Machine Learning Classification of Body Part, Imaging Axis, and Intravenous Contrast Enhancement on CT Imaging

Canadian Association of Radiologists' Journal
2023, Vol. 0(0) 1–10

DOI: 10.1177/08465371231180844
journals.sagepub.com/home/caj
Sage

Wuqi Li[1,*], Hui-Ming Lin[2,*], Amy Lin[2,3], Marc Napoleone[2,3], Robert Moreland[2,3], Alexis Murari[1], Maxim Stepanov[1], Eric Ivanov[1], Abhinav Sanjeeva Prasad[1], George Shih[4], Zixuan Hu[1], Suvd Zulbayar[5], Ervin Sejdić[1,6,], and Errol Colak[2,3,7,**]**

Abstract

Purpose: The development and evaluation of machine learning models that automatically identify the body part(s) imaged, axis of imaging, and the presence of intravenous contrast material of a CT series of images. **Methods:** This retrospective study included 6955 series from 1198 studies (501 female, 697 males, mean age 56.5 years) obtained between January 2010 and September 2021. Each series was annotated by a trained board-certified radiologist with labels consisting of 16 body parts, 3 imaging axes, and whether an intravenous contrast agent was used. The studies were randomly assigned to the training, validation and testing sets with a proportion of 70%, 20% and 10%, respectively, to develop a 3D deep neural network for each classification task. External validation was conducted with a total of 35,272 series from 7 publicly available datasets. The classification accuracy for each series was independently assessed for each task to evaluate model performance. **Results:** The accuracies for identifying the body parts, imaging axes, and the presence of intravenous contrast were 96.0% (95% CI: 94.6%, 97.2%), 99.2% (95% CI: 98.5%, 99.7%), and 97.5% (95% CI: 96.4%, 98.5%) respectively. The generalizability of the models was demonstrated through external validation with accuracies of 89.7 - 97.8%, 98.6 - 100%, and 87.8 - 98.6% for the same tasks. **Conclusions:** The developed models demonstrated high performance on both internal and external testing in identifying key aspects of a CT series.

Résumé

Objectif : La mise au point et l'évaluation de modèles d'apprentissage machine identifiant automatiquement les parties du corps sur une image, l'axe de l'imagerie et la présence de produit de contraste intraveineux dans une série d'images de TDM. **Méthodes :** Cette étude rétrospective a inclus 6955 séries tirées de 1198 études (501 femmes, 697 hommes, âge moyen : 56,5 ans) obtenues entre janvier 2010 et septembre 2021. Chaque série a été annotée par un radiologiste certifié et entraîné avec des étiquettes indiquant 16 parties du corps, 3 axes d'imagerie et la présence ou l'absence d'utilisation de produit de contraste. Les études ont été affectées de manière aléatoire aux ensembles de données d'entraînement, de validation et de tests dans les proportions respectives de 70 %, 20 % et 10 % dans le but de développer un réseau neuronal profond 3D pour chaque tâche de classification. Une validation externe a été menée avec un total de 35 272 séries issues de 7 ensembles publics de données disponibles. La précision de la classification de chaque série a été évaluée de façon indépendante pour chaque tâche afin d'évaluer les performances du modèle. **Résultats :** L'exactitude de l'identification des parties du corps, des axes d'imagerie et de la présence de produit de contraste a été, respectivement, de 96,0 % (IC à 95 % : 94,6 % à 97,2 %), 99,2 % (IC à 95 % : 98,5 % à

AQ1

[1] Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada
[2] Department of Medical Imaging, Unity Health Toronto, Toronto, ON, Canada
[3] Department of Medical Imaging, Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[4] Department of Radiology, Weill Cornell Medicine, New York, NY, USA
[5] Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[6] North York General Hospital, Toronto, ON, Canada
[7] Li Ka Shing Knowledge Institute, St Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

*These authors contributed equally to this work. **These authors jointly supervised this work.

Corresponding Author:
Errol Colak, Department of Medical Imaging, Unity Health Toronto, University of Toronto, 30 Bond Street Toronto, ON M5B 1W8, Canada.
Email: Errol.Colak@unityhealth.to

99,7 %) et 97,5 % (IC à 95 % : 96,4 % à 98,5 %). La possibilité de généraliser les modèles a été démontrée par validation externe avec une précision respective de 89,7 % à 97,8 %, 98,6 % à 100 % et 87,8 % à 98,6 % pour les mêmes tâches. **Conclusions :** Les modèles mis au point ont démontré une haute performance en ce qui concerne autant les tests internes que les tests externes pour l'identification des aspects principaux des séries de TDM.

Keywords
computed tomography, CT, computed tomography series, series categorization, body part recognition, intravenous contrast, machine learning, automation

Introduction

A fundamental challenge facing machine learning (ML) model development is the acquisition and processing of large, high-quality datasets of representative real-world data necessary for training.[1] In medical imaging projects, much of this labor-intensive work has been done manually by radiologists or trained researchers and is often the rate limiting step in the development pipeline.[2] In order to expand the scale, streamline dataset curation, and accelerate model development, automating aspects of dataset curation are becoming increasingly important.

Medical imaging studies that meet inclusion criteria are downloaded from a Picture Archiving and Communication Systems (PACS) in Digital Imaging and Communications in Medicine (DICOM) format for subsequent processing. Each study consists of one or more series of images, which can differ in the anatomical region imaged, imaging axis, convolutional kernel, slice thickness, and presence and timing of intravenous (IV) contrast. Developers have an explicit definition of the imaging that needs to be included in terms of these basic parameters based on the target input of the final model. In principle, data extraction should be a straightforward process of using DICOM metadata to select desired imaging data but in practice this is very often a time consuming, frustrating, and highly manual selection process due to the heterogeneity and inconsistency of metadata.[3] This process is only worsened when the imaging data is derived from equipment of multiple vendors or different institutions.

A system that operates independently of DICOM metadata and accurately categorizes imaging along key parameters would be valuable in both model development and deployment. To our knowledge, there has been limited prior investigation of ML mediated CT series classification models. Several ML models for classifying body parts and contrast enhancement have been developed but these handled a limited number of body parts and phases of post-contrast imaging.[4-8]

The aim of this study was the development and evaluation of ML models that automatically classify CT series by the body part(s) imaged, axis of imaging, and IV contrast enhancement.

Methods

This retrospective study was approved by a Hospital Research Ethics Review Board with a waiver for informed consent.

Dataset Curation

The radiology information system (syngo, Siemens Medical Solutions) was searched using Nuance mPower (Nuance Communications) to determine the names and frequency of CT protocols performed between January 1, 2010 and September 30, 2021. These protocols are indication based and define the anatomical region to be imaged and specific technical parameters used for image acquisition. The most common and representative protocols for neuro, thoracic, and abdominopelvic imaging were identified. These 83 protocols accounted for 79.7% of all CT imaging performed during this period. A random sample of 15 studies from each of these protocols was performed. Images in DICOM format were downloaded from Philips Vue PACS (Philips) using RSNA Anonymizer and underwent de-identification using SapienSecure (Sapien). Scout images, scanned documents, dose reports, and post-processed images were removed. A custom Python (Python Software Foundation) script further limited DICOM metadata fields to a "white-list". Manual review of images ensured no private health information remained. The CT studies were performed on a multi-detector CT scanner (Revolution, LightSpeed 64, or Optima 64; General Electric Medical Systems).

Two fellowship trained radiologists (A.L. in neuroradiology and M.N. in body imaging) served as annotators. Annotation was performed using a web-based annotation platform (MD.ai). Each series was labelled for body part, imaging axis, and IV contrast. The body part label refers to the specific anatomical region(s) being imaged as part of a CT series. A total of 1198 examinations (age 18-98 years, mean age 56.5 ± 18.0 years; 697 males, age 18-98 years, mean age 55.1 ± 18.0 years; 501 females, age 18-96 years, mean age 58.5 ± 17.9 years) comprising of 6955 series (4098 male, 2857 female) were included. Forty-seven studies could not be downloaded successfully from the PACS. Dataset characteristics are summarized in Table 1. No patient was represented more than once. Each CT study was present in only a single dataset partition to prevent data leakage.

Table 1. Data Composition of the Training, Validation, and Testing Sets. The Number in each Cell Represents the Number Series of the Corresponding Class. The Numbers in Parentheses Indicate the Proportion in the Dataset.

	Total	Training	Validation	Testing
Body part				
Head	1054 (15.2%)	726 (15.1%)	212 (15.3%)	116 (15.4%)
Head and face	149 (2.1%)	104 (2.2%)	32 (2.3%)	13 (1.7%)
Face/Orbits/Sinus	744 (10.7%)	517 (10.7%)	153 (11.0%)	74 (9.8%)
Temporal bone	175 (2.5%)	126 (2.6%)	34 (2.4%)	15 (2.0%)
Head and neck	349 (5.0%)	237 (4.9%)	76 (5.5%)	36 (4.8%)
Neck	218 (3.1%)	150 (3.1%)	43 (3.1%)	25 (3.3%)
C-spine	806 (11.6%)	531 (11.0%)	172 (12.4%)	103 (13.6%)
C and T spine	99 (1.4%)	71 (1.5%)	18 (1.3%)	10 (1.3%)
T-spine	195 (2.8%)	134 (2.8%)	37 (2.7%)	24 (3.2%)
T and L-spine	419 (6.0%)	293 (6.1%)	87 (6.3%)	39 (5.2%)
L-spine	183 (2.6%)	131 (2.7%)	30 (2.2%)	22 (2.9%)
Chest	749 (10.8%)	502 (10.4%)	136 (9.8%)	111 (14.7%)
Chest-abdomen-pelvis	155 (2.2%)	108 (2.2%)	31 (2.2%)	16 (2.1%)
Abdomen	409 (5.9%)	285 (5.9%)	82 (5.9%)	42 (5.6%)
Abdomen-pelvis	983 (14.1%)	710 (14.8%)	194 (14.0%)	79 (10.5%)
Pelvis	268 (3.9%)	186 (3.9%)	52 (3.7%)	30 (4.0%)
Axis				
Axial	3947 (56.8%)	2758 (57.3%)	761 (54.8%)	428 (56.7%)
Coronal	1551 (22.3%)	1065 (22.1%)	315 (22.7%)	171 (22.6%)
Sagittal	1457 (20.9%)	988 (20.5%)	313 (22.5%)	156 (20.7%)
IV Contrast				
No	3208 (46.1%)	2186 (45.4%)	657 (47.3%)	365 (48.3%)
Yes	3747 (53.9%)	2625 (54.6%)	732 (52.7%)	390 (51.7%)
Kernel				
Soft tissue	4470 (64.3%)	3112 (64.7%)	884 (63.6%)	474 (62.8%)
Bone	2165 (31.1%)	1480 (30.8%)	448 (32.3%)	237 (31.4%)
Lung	320 (4.6%)	219 (4.6%)	57 (4.1%)	44 (5.8%)
Total	6955 (1198 studies)	4811 (833 studies)	1389 (237 studies)	755 (128 studies)

Data Preprocessing

Hounsfield unit (HU) values were extracted from each DICOM file and transformed from 512 × 512 into 224 × 224 8-bit grayscale images. Hounsfield unit values ranging between −1000 and 400 were converted to a [0, 255] scale while values below −1000 HU were assigned 0 and those above 400 HU were assigned 255. Image pixel values were normalized to the range of [0, 1]. For each series, images were stacked to construct a 3D matrix with dimension Nx224 × 224, where N is the total number of images in the series. This matrix was resized to 128 × 224 × 224 by down-sampling and up-sampling operations to match the model's input.

Model Development

The 3D residual neural network (3D ResNet)[9] was utilized to perform the CT series-level classification tasks in our study. The 3D ResNet is composed of 3D convolution blocks and pooling layers. Shortcut connections were included to link early to late layers and skip layers in between, making the networks easier to train, especially for very deep neural networks. The 3D ResNet was adjusted to accept inputs of size 128 × 224 × 224 with 50 layers. The model output was controlled by a fully connected layer that transformed the extracted features into classification scores, normalized into positive values that sum to 1 using a SoftMax function. The size of the fully connected layer was adjusted in each task to map to the number of classes. Three models were constructed and separately trained, where each model corresponded to a classification task (Figure 1). Model outputs for each CT series were compared to ground truth labels to evaluate comprehensive predictions.

During model development, data was divided at the study-level with a proportion of 70%, 20% and 10%, corresponding to the training, validation, and testing sets, respectively. The models were trained from scratch with a stochastic gradient descent optimizer and initial learning rate of .001. The learning rate was decayed every 7 epochs by a factor of .1. Validation loss (cross-entropy) was evaluated at the end of

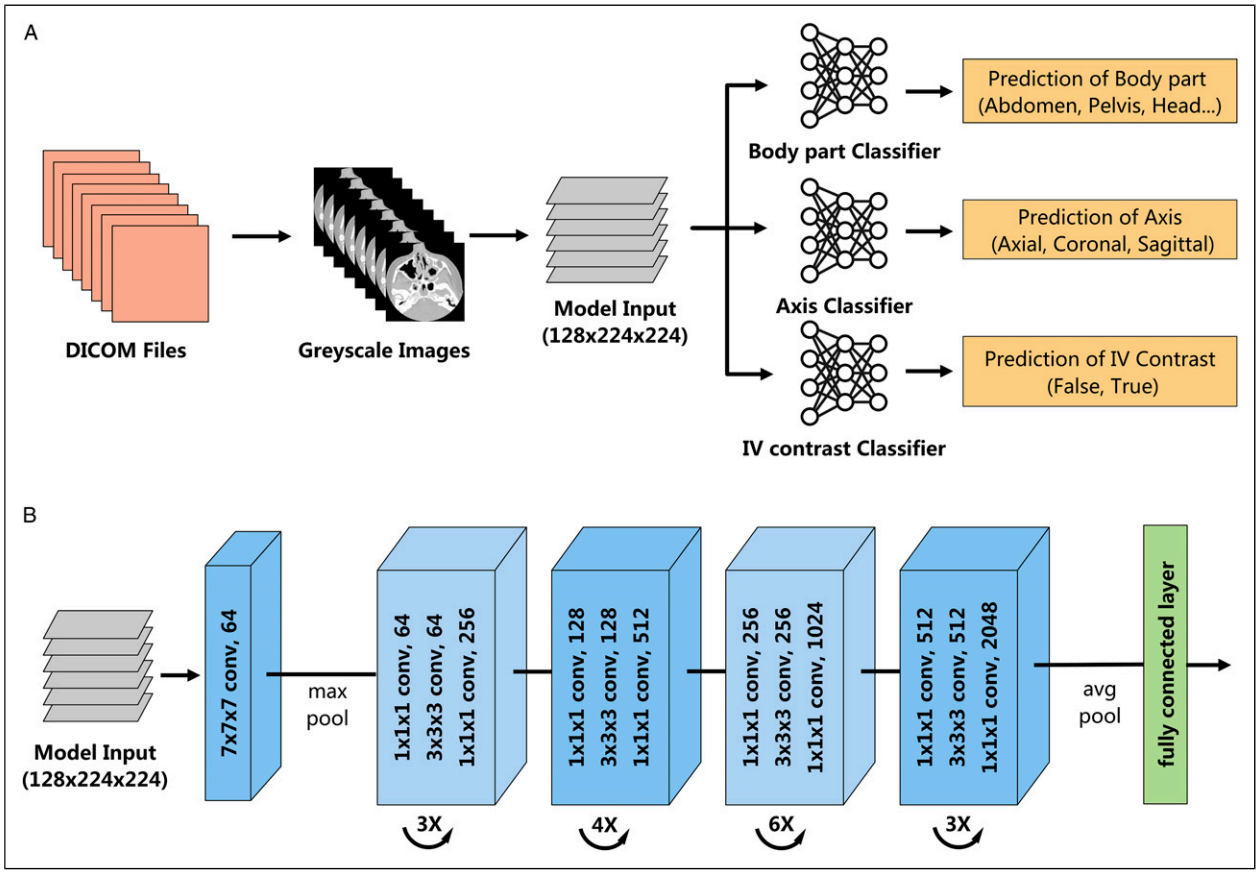

Figure 1. Overview of the system architecture. (A) Classification pipeline. Three models were employed, and each model corresponds to a classification task. (B) Network structure of the utilized ResNet50. DICOM = Digital Imaging and Communications in Medicine.

each epoch, and an early stopping strategy was utilized with the patience of ten epochs. The training and testing were constructed in the PyTorch framework (The Linux Foundation) on an Nvidia Quadro RTX 8000 48G GPU (NVIDIA Corporation).

Evaluation of Model Generalizability

To demonstrate generalizability of model performance, external validation was performed using 7 publicly available medical imaging datasets. These datasets include the Radiological Society of North America (RSNA) Intracranial Hemorrhage (ICH) Detection,[10] RSNA 2022 Cervical Spine Fracture Detection,[11] RSNA Pulmonary Embolism CT,[12] SPIE-AAPM Lung CT Challenge,[13] StageII-Colorectal-CT,[14] CPTAC-LSCC,[15] and C4KC-KiTS[16] datasets. The axis of imaging, body part, kernel, and presence of IV contrast is homogeneous for the RSNA datasets. The non-RSNA datasets were annotated by a board-certified radiologist.

Statistical Analysis

For each input series, the final prediction aligned with the class that had the maximum score. In each task, model performance was evaluated by determining accuracy, top-2 accuracy, micro averaged F1 score and confusion matrices. Furthermore, area under the receiver operating characteristic curve (AUC), accuracy, balanced accuracy, F1 score, positive and negative predictive values, sensitivity, specificity for each class were also calculated. Bootstrapping experiments were conducted with 1000 iterations and 95% CIs were calculated with .025 and .975 percentiles. Analyses were performed with metrics API from Scikit-learn version 1.1.1.

Results

Performance of Body Part Classification

There were 16 different body part labels included in the model's input covering anatomy from head to pelvis. Overall body part classification accuracy was 96.0% (95% CI: 94.6%, 97.2%) with a top-2 accuracy of 99.1% (95 CI%: 98.3%, 99.7%) (Table 2). The sensitivity ranges from a low of .50 for "cervical and thoracic spine" to a high of 1.0 for several other anatomical ranges (Table 3 and Figure 2C). For the body parts with a lower sensitivity, the model tended to select an anatomically similar body range. For instance, when the model failed to correctly classify the "cervical and thoracic spine", it

Table 2. Evaluation Results of Accuracy, top-2 Accuracy, Weighted Sensitivity, and Micro Averaged F1 Score for all Tasks in the Test Set. The Numbers in Parentheses Indicate the 95% Confidence Interval.

Task	Accuracy (%)	Top-2 Accuracy (%)	Sensitivity, %	F1 Score, %
Body part	96.0 (94.6, 97.2)	99.1 (98.3, 99.7)	96.0	96.0
Axis	99.2 (98.5, 99.7)	99.9 (99.6, 100.0)	99.2	99.2
IV Contrast	97.5 (96.4, 98.5)	100.0 (100.0, 100.0)	97.5	97.5

assigned a cervical or thoracic spine label instead. Similarly, a sensitivity of .80 for neck is a result of misallocation to the cervical spine. For all misclassified body part in the testing set, the predicted anatomy always included the actual body part of the input series.

Subsequent external validation was performed on multiple publicly available datasets (Table 4). Classification accuracy ranged between 89.7% (95% CI: 88.5%, 91.0%) and 97.8% (95% CI: 97.5%, 98.1%). Classification accuracy on the pooled external dataset was 97.3% (95% CI: 97.1%, 97.5%). Like the internal test set, almost all classification errors were the result of misallocation to a similar body part or one with overlapping coverage (Figure 3C). For example, cervical spine was misclassified 9% of the time as neck. Detailed performance metrics for each external validation dataset are available in the supplemental materials.

Performance of Axis of Imaging Classification

Three standard CT axes of imaging (axial, coronal and sagittal) were used as classification model outputs. Overall accuracy on the test set was 99.2% (95% CI: 98.5%, 99.7%) with a top-2 accuracy of 99.9% (95% CI: 99.6%, 100%) (Table 2). Only the sagittal orientation was ever misclassified while the other orientations were always correctly identified (Figure 2A). Model performance was maintained on external validation with classification accuracy ranging between 98.6% (95% CI: 95.7%, 100%) and 100% (95% CI: 100%, 100%) (Table 4).

Performance of Intravenous Contrast Classification

CT imaging can be performed with or without IV contrast agent. Classification accuracy for IV contrast was 97.5% (95% CI: 96.4%, 98.5%) (Table 2). On external validation, model accuracy ranges between 87.8% (95% CI: 86.8%, 88.8%) and 98.6% (95% CI: 95.7%, 100.0%) (Table 4). A notable drop in performance was noted when external validation was performed with the RSNA ICH datasets. This is attributable to CT studies with intracranial hemorrhage (ICH) that were misinterpreted as being performed with IV contrast agent. This suggests under-representation in our original training set of cases with ICH which can be a mimic of IV contrast.

Performance of Multi-Class Classification

The results of the 3 prior tasks were compiled to determine simultaneous overall performance of the models on a complete classification task. On the testing set, the model successfully classified a CT series in all 3 tasks correctly in 700 out of 755 (92.7%), incorrect on only one of the parameters on 55 series (7.3%), and all series had at least 2 labels correctly classified. There was similar performance on the pooled external validation dataset with accuracy ranging between 93.7% (95% CI: 93.1%, 94.3%) and 98.1% (95% CI: 96.2%, 99.5%) (Table 4).

Discussion

A key challenge in model development and deployment is ensuring that the correct studies are identified for use in training or as input to the final system. The use of DICOM metadata to filter imaging studies for these tasks is often employed, which should work in theory but is frequently problematic in practice due to erroneous or inconsistent data.[3,17] Reliance on metadata can result in the need for a resource intensive manual review of the dataset prior to training or model failure in clinical practice when there is a mismatch between expected and actual studies used as model inputs. To help address both problems, we have developed ML models that can classify the body part(s) imaged, imaging axis, and IV contrast enhancement for a series of CT images based on images alone.

A small number of published studies have investigated the use of ML for the classification of body parts imaged by CT.[4,5] Na et al[4] reported a model that could classify a CT into 5 anatomical locations with an accuracy of 100% on internal and 99.8% on external validation sets. Recently, Raffy et al[18] proposed a model that could classify 17 body parts with a sensitivity of 92.5% at the image-level and 1.1% improvement at the series-level. Our study differs from previous work by providing more detailed body part classification by including 16 labels, many of which cover multiple anatomical regions. This granular label schema better reflects real-world clinical practice in which CT studies may have overlapping coverage and different scanning parameters such as the field-of-view, radiation dose, slice thickness, and phase of post-contrast imaging. Instead of processing each image sequentially in a CT series, we used 3D models to capture the features of an entire CT series simultaneously for the classification tasks. As a result, our models demonstrated high accuracy in both

Table 3. Evaluation Results of Area Under the Receiver Operating Characteristic Curve (AUC), Accuracy, Balanced Accuracy (Bal. Acc), F1 Score, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, and Specificity for Each Class on the Test Dataset. The Numbers in Parentheses Indicate the 95% Confidence Interval.

Classification Class	AUC	Accuracy (%)	Bal. Acc (%)	F1 Score (%)	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)
Body part								
Head	.999 (.996, 1.000)	99.7 (99.3, 100.0)	99.5	99.1	99.1 (97.4, 100.0)	99.8 (99.5, 100.0)	99.1 (97.4, 100.0)	99.8 (99.5, 100.0)
Head and face	.999 (.996, 1.000)	99.6 (99.1, 100.0)	96.0	88.9	85.7 (64.3, 100.0)	99.9 (99.6, 100.0)	92.3 (76.9, 100.0)	99.7 (99.3, 100.0)
Face/Orbits/Sinus	1.000 (1.000, 1.000)	99.7 (99.3, 100.0)	98.6	98.6	100.0 (100.0, 100.0)	99.7 (99.3, 100.0)	97.3 (93.2, 100.0)	100.0 (100.0, 100.0)
Temporal bone	1.000 (1.000, 1.000)	100.0 (100.0, 100.0)	100.0	100.0	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Head and neck	1.000 (1.000, 1.000)	99.9 (99.6, 100.0)	99.9	98.6	97.3 (91.9, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	99.9 (99.6, 100.0)
Neck	.997 (.993, 1.000)	98.8 (97.9, 99.5)	89.7	81.6	83.3 (66.7, 95.8)	99.3 (98.6, 99.9)	80.0 (64.0, 96.0)	99.5 (98.9, 99.9)
C-spine	.999 (.996, 1.000)	98.5 (97.5, 99.3)	97.5	94.7	93.4 (87.7, 97.2)	99.4 (98.6, 99.8)	96.1 (91.3, 99.0)	98.9 (98.0, 99.7)
C and T spine	.997 (.993, 1.000)	99.2 (98.5, 99.7)	74.9	62.5	83.3 (50.0, 100.0)	99.3 (98.8, 99.9)	50.0 (20.0, 80.0)	99.9 (99.6, 100.0)
T-spine	.999 (.996, 1.000)	99.7 (99.3, 100.0)	99.9	96.0	92.3 (80.8, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	99.7 (99.3, 100.0)
T and L-spine	1.000 (1.000, 1.000)	99.9 (99.6, 100.0)	99.9	98.7	97.5 (92.5, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	99.9 (99.6, 100.0)
L-spine	.996 (.991, 1.000)	99.5 (98.9, 99.9)	90.9	90.0	100.0 (100.0, 100.0)	99.5 (98.9, 99.9)	81.8 (63.6, 95.5)	100.0 (100.0, 100.0)
Chest	1.000 (1.000, 1.000)	99.9 (99.6, 100.0)	99.5	99.5	100.0 (100.0, 100.0)	99.8 (99.5, 100.0)	99.1 (97.3, 100.0)	100.0 (100.0, 100.0)
Chest-abdomen-pelvis	.999 (.996, 1.000)	99.7 (99.3, 100.0)	96.8	93.8	93.8 (81.2, 100.0)	99.9 (99.6, 100.0)	93.8 (81.2, 100.0)	99.9 (99.6, 100.0)
Abdomen	.993 (.987, .999)	98.7 (97.7, 99.5)	92.6	87.8	90.0 (80.0, 97.5)	99.2 (98.5, 99.7)	85.7 (73.8, 95.2)	99.4 (98.9, 99.9)
Abdomen-pelvis	.999 (.996, 1.000)	99.2 (98.5, 99.7)	99.6	96.3	92.9 (87.1, 97.6)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	99.1 (98.4, 99.7)
Pelvis	1.000 (1.000, 1.000)	100.0 (100.0, 100.0)	100	100	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Axis								
Axial	.999 (.996, 1.000)	99.6 (99.1, 100.0)	99.6	99.6	99.8 (99.3, 100.0)	99.4 (98.5, 100.0)	99.5 (98.8, 100.0)	99.7 (99.1, 100.0)
Coronal	1.000 (1.000, 1.000)	99.3 (98.7, 99.9)	99.6	98.6	97.2 (94.3, 99.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	99.1 (98.3, 99.8)
Sagittal	1.000 (1.000, 1.000)	99.5 (98.9, 99.9)	98.7	98.7	100.0 (100.0, 100.0)	99.3 (98.7, 99.8)	97.4 (94.9, 99.4)	100.0 (100.0, 100.0)
IV Contrast								
No	.993 (.987, .999)	97.5 (96.4, 98.5)	97.4	97.4	99.1 (98.0, 100.0)	96.0 (93.8, 97.8)	95.6 (93.4, 97.5)	99.2 (98.2, 100.0)
Yes	.993 (.987, .999)	97.5 (96.4, 98.5)	97.4	97.6	96.0 (93.8, 97.8)	99.1 (98.0, 100.0)	99.2 (98.2, 100.0)	95.6 (93.4, 97.5)

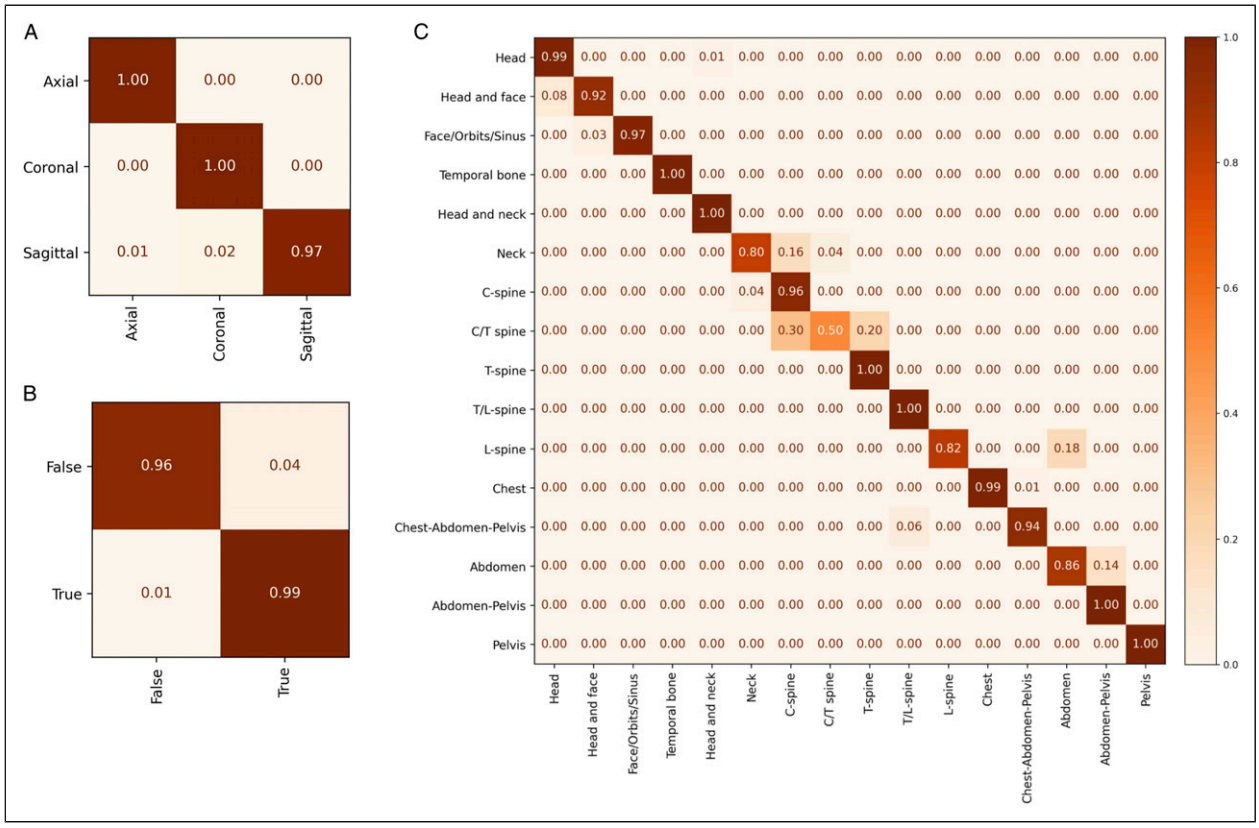

Figure 2. Confusion matrices for all tasks in the test dataset. (A) axis of imaging, (B) presence of intravenous contrast, and (C) body part. The *x*-axis represents the predicted labels while the *y*-axis indicates the ground truth.

Table 4. Evaluation Results from External Validation. The Numbers in Parentheses Indicate the 95% Confidence Interval.

Dataset	Number of Series	Body Part Accuracy (%)	Axis Accuracy (%)	IV Contrast Accuracy (%)	Multi-Class Accuracy (%)	All Labels Accuracy (%)
SPIE-AAPM Lung CT Challenge	70	97.1 (92.9, 100.0)	98.6 (95.7, 100.0)	98.6 (95.7, 100.0)	94.3 (88.6, 98.6)	98.1 (96.2, 99.5)
Stagell-Colorectal-CT	230	97.0 (94.8, 98.7)	100.0 (100.0, 100.0)	94.3 (91.3, 97.0)	91.3 (87.8, 94.8)	97.1 (95.8, 98.3)
C4KC-KiTS	411	95.1 (92.9, 97.1)	100.0 (100.0, 100.0)	97.8 (96.4, 99.0)	93.2 (90.5, 95.4)	97.6 (96.8, 98.5)
CPTAC-LSCC	146	93.8 (89.7, 97.9)	100.0 (100.0, 100.0)	95.2 (91.8, 98.6)	89.0 (84.2, 93.8)	96.3 (94.5, 97.9)
RSNA ICH train	21,615	97.8 (97.6, 98.0)	99.7 (99.6, 99.8)	88.3 (87.8, 88.7)	86.0 (85.6, 86.4)	95.3 (95.1, 95.4)
RSNA ICH test	3502	97.5 (96.9, 98.0)	99.7 (99.5, 99.9)	87.8 (86.8, 88.8)	85.2 (84.2, 86.4)	95.0 (94.6, 95.4)
RSNA C-spine	2019	89.7 (88.5, 91.0)	99.0 (98.6, 99.4)	92.4 (91.3, 93.6)	82.3 (80.6, 83.9)	93.7 (93.1, 94.3)
RSNA CTPA	7279	97.8 (97.5, 98.1)	99.4 (99.2, 99.5)	94.9 (94.4, 95.4)	92.3 (91.7, 92.9)	97.4 (97.2, 97.6)
Pooled dataset	35,272	97.3 (97.1, 97.5)	99.6 (99.5, 99.7)	90.0 (89.8, 90.4)	87.2 (86.8, 87.5)	95.6 (95.5, 95.8)

internal and external testing despite the greater number of body parts and variety of imaging axes, kernels, and phases of post-contrast imaging.

When the model failed at predicting the body part it typically occurred under 2 conditions. The first is when the ground truth and predicted body part anatomy overlapped such as “C spine” and “C and T spine”. This misclassification is likely accounted for by variations in *Z*-axis scan coverage which is manually defined by a CT technologist at the time of imaging. For example, a technologist may extend the field of view of a cervical spine CT to the mid-thoracic spine, or a substantial portion of the chest may be included in an abdominal CT. The second condition is when the same anatomy was imaged but with different acquisition parameters. For example, a CT of the cervical spine is optimized to assess bony structures and uses different scanning parameters than a CT of

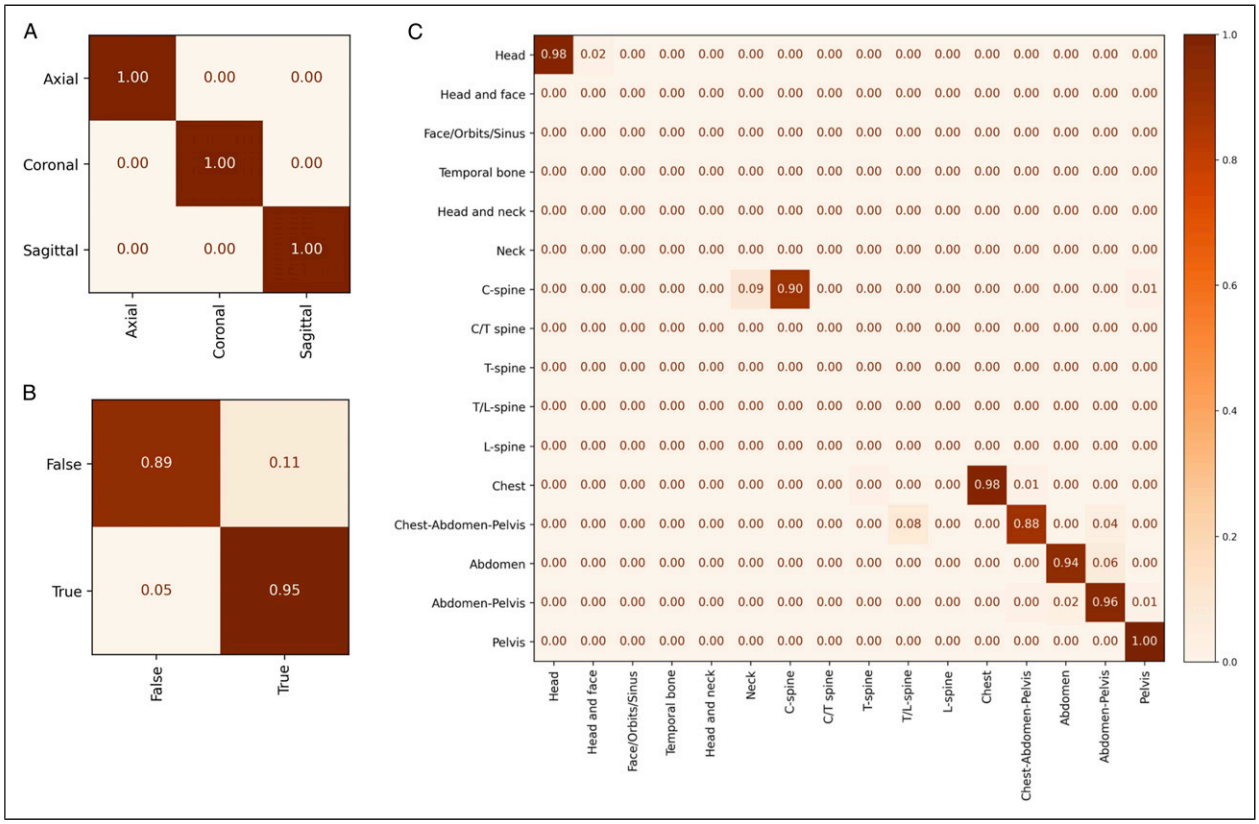

Figure 3. Confusion Matrices from the pooled external validation datasets. (A) axis of imaging, (B) presence of intravenous contrast, and (C) body part. The *x*-axis represents the predicted labels while the *y*-axis indicates the ground truth.

the neck which is optimized for soft tissue structures even though they cover mostly the same anatomical area. High performance may be possible if related labels were collapsed but we felt granular labels would better filter data for use by dataset curators.

The model demonstrated high performance in imaging axis classification on both internal and external validation. The model for IV contrast detection showed an AUC of .99 on the internal test set and ranged between .96 and 1.0 on external validation. This performance is slightly higher than a recent study that reported models with AUCs ranging between .95 and 1.0.[5] Moreover, that study included studies from 2 body parts each with a single phase of contrast enhancement, whereas our dataset included 15 phases and 16 body part labels.

External validation was performed on 7 publicly available datasets with representation of multiple body parts, kernels, and non/post-contrast imaging. External datasets ranged in size from 70 to 21,615 series with a pooled dataset size of 35,272 series. Models generalized well with high performance in all 3 tasks. Like the internal validation, incorrect body part predictions typically occurred when the ground truth label and prediction were similar. Examples include "Head" and "Head and face", "C-spine" and "Neck", and "Abdomen" and "Abdomen-Pelvis". The lower accuracy in detecting IV contrast on the RSNA ICH dataset is accounted for by the high prevalence of studies with ICH which can resemble IV contrast within cerebral blood vessels.

The automated classification of CT series has uses beyond dataset curation. Similar problems can exist during clinical deployment in correctly identifying appropriate imaging studies to serve as model input. With deployed models, there is no longer any realistic possibility for manual study assignment and there are significant consequences of failing to apply the model to a valid imaging study or applying it to an invalid one. ML models are inherently fragile when exposed to data dissimilar to their training data, so ensuring this does not occur is a necessary practical step before safe deployment can occur. In addition, the automated classification of CT series could help optimize image display organization in PACS workstations. Hanging protocols based on DICOM metadata are unreliable and radiologists are forced to constantly spend unproductive time organizing their work environments.[19,20] ML models hold promise in improving the reliability and robustness of hanging protocols when compared to using DICOM metadata.[21,22] In addition, automatic series classification could potentially map similar series and protocols to a standardized schema and help facilitate the curation of multi-institutional datasets. Accurate series classification may also result in more accurate clinical

ML outputs by preventing suboptimal or incorrect series from being sent to ML models, thereby potentially improving patient care.

This study has several limitations. The dataset spanned over a 10-year period during which CT technology consistently improved. It is possible that more remote CT scans may be less relevant for model development given the trend towards thinner acquisitions and recent innovations such as dual energy CT. While the data used to train our models was generated from 3 different CT scanners, they were from a single manufacturer. A broader representation of scanners and manufacturers may further improve model generalizability.

Conclusion

ML models can accurately classify a CT series along 3 important parameters (body part, imaging axis, and IV contrast enhancement) commonly used in model development and deployment. The models demonstrated efficacy and generalizability on a large-scale pooled dataset of numerous well-curated external datasets covering all commonly used study types and are ready for deployment in data processing pipelines.

Acknowledgments

The authors would like to acknowledge the contributions and support of Derek Beaton, Blair Jones, Kate MacGregor, William Parker, and Colin Purcell.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Errol Colak received funding from the Odette Professorship in Artificial Intelligence for Medical Imaging, St. Michael's Hospital, Unity Health Toronto.

ORCID iDs

Hui-Ming Lin https://orcid.org/0000-0002-9598-3921
Alexis Murari https://orcid.org/0009-0009-1550-6147
Eric Ivanov https://orcid.org/0009-0008-7069-0453
Suvd Zulbayar https://orcid.org/0009-0008-0217-9629

Supplemental Material

Supplemental material for this article is available online.

References

1. Thirumuruganathan S, Tang N, Ouzzani M, Doan A. Data curation with deep learning. *InEDBT*. 2020:277-286.
2. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295(1):4-15. doi:10.1148/radiol.2020192224.
3. Gueld MO, Kohnen M, Keysers D, et al. Quality of DICOM header information for image categorization. In: Siegel EL, Huang HK, eds. *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*. ■■■: SPIE; 2002. doi:10.1117/12.467017 AQ2
4. Na S, Sung YS, Ko Y, et al. Development and validation of an ensemble artificial intelligence model for comprehensive imaging quality check to classify body parts and contrast enhancement. *BMC Med Imag*. 2022;22(1):87. doi:10.1186/s12880-022-00815-4
5. Ye Z, Qian JM, Hosny A, et al. Deep learning-based detection of intravenous contrast enhancement on CT scans. *Radiol Artif Intell*. 2022;4(3):e210285. doi:10.1148/ryai.210285
6. Ouyang Z, Zhang P, Pan W, Li Q. Deep learning-based body part recognition algorithm for three-dimensional medical images. *Med Phys*. 2022;49(5):3067-3079. doi:10.1002/mp.15536
7. Chiang CH, Weng CL, Chiu HW. Automatic classification of medical image modality and anatomical location using convolutional neural network. *PLoS One*. 2021;16(6):e0253205. doi:10.1371/journal.pone.0253205
8. Manabe K, Asami Y, Yamada T, Sugimori H. Improvement in the convolutional neural network for computed tomography images. *Appl Sci (Basel)*. 2021;11(4):1505. doi:10.3390/app11041505
9. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. Published online; 2017. doi:10.48550/ARXIV.1708.07632
10. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge. *Radiol Artif Intell*. 2020;2(3):e190211. doi:10.1148/ryai.2020190211
11. RSNA 2022 Cervical spine fracture detection. Kaggle. https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection. Accessed September 17, 2022.
12. Colak E, Kitamura FC, Hobbs SB, et al. The RSNA pulmonary embolism CT dataset. *Radiol Artif Intell*. 2021;3(2):e200254. doi:10.1148/ryai.2021200254
13. Armato III, Samuel G, Hadjiiski L, et al. *SPIE-AAPM-NCI Lung Nodule Classification Challenge Dataset*. ■■■: UZLSU3FL. Published online 2015. doi:10.7937/K9/TCIA.2015
14. Tong T, Li M. Abdominal or pelvic enhanced CT images within 10 Days before surgery of 230 patients with stage II Colorectal Cancer; 2022. Published online. doi:10.7937/P5K5-TG43
15. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Lung Squamous Cell Carcinoma collection (CPTAC-LSCC). Published online 2018. doi:10.7937/K9/TCIA.2018.6EMUB5L2
16. Heller N, Sathianathen N, Kalapara A, et al. C4KC KiTS challenge Kidney Tumor segmentation dataset. Published online 2019. doi:10.7937/TCIA.2019.IX49E8NX

17. Gauriau R, Bridge C, Chen L, et al. Using DICOM metadata for radiological image series categorization: A feasibility study on large clinical brain MRI datasets. *J Digit Imag*. 2020;33(3): 747-762. doi:10.1007/s10278-019-00308-x
18. Raffy P, Pambrun JF, Kumar A, et al. Deep learning body region classification of MRI and CT examinations. *J Digit Imag*. 2023. Published online 9 March 2023. doi:10.1007/s10278-022-00767-9
19. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol*. 2021;28(9): 1225-1235. doi:10.1016/j.acra.2020.01.012
20. Moise A, Atkins MS. Design requirements for radiology workstations. *J Digit Imag*. 2004;17(2):92-99. doi:10.1007/s10278-004-1003-9
21. Filice RW, Stein A, Pan I, Shih G. Federated deep learning to more reliably detect body part for hanging protocols, relevant priors, and workflow optimization. *J Digit Imag*. 2022;35(2): 335-339. doi:10.1007/s10278-021-00547-x
22. Kitamura G. Hanging protocol optimization of lumbar spine radiographs with machine learning. *Skeletal Radiol*. 2021;50(9): 1809-1819. doi:10.1007/s00256-021-03733-8