Toward understanding deep learning classification of anatomic sites: lessons from the development of a CBCT projection classifier

Juan P. Cruz-Bastida,[a,b,*] **Erik Pearson**,[b] **and Hania Al-Hallaq**[b]
[a]University of Chicago, Department of Radiology, Chicago, Illinois, United States
[b]University of Chicago, Department of Radiation and Cellular Oncology, Chicago, Illinois, United States

Abstract

Purpose: Deep learning (DL) applications strongly depend on the training dataset and convolutional neural network architecture; however, it is unclear how to objectively select such parameters. We investigate the classification performance of different DL models and training schemes for the anatomic classification of cone-beam computed tomography (CBCT) projections.

Approach: CBCT scans from 1055 patients were collected and manually classified into five anatomic classes and used to develop DL models to predict the anatomic class from single x-ray projections. VGG-16, Xception, and Inception v3 architectures were trained with 75% of the data, and the remaining 25% was used for testing and evaluation. To study the dependence of the classification performance on dataset size, training data was downsampled to various dataset sizes. Gradient-weighted class activation maps (grad-CAM) were generated using the model with highest classification performance, to identify regions with strong influence on CNN decisions.

Results: The highest precision and recall values were achieved with VGG-16. One of the best performing combinations was the VGG-16 trained with 90 deg projections (mean class precision = 0.87). The training dataset size could be reduced to ~50% of its initial size, without compromising the classification performance. For correctly classified cases, Grad-CAM were more heavily weighted for anatomically relevant regions.

Conclusions: It was possible to determine those dependencies with a higher influence on the classification performance of DL models for the studied task. Grad-CAM enabled the identification of possible sources of class confusion.

 [DOI: 10.1117/1.JMI.9.4.045002]

Keywords: convolutional neural network; deep learning; transfer learning; explainable AI; gradient weighted class activation map; anatomic classification; cone beam computed tomography.

Paper 21294GRR received Nov. 1, 2021; accepted for publication Jun. 16, 2022; published online Jul. 25, 2022.

1 Introduction

Cone beam-computed tomography (CBCT) is the most commonly used three-dimensional image-guided radiotherapy (IGRT) modality.[1] Because it enables imaging of the patient on the treatment table, it is used on a daily or weekly basis to ensure accurate positioning of the target and sufficient avoidance of the nearby organs-at-risk prior to delivering radiation.[1] The scanning protocol used for a CBCT acquisition is selected manually at the time of the scan and is typically determined by the general anatomic location of the scan. Automatically identifying the anatomic location of a CBCT scan from a single x-ray projection image using deep learning (DL) methods could be the first step in developing a framework to guide CBCT protocol selection to ensure

*Address all correspondence to Juan P Cruz-Bastida, cruzbastida@uchicago.edu

2329-4302/2022/$28.00 © 2022 SPIE

consistent image quality and proper patient exposure. It could also be used to alert the treatment team to image quality degradations prior to the completion of the scan. Additional considerations, such as patient size and presence of implants (e.g., hip prothesis, pacemakers), could be added to the framework to increase robustness; for example, an average-sized prostate cancer patient with a hip prothesis may benefit from a more penetrating CBCT protocol.

The implementation of DL methods to CBCT data introduces methodological questions that have not been addressed to the best of our knowledge. For general image classification tasks, the performance has been shown to depend in a complex manner on the CNN architecture[2] and dataset size.[3] Many standard classifier architectures, including VGG-16,[4] are developed for and trained with ImageNet data,[5] which contains some 14 million 224×224 RGB color photographs of objects belonging to 1000 classes. In comparison, the CBCT images are 1024×768, grayscale radiographs of a 20-cm section of patients in standard treatment positions. The visual information, content, and similarity between images may be substantially different between CBCT and ImageNet datasets. Furthermore, each CBCT scan contains hundreds of projections acquired approximately every 0.4 deg as the imaging system rotates around the patient. For the task of utilizing anatomic classification to guide acquisition protocols or to identify potential exposure errors, the classification should utilize the first projection images regardless of the orientation of the imaging system. Experience suggests that human observers find classifying images from arbitrary, oblique angles more challenging than standard AP or lateral projections.[6,7]

Alternatively, DL models are often a subject of epistemological criticism due to lack of interpretability of the CNN performance.[8,9] The interpretability of CNNs is especially important in medical applications, due to the impact of medical tasks on patient health. To address these concerns, several techniques have been developed to explain or visualize key elements involved in the CNN decision process;[10,11] however, such techniques are not always integrated to DL model development.

Therefore, the objective of this work is to study the impact of (1) CNN architecture, (2) training dataset size, and (3) projection angle on the development of DL solutions for anatomic classification of CBCT projection images, and (4) to investigate the CNN performance with aid of interpretability tools. By investigating each of these matters on the straightforward task of classifying CBCT projection images, we hope to gain insight into the strengths and limitations of the DL models for medical imaging applications.

2 Methods

2.1 *CBCT Data Acquisition and Preprocessing*

All anonymized projection images from routine clinical radiotherapy CBCT scans from 1055 patients were collected over 24 nonconsecutive months. Because each patient can contribute scans from multiple treatment fractions, this resulted in a total of 6850 scans. The retrospective use of the data for this study was reviewed and approved by the institutional review board. All scans were acquired on TrueBeam linear accelerators (Varian Medical Systems, Palo Alto, California) using routine clinical protocols named for the anatomic class for which they are intended (see Appendix A). While some institutions manually alter protocol parameters for each patient, this is not our institutional practice.

Raw projections from 0 deg to 360 deg, at 45 deg intervals, were log normalized, then truncated to a given pixel value range (from 0.5 to 7.5) and mapped to 8-bit PNG images, using a MATLAB routine. The dynamic range was selected to include the majority of pixel values across the anatomic sites while excluding those outside the patient (i.e., <0.5). PNG images were used for labeling (Sec. 2.2) and for CNN training (Sec. 2.4). As others have done for medical image applications,[12,13] PNG images were rebinned to match the CNN architecture input size.[14]

2.2 *Image Labeling*

CBCT scans were manually classified into five anatomic classes by a medical physicist with 20 years of experience in the practice of radiotherapy. A MATLAB interface was used to review

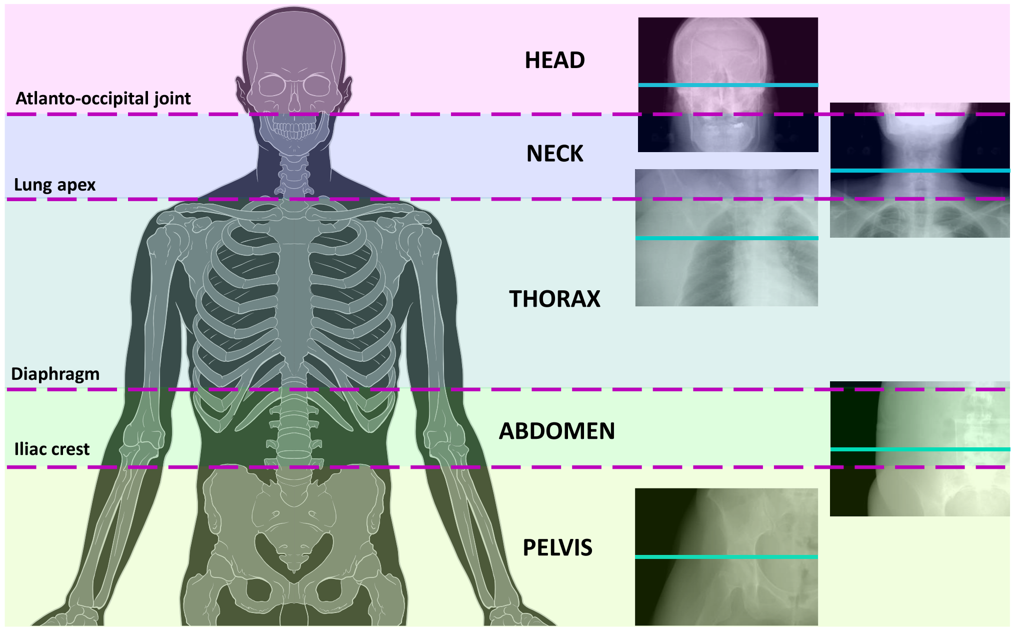

Fig. 1 Examples of projection images in each of the five classes with the mid-plane marked by a cyan line. Magenta lines indicate the boundaries between classes. Upper and lower anatomic landmarks used to determine the boundaries for each of the classes are indicated.

all available projection images from 0 deg to 360 deg at 90 deg intervals, providing up to five projections per scan. Orthogonal views were used since they are more intuitive for a human observer and are thus standard in IGRT clinical workflow. Examples of projection images in each of the five classes with the midplane marked by a cyan line are shown in Fig. 1. The class selection was based on the mid-plane location relative to anatomical boundary lines. Upper and lower boundaries for each class are also listed in Fig. 1.

2.3 *Data Curation*

There can be significant variability in the number of CBCT scans each patient receives based on the fractionation schedule and/or image frequency determined by their treatment protocol, e.g., a patient treated with hypofractionated stereotactic body radiotherapy may have only a few scans, whereas a prostate patient may have 10 times as many.[6] This poses several challenges: (a) the similarity of the images from fraction to fraction may result in data that is highly redundant and (b) an imbalance in the proportion of data between classes may result. To mitigate this issue, a maximum number of scans per patient were included in the dataset used for model development: from 2 to 10 depending on the anatomic class. Ideally, 10 scans per patient would have been included for all classes but this was not possible for hypofractionated or single-fraction (i.e., stereotactic brain radiosurgery) treatments.

From each CBCT scan, projection images at every 45 deg (from 0 deg to 360 deg) were gathered. Projection data were grouped by view angle, resulting in eight different subsets. Each subset was divided as follows: 75% to train the model and 25% for testing and evaluation (equally partitioned), with no overlap of patients across training, testing, and evaluation datasets. Table 1 shows the resulting number of images per angular subset for each anatomic class as well as the total number of patients contributing to each class. Note that a range of images is given for head and neck classes because such scans were typically acquired with a head protocol, which utilizes a short scan (i.e., scan range ≈ 200 deg) resulting in an angularly asymmetric dataset size.

2.4 *DL Model Development*

Three CNNs were built using the Keras Deep Learning[15] library with a TensorFlow[16] backend engine (version 2.1.0, TensorFlow) to predict the anatomic class from a single projection image,

Table 1 Total number of patients and images in the training, testing, and evaluation datasets for each anatomic class, given a projection angle. A range is reported for classes imaged with a shortened CBCT scan rotation (<360 deg).

Class	Number of patients	Number of projections		
		Training	Testing	Evaluation
Head	117	[90, 280]	[20, 50]	[20, 50]
Neck	268	[100, 370]	[20, 70]	[20, 70]
Thorax	213	300	50	50
Abdomen	169	340	60	60
Pelvis	288	410	70	70

given a fixed projection angle. These architectures, VGG-16,[4] Xception,[17] and Inception v3,[18] were selected because they have been proven to be effective in the classification of natural[19–21] and medical images.[12,20,22,23] The classification layer of each network was adjusted to the number of classes in our classification task. All convolutional layers used rectified linear unit (ReLU) activation functions. CNNs were trained using the stochastic gradient descent optimizer, with a learning rate of 0.001, and transfer learning; that is, using the weights of a pretrained CNN (with the ImageNet dataset) as initial weights. Training was performed in two stages: (1) a warm-up stage (50 epochs), where all the weights of the networks were kept fixed, except for those of the last layer (classification layer), and (2) the training stage (100 epochs), where all weights were updated. Keras built-in data augmentation tools were used to prevent overfitting, allowing shifts (up to 10%) and zoom (up to 20%).

To study the dependence of the classification performance on dataset size, the training subset was downsampled to 5%, 10%, 15%, 20%, 30%, 50%, and 70% of its original size. For each training scheme, CNN training and evaluation was repeated 10 times to obtain the performance statistics. CNN training was performed with an NVIDIA GeForce RTX 2070 GPU and required about 1 h per repetition.

2.5 *Quantification of Classification Performance*

The classification performance of the DL model was quantified in terms of precision (i.e., positive predictive value) and recall (i.e., sensitivity). A precision score of 1.0 means every item labeled as belonging to a class C does indeed belong to class C, whereas a recall of 1.0 means every item from class C was labeled as belonging to class C.

2.6 *Visualization of Features Identified by CNNs*

Gradient-weighted class activation maps (grad-CAM)[24] were used to identify key features involved in the CNN classification process. A number of methods have been developed for visualizing pixel attribution (saliency maps), i.e., the regions of an image that are relevant for a CNN classification.[10,11,25] Grad-CAM was used here as it has been previously shown to be useful in the interpretation of medical image classification models.[10,11,25] Grad-CAM of images in the evaluation dataset were generated using the DL model with the highest precision and recall values. In brief, Grad-CAM illustrates the relative spatial activation of the final CNN layer (before classification) with respect to the network output. Grad-CAM were computed as ReLU activations of a weighted sum of feature maps obtained from the final CNN layer, where sum weights were given by gradients of the class score with respect to each feature map.[24] The generated Grad-CAM (evaluation dataset) were visually inspected by a human observer, to identify regions of the image (if any) with a strong influence in the CNN decision.

3 Results

3.1 *Dependence of the Classification Performance on CNN Architecture*

Table 2 shows the precision and recall mean values obtained from 10 repetitions for the three CNN architectures. In this case, the models were trained using 90 deg projections [i.e., antero-posterior (AP) orientation for a typical head-first supine patient positioning], with a 75%/25% split between training and testing/evaluation datasets. These results show that the highest precision and recall averaged over all classes were achieved with VGG-16. Results from other view angles similarly showed VGG-16 had the highest performance. Therefore, only VGG-16 is considered hereafter. The head and neck classes showed high precision values regardless of the CNN architecture.

3.2 *Dependence of the Classification Performance on X-Ray Projection Angle*

Next, the ability of VGG16 to classify the projections from different projection angles at 45 deg intervals was evaluated by independently retraining and testing the network as described in the preceding section. Figure 2 shows precision (magenta) and recall (cyan) mean values obtained from 10 repetitions as polar plots at various projection angles. The dataset size, normalized to the maximum number of images per anatomic class, is also displayed in yellow to demonstrate the angular symmetry of the data. The highest precision ranged from 1.00 to 0.76 across classes, whereas the lowest precision ranged from 0.93 to 0.66. As for recall, the highest and lowest values ranged from 0.98 to 0.81 and 0.87 to 0.66, respectively. In most cases, the highest precision/recall values were obtained for 90 deg or 135 deg. Even though the head and neck dataset sizes were angularly asymmetric due to the short-scan CBCT acquisition, the classification performance for these classes does not reflect this asymmetry, which suggests that the performance of the model was not data limited.

3.3 *Dependence of the Classification Performance on Training Dataset Size*

The final factor considered in this study was the size of the training dataset. The VGG-16 CNN architecture was used with 90 deg projection images and the training data set size was down-sampled as described in Sec. 2.4. Figure 3 shows precision and recall mean values obtained from 10 repetitions for different training dataset sizes and demonstrates that the classification performance plateaus around 750 images (∼150 images per class), which corresponds to 50% of the

Table 2 Classification precision and recall for 90 deg projection images averaged over 10 repetitions per anatomic class for three CNN architectures using a 75%/25% split between training and testing/evaluation data. Error estimates correspond to standard deviation values, except for the class-mean, for which error estimates correspond to the square sum of standard deviation values from all classes.

	Xception		Inception v3		VGG-16	
	Precision	Recall	Precision	Recall	Precision	Recall
Head	0.97 ± 0.03	0.76 ± 0.08	0.99 ± 0.03	0.76 ± 0.27	0.96 ± 0.03	0.97 ± 0.00
Neck	0.91 ± 0.04	0.94 ± 0.02	0.98 ± 0.00	0.48 ± 0.19	1.00 ± 0.01	0.98 ± 0.01
Thorax	0.71 ± 0.07	0.16 ± 0.04	0.49 ± 0.29	0.09 ± 0.03	0.73 ± 0.06	0.78 ± 0.08
Abdomen	0.55 ± 0.03	0.92 ± 0.04	0.36 ± 0.04	0.96 ± 0.11	0.76 ± 0.07	0.73 ± 0.09
Pelvis	0.83 ± 0.06	0.80 ± 0.04	0.93 ± 0.14	0.59 ± 0.36	0.89 ± 0.03	0.87 ± 0.07
Mean	0.80 ± 0.11	0.72 ± 0.11	0.75 ± 0.33	0.58 ± 0.50	0.87 ± 0.10	0.86 ± 0.14

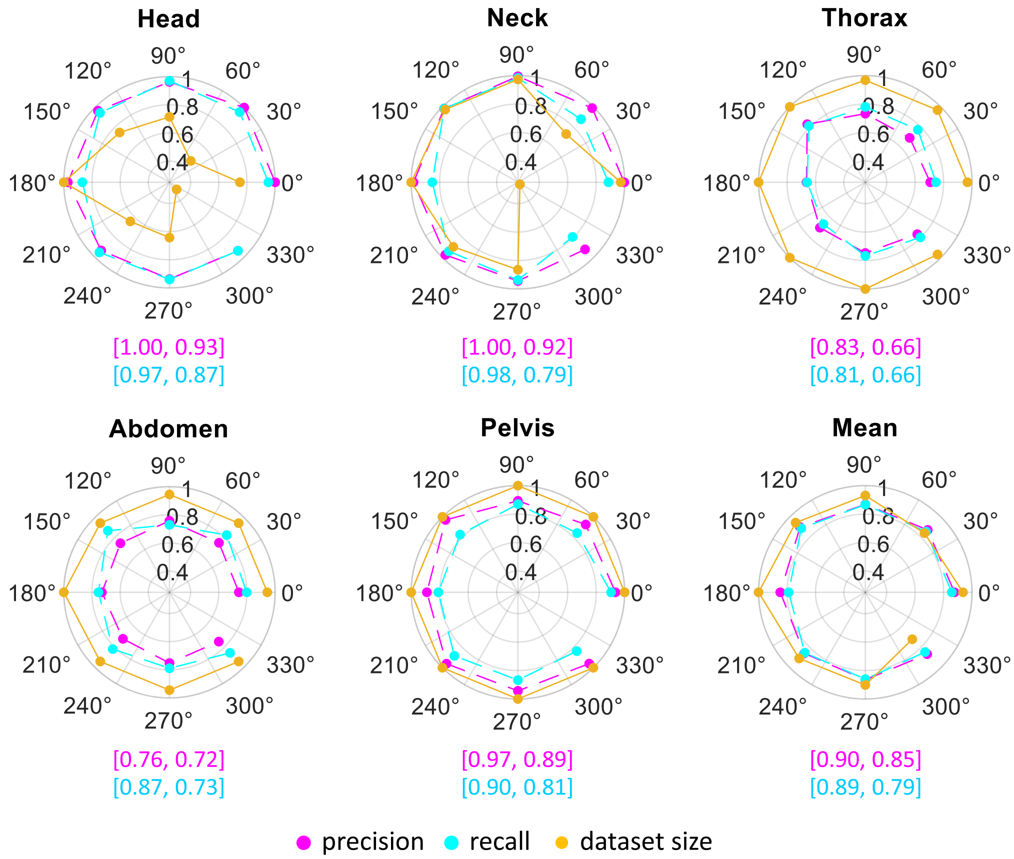

Fig. 2 Mean classification precision (magenta) and recall (cyan) per anatomic class as a function of projection angle calculated using the VGG-16 CNN architecture and a 75%/25% split between training and testing/evaluation. Mean precision and recall ranges are provided below each class subplot. The training dataset size (yellow) was normalized to the maximum number of images per anatomic class. Standard deviation values are between 0.01 and 0.06 for precision and between 0.04 and 0.12 for recall.

original size of the training dataset. These results suggest that, for this classification task and DL model, the dataset size was sufficient for the analysis described in the preceding sections and that this architecture could be used with a smaller dataset without compromising the overall classification performance. Also, note that for head and neck classes, the plateau was reached even faster and with a higher precision, which is consistent with the robustness reported for these classes in the previous subsections.

3.4 *Overall Classification Performance*

From Secs. 3.1 to 3.3, it was observed that one of the best performing combinations was the VGG-16 architecture trained with 90 deg projection images. In addition, it was shown that the training dataset size could be reduced to about 50% of the size of the initial training dataset, without compromising the classification performance. Table 3 shows the confusion matrix for this model implementation, assessed only on the evaluation datasets, using a 50%/50% split between training and testing/evaluation data. However, as discussed in Sec. 3.2, other projection angles achieve similar classification performance, which means that high precision/recall values were not limited to 90 deg projection images.

From Table 3, it can be observed that the classifier has high sensitivity (≥91%) for head, neck, and thorax but was lower for abdomen and pelvis classes. Note that, misclassified images were most frequently confounded with a neighboring class, e.g., abdomen was misclassified most frequently as thorax and pelvis as abdomen.

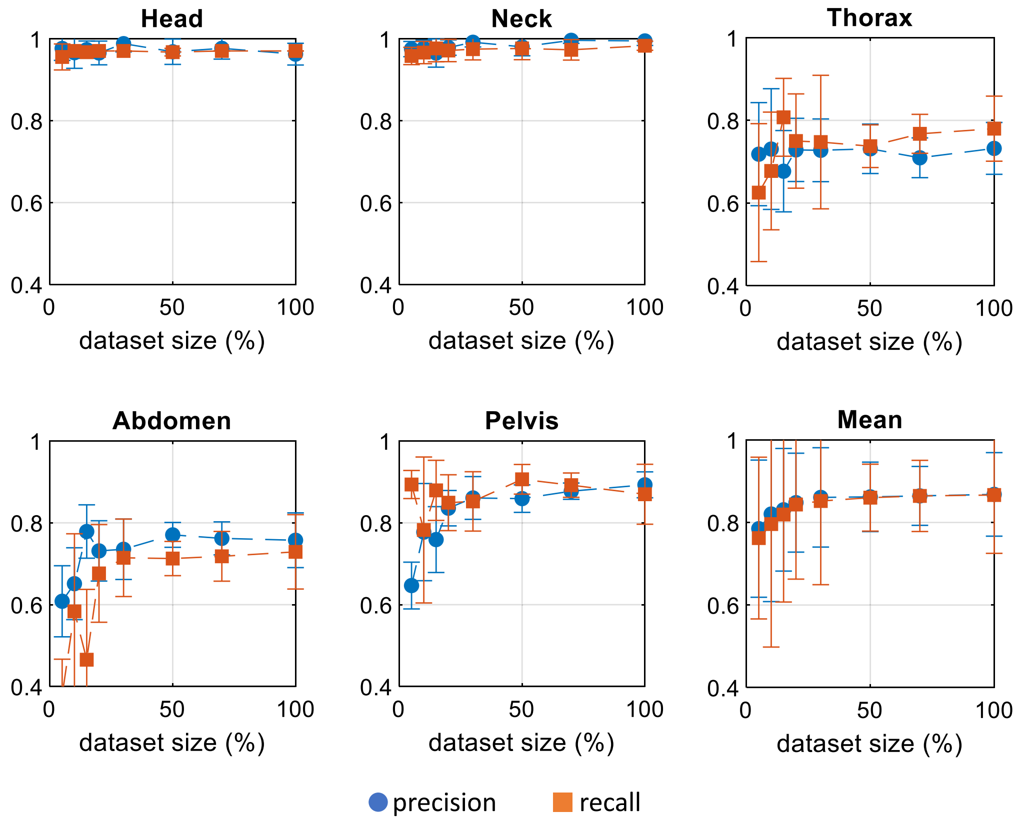

Fig. 3 Mean classification precision (blue) and recall (orange) per anatomic class at 90 deg using VGG-16 as the training dataset size was reduced from the original 75%, and the evaluation and testing datasets sizes were kept constant. Error bars correspond to standard deviation values, except for the class-mean plot, for which error bars correspond to the square sum of standard deviation values from all classes.

Table 3 Confusion matrix for VGG-16 CNN architecture trained with 90 deg (AP projection). Each cell contains the number of images categorized into each class and its corresponding percentage, relative to the number of images of a given class (zero values omitted for clarity). In this case, 50% of the data were used for training, while the remaining 50% were equally split into evaluation and testing.

	Prediction					
Class	Head	Neck	Thorax	Abdomen	Pelvis	Total
Head	70	—	—	—	—	70
	100%	—	—	—	—	100%
Neck	1	106	7	—	—	114
	1%	**93%**	6%	—	—	100%
Thorax	—	3	81	3	2	89
	—	3%	**91%**	3%	2%	100%
Abdomen	—	—	13	94	8	115
	—	—	11%	**82%**	7%	100%
Pelvis	—	—	4	11	109	124
	—	—	3%	9%	**88%**	100%

Note: Those events where the prediction matches the nominal class (correctly classified) are highlighted in bold.

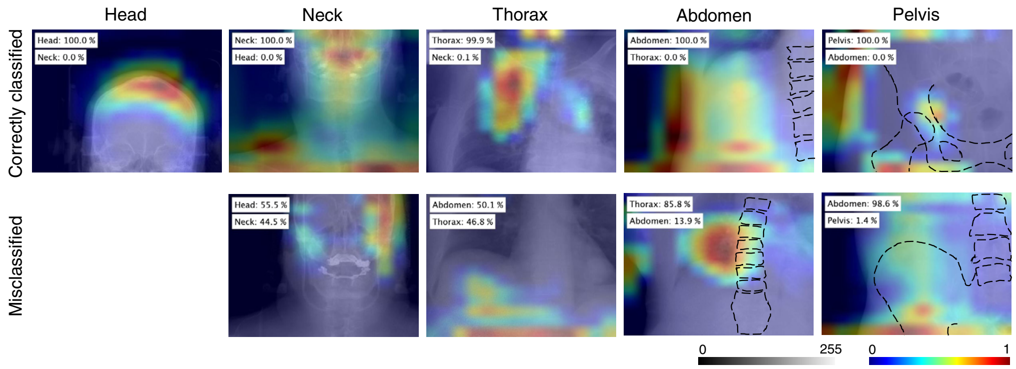

Fig. 4 Representative cases of correctly classified and misclassified 90 deg projection images for each anatomic class; all head projections were correctly classified. The Grad-CAM is displayed in a normalized "jet" color scale, and the projection image is displayed in an 8-bit grayscale. Reference boney structures for abdomen and pelvis images are indicated with dashed lines, for ease for visualization. The two most probable classes predicted by the CNN are indicated with corresponding numerical probability values, in the upper left-hand corner of each image. For this model, VGG-16 was used along with a 50%/50% split of the data for training and testing/evaluation.

3.5 *Visualization of Features Identified by CNNs*

Figure 4 shows representative Grad-CAMs overlayed on their corresponding 90 deg projections. Two examples are shown for each anatomic class: one correctly classified and the other incorrectly classified by the DL model summarized in Sec. 3.4, with an exception in the head class for which all cases in the evaluation dataset were correctly classified. For each image, the two most probable classes predicted by the DL model are indicated on the image along with numerical probability values output by the CNN. These images demonstrate that the activation for correctly classified cases was more heavily weighted for anatomically relevant regions, except for abdomen, where the object of focus was not entirely clear. In the case of head, we identified a higher emphasis on the cranial bones. For neck class, high weights were located close to the mandible and clavicle. For thorax, the highest weights were in the lung. For abdomen, the high weights do not seem to have a correlated landmark in the patient but were located at the borders of the body habitus. This was also true for pelvis, but in this case a high emphasis was additionally identified close to the hip joint. These trends were consistent for most of the reviewed cases (other examples are provided as supplementary material) and suggest that the CNN was learning anatomic features to perform the classification correctly. In general, the CNN did not fixate on image artifacts, metal implants (e.g., hip prostheses), or medical devices (e.g., pacemakers). In some projections, high weights were located at the inferior border of the images as shown by the red bands in some of the images in Fig. 4.

Another interesting feature of Grad-CAM was the weight distribution for misclassified cases. Even though anatomic structures, such as cranial bones or lungs, were also emphasized in misclassified cases, such emphasis does not match with that of correctly classified cases. In other words, while the CNN was still successful in identifying key structures, the network's attention was shifted to a nearby anatomic landmark (i.e., lung identified in an abdomen projection).

4 Discussion

A DL model was developed for the anatomic classification of single x-ray projection images. In summary, the classification performance was evaluated for different CNN architectures, dataset sizes, and projection angles. The results were interpreted with aid of Grad-CAM. The results indicate that a VGG16 architecture can determine the anatomic class of a 90 deg x-ray projection image with sensitivity $\geq$91% for head, neck, and thorax classes, and $\geq$82% for abdomen and pelvis classes. In general, misclassified images were "near misses"; that is, the image was classified into the neighboring classes. In terms of dataset size, the classification performance plateaued around 750 images (~150 images per class), which corresponds to 50% of the original size of the training dataset. The classifier performed marginally better for 90 deg and 135 deg projection angles; however, there was not a substantial nor a systematic change in performance versus angle across classes. Grad-CAM were used to visualize those image features with a higher influence in the CNN performance. Visual inspection of Grad-CAM suggests that the CNN was learning to identify anatomic structures (i.e., cranial bones, lungs, etc.) and patient borders to perform a correct classification.

It is worth emphasizing that the CNN performance metrics reported in this work correspond to mean values of multiple training/evaluation realizations. While this is time consuming, it allowed us to make stronger conclusions about our results, as CNN training is subject to random variations introduced by most error function optimizers.[26] To the best of our knowledge, this is not a common practice and should be adopted when feasible.

It is not clear why VGG16 outperforms the other CNN architectures (Xception and Inception v3), especially since these other architectures are considerably deeper, which is often believed to be beneficial, and show a better classification performance for natural images.[14] A possible explanation could be the number of trainable parameters (determined by both the CNN depth and the size of the convolutional kernels), which is much larger for VGG16 (~138 M) than for Xception and Inception v3 (~22 and 23 M, respectively).[14] If so, it would suggest that the interplay between depth and kernel design is more important than CNN depth itself for this type of classification task and needs to be explicitly investigated.

The number of training images per class for a given view angle was less than 500 (see Table 1), which is relatively small compared with reference datasets used to train models from scratch.[5,13] Therefore, transfer learning was utilized in this work. CNN hyperparameters, such as the optimizer, number of epochs, etc., were arbitrarily selected and kept fixed in this study. Given that CCN weights were initialized with pretrained values, it is unlikely that hyperparametric optimization would lead to major performance improvements[27] but this would need to be explicitly validated in future work.

Even though the training dataset contained less head and neck images than from any other class, overall, the classification performance for head and neck was the highest in all cases. Also, a faster convergence was observed for these two classes when varying the training dataset size. A possible explanation could be that head and neck have smaller variations in size and shape across the population, compared with other classes. If this is the case, it would mean that adequate class balancing, for applications where settings can be controlled, should be weighted by intraclass variability. This will be subject of future work.

The Grad-CAM proved to be useful to interpret CNN performance and to identify potential sources of class confusion. Based on our results, most of the misclassified cases occurred when the mid-plane of a given image was close to class boundaries, in which case, the attention of the CNN seemed to arbitrarily shift to anatomic features with a higher importance for neighboring classes. While undesirable, this error is understandable as it also observed for human readers. This type of analysis made it easier to understand the limitations of a DL model and to make decisions; for example, to improve the accuracy of our model, it would be reasonable to expand the training data set with cases (correctly labeled) where the mid-plane was close to anatomic boundaries. Such a hypothesis would not be immediately obvious by training the model as a "black box."

In several cases, Grad-CAM showed bands of higher weights at the inferior borders of the images (see Fig. 4). While these features do not have an obvious interpretation, they could be due to subtle correlations in the training data, which may not be indicative of the class globally.[13,28,29] For example, it could be possible that image edges contain fingerprints of the hardware where the patient was scanned, which may correlate with the anatomic class since patients of the same treatment site are often grouped on the same treatment machine. Retrospective inspection of the projection headers ruled out the presence of imaging blades at the border of the projection images. While there may not be an obvious landmark at the inferior image borders, the CNN could be identifying spurious features that were common to the images in our data sets. Others have shown similar findings noting that this may make translation to other data sets less effective.[13,28] Thus, to increase generalizability, the model development would, ideally, include multi-institutional data to reduce the impact of hardware or site-specific factors.

Limitations of our study were identified and are discussed as follows. The transformation of raw projections to 8-bit PNG images (see Sec. 2.1) and subsequent data rebinning reduced the image quality of projection data, which could have restrained the classification performance of the developed models.[30,31] Although utilizing alternate dynamic range settings (e.g., bone window) when transforming projection images to PNGs had a small effect on CNN performance (∼3%), the effect of dynamic range selection on performance should be further investigated. Training the CNN with multiple views, instead of a single projection view as we chose to do in this work, should also be further investigated. The manual classification of projection images was performed by a single individual and errors in classification did occur, although infrequently, which affected the definition of ground truth. Each of the studied parameters was evaluated independently, regarding the impact on model performance, future work will evaluate the interdependence of these parameters, as well as additional characteristics that may impact the classification performance; for example, those associated with data augmentation and error function optimization. Another limitation is the number of CNN architectures tested. The use of a single method (Grad-CAM) to explain the CNN performance is also a limiting factor; future work will focus in corroborating the consistency of Grad-CAM with other visualization techniques.[10]

Even though our motivation and final goal is to use DL techniques to automate the selection of CBCT acquisition and reconstruction parameters for IGRT applications, our conclusions are not limited to CBCT or IGRT. Automatic anatomic classification could be a relevant task for other imaging techniques based on two-dimensional x-ray projections, such as conventional radiography.[32]

5 Conclusion

DL applications strongly depend on the characteristics of the training dataset and CNN model; however, quite often, it is unclear how to objectively select such parameters. This work thoroughly investigates the classification performance of different DL models and training schemes, given an anatomic classification task. From the results, it was possible to determine and better understand those dependencies with a higher influence on the classification performance for this type of task. The use of Grad-CAM enabled identification of possible sources of class confusion, which could help to improve future developments.

6 Appendix A: CBCT Acquisition Protocols

Depending on the anatomic class, different routine clinical protocols were used for CBCT acquisition. These protocols are summarized in Table 4. Further details on acquisition protocols and hardware can be consulted elsewhere.[33]

Table 4 Routine clinical protocols to acquire CBCT data (for IGRT purposes) with TrueBeam systems at our institution.

Mode	Voltage (kV)	Current (mA)
Head	100	20
Thorax	125	20
Pelvis	125	80
Pelvis obese	140	99

Disclosures

HA receives royalties and licensing fees for computer-aided diagnosis technology through the University of Chicago.

Acknowledgments

The authors would like to acknowledge Shumin Li for her contribution to the development of the image-review GUI. This research was funded by Varian Medical Systems. The portion of this work regarding the effect of training dataset size on the classification performance was presented at SPIE Medical Imaging 2022 and is the subject of the conference proceedings paper in Ref. 34.

References

1. S. Korreman et al., "The European Society of Therapeutic Radiology and Oncology–European Institute of Radiotherapy (ESTRO–EIR) report on 3D CT-based in-room image guidance systems: a practical and technical review and guide," *Radiother. Oncol.* **94**(2), 129–144 (2010).
2. B. Saiharsha et al., "Evaluating performance of deep learning architectures for image classification," in *5th Int. Conf. Commun. and Electron. Syst. (ICCES)*, pp. 917–922 (2020).
3. C. Luo et al., "How does the data set affect CNN-based image classification performance?" in *5th Int. Conf. Syst. and Inf. (ICSAI)*, pp. 361–366 (2018).
4. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv14091556 Cs (2015).
5. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 248–255 (2009).
6. Z. Chang et al., "6D image guidance for spinal non-invasive stereotactic body radiation therapy: comparison between ExacTrac X-ray 6D with kilo-voltage cone-beam CT," *Radiother. Oncol.* **95**(1), 116–121 (2010).
7. J. Park et al., "Evaluation of the setup discrepancy between 6D ExacTrac and cone beam computed tomography in spine stereotactic body radiation therapy," *PLoS One* **16**(5), e0252234 (2021).
8. X. Jia, L. Ren, and J. Cai, "Clinical implementation of AI technologies will require interpretable AI models," *Med. Phys.* **47**(1), 1–4 (2020).
9. M. Reyes et al., "On the interpretability of artificial intelligence in radiology: challenges and opportunities," *Radiol. Artif. Intell.* **2**(3), e190043 (2020).
10. D. T. Huff, A. J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," *Phys. Med. Biol.* **66**(4), 04TR01 (2021).
11. A. Singh, S. Sengupta, and V. Lakshminarayanan, *Explainable Deep Learning Models in Medical Image Analysis* (2020).
12. K. Mendel et al., "Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography," *Acad. Radiol.* **26**(6), 735–743 (2019).
13. J. Crosby et al., "Deep convolutional neural networks in the classification of dual-energy thoracic radiographic views for efficient workflow: analysis on over 6500 clinical radiographs," *J. Med. Imaging* **7**(1), 016501 (2020).
14. K. Team, "Keras documentation: keras applications," https://keras.io/api/applications/ (accessed 13 September 2021).
15. F. Chollet, *Keras*, GitHub (2015).
16. M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 19 (2015).
17. F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 1800–1807 (2017).
18. C. Szegedy et al., "Rethinking the inception architecture for computer vision," arXiv151200567 Cs (2015).
19. M. Pak and S. Kim, "A review of deep learning in image recognition," in *4th Int. Conf. Comput. Appl. and Inf. Process. Technol. (CAIPT)*, pp. 1–3 (2017).
20. H. Wu, Q. Liu, and X. Liu, "A review on deep learning approaches to image classification and object segmentation," *Comput. Mater. Contin.* **60**(2), 575–597 (2019).
21. Ò. Lorente, I. Riera, and A. Rana, *Image Classification with Classic and Deep Learning Techniques* (2021).

22. B. Sahiner et al., "Deep learning in medical imaging and radiation therapy," *Med. Phys.* **46**(1), e1–e36 (2019).
23. B. J. Erickson et al., "Machine learning for medical imaging," *RadioGraphics* **37**(2), 505–515 (2017).
24. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 618–626 (2017).
25. K. A. Philbrick et al., "What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images," *Am. J. Roentgenol.* **211**(6), 1184–1193 (2018).
26. S. Ruder, "An overview of gradient descent optimization algorithms," arXiv160904747 Cs (2017).
27. R. Ribani and M. Marengoni, "A survey of transfer learning for convolutional neural networks," in *32nd SIBGRAPI Conf. Graph., Patterns and Images Tutorials (SIBGRAPI-T)*, pp. 47–57 (2019).
28. J. P. Cohen et al., "Problems in the deployment of machine-learned models in health care," *CMAJ* **193**(35), E1391–E1394 (2021).
29. J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLOS Med.* **15**(11), e1002683 (2018).
30. L. Zheng et al., "Good practice in CNN feature transfer," arXiv160400133 Cs (2016).
31. C. F. Sabottke and B. M. Spieler, "The effect of image resolution on deep learning in radiography," *Radiol. Artif. Intell.* **2**(1), e190015 (2020).
32. P. H. Yi et al., "Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning," *Pediatr. Radiol.* **49**(8), 1066–1070 (2019).
33. Varian Medical Systems, "On-Board Imager (OBI) advanced imaging reference guide," Document ID B502202R01C, Palo Alto (2013).
34. J. P. Cruz-Bastida, E. Pearson, and H. Al-Hallaq, "Towards understanding the dependencies of deep learning classification of anatomical sites from CBCT x-ray projections," *Proc. SPIE* **12031**, 1203124 (2022).

Juan P. Cruz-Bastida is a postdoctoral scholar at the University of Chicago. He received his PhD in medical physics from the University of Wisconsin-Madison in 2019. His current research interests include the objective assessment of image quality and AI applications to x-ray imaging.

Erik Pearson received his PhD in medical physics from the University of Chicago. His research focuses on: intrafraction kV imaging, electron paramagnetic resonance oximetry, and advanced CT reconstruction methods combined with image acquisition strategies for CBCT imaging in IGRT applications.

Hania Al-Hallaq investigates the use of medical images to inform treatment selection, guide treatment positioning, and assess treatment response following radiotherapy. Her research background in texture analysis and expertise as a clinical radiotherapy physicist has enabled her to contribute significantly to translational cancer research in several institutional and national collaborative efforts.