Using Deep Learning to Aid Brown Fat Detection in 18F FDG PET/CT

Ellen X. Sun, MD, Brigham and Women's Hospital; Christopher P. Bridge, PhD; Ryan C. King, MD; Richard Thomas, MD; Florence C. Ubah; Hyewon Hyun, MD; Katherine P. Andriole, PhD, FSIIM

Background/Problem Being Solved

Brown fat is normal tissue that when activated can exhibit high glucose metabolism and therefore increased uptake on fluorine-18 fluorodeoxyglucose positron emission tomography (^{18}F-FDG PET). As brown fat shares common locations with nodal disease such as in the neck, supraclavicular, axillary, mediastinum, and paravertebral regions, differentiating metabolically active brown fat from malignant or abnormal tissue on oncologic PET/CT can be a difficult and time consuming process.

Interventions

We developed a deep learning model using a 3D UNet convolutional neural network with the segmentation task of classifying each pixel as brown fat or not brown fat. 277 PET/CT studies with reported metabolic active brown fat were retrospectively reviewed by one reader, who used a web-based annotation tool to manually label semi-automatically detected FDG-avid regions as normal uptake, brown fat, malignancy or other. Only image slices from the skull base to diaphragm were examined. 50 randomly selected studies were chosen as a test set for the model and also separately reviewed and annotated by a second reader. 181 studies annotated by reader 1 were used to train the model via supervised learning. The model was then compared to the ground truth annotation by reader 1 and a simpler automatic method based on CT density and PET uptake thresholding. Inter-reader variability was also analyzed.

Barriers/Challenges

Inter-reader variability between the two readers was high. This was attributed to 1) the inherent and subjective uncertainty in the image analysis task, and 2) limitation of the annotation tool that uses a single user-selected PET threshold for an entire image series to delineate regions of high uptake, for which there was considerable disagreement between the thresholds selected by the two readers.

Outcome

Compared to the ground truth, the model yielded an average Dice overlap score of 0.53 and a sensitivity of 59% for detecting brown fat voxels. Review of studies that had low Dice scores showed that the dissimilarity could be attributed to differences in the extent of the semi-automatically detected FDG-avid regions due in turn to different PET thresholds used rather than differences in labels, suggesting that the actual qualitative performance of the model was higher. The image thresholding method had a poorer performance with lower Dice score (0.30) and more frequently mislabeled malignancy as brown fat. Attempts to re-train the model for detecting malignancy yielded a poor Dice score of 0.1, likely resulting from the large variation in pathologies in the training dataset.

Conclusion/Statement of Impact

We demonstrated a proof-of-concept method for brown fat detection on ^{18}F-FDG PET/CT using a UNet convolutional neural network, which can automatically detect approximately 60% of the areas marked as brown fat by a human expert while avoiding areas that represent malignancy. When fully developed, this application has the potential to significantly shorten interpretation time of oncologic PET/CT and decrease diagnostic errors.

Lessons Learned

The manual annotation process is highly subjective. However, inter-reader variation can be reduced by accounting for differences in PET thresholding, suggesting that a more standardized PET threshold is required to refine the model. For specific malignancy detection, a larger and more focused training dataset is needed.

Keywords
deep learning, convolutional neural network, brown fat, PET/CT

Figures and Tables

	Brown Fat Dice Score[#]	Brown Fat Sensitivity[&]	Lesion Confusion Rate[§]
Deep Learning Model	0.53	0.59	0.005
Thresholding Method	0.30	0.47	0.079
Reader 2	0.43	0.48	0.014

Table 1: Comparison of results of the deep learning model, the automatic thresholding-based method, and a second human reader. In each case reader 1's annotations are considered the ground truth.
[#] Dice score is a measure of overlap between the tested method (or annotation by reader 2) and "ground truth" manual annotation by reader 1.
[&] Brown fat sensitivity is the proportion of manually-labeled brown fat voxels that were detected as brown fat by the model, or true positive rate.
[§] Lesion confusion rate is the proportion of manually-labeled lesion/pathology voxels that were mislabeled as brown fat by the model.

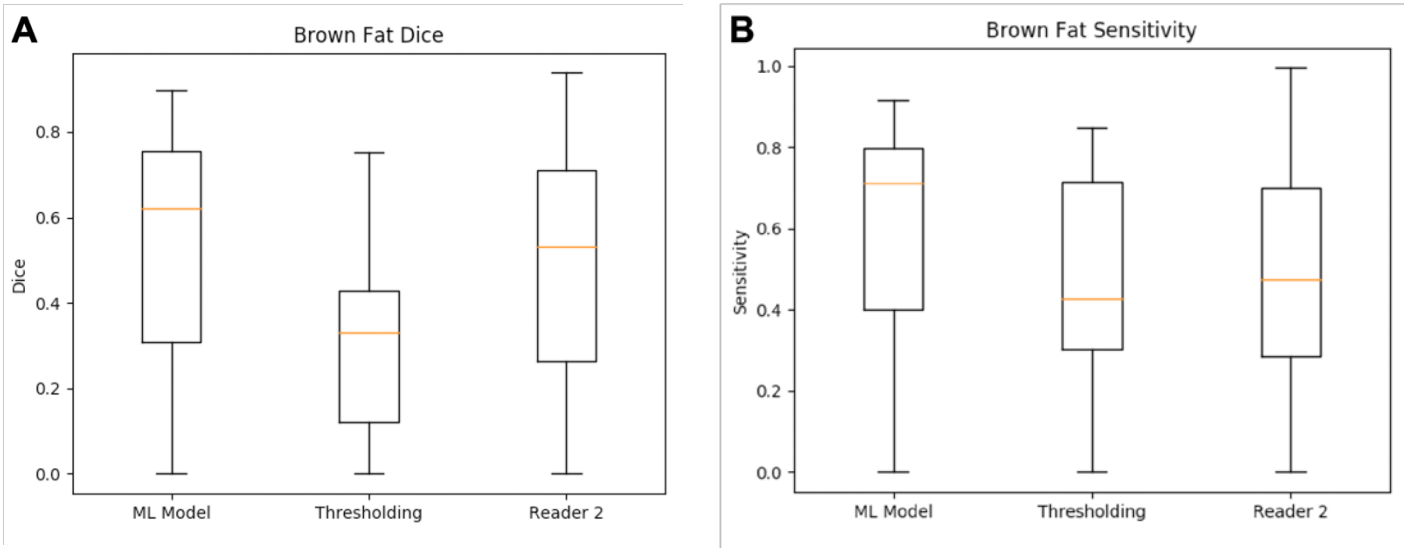

Figure 1: Box-and-whisker plots comparing brown fat dice scores (A), brown fat sensitivity (B), and lesion confusion rate (C) of the deep learning model (ML model), the thresholding-based method, and reader 2. The boxes show the upper and lower quartiles. Horizontal lines within the boxes represent the median values. The whiskers represent the maximum and minimum values, with the exception of outliers which are depicted by the plotted circles. An outlier is defined as a value less than the lower or upper quartile by more than 1.5 times the interquartile range.

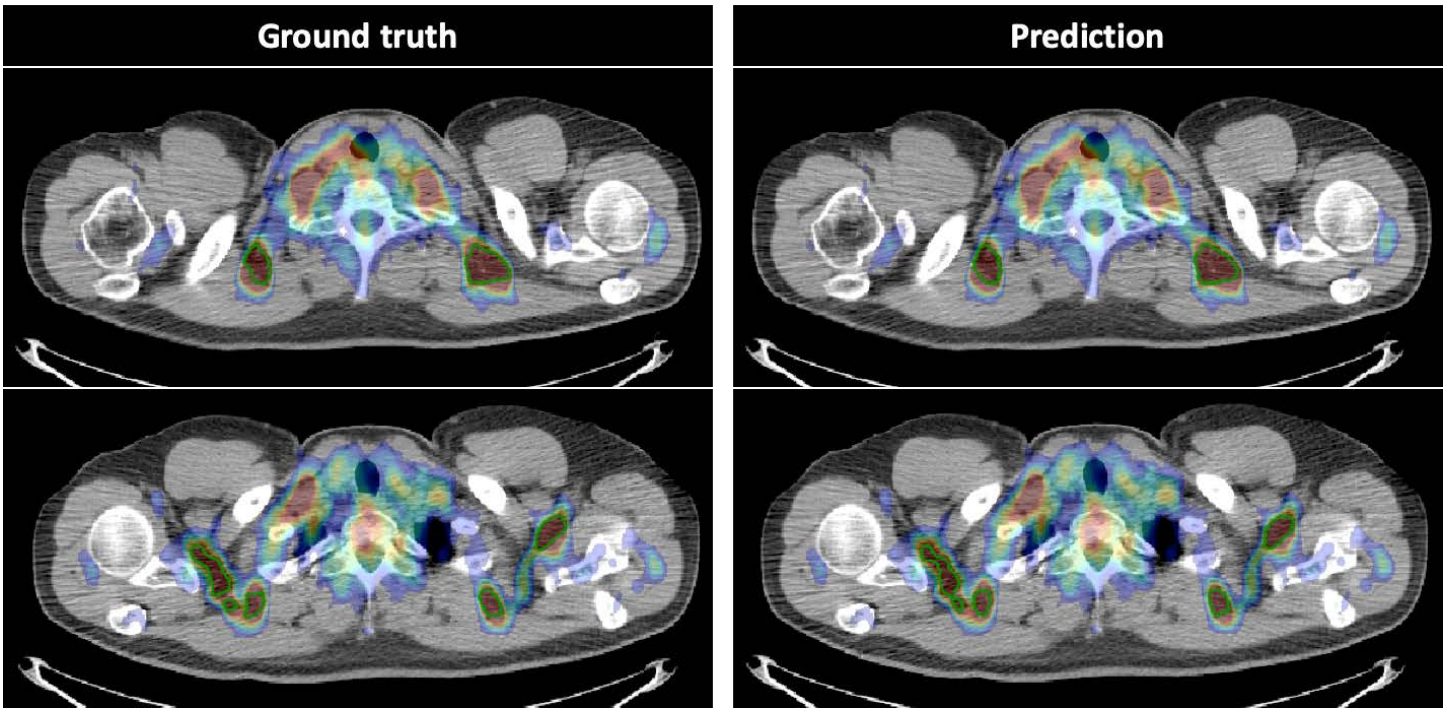

Figure 2: Example from the test set with the best brown fat Dice score (0.90). Two axial image slices are shown from the ground truth annotation (left column) and prediction (right column) by the deep learning model. Labeled areas of brown fat are indicated with a green boundary.