with the DICOM images was removed. Specifically, a Python script was written to remove all DICOM tags of protected health information such as patient's name, patient's date of birth, and patient ID, etc. We retained only a limited number of DICOM attributes that are necessary for processing raw images. The full list of these DICOM attributes is provided in Table 9 in the Supporting Information.

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

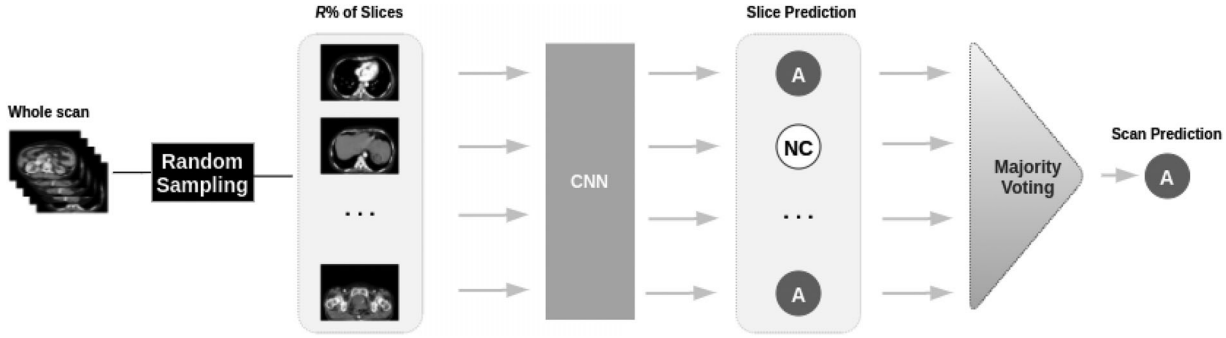


**FIGURE 2** The overall pipeline for the phase prediction from abdominal CT scans. The slices are passed sequentially through a single CNN model. Using predicted labels produced by the CNNs, a majority voting is performed to boost system performance

**TABLE 1** Characteristics of patients in the training and test datasets

| | Characteristics | Training set | Test set | Total |
|---|---|---|---|---|
| Statistics | Acquisition time (years) | 2015–2020 | 2015–2020 | 2015–2020 |
| | Number of scans | 830 | 358 | 1188 |
| | Number of slices | 271 426 | 121 134 | 392 560 |
| | Image size (slice, pixel × pixel) | 512 × 512 | 512 × 512 | 512 × 512 |
| | Male (%) | 53.76 | 56.23 | 55.21 |
| | Female (%) | 23.12 | 24.15 | 23.73 |
| | Unidentified sex (%) | 23.12 | 19.62 | 21.06 |
| | Data size (GB) | 131.1 | 56.7 | 187.8 |
| Number of slices | Noncontrast | 67 250 (24.78%) | 27 906 (23.03%) | 95 156 (24.24%) |
| | Venous | 101 040 (37.22%) | 47 865 (39.51%) | 148 905 (37.93%) |
| | Arterial | 90 058 (33.18%) | 40 811 (33.69%) | 130 869 (33.34%) |
| | Others | 13 078 (4.82%) | 4552 (3.75%) | 17 630 (4.49%) |
| Number of scans | Noncontrast | 138 (16.63%) | 45 (12.57%) | 183 (15.40%) |
| | Venous | 279 (33.61%) | 133 (37.15%) | 412 (34.68%) |
| | Arterial | 340 (40.96%) | 151 (42.18%) | 491 (41.33%) |
| | Others | 73 (8.80%) | 29 (8.10%) | 102 (8.59%) |

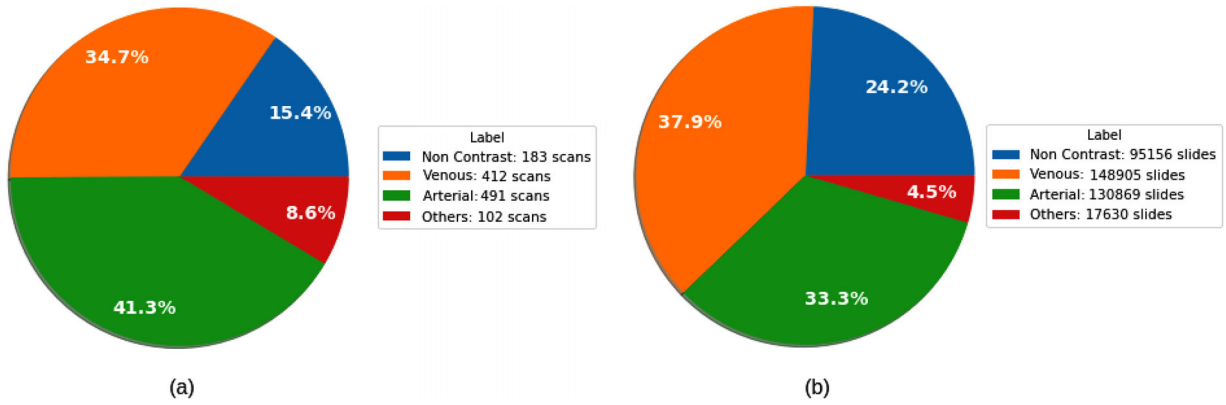


**FIGURE 3** Illustration of the scan ratio (a) and the slice ratio (b) among the 4 categories from the whole dataset

**TABLE 2** Distribution of the slice thickness over the whole dataset of 1188 scans

| Slice thickness (mm) | 0.5 | 0.625 | 1.25 | 2.0 | 2.5 | 5.0 |
|---|---|---|---|---|---|---|
| Number of scans | 97 | 3 | 727 | 22 | 2 | 337 |

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

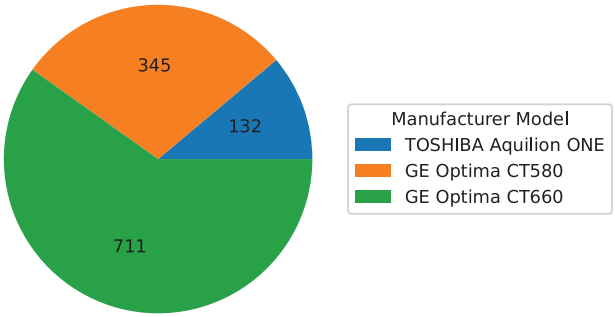


**FIGURE 4** The distribution of the CT scanner models and their manufacturers over the whole dataset of 1,188

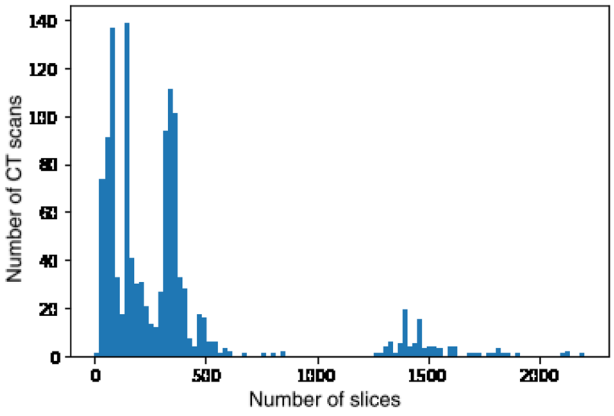


**FIGURE 5** The distribution of the number of slices per CT scan in our entire dataset

### 2.2.3 | Data annotation

The dataset was labeled for a total of four contrast-phase classes: (1) noncontrast, (2) venous, (3) arterial, and (4) others. Here the others category refers to all scans that cannot be correctly classified as either of three phases noncontrast, venous, and arterial. They may include scans of the delay phase or scans that belong to a transitional state between two phases. To annotate the imaging data, we designed and built a web-based labeling framework called VinDr Lab (https://vindr.ai/vindr-lab).[18] Two radiologists were hired to remotely annotate the data. Once the labeling was completed, the labels were exported in comma-separated values (CSV) format and used for training deep learning algorithms.

In total, 265 studies were annotated. The whole dataset was then divided into training and validation sets by a ratio of 70%/30% accordingly. Since each study usually contains multiple scans of the same patient, the train-test split was stratified by the study level to avoid data leakage. As a result, our training set consists of 271 426 slices from 830 scans (186 studies), while our validation set contains 121 134 slices from 358 scans (79 studies).

### 2.2.4 | Data records

To encourage new advances in this research direction, we will make the dataset freely accessible via our project website at https://vindr.ai/datasets/abdomen-phases. Specifically, all imaging data and the corresponding ground truth labels for the training and validation sets will be provided. The images are organized into two folders, one for training and the other for validation in which each image has a unique, anonymous identifier.

## 2.3 | Model development

This section describes in detail our model development method. We exploit state-of-the-art, high-performing deep CNN architectures for the task of recognizing multiphase in contrast-enhanced CT scans. We describe our network architecture choice and training methodology as the following.

### 2.3.1 | Network architecture

A set of state-of-the-art deep CNN models has been deployed and evaluated on the collected dataset, including ResNet-18,[19] ResNet-34,[19] SEResNet-18, ResNext-50,[20] EfficientNet-B0,[21] EfficientNet-B2,[21] GhostNet,[22] and CD-GAN.[23] These deep networks are well known to be effective for image recognition tasks. Each network accepts a CT scan as input and predicts the corresponding contrast phase label. For implementation, we followed the same instructions and recommendations from the original papers.[19–23] We considered the EfficientNet[21] model as our main network architecture choice due to the high level of accuracy and efficiency of this architecture compared to previous deep CNNs. Details of the EfficientNet[21] architecture are provided in Section VIII.A in the Supporting Information.

### 2.3.2 | Training methodology

In the training stage, all images were fed into the networks with a size of 224×224 pixels. Input images extracted from raw DICOM files were, first, converted to standard Hounsfield units (HU), using Rescale Slope and Rescale Intercept from DICOM headers. Afterward, we applied the HU window with a window center of 50 and the window width of 400 to the image. During the training process, we used the Adam optimizer[24] with an initial learning rate of $10^{-2}$ and the cosine annealing scheduler[25] with a linear warm-up.[26] Each network was trained end-to-end for 15 epochs. To this end, we minimized the binary cross-entropy loss function between the ground-truth labels and the predicted label by the network over the training samples. The proposed

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

deep network was implemented in Python using PyTorch version 1.7.1 (https://pytorch.org/). All experiments were conducted on a Ubuntu 18.04 machine with a single NVIDIA Geforce RTX 2080 Ti with 11 GB memory.

## 3 | EXPERIMENTS AND RESULTS

### 3.1 | Experimental setup

#### 3.1.1 | Internal validation

Extensive experiments were conducted to evaluate the performance of the proposed method. Specifically, we first evaluated the slicewise classification performance of trained CNN models (i.e., ResNet-18,[19] ResNet-34,[19] SEResNet-18, ResNext-50,[20] EfficientNet-B0,[21] EfficientNet-B2,[21] GhostNet,[22] and CD-GAN[23]) on the validation set of 121 134 slices. Next, we reported the classification performance of the best performing network at the scanwise level by applying the majority voting on *R*% of the slices selected from each scan. We experimented with *R* ranging from 1 to 20 at an interval of 5, and then 20–100 at an interval of 10. Finally, to compare the proposed 2D approach with previous 3D state-of-the-art approaches, we reimplemented two-phase recognition approaches on CT scans including 3DSE[14] and CD-GAN.[23] These approaches were trained on the training dataset using the same hyperparameter settings as described in the original papers.[14,23] We also measured the average inference time (second) per scan for each approach and compared it with our proposed 2D method.

#### 3.1.2 | External validation

To verify the generalization ability of the proposed deep learning model, we evaluated it on two external datasets, including LiTS[17] and CPTAC-CCRCC.[16] The LiTS[17] dataset contains 131 CT scans in the training set and 70 CT scans in the test set. It was originally developed for the development of liver segmentation methods. The CPTAC-CCRCC[16] was introduced by the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) and was developed for investigating clear cell renal cell carcinoma (CCRCC). We utilized the imaging data for CCRCC tumors, containing 242 CT scans for our external testing. However, the original aim of these datasets did not match the purpose of this study. As a result, there were no target labels for this dataset. Our radiologist team, therefore, classified scans from these datasets into four phase categories. As a result, the LiTS[17] dataset has eight scans of the arterial phase and 123 scans of the venous phase. Meanwhile, CPTAC-CCRCC[16] contains 57, 69, 53, and 63 scans from four categories noncontrast, venous, arterial, and others, respectively.

### 3.2 | Evaluation metrics

We report the classification performance using mean accuracy, macroaverage precision, macroaverage precision recall, and macroaverage F1 score. These performance indicators are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

Here, TP, FP, and FN are the number of true-positive, false-positive, and false-negative samples accordingly.

### 3.3 | Experimental results

#### 3.3.1 | Model performance on the internal test set

Table 3 summarizes quantitative results for several state-of-the-art CNN classification models on the internal test set of 121 134 slices. Note that, while training and benchmarking those models, we chose to fix the input image size to 128×128 for the sake of saving computations. It can be seen that EfficientNet-B2 achieved the best performance with a macroaveraged recall of 85.92%, a macroaveraged precision of 84.70%, and a macroaveraged F1 score of 85.26%. This architecture was then selected to conduct all the remaining experiments.

We further investigated the impact of different image input sizes on the performance of the selected model, that is, EfficientNet-B2, as shown in Table 4. We observed that using the input images with a size of 224×224 for training gave us the best result: a macroaveraged accuracy of 93.51%, a macroaveraged recall of 85.46%, a macroaveraged precision of 87.45%, and a macroaveraged F1 score of 86.43%. In addition, training the model with the 224×224 images only took 10 min for each epoch instead of 60 min when using the input images of size 512×512.

The classification performance of EfficientNet-B2 on our test set is shown in Table 5. We computed the

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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 3** Experimental results across CNN models on the slice-level evaluation when trained with input images of size 128 × 128

| Network architecture | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| ResNet-18 | 0.8964 | 0.8249 | 0.8152 | 0.8198 |
| ResNet-34 | 0.9095 | 0.8456 | 0.8172 | 0.8271 |
| SEResNet-18 | 0.8957 | 0.8192 | 0.8095 | 0.8141 |
| ResNext-50 | 0.9159 | 0.8515 | 0.8444 | 0.8475 |
| EfficientNet-B0 | 0.9229 | 0.8486 | 0.8483 | 0.8484 |
| EfficientNet-B2 | 0.9215 | 0.8470 | 0.8592 | **0.8526** |
| GhostNet | 0.9151 | 0.8397 | 0.8398 | 0.8397 |
| CD-GAN | 0.8979 | 0.8093 | 0.8526 | 0.8219 |

*Note*: The best F1 score is in bold.

**TABLE 4** Performance of EfficientNet-B2 on the slice-level evaluation with different input image sizes

| Image size | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| 128×128 | 0.9215 | 0.8470 | 0.8592 | 0.8526 |
| **224×224** | **0.9351** | **0.8546** | **0.8745** | **0.8643** |
| 512×512 | 0.9298 | 0.8400 | 0.8466 | 0.8431 |

*Note*: Best results are in bold.

**TABLE 5** Across-class quantitative results of the proposed method on the internal test set for both the slice-level and scan-level predictions. The fraction of randomly sampled slices used in majority voting was $R\% = 30\%$

| | Categories | Precision (95% CI) | Recall (95% CI) | F1 score (95% CI) |
|---|---|---|---|---|
| Slicewise | Noncontrast | 0.9982 (0.9976, 0.9986) | 0.9937 (0.9927, 0.9946) | 0.9959 (0.9953, 0.9964) |
| | Venous | 0.9195 (0.9169, 0.9218) | 0.9185 (0.9160, 0.9218) | .9190 (0.9171, 0.9207) |
| | Arterial | 0.9403 (0.9379, 0.9426) | 0.9208 (0.9181, 0.9234) | .9305 (0.9286, 0.9322) |
| | Others | 0.5523 (0.5390, 0.5656) | 0.6760 (0.6623, 0.6893) | .6079 (0.5967, 0.6192) |
| | Mean | 0.8546 (0.8491, 0.8560) | 0.8745 (0.8737, .8806) | .8643 (0.8602, 0.8664) |
| Scanwise | Noncontrast | 1.0 (0.9989, 1.0) | 1.0 (0.9897, 1.0) | 1.0 (0.9887, 1.0) |
| | Venous | 0.9124 (0.8944, 0.9259) | 0.9741 (0.9591, 0.9816) | 0.9396 (0.9286, 0.9498) |
| | Arterial | 0.9977 (0.9843, 1.0) | 0.9454 (0.9322, 0.9587) | 0.9708 (0.9637, 0.9811) |
| | Others | 0.7809 (0.7213, 0.8331) | 0.7358 (0.6668, 0.8048) | 0.7617 (0.7026, 0.8187) |
| | Mean | 0.9247 (0.9083, 0.9408) | 0.9180 (0.8976, 0.9359) | 0.9209 (0.9033, 0.9374) |

precision, recall, and F1 score for both the slice and scan levels along with their 95% confidence interval (CI) using bootstrapping over 5000 resamples of the test set. Our proposed model achieved a mean F1 score of 0.8643 (95% CI (0.8602, 0.8664)) for the slice-level prediction and a mean F1 score of 0.9209 (95% CI (0.9033, 0.9374)) for the scan-level prediction, with the majority voting on 30% of the total slices. We observed that the reported performance remained consistent between our three main classes: noncontrast, venous, and arterial, while there was a visible gap between others and the rest. Additionally, the ROC (receiver operating characteristic) curves of the proposed model for the four classes are plotted in Figure 6 along with their corresponding AUC (area under the ROC curve) scores on the slice-level test set. Unlike the F1 score, AUC is a threshold-independent metric. Nevertheless, it can be seen that the AUC scores reported in Figure 6 are strongly correlated with the slicewise F1 scores given in Table 5.

The effect of using different values of $R$ when performing random sampling with the majority voting for scan-level prediction is illustrated in Figure 7. It can be clearly seen that the macroaverage F1 score increased as $R$ (the percentage of slices in each scan to be selected randomly for inference) approaches 20% and leveled off as $R$ increased. By applying $R = 30$, we observed a 5.66% increase in the macroaveraged F1 score compared to the slicewise performance.

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

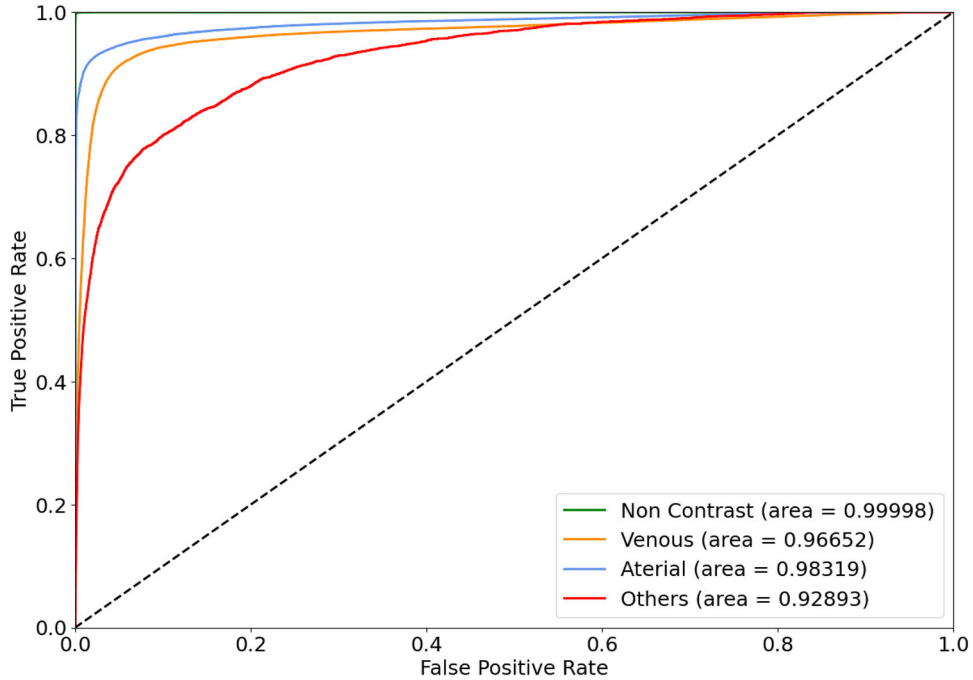


**FIGURE 6** ROC curves of the trained EfficientNet-B2 for the 4 different classes on the slice level

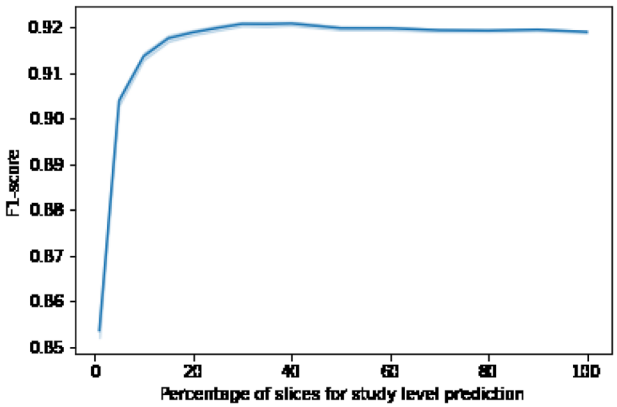


**FIGURE 7** Scan-wise performance (mean F1-score) of the trained EfficientNet-B2 on the internal test set is plotted against the percentage R of randomly sampled slices per scan used in majority voting. The shadow strip depicts the 95% confidence intervals of these F1-scores

### 3.3.2 | Comparison to state-of-the-art methods

To demonstrate the effectiveness of the proposed 2D approach, we compared our result with recent state-of-the-art methods[14,23] for the recognition of multiphase in contrast-enhanced CT scans. To this end, we reproduced the 3DSE by Zhou et al.[14] and the CD-GAN[23] by Tang et al. and reported their performance of these approaches using the F1 score on the test set. For a fair comparison, we applied the same training methodologies and hyperparameter settings as reported in the original papers.[14,23] In particular, the input image size to the model was fixed to 128×128 when compared to CD-GAN. The experimental results are provided in Table 6. We found that the proposed 2D approach significantly surpassed the previous state-of-the-art approaches (an improvement of 6.09% compared to the 3DSE[14] and 3.07% compared to CD-GAN[23]), while requiring less time for inference.

### 3.3.3 | Model performance on the external test set

Table 7 presents the experimental results on two external test sets LiTS and CPTAC-CCRCC. The average F1 score on the LiTS was 86.94%, while the average F1 score on the CPTAC-CCRCC was 76.79%. We found that the proposed method suffered from covariate shift; however, it still remains at a high level of F1 score.

## 4 | DISCUSSION

### 4.1 | Key findings

The phase recognition is important for medical imaging data collection and the deployment of machine learning models in practice. From the clinical perspective, a method for fast and precise recognition of CT phases can effectively aid in the diagnosis of abdominal pathologies.[27] By training a set of strong deep CNN models on a large-scale, annotated dataset, we built

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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 6** Comparison of state-of-the-art approaches

| | Method | Precision | Recall | F1 score | Inference time (s) |
|---|---|---|---|---|---|
| Scanwise | EfficientNet-B2 + sampling (ours) | **0.9209** | **0.9220** | **0.9209** | **6.87 × 1e–4** |
| | 3DSENet | 0.8288 | 0.9092 | 0.8600 | 2.12 × 1e–3 |
| Slicewise | EfficientNet-B2 (ours) | **0.8470** | **0.8592** | **0.8526** | 8.98 × 1e–5 |
| | CD-GAN | 0.8093 | 0.8526 | 0.8219 | 1.60 × 1e–5 |

*Note*: Best results are in bold.

**TABLE 7** Across-class quantitative results on the external datasets

| | Categories | Precision | Recall | F1 score | Number of samples |
|---|---|---|---|---|---|
| LiTS | Noncontrast | N/A | N/A N/A | N/A | 0 |
| | Venous | 0.9868 | 0.9763 | 0.9804 | 124 |
| | Arterial | 0.7312 | 0.7987 | 0.7584 | 7 |
| | Others | N/A | N/A N/A | N/A | 0 |
| CPTAC-CCRCC | Non-Contrast | 0.7728 | 0.9140 | 0.8374 | 57 |
| | Venous | 0.7018 | 0.8833 | 0.7829 | 69 |
| | Arterial | 0.9077 | 0.8191 | 0.8609 | 53 |
| | Others | 0.7688 | 0.4858 | 0.5905 | 63 |

an automated system that is able to accurately recognize contrast multiphases from CT scans. In particular, we empirically showed that a major improvement has been achieved, in terms of the F1 score and inference time by applying the proposed random sampling and majority voting. Compared to previous state-of-the-art 3D approaches, our model showed 30 times improved inference time and a nearly 6% improvement in the F1 score on our dataset.

Although a highly accurate performance has been achieved across three classes: noncontrast, arterial, and venous, we acknowledge that the proposed method reveals some limitations. To make a correct classification of contrast phases, experts often rely on multiple slices containing arteries, veins, and parenchyma. However, in our method, slices from each scan are predicted independently, without incorporating information from other regions. In addition, since contrast materials are absorbed differently for each individual, slices of the same regions from two different phases, such as the arterial and venous phases, could have similar brightness in the arteries. An example is demonstrated in Figure 8: there is a clear brightness difference between the two images in the arterial phase and the top left arterial image resembles slices from the venous phase. For this reason, our slice-level predictions are prone to errors. Another challenge is related to the nature of our dataset. Our samples vary in the scan range, and some abdominal CT scans can include neck or thighs where contrast material does not pass through, making these slices indifferentiable across our four classes. Therefore, our model is likely to produce a false prediction on these slices, which contributes negatively to the performance of the scan-level prediction.

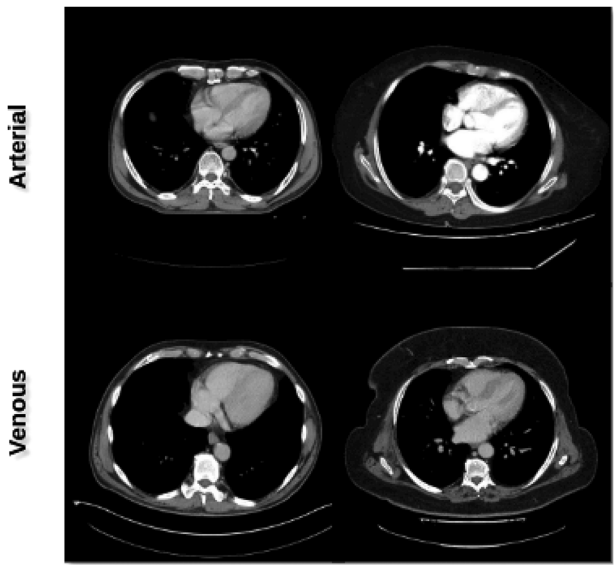


**FIGURE 8** Aortic area images from scans of 4 different patients demonstrate the visual variance between images of the same categories

The low generalizability of deep learning-based diagnostic systems[28–31] to datasets and scanners beyond the ones they have been trained with, has been limiting the use of such methods in real-world clinical settings. We showed that the proposed deep learning

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

method was successfully generalized to two different datasets from other hospital sites, each with a different CT scanner.

## 4.2 | Future work

There are several possible mechanisms to improve our current method. The most promising direction is to eliminate nonaffected contrast-enhanced regions of a scan, such as pelvis areas. This would improve slice-level prediction since the model is forced to learn and predict images with clearer features. In addition, due to the performance drop-in "others," future work includes applying techniques for reducing the impact of imbalanced data. For example, weighted Binary Cross-Entropy (BCE) losses,[32] which directly penalize probabilistic false positives, can be used. We also plan to experiment with training and testing the proposed method on the coronal and sagittal projections of the CT scans, so that each input image could contain all necessary components: arteries, veins, and parenchyma, which are used to identify the correct contrast phases. Moreover, we will conduct additional experiments, incorporating training procedure refinements[33] such as data augmentation methods to further improve the generalization of our method. Lastly, it is worth investigating more sophisticated methods for sampling and synthesizing slice-level predictions, such as the multi-instance learning paradigm,[34] rather than the straightforward random sampling and majority voting strategies discussed in the paper.

## 5 | CONCLUSION

In this study, we developed a 2D deep learning–based approach for the recognition of contrast phases in abdominal CT scans. We adopted a random sampling strategy to improve the classification performance and reduce inference time. The introduction of a random sampling mechanism helps avoid training and inferring on 3D data, which are usually much more costly, while still attaining impressive performances. Extensive experimental results on both the internal and external datasets have demonstrated that the proposed approach significantly outperformed previous state-of-the-art 3D approaches.

### ACKNOWLEDGMENTS

This work was supported by VinBigData JSC. We are especially thankful to all of our collaborators, including radiologists, physicians, and technicians, who participated in the data collection and labeling process.

### CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

### REFERENCES

1. Brancatelli G, Federle MP, Grazioli L, Blachar A, Peterson MS, Thaete L. Focal nodular hyperplasia: CT findings with emphasis on multiphasic helical CT in 78 patients. *Radiology*. 2001;219:61-68.
2. Bronstein YL, Loyer EM, Kaur H, et al. Detection of small pancreatic tumors with multiphasic helical CT. *Am J Roentgenol*. 2004;182:619-623.
3. Smithuis R. CT contrast injection and protocols. 2014. Accessed Dec 1, 2021. https://radiologyassistant.nl/more/ct-protocols/ct-contrast-injection-and-protocols
4. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
5. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology*. 2018;286:887-896.
6. Wang W, Iwamoto Y, Han X, et al. Classification of focal liver lesions using deep learning with fine-tuning. In: *Proceedings of the International Conference on Digital Medicine and Image Processing*. ACM; 2018:56-60.
7. Yoshinobu Y, Iwamoto Y, Xianhua H, et al. Deep learning method for content-based retrieval of focal liver lesions using multi-phase contrast-enhanced computer tomography images. In: *IEEE International Confer- ence on Consumer Electronics (ICCE)*. IEEE; 2020:1-4.
8. Gao R, et al. Deep learning for differential diagnosis of malignant hepatic tumorsbased on multi-phase contrast-enhanced CT and clinical data. *J Hematol Oncol*. 2021;14:1-7.
9. Nayak A, Kayal EB, Arya M, et al. Computer-aided diagnosis of cirrhosis and hepatocellular carcinoma using multi-phase abdomen CT. *Int J Comput Assist Radiol Surg*. 2019;14:1341-1352.
10. Park S, et al. Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation. *Diagn Interv Imaging*. 2020;101:35-44.
11. Harvey H, Glocker B. A standardised approach for preparing imaging data for machine learning tasks in radiology. In: Ranschaert E, Morozov S, Algra P (eds.). *Artificial Intelligence in Medical Imaging*. Springer; 2019:61-72.
12. Gueld MO, Kohnen M, Keysers D, et al. Quality of DICOM header information for image categorization. In: *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*. Proceedings of SPIE, Vol. 4685. SPIE; 2002:280-287.
13. Sun C, Guo S, Zhang H, et al. Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs. *Artif Intell Med*. 2017;83:58-66.
14. Zhou B, Harrison AP, Yao J, et al. CT data curation for liver patients: phase recognition in dynamic contrast-enhanced CT. arXiv:1911.06395 [eess.IV]. 2019.
15. Tang Y, et al. Contrast phase classification with a generative adversarial network. In *Medical Imaging 2020: Image Processing*. Proceedings of SPIE, Vol. 11313. SPIE; 2020:1131310.
16. Kalayci S, Petralia F, Wang P, Gümüş ZH. ProNetView-ccRCC: A web-based portal to interactively explore clear cell renal cell carcinoma proteogenomics networks. *Proteomics*. 2020;20:2000043.
17. Bilic P, et al. The liver tumor segmentation benchmark (LITS). arXiv preprint. arXiv:1901.04056. 2019.
18. Nguyen NT, Truong PT, Ho VT, et al. VinDr lab: a data platform for medical AI. 2021. Accessed Dec 1, 2021. https://github.com/vinbigdata-medical/vindr-lab
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2016:770-778.
20. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2017:1492-1500.

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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

21. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: *The 36th International Conference on Machine Learning*. PMLR; 2019:6105-6114.
22. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C, Ghostnet: More features from cheap operations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2020:1580-1589.
23. Tang Y, Lee HH, Xu Y, et al. Contrast phase classification with a generative adversarial network. arXiv:1911.06395 [eess.IV]. 2019.
24. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.
25. Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983. 2016.
26. Ma J, Yarats D. On the adequacy of untuned warmup for adaptive optimization. arXiv preprint arXiv:1910.04209. 2019:7.
27. Guite K, Hinshaw L, Lee F. Computed tomography in abdominal imaging: how to gain maximum diagnostic information at the lowest radiation dose. In: Wang D (ed). *Selected Topics on Computed Tomography*. IntechOpen. 2013. [Online]. Available: https://intechopen.com/chapters/43704 doi: https://doi.org/10.5772/55903
28. Therrien R, Doyle S. Role of training data variability on classifier performance and generalizability. In: *Medical Imaging 2018: Digital Pathology*. Proceedings of SPIE, Vol. 10581. SPIE; 2018:1058109.
29. Liang X, Nguyen D, Jiang SB. Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion. *Mach Learn: Sci Technol*. 2020;2:015007.
30. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295:4-15.
31. Nadeem SA, Comellas AP, Hoffman EA, Saha PK. Generalizability of a deep learning airway segmentation algorithm to a blinded low-dose CT dataset. In: *Medical Imaging 2021: Image Processing*. Proceedings of SPIE, Vol. 11596. SPIE; 2021: 115963I.
32. Ho Y, Wookey S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*. 2019;8:4806-4813.
33. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M. Bag of tricks for image classification with convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2019:558-567.
34. Yan Z, Zhan Y, Peng Z, et al. Multi-instance deep learning: discover discriminative local anatomies for body part recognition. *IEEE Trans Med Imaging*. 2016;35:1332-1343.
35. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2009:248-255.


## 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:** 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;49:4518-4528. https://doi.org/10.1002/mp.15551

24734209, 2022, 7, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15551 by Hong Kong University Of, Wiley Online Library on [28/01/2024]. 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