nature portfolio

Peer Review File

Predicting Standardized Uptake Value of Brown Adipose Tissue from CT scans using convolutional neural networks

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Erdil E et al investigated the potential of training convolutional neural network (CNN) for estimating brown adipose tissue (BAT) activity from CT data. Because FDG-PET/CT, a gold standard method for estimating BAT activity, exposes participants to radiation, repeated scanning is not recommended basically although the repeated scanning, e.g. before and after any treatment or intervention, is required to investigate changes in BAT activity. Moreover, cold exposure for a few hours prior to scanning is required to activate and visualize BAT; clinical data of FDG-PET/CT scan for cancer diagnosis could not be an alternative. The authors here proposed that BAT activity can be estimated with a certain accuracy by CNN when trained on cold stimulated cohort data. This is important and interesting challenge because if doable, researchers can refer so many clinical data and perform BAT study with a huge number of participants. The study concept is highly important to the research field However, I have some concerns to be addressed.

Specific comments
Comment-1. The manuscript needs to be reconstructed to be reviewed and there are many redundant sentences. For example, the manuscript is constructed with 1) Introduction, 2) Datasets which includes results, 3) Results introducing Fig 1, 2, 3-5) Discussion introducing Fig 3-8, 4) Methods introducing 8-11. The authors should reconstruct the manuscript and delete unnecessary words and sentences.

Comment-2. The major concern of this reviewer is that HU (or adipocyte lipid content) is highly variable dependent on adiposity. It is thus very important if correlation between HU-derived index and SUV-derived index is significant in very lean population and in very obese population. Also, it dice score correlate with BMI?

Comment-3. Why the improvement of accuracy by CNN compared to UT HU is much different between Basal (75%) and Granada cohorts (23%)? The reason might be important to find potential limitation of CNN.

Comment-4. Is there any correlation between the Dice sore and age of participants? As this is very useful for investigation of BAT in clinical data in which the most of patients are elderly or middle age-adults, it is very important to test if CNN can estimate BAT activity even in older populations.

Reviewer #2 (Remarks to the Author):

Remarks to authors:
- While the application of CNNs represents a significant advancement over conventional CT thresholding methods, there appears to be room for enhancement. The authors focused exclusively on one region of BAT despite the presence of BAT in various areas in human adults. Extending the analysis to at least one additional region, such as cervical, paraspinal, or mediastinal regions, would

be beneficial. Incorporating multiple regions would significantly improve the robustness of their CNNs application. Given the considerable variability in human BAT prevalence among individuals and patients, the CNNs method appears to be difficult to use in general under current clinical cohorts. Therefore, CNNs could be a more precise and robust arm for rodent BAT research. It might be relevant to know if the authors conducted BAT imaging analysis using rodent samples. Data pre-processing (4.1) and training using the processed data (4.2) are paramount procedures of CNNs development. While the authors intended to describe these processes, there is a lack of transparency. It is unclear how the authors determined voxel and cropped sizes. Were these determinations based on scientific or mathematical principles? What criteria were used for manual cropping? How many images did authors use for the training for intra- and inter-cohort analysis? To improve accessibility for non-deep learning researchers, providing a visual and informative segmentation flow, potentially through modifications and clarifications to Figure 10, would be highly informative. It should be explicitly stated that input CT images were cropped in the supraclavicular region with a size of 320 x 480, as opposed to the whole CT image shown in Figure 10.

- On line 405, page 15, it is keen to know how many false positives are produced by the CT thresholding method that does not correspond to the BAT region.

- Including definitions and illustrations of Dice scores in the manuscript would enhance its accessibility for non-machine learning experts. Explaining how to interpret these scores would be instructive and improve overall clarity.

- While the authors intended to visualize the predicted PET using their CNNs in Figures 3 and 4, these figures appear to be raw image alignments with insufficient explanations. To enhance their utility, it is advisable to include scale bars, patient IDs, and clear demarcation of the target area.

- Determining dataset bias, as demonstrated in Figures 5 to 7, is a potential critical experiment. These figures could potentially be combined into one for improved clarity. To enhance their informative value, consider adding an experimental flow from the training process, providing detailed explanations (e.g., clarifying what "mean Slice 1 or 2" refers to and specifying the source of these slices and the associated patients), and conducting statistical analyses between trials.

- On line 331, page 11, the authors suggest a more balanced training cohort with equal numbers of subjects. Do authors have tangible criteria based on their study?

- While 2D imaging analysis would be concise and clear, analysis of 3D images could be ideal for CT imaging, as mentioned on line 349, page 12, deserves more comprehensive exploration. This could involve a thorough comparison with 2D, including a discussion of the number of samples used for evaluating 3D U-Net architecture and its implications for active BAT detection accuracy. The description and results presented in the manuscript require further elaboration and additional evidence to substantiate the authors' claims regarding the utility of 2D CNNs in the current setting.

- In light of recent findings (e.g., Wagner et al., Cancer Cell 2023), which suggest the potential of Transformers as next-generation imaging analysis calivers, it might be worth discussing how the utilization of Vision Transformers (ViT) could address limitations in the current CNN approach.

- While it is reasonable to suggest that CNNs offer cost-efficiency, the authors repeatedly emphasize the cost-benefit of CT using CNNs without providing actual mathematical calculations or a simulation model. The authors have to provide these calculations and models to substantiate this claim. Without such evidence, the cost-efficiency argument lacks a strong foundation.

- The Abstract and Conclusion mention this study in personalized medicine/treatment, but how the authors justify this claim remains unclear. To strengthen their assertion, the authors should provide concrete evidence of how their CNNs application is linked to personalized medicine.

- Only Neural Networks on the title would be overstated due to containing various models, including ANN, RNN, and Transformers. It is essential that the title accurately reflects the content of the manuscript.

- The description of adipose biology in the abstract appears outdated, particularly regarding cell types in adipose tissue. Mounting evidence shows that beige adipocytes are a distinct cell type in adipose tissues (Wu et al., Cell 2012; Ikeda et al., Trends Endocrinol Metab 2018).

- It might not be the time to organize the manuscript, but the manuscript contains numerous minor yet critical editorial errors, including typos, an atypical style of life scientific terminology (e.g., "fat cells," which is more suitable for popular science articles), improper usage of gene and protein symbols, and misquotation of reference results (e.g., "R = 0.66" instead of "R > 0.66"). The random numerical order for the citations and the small font size in the tables and Figures make readers and reviewers unfriendly. Additionally, the manuscript exhibits inconsistent article styles and formats compared to this journal format. All abbreviations (e.g., CT, GLP-1, and ACE) should be spelled out, and line plots should be presented with appropriate dispersion. Furthermore, all CT or PET images should include a size bar, even for processed images.

Reviewer #3 (Remarks to the Author):

The reviewer questions the novelty of the work, suggesting it may not be sufficiently innovative.
The reviewer is wondering about the underlying theory that PET images can be predicted from CT scans. There can be two subjects that have the same CT but different PET images.
CT reflects stable structural images, whereas PET images can vary due to various factors like health conditions and test circumstances. A more detailed explanation of the underlying theory is needed for the high level journal.
The reviewer recommends including information about the Attention U-Net in the abstract, not just CNN.
Can the technique be used for predicting tumor images?

Dear Reviewers,

We would like to thank you for the time, effort, and expertise you have dedicated to evaluating our manuscript, titled "Predicting Standardized Uptake Value of Brown Adipose Tissue from CT scans using convolutional neural networks." Your constructive critiques and insightful suggestions have been invaluable in guiding our revisions and significantly improving the quality of our work.

We have thoroughly reviewed and considered each comment and suggestion provided, and have endeavored to address them comprehensively in our point-by-point response below and in our revised manuscript. In the below point-by-point response, we indicated the reviewers' comments with the blue color, our responses to the reviews with the green color, the changes that we added to the manuscript with the red color, and the quotations from the first version of the manuscript with the **black** color. Additionally, all additions in the revised manuscript are indicated by the red color and crossed out the text that we removed.

We are grateful for the opportunity to revise our work through your thoughtful and detailed feedback. We are readily available to offer additional information should further concerns or comments emerge.

Thank you once again for your invaluable feedback and guidance.

Authors

Reviewer 1

R1.1 The manuscript needs to be reconstructed to be reviewed and there are many redundant sentences. For example, the manuscript is constructed with 1) Introduction, 2) Datasets which includes results, 3) Results introducing Fig 1, 2, 3-5) Discussion introducing Fig 3-8, 4) Methods introducing 8-11. The authors should reconstruct the manuscript and delete unnecessary words and sentences.

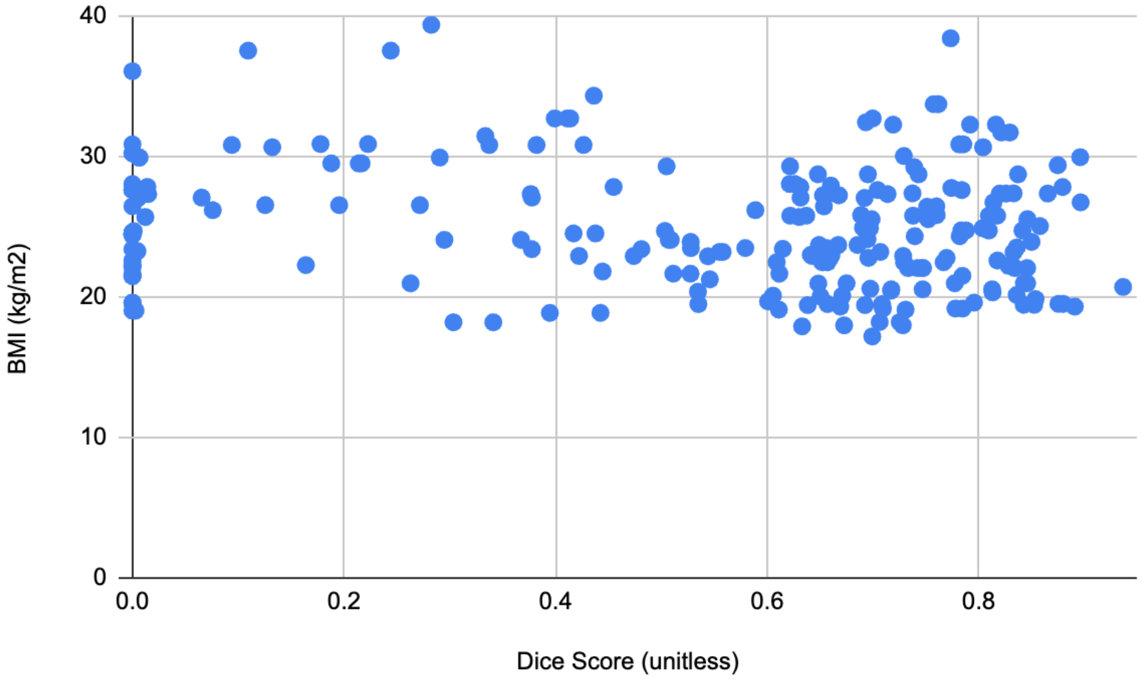
We reconstructed the revised manuscript as follows:

- Section 1 is Introduction where we introduced the problem and motivated our paper.
- Section 2 is Datasets which contains dataset specific information as requested by the Editor. This section does not contain any result.
- Section 3 is Results and Discussion where we present the results of different experiments in different subsections and discuss the potential limitations and future directions at the final subsection. In Section 3, we also introduce Figures 1 to 8.
- Section 4 is Method introducing the details about data pre-processing, network architecture, training details, the full pipeline, and the baseline method. In this section, we introduce Figures 9 and 10.
- Section 5 is Conclusion and Future Work where we summarize our findings and suggest future directions.

Additionally, we carefully read the manuscript and deleted parts that might be redundant. We would be happy to incorporate additional structural changes in the final manuscript based on your suggestion.

R1.2 The major concern of this reviewer is that HU (or adipocyte lipid content) is highly variable dependent on adiposity. It is thus very important if correlation between HU-derived index and SUV-derived index is significant in very lean population and in very obese population. Also, it dice score correlate with BMI?

Thanks to the reviewer for raising this point and we agree that investigating the significance of the correlation between HU-derived index and SUV-derived index in a population with different BMI (e.g. lean or obese population) is interesting. In our experiments, we measured the agreement between the HU-derived index (CNN's predictions) and SUV-derived index (ground truth) using Dice score, which can serve as a measure of correlation between both indices. Therefore, we measured the correlation between Dice score and BMI, and obtained Pearson's correlation coefficient of -0.131. We present the scatter plot of Dice score vs BMI for all folds below:



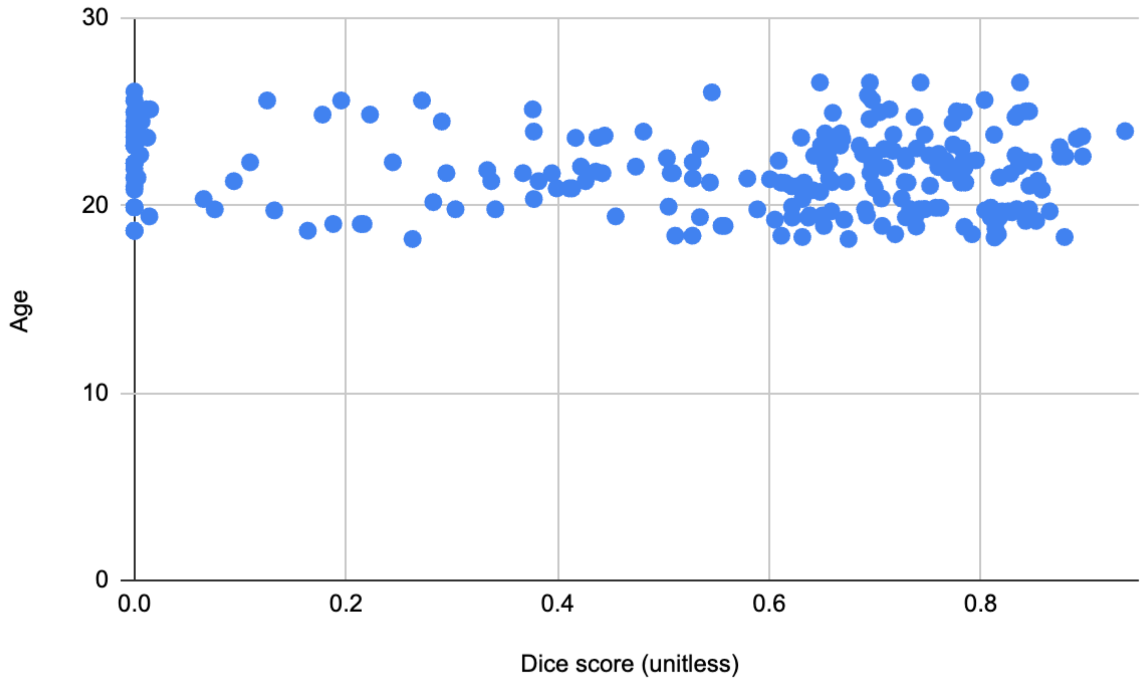
R1.3 Why the improvement of accuracy by CNN compared to UT HU is much different between Basal (75%) and Granada cohorts (23%)? The reason might be important to find potential limitation of CNN.

As we discussed in Section 3.5, both the Granada and the Basel models perform worse on subjects having low BAT activity since the majority of the samples have high BAT activity in both datasets. We observe from the scatter plots of Target BAT volume (ml) vs Dice score (unitless) and Predicted BAT volume (ml) vs Dice score (unitless), Granada cohort has more samples with low BAT activity than the Basel cohort. This means that there are more samples overpredicted by the corresponding model in the Granada cohort compared to Basel cohort. Such over-predictions lead to low Dice scores and reduce the average Dice results. Additionally, we observe that there are many samples with zero activity in the Granada cohort, and overpredicting such samples leads to zero Dice score. Such samples were taken into account when computing the average Dice score results shown in Table 1, which reduced the average Dice scores for the Granada cohort significantly. Whereas, we do not observe the same picture in the Basel cohort. We discussed such overpredictions as a limitation of the CNN in Section 3.5, under the Dataset bias subsection, in the manuscript.

To quantify the effect of samples with zero Dice scores in the Granada cohort to the results in Table 1, we computed the average Dice scores by not considering zero Dice scores. We observe that the results increase from 0.512 to 0.632 leading to an increase of approximately 50% from 23% compared to HU thresholding method.

R1.4 Is there any correlation between the Dice sore and age of participants? As this is very useful for investigation of BAT in clinical data in which the most of patients are elderly or middle age-adults, it is very important to test if CNN can estimate BAT activity even in older populations.

We measured the correlation between the age of the participants and Dice score on the Granada cohort, our largest research cohort obtained with cold-exposure, and obtained Pearson's correlation coefficient of -0.142. This result shows that there is no correlation between these two quantities on this dataset. We present the scatter plot of Dice score vs Age for all folds below:



Reviewer 2

R2.1 The authors focused exclusively on one region of BAT despite the presence of BAT in various areas in human adults. Extending the analysis to at least one additional region, such as cervical, paraspinal, or mediastinal regions, would be beneficial. Incorporating multiple regions would significantly improve the robustness of their CNNs application.

Thanks to the reviewer for the valuable suggestion. We concur that extending the analysis to encompass multiple BAT regions would enrich the study. To this end, we conducted additional experiments within the Granada cohort, which represents the largest cold-exposure cohort in our study. These additional experiments aimed to assess the CNN's efficacy in identifying BAT across various depots, including the cervical, supraclavicular, and paraspinal areas. For this purpose, we trained the CNN using the original images that encompass the cervical, supraclavicular, and paraspinal regions, diverging from our initial approach of focusing solely on images cropped around the supraclavicular area, as described in our main manuscript.

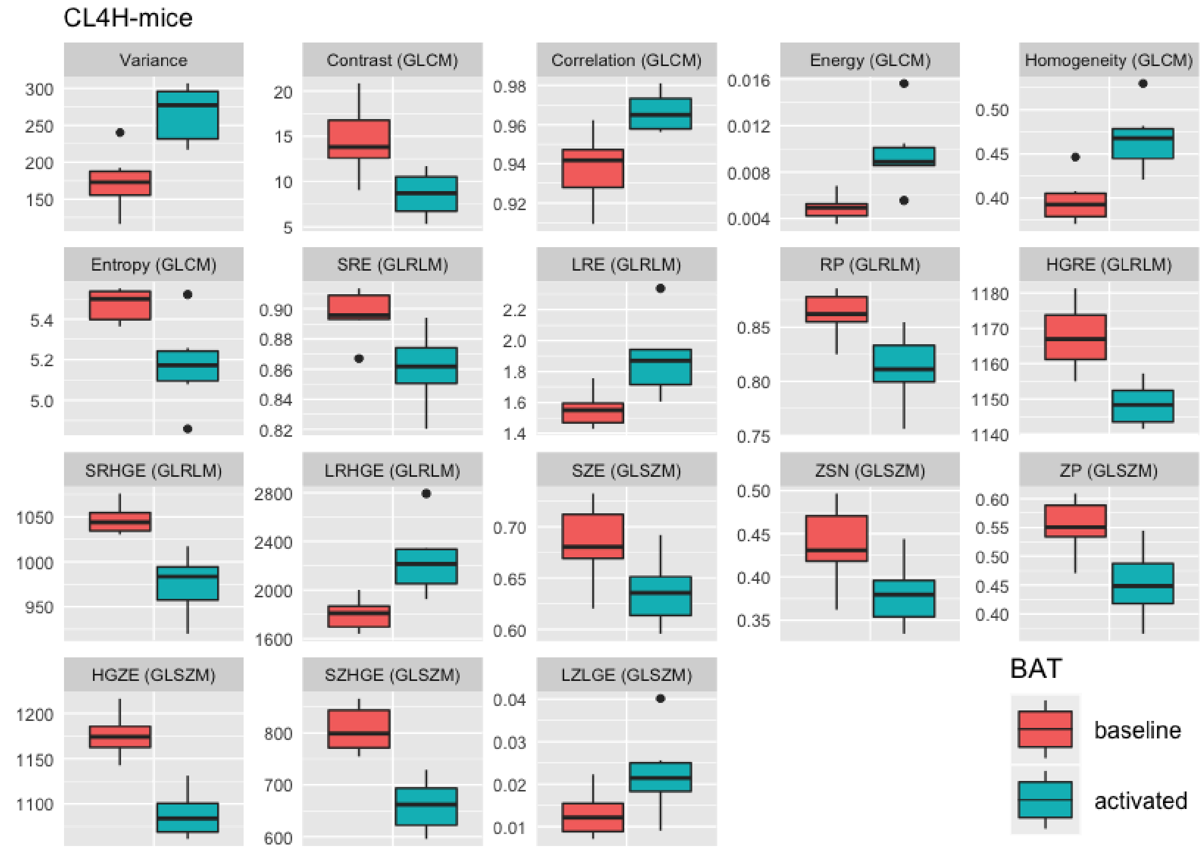
For ground truth segmentation, we identified brown adipose tissue (BAT) by the intersection of thresholded original CT and PET scans. Specifically, we applied Hounsfield Unit (HU) thresholds ranging from -180 to -10 for CT scans and a standardized uptake value (SUV) threshold of 1.5 for PET scans. The predicted segmentations were similarly derived by intersecting thresholded original CT scans with thresholded predicted PET scans. We quantified the similarity between the ground truth and predicted segmentations using the Dice score, which resulted in an average score of 0.519 across the test sets in a 5-fold cross-validation, whereas the Dice score for segmenting the supraclavicular region is 0.521, as shown in Table 1 in the manuscript. These results demonstrate that CNNs are capable of predicting BAT activity across various BAT regions with consistent accuracy.

Additionally, we assessed the efficacy of the HU thresholding-based method for segmenting active brown adipose tissue (BAT) across various depots, including cervical, supraclavicular, and paraspinal regions. For this, we applied HU thresholds ranging from -180 to -10 for CT scans. It is important to note that we did not apply any standardized uptake value (SUV) threshold, as the method does not predict PET activations of BAT. The segmentations predicted by the HU thresholding-based method were compared with the ground truth using the Dice score, resulting in a notably low score of 0.098. This outcome highlights a significant disparity in the effectiveness of the HU thresholding-based method across different experiments. Specifically, the Dice score for predicting BAT activity in multiple regions is substantially lower than the score reported in Table 1 for the supraclavicular region alone. This discrepancy arises because, in the manuscript's experiments, we enhanced the accuracy of the HU thresholding method by intersecting its segmentations with manually segmented supraclavicular regions, effectively minimizing false positives. Please also see our response to your comment R2.4 below about the quantification of the false positives produced by the solely HU thresholding without applying any manual segmentation mask. However, in our broader assessment of BAT activity across multiple regions, we did not apply manual segmentation masks to neither the CNN evaluations nor the HU thresholding method, due to the lack of manual segmentations for all examined areas.

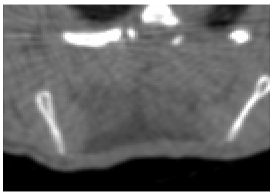
We would be happy to include the results of this experiments to the main paper or appendix in the final version based on the feedback of the reviewer.

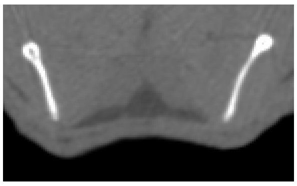
R2.2 Given the considerable variability in human BAT prevalence among individuals and patients, the CNNs method appears to be difficult to use in general under current clinical cohorts. Therefore, CNNs could be a more precise and robust arm for rodent BAT research. It might be relevant to know if the authors conducted BAT imaging analysis using rodent samples

We fully agree with the reviewer that imaging is an interesting modality to evaluate BAT in rodent samples. Indeed, in an experiment preceding the current analysis by about two years, a significant association was found between texture analysis features (a simpler computer vision tool) and BAT activation in mice.



However, it should be noted that the brown fat in these mice is morphologically quite different from the depots in humans (which are probably more akin to the mice's beige fat depots). In fact, the activation can be appreciated by simple change of mean CT attenuation of the brown fat depot, presumably due to increased perfusion/water content:


Inactive


Active (higher CT attenuation)

R2.3 Data pre-processing (4.1) and training using the processed data (4.2) are paramount procedures of CNNs development. While the authors intended to describe these processes, there is a lack of transparency.

Thanks for the review and suggestions. Please see our responses to different points in this review below:

R2.3-1 It is unclear how the authors determined voxel and cropped sizes. Were these determinations based on scientific or mathematical principles? What criteria were used for manual cropping?

We set the voxel sizes in alignment with the default voxel size of the Granada cohort, which is the largest cold exposure cohort in our study. Images from other datasets were resampled to match the voxel size of this primary cohort. We added the following explanation shown in red in the revised manuscript.

“**We resampled all volumes, PET and CT, from all cohorts to voxel size of 0.976X0.976X1.5 mm^3**, the default voxel size of the Granada cohort.”

The size of the crop is determined based on two criteria: 1) ensuring the cropped region encompasses the supraclavicular area, and 2) maintaining the cropped image's dimensions as divisible by 2^3 where 3 represents the number of downsampling operations within our network architecture as illustrated in Figure 9b of the revised manuscript. We added the following explanation shown in red in the revised manuscript.

“The crop size of 320x480 was determined such that the cropped region encompasses the supraclavicular area and the cropped image's dimensions are divisible by 2^3 where 3 is the number of max pooling operations used for downsampling the image size in our network architecture as shown in Figure 9b.”

R2.3-2 How many images did authors use for the training for intra- and inter-cohort analysis?

The distribution of train, validation, and test volumes is detailed in Table 2 for each dataset. It is important to note that the same models were employed for both intra- and inter-cohort analyses. In the intra-cohort analysis, the model was evaluated on the test volumes from the same cohort it was trained on. Conversely, in the inter-cohort analysis, the model was tested across the total number of volumes from different datasets. For instance, in the Granada cohort's intra-cohort analysis, the CNN was trained on 148 training volumes from this cohort and evaluated on its 48 test volumes. For the inter-cohort analysis of the Granada cohort, we applied the same model, originally trained for the intra-cohort experiment, to test it on the 32, 480, and 85 test volumes from the Basel, Zurich, and MSKCC cohorts, respectively.

R2.3-3 To improve accessibility for non-deep learning researchers, providing a visual and informative segmentation flow, potentially through modifications and clarifications to Figure 10, would be highly informative. It should be explicitly stated that input CT images were cropped in the supraclavicular region with a size of 320 x 480, as opposed to the whole CT image shown in Figure 10.

We revised Figure 10 (which is Figure 9 in the revised version) to visually illustrate the preprocessing step that involves cropping a region of interest around the supraclavicular area.

R2.4 - On line 405, page 15, it is keen to know how many false positives are produced by the CT thresholding method that does not correspond to the BAT region.

To quantify the false positives that do not correspond to brown adipose tissue (BAT) regions produced by the Hounsfield Unit (HU) thresholding-based method, we computed the Precision of HU thresholding-based segmentations against the ground truth segmentation without applying the manual segmentation of the supraclavicular region as conducted in our manuscript. We obtained the segmentations by applying HU thresholds of -180 and -10 to the CT scans. The ground truth segmentations were derived by intersecting thresholded original CT and PET scans, specifically applying HU thresholds of -180 to -10 for CT scans and a standardized uptake value (SUV) threshold of 1.5 for PET scans.

The formula for computing Precision is:

Precision = (True Positives)/(True Positives + False Positives)

This metric ranges from 0 to 1, where higher scores indicate fewer false positives. In this experiment, we obtained a Precision of 0.084, which increased to 0.486 when evaluations were conducted solely within the supraclavicular region delineated by manual segmentations, as described in our manuscript. These results demonstrate a significant advantage of manual segmentation in enhancing the precision of the HU thresholding method.

Furthermore, we measured Precision with and without manual segmentation of the supraclavicular region for the CNN to compare with the HU thresholding-based method. In this context, Precision was 0.556 without manual segmentation and increased to 0.573 with manual segmentation. These findings indicate that the CNN-based method benefits slightly from manual segmentation, as it generates considerably fewer false positives compared to the HU thresholding-based method. We applied the manual segmentation mask in our experiments because such manual intervention is also utilized by researchers analyzing BAT in their studies.

R2.5 - Including definitions and illustrations of Dice scores in the manuscript would enhance its accessibility for non-machine learning experts. Explaining how to interpret these scores would be instructive and improve overall clarity.

Thanks to reviewer for the suggestion. We added a sentence in Section 3.1. to define the Dice Score and interpret its results.

“Dice score measures the degree of overlap between two sets X and Y as $2|X \cap Y|/(|X|+|Y|)$ and takes a value between 0 and 1, where higher scores indicate better segmentation performance.”

R2.6 - While the authors intended to visualize the predicted PET using their CNNs in Figures 3 and 4, these figures appear to be raw image alignments with insufficient explanations. To enhance their utility, it is advisable to include scale bars, patient IDs, and clear demarcation of the target area.

Thanks to reviewer for the suggestion. We updated Figures 3 and 4 as suggested by the reviewer. We also updated the other figures which contain PET visualizations, accordingly.

R2.7 - Determining dataset bias, as demonstrated in Figures 5 to 7, is a potential critical experiment. These figures could potentially be combined into one for improved clarity. To enhance their informative value, consider adding an experimental flow from the training process, providing detailed explanations (e.g., clarifying what "mean Slice 1 or 2" refers to and specifying the source of these slices and the associated patients), and conducting statistical analyses between trials.

Thanks for suggesting to combine Figures 5 to 7 into one big figure for clarity. We would be happy to do this in the final version of the paper upon acceptance, using Nature Communications journal's formatting, which will be provided at that stage.

We updated Figure 7 to clarify different slices and subjects.

In the manuscript, we performed statistical analysis via permutation test between the results of CNNs and HU thresholding method. The results show that CNNs achieve significantly higher Dice scores compared to the HU thresholding method. Please see Table 1 and the first paragraph of Section 3.2. for these analysis.

As the reviewer rightfully requested, we also conducted a similar statistical analysis for the inter-cohort experiments, specifically comparing the intra- and inter-cohort performances of both Basel and Granada models. Our analysis revealed no statistically significant difference in the Granada model's performance across cohorts, with a p-value of 0.77. Conversely, we observed a significant difference in the Basel model's performance, with a p-value well below 1%. These results support our previous observations in Section 3.5, under the Generalization subsection, that the generalization ability of the Granada model is better than the Basel model. We included the results of these statistical analysis in Section 3.2 as follows:

”We conducted statistical analysis to measure the generalization performance of the Basel and Granada models. In particular, we applied permutation test [40] between the intra and inter-cohort performance of both models. Our analysis revealed no statistically significant difference in the Granada model's performance across cohorts, with a p-value of 0.77. Conversely, we observed a significant difference in the Basel model's performance, with a p << 1%. **We discuss the generalization performance of CNNs further in Sec. 3.5.**”

R2.8 - On line 331, page 11, the authors suggest a more balanced training cohort with equal numbers of subjects. Do authors have tangible criteria based on their study?

Handling data imbalance is a very challenging and common problem in medical imaging. It is still an ongoing research area and very difficult to come up with a tangible criteria (Litjens et al., 2017). Our experiments involve two distinct datasets characterized by varying levels of data imbalance, as illustrated in Figure 8. Specifically, the Granada cohort includes a higher number of subjects exhibiting very low BAT activity in comparison to the Basel cohort.

As discussed in the 'Dataset Bias' section (Section 3.5), the model trained on the Granada cohort demonstrates superior accuracy in predicting lower BAT activity levels compared to the model trained on the Basel cohort. Based on these empirical observations, we hypothesize that training models with datasets comprising an equal number of subjects across various BAT activity levels would yield more accurate results across different BAT activity levels. Similar hypotheses have been validated by previous studies, which demonstrate diminishing accuracy as the data imbalance ratio increases artificially (Sangalli et al., 2021). However, this should still be validated specifically for the dataset.

Litjens, Geert, et al. "A survey on deep learning in medical image analysis." Medical image analysis 42 (2017): 60-88.

Sangalli, Sara, et al. "Constrained optimization to train neural networks on critical and under-represented classes." Advances in Neural Information Processing Systems 34 (2021): 25400-25411.

R2.9 - While 2D imaging analysis would be concise and clear, analysis of 3D images could be ideal for CT imaging, as mentioned on line 349, page 12, deserves more comprehensive exploration. This could involve a thorough comparison with 2D, including a discussion of the number of samples used for evaluating 3D U-Net architecture and its implications for active BAT detection accuracy. The description and results presented in the manuscript require further elaboration and additional evidence to substantiate the authors' claims regarding the utility of 2D CNNs in the current setting.

Thanks to the reviewer for suggesting experiments with a 3D network architecture. We agree that quantitative comparisons between 2D and 3D architectures for BAT activity prediction would strengthen the claim to justify our choice of using a 2D architecture. To support our claim with quantitative results as suggested by the reviewer, we trained a 3D Attention U-Net on the Granada cohort, the largest cold-exposure cohort in our experiments. We evaluated the BAT activity prediction performance of the 3D network on multiple BAT depots in cervical, supraclavicular, and paraspinal regions, as also suggested by the reviewer in review R2.1 for the 2D networks. In this experiment, we obtained an average Dice score of 0.407 across the test sets of 5-folds using the 3D model, which is significantly lower than the Dice score of 0.519 we obtained when evaluating the 2D network on the same BAT depots as detailed in R2.1. Note that we used the same training, test, and validation splits in both 2D and 3D experiments and the number of volumes in each split is given in Table 2.

R2.10 - In light of recent findings (e.g., Wagner et al., Cancer Cell 2023), which suggest the potential of Transformers as next-generation imaging analysis calivers, it might be worth discussing how the utilization of Vision Transformers (ViT) could address limitations in the current CNN approach.

Thanks to the reviewer for highlighting the intriguing paper by Wagner et al. in Cancer Cell 2023, which demonstrates a successful application of Vision Transformers (ViT) for predicting biomarkers from histology images.

ViTs and other transformer-based methods proposed for image analysis tasks hold significant potential to enhance the performance of CNNs (Liu et al., 2021; Krillov et al., 2023). Transformers excel over CNNs in modeling long-range dependencies and learning global representations. Moreover, they have been claimed to be more robust against shifts between training and test distributions compared to CNNs (Paul et al., 2022). However, the advantages of transformer-based networks are contingent on the availability of large amounts of labeled data. In scenarios with small or medium-sized datasets, their performance is often inferior to that of CNNs (Dosovitskiy et al., 2020). This is because CNNs benefit from inherent inductive biases such as translation equivariance and locality, which are not as pronounced in transformers (Dosovitskiy et al., 2020).
Given the limited size of our dataset compared to the large datasets used in successful applications of ViTs, such as the study by Wagner et al. that utilized data from 30,000 patients, we chose to employ CNNs for our research.

We concur with the reviewer that transformer-based architectures could potentially overcome the limitations of CNNs for BAT activity prediction from CT scans, particularly by enhancing inter-cohort performance results through their superior generalization ability when trained on very large datasets. To highlight the prospective utility of ViTs in future work, we have added the following paragraph to Section 5 of the revised manuscript:

“In this paper, we employed a CNN-based neural network architecture, specifically the Attention U-Net. Recently, transformer-based models, such as Vision Transformers (ViTs) [51], have achieved significant improvements over CNNs in terms of accuracy and generalization across a variety of image analysis tasks [52]. However, it’s important to note that the superior performance of transformer-based networks largely depends on the availability of extensive labeled datasets. In contexts where only small or medium-sized datasets are available, their performance tends to be less effective compared to that of CNNs [51]. Given that our study relies on relatively small datasets, we opted for CNNs. Nevertheless, should large cold-exposure [18F]-FDG PET/CT datasets become available for future research, transformer-based architectures have the potential to significantly enhance performance over CNNs.”

Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).

Paul, Sayak, and Pin-Yu Chen. "Vision transformers are robust learners." Proceedings of the AAAI conference on Artificial Intelligence. Vol. 36. No. 2. 2022.

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

R2.11 - While it is reasonable to suggest that CNNs offer cost-efficiency, the authors repeatedly emphasize the cost-benefit of CT using CNNs without providing actual mathematical calculations or a simulation model. The authors have to provide these calculations and models to substantiate this claim. Without such evidence, the cost-efficiency argument lacks a strong foundation.

Thanks to the reviewer for raising this important point. It is very difficult to analyze the cost-efficiency of using CNNs compared to the other approaches because the cost does not only include the financial cost of the imaging but also include the cost of radiation exposure.

In our simulation presented in the paper in Section 3.3, in order to create a cohort consisting of only BAT+ subjects, we selected 50 BAT+ subjects from a population of 240 subjects based on their (i) predicted BAT activities from CT scans using CNNs, (ii) HU thresholding-based method and (iii) randomly choosing the 50 subjects without using any information from CT scans. The results showed that the number of false positives (number of subjects identified as BAT+ while they are BAT-) obtained by each method are 17, 27, and 33, for CNN, HU thresholding, and random selection, respectively. These results show that using CNN reduces the cost of unnecessary radiation exposure by 10 and 16 subjects compared to the HU thresholding and random methods.

We can extend this simulation to calculate the financial imaging costs of these three approaches.

1- Both CNN and HU thresholding method require CT scans. So, the whole population of 240 subjects should undergo CT imaging for these methods while such cost does not exist in the random method. Assuming that a single CT scan costs $M, the associated costs for these methods as as follows:

CNN:	240 x $M
HU thresholding:	240 x $M
Random:	$0

2- To obtain 50 BAT+ subjects in the stratified cohort, CNN requires 75 PET/CT scans (efficiency 66%), HU thresholding requires 92 PET/CT scans (efficiency 54%), and random method requires 147

PET/CT scans (efficiency 34%). Assuming that a single PET/CT scan costs $N, the associated costs for these methods as as follows:

- CNN:	75 x $N
- HU thresholding:	92 x $N
- Random:	147 x $N

So the total costs of these methods are as follows:

- CNN:	240 x $M + 75 x $N
- HU thresholding:	240 x $M + 92 x $N
- Random:	147 x $N

From the total financial costs, it is obvious that CNN has much lower cost compared to the HU thresholding method by saving $17N. Comparing CNN and the random method shows that CNN have lower financial cost as long as $M < 0.3 x $N, which is usually the case in many countries. For example in Switzerland, the cost of PET/CT imaging N = 2'100 CHF while the CT imaging costs M = 600 CHF, resulting M/N = 0.285.

R2.12 - The Abstract and Conclusion mention this study in personalized medicine/treatment, but how the authors justify this claim remains unclear. To strengthen their assertion, the authors should provide concrete evidence of how their CNNs application is linked to personalized medicine.

We thank the reviewer for raising this point. We have altered the manuscript in several key locations to address the reviewer's comment.

The large impact our results could have lies in the potential to perform large-scale opportunistic screening for BAT, which may be helpful for stratifying cohorts and developing personalized treatments. We elaborated on these points in the Introduction between lines 65-71 as follows:

"**One example of a personalized strategy is to activate BAT by the use of oral intake of bile acids [17] or selective β_2 [18] or β_3-agonists [19]** in the context of obesity, **or suppress it by propranolol, a non-selective β-blocker [20]** in the context of cancer-cachexia. By gaining a deeper understanding of BAT through large-scale population studies, it may become possible to create stratified cohorts and develop further personalized treatment strategies targeting BAT."

The explanation above is also linked to the particular contribution of CNNs for tailoring treatments and interventions to specific patient subgroups in the last paragraph of the Introduction:

"... **The results suggest that CNNs trained on the cold exposure cohorts can be used to classify the subjects into high or low-activity categories using the predicted SUV of BAT only from CT scans. Accurately identifying these classes allows patient stratification by creating cohorts with the desired number of subjects from each category. Obtaining stratified cohorts is extremely useful for researchers to tailor treatments and interventions to specific patient subgroups. We show that CNNs can serve as a useful tool to this end by preselecting cohorts with the desired number of subjects from each BAT activity class, significantly reducing the number of subjects mistakenly included in the selected cohort. The preselected cohort can then undergo actual [18F]-PET/CT scans for more accurate quantification. Thus the actual [18F]-PET/CT scans are only obtained from subjects most relevant to the cohort, substantially lowering the PET acquisition cost for creating such a cohort**."

Lastly, we updated the Conclusion to restate the contribution of using CNN for developing personalized treatment strategies:

"**... Additionally, we demonstrated that this ability of CNNs can be used to create stratified cohorts of BAT+ and/or BAT- subjects, thereby reducing the number of subjects mistakenly assigned to a wrong cohort. This allows researchers to conduct experiments on cohorts with desired characteristics which is extremely useful, e.g., when investigating the effect of a drug on a certain population with large BAT activity. This research represents a significant step towards** opportunistic screening for BAT activity, which may be helpful in future research on **personalized treatment strategies for metabolism-related diseases by developing predictive models using neural networks with high-quality datasets from sizeable cohorts. By enabling extensive population analyses with reduced radiation exposure and cost, CNNs provide a promising path toward more effective diagnoses and treatments in the future..**"

R2.13 - Only Neural Networks on the title would be overstated due to containing various models, including ANN, RNN, and Transformers. It is essential that the title accurately reflects the content of the manuscript.

We changed the title to the following one to emphasize that we used convolutional neural networks:

"Predicting Standardized Uptake Value of Brown Adipose Tissue from CT scans using convolutional neural networks"

R2.14 - The description of adipose biology in the abstract appears outdated, particularly regarding cell types in adipose tissue. Mounting evidence shows that beige adipocytes are a distinct cell type in adipose tissues (Wu et al., Cell 2012; Ikeda et al., Trends Endocrinol Metab 2018).

Thanks to the reviewer for raising an important point. The introduction has been too simplified with the dichotomy of white and brown adipocytes. We added beige adipocytes as a distinct population of energy expending adipocytes to the introduction. We updated the introduction in the revised manuscript as follows:

"**Recent research on adipose tissue has yielded promising results toward possible personalized strategies. Adipose tissue can be subdivided into ~~distinct~~** different **types. White adipose tissue (WAT) is specialized for storing chemical energy in the form of triglycerides [12]. In contrast, brown adipose tissue (BAT) dissipates energy in the form of heat in a process called non-shivering thermogenesis through uncoupling protein** 1 **(Ucp1) [33]**. In the past decade it has become obvious that white adipocytes can transform into another distinct type of energy expending adipocytes which have been called beige adipocytes. Moreover, energy expending adipose tissue depots usually consist of a mixture of brown, beige and white adipocytes [47]. To improve readability, we will use the term BAT for these depots throughout this paper. **Since BAT is an energy-dissipating organ, its activation or deactivation could potentially be used to promote weight loss or gain and improve metabolic control. Over the last decade, researchers have been working on understanding the connection between BAT activity and metabolism, which may contribute to developing personalized treatment strategies for various metabolic diseases [38, 2, 17, 11]**"

R2.15 - It might not be the time to organize the manuscript, but the manuscript contains numerous minor yet critical editorial errors, including typos, an atypical style of life scientific terminology (e.g., "fat cells," which is more suitable for popular science articles), improper usage of gene and protein symbols, and misquotation of reference results (e.g., "R = 0.66" instead of "R > 0.66"). The random numerical order for the citations and the small font size in the tables and Figures make readers and reviewers unfriendly. Additionally, the manuscript exhibits inconsistent article styles and formats compared to this journal format. All abbreviations (e.g., CT, GLP-1, and ACE) should be spelled out, and line plots should be presented with appropriate dispersion. Furthermore, all CT or PET images should include a size bar, even for processed images.

Thanks for the suggestions for organizing the manuscript and correcting the errors. We have made the following changes in the revised manuscript based on your suggestion:

- Replaced “fat cells” with “adipoctyes”
- We changed "R > 0.66" to "R = 0.66"
- We changed the citations style to the ordered one
- We increased the font sizes in figures
- We wrote the expanded forms of abbreviations.
- We include colorbar for all PET images
- We read the text and corrected typos.

Additionally, we acknowledge the reviewer's point regarding the inconsistency of the manuscript's style with the standard format of Nature Communications. This is primarily due to the journal's policy for initial submissions, which states: “Nature Communications is flexible with regard to the format of initial submissions. Within reason, style and length will not directly influence consideration of a manuscript. We also do not require a particular structure or format at first submission. If and when revisions are required, the editor will provide detailed formatting instructions at that time.” (source: https://www.nature.com/ncomms/submit/article). We are happy to adjust the manuscript's style to meet the editor's requirements upon acceptance.

Reviewer 3

R3.1 - The reviewer questions the novelty of the work, suggesting it may not be sufficiently innovative.

Thanks to the reviewer for the comment.

We would like to start answering this comment by restating the main motivation of this manuscript. The conventional approach to detect active Brown Adipose Tissue (BAT) involves [18F]-Fluorodeoxyglucose ([18F]-FDG) PET/CT scans. However, this technique is expensive and involves radiation exposure, limiting its suitability for large-scale population studies. These issues can be mitigated by developing methods that predict [18F]-FDG uptake by BAT from CT scans. Some papers investigates if this is a realistic aim and show some correlation between Hounsfield Units (HU) in CT and the Standardized Uptake Value (SUV) in PET scans in BAT depot. There are two main shortcomings of these existing works: 1) mean HU is a simple statistics that is used to predict [18F]-FDG uptake by BAT and there can be more complicated features that can lead to more accurate predictions, 2) the findings are based on small cohorts and require validation on larger cohorts

Our manuscript advances the existing works by using a learning-based computational method based on convolutional neural networks (CNNs), a methodology that significantly diverges from the reliance on hand-crafted features like mean HU and automatically learns features that are useful for the task from the data in BAT research. Additionally, we validate the effectiveness of CNNs on 4 different cohorts (including both clinical and research cohorts), the largest and the most diverse cohorts to date, and discuss the limitations. To our knowledge, our work represents the first application of neural network-based methodology for predicting [18F]-FDG uptake in BAT, as well as the most extensive cohort analysis conducted in this domain.

We acknowledge the reviewer's concerns regarding the methodological innovation of our work. While we concede that the core methodology—employing CNNs—may not in itself constitute a novel approach, the application of such a methodology in this context, combined with the scale and breadth of our validation efforts, offers substantial and novel insights into BAT research. We believe these contributions warrant the attention of the research community and find Nature Communications an apt platform for disseminating our findings.

R3.2 - The reviewer is wondering about the underlying theory that PET images can be predicted from CT scans. There can be two subjects that have the same CT but different PET images. CT reflects stable structural images, whereas PET images can vary due to various factors like health conditions and test circumstances. A more detailed explanation of the underlying theory is needed for the high level journal.

Thank to the reviewer for the comment which is very critical to understand the main motivation of our manuscript. We understand and acknowledge the crucial distinction that the reviewer highlighted between CT and PET imaging modalities. CT scans provide detailed structural information, representing stable anatomical details, whereas PET images capture dynamic metabolic and biochemical processes, which can indeed vary significantly due to various factors such as health conditions and testing circumstances.

We would like to clarify that our work does not claim the general feasibility of predicting complete PET scans from CT images across all tissues and conditions, recognizing the inherent limitations and differences between these imaging modalities. Instead, our research is specifically focused on the prediction of PET activity within brown adipose tissue (BAT) depots from CT scans. This approach is grounded in existing literature that indicates some correlation between the average Hounsfield Units (HU) in CT and the Standardized Uptake Value (SUV) in PET scans in BAT depots as mentioned between lines 73-76 of the revised manuscript. Our manuscript aims to leverage convolutional neural networks to predict PET activity of BAT in CT images, specifically within BAT depots, rather than attempting to predict the entirety of PET scans from CT data.

R3.3 - The reviewer recommends including information about the Attention U-Net in the abstract, not just CNN.

Thanks to the reviewer for the recommendation to include the specific architecture that we used in the abstract. We added the following sentence in the abstract of the revised manuscript:

“In our experiments, we adopted the widely used CNN architecture known as Attention U-Net.”

R3.4 - Can the technique be used for predicting tumor images?

Thanks for the review that is also related to reviewer’s previous review in R3.2. The answer to this question is definitely NO! We cannot say that CNNs can predict PET activity of tumors from CT scans without having any evidence that CT has any signal about the activity. We only aim to predict PET activity of BAT in the regions corresponding to BAT depots from CT scans since the earlier studies empirically show that CT carries such signal for BAT activity.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

R1.1: The authors improved construction in the revised manuscript.

R1.2: No correlation between BMI and Dice score is important as it suggests that observed correlation between target BAT volume and dice score is not spurious correlation because of possible confounding factors such as adiposity (adiposity and BAT functionality are negatively correlated; moreover, an increased adipocyte size and CT HU in fat tissue are also negatively correlated) . This result should be reported and be discussed in the manuscript.

R1.3. Thank you for the explanation.

R1.4. Thank you for the additional analysis of correlation between age and Dice score. According to the correlation, I realized that all the participants of this cohort are young (look less than 28 years old). As I explained in the original comment, it is very important if CNN can estimate BAT activity even in older populations. The limited range of age could be a major limitation of the study while I agree that this is the variable first evidence showing CNN can predict BAT functionality. Then, the authors should report with integrity participant's age (the mean, sd, and range) as well as BMI (the mean, sd, and range) for all cohorts as new supplemental tables or alternatives.

Reviewer #2 (Remarks to the Author):

Considering the robustness of this study, results from the experiment assessing various BAT depots in different locations should be included in the manuscript, preferably a main figure. I appreciate the author's dedication to satisfactorily responding to all queries posed following the review of the initial manuscript.

Reviewer #3 (Remarks to the Author):

I previously asked if the technique could be used for predicting tumor images? The authors answered as follows:
The answer to this question is definitely NO! We cannot say that CNNs can predict PET activity of tumors from CT scans without having any evidence that CT has any signal about the activity. We only aim to predict PET activity of BAT in the regions corresponding to BAT depots from CT scans since the earlier studies empirically show that CT carries such signal for BAT activity

However, I do not quite understand why "The answer to this question is definitely NO!". Upon looking at the provided image, it appears that AI predicted not only FDG uptake in BAT but also non-BAT physiological uptakes across the entire slice. If the prediction is indeed based on changes in the

CT values of fat, the question arises as to why this method cannot similarly apply to tumors, which also exhibit changes in CT values. Is it only due to the lack of oncology patients in the current cohorts?
For other questions, I think the authors appropriately responded.

Dear Reviewers,

We would like to thank you for the time, effort, and expertise you have dedicated to evaluating our manuscript, titled "Predicting Standardized Uptake Value of Brown Adipose Tissue from CT scans using convolutional neural networks." Your constructive critiques and insightful suggestions have been invaluable in guiding our revisions and significantly improving the quality of our work.

We have thoroughly reviewed and considered each comment and suggestion provided, and have endeavored to address them comprehensively in our point-by-point response below and in our revised manuscript. In the below point-by-point response, we indicated the reviewers' comments with the blue color, our responses to the reviewer's comments with the green color and, the changes that we added to the manuscript with the red color. Additionally, all additions in the second revision of the manuscript are indicated by the red color.

We are grateful for the opportunity to revise our work through your thoughtful and detailed feedback. We are readily available to offer additional information should further concerns or comments emerge.

Thank you once again for your invaluable feedback and guidance.

Authors

Reviewer 1

R1.2 No correlation between BMI and Dice score is important as it suggests that observed correlation between target BAT volume and dice score is not spurious correlation because of possible confounding factors such as adiposity (adiposity and BAT functionality are negatively correlated; moreover, an increased adipocyte size and CT HU in fat tissue are also negatively correlated) . This result should be reported and be discussed in the manuscript.

Following the reviewer's suggestion, we added the following paragraph in the Discussion section:

Correlation between Dice score and BMI: In the assessment of the Dice score between the CNN-predicted HU-derived indices (predicted BAT activity) and the actual SUV-derived indices (ground truth BAT activity), an important consideration is the potential influence of confounding factors, such as adiposity, on the predictive accuracy of our model. Adiposity is inversely related to BAT functionality, and increased adipocyte size is associated with higher BMI levels. Thus, we investigated whether the predictive accuracy of our model - as measured by the Dice score - was influenced by the BMI of the subjects in our cohorts.

Our analysis revealed a Pearson's correlation coefficient of -0.131 and -0.076 between the Dice score and BMI for the Granada and Basel cohorts, respectively - suggesting a negligible inverse relationship. This finding is particularly significant as it indicates that the predictive accuracy of our model is not confounded by variations in adiposity levels among subjects. The absence of a significant correlation between BMI and the Dice score substantiates the reliability of the observed agreement between the predicted and actual BAT volumes, affirming that this agreement is not a spurious correlation attributable to adiposity. This result underlines the robustness of our CNN-based predictive model, demonstrating its applicability across a diverse population with varying BMI levels.

R1.4 Thank you for the additional analysis of correlation between age and Dice score. According to the correlation, I realized that all the participants of this cohort are young (look less than 28 years old). As I explained in the original comment, it is very important if CNN can estimate BAT activity even in older populations. The limited range of age could be a major limitation of the study while I agree that this is the variable first evidence showing CNN can predict BAT functionality. Then, the authors should report with integrity participant’s age (the mean, sd, and range) as well as BMI (the mean, sd, and range) for all cohorts as new supplemental tables or alternatives.

We added the mean, std, and range values for the age and BMI of the subjects in the Granada and Basel cohorts in Section 2.

For the Basel cohort:
The subjects in this cohort are between 19 and 33 years old, with an average age of 24.45 years and a standard deviation of 4.38 years. Their BMI values range from 18.6 to 27.5 kg/m2, with a mean BMI of 22.5 and a standard deviation of 2.31 kg/m2.

For the Granada cohort:
The subjects in this cohort are between 18 and 27 years old, with an average age of 22.07 years and a standard deviation of 2.23 years. Their BMI values range from 17.2 to 39.40 kg/m2, with a mean BMI of 24.91 and a standard deviation of 4.62 kg/m2.

Reviewer 2

R2.1 Considering the robustness of this study, results from the experiment assessing various BAT depots in different locations should be included in the manuscript, preferably a main figure. I appreciate the author's dedication to satisfactorily responding to all queries posed following the review of the initial manuscript.

Following the reviewer's suggestion, we have incorporated the results of the experiment that assesses the performance of CNNs in predicting the PET activity of BAT across various BAT depots into the paper. We summarized the key findings briefly in the main text within the Discussion section as follows:

Predicting the PET activity of BAT in other BAT depots: In this paper, we primarily focus on predicting the PET activity of BAT from CT scans in the supraclavicular region, as it is one of the largest BAT depots in humans. We then extended our analysis to include multiple BAT regions, aiming to determine whether the capabilities of CNNs can be generalized to predict BAT activity in additional depots, such as the cervical, supraclavicular, and paraspinal areas. The results showed that CNNs can predict BAT activity in other BAT depots without loss in accuracy; significantly improving the HU thresholding-based method. The details of this experiment are provided in Appendix A.

Then, we presented the details of the experiment in Appendix A.

Appendix A: Predicting the PET activity of BAT in other BAT depots

In this experiment, we trained the CNN using the original images that encompass the cervical, supraclavicular, and paraspinal regions, diverging from our initial approach of focusing solely on images cropped around the supraclavicular area, as described in Section 4. For ground truth segmentation, we identified brown adipose tissue (BAT) by the intersection of thresholded original CT and PET scans. Specifically, we applied Hounsfield Unit (HU) thresholds ranging from -180 to -10 for CT scans and a standardized uptake value (SUV) threshold of 1.5 for PET scans. The predicted segmentations were similarly derived by intersecting thresholded original CT scans with thresholded predicted PET scans. The predicted segmentations were similarly derived by intersecting thresholded original CT scans with the thresholded predicted PET scans. We quantified the similarity between the ground truth and predicted segmentations using the Dice score, which resulted in an average score of 0.519 across the test sets in a 5-fold cross-validation, whereas the Dice score for segmenting the supraclavicular region is 0.521, as shown in Table 1. These results demonstrate that CNNs are capable of predicting BAT activity across various BAT regions with consistent accuracy.

Additionally, we assessed the efficacy of the HU thresholding-based method for segmenting active brown adipose tissue (BAT) across various depots, including cervical, supraclavicular, and paraspinal regions. For this, we applied HU thresholds ranging from -180 to -10 for CT scans. It is important to note that we did not apply any standardized uptake value (SUV) threshold, as the method does not predict PET activations of BAT. The segmentations predicted by the HU thresholding-based method were compared with the ground truth using the Dice score, resulting in a notably low score of 0.098. This outcome highlights a significant disparity in the effectiveness of the HU thresholding-based method across different experiments. Specifically, the Dice score for predicting BAT activity in multiple regions is substantially lower than the score reported in Table 1 for the supraclavicular region alone. This discrepancy arises because, in the manuscript's experiments, we enhanced the accuracy of the HU thresholding method by intersecting its segmentations with manually segmented supraclavicular regions, effectively minimizing false positives. However, in our broader assessment of BAT activity across multiple regions, we did not apply manual segmentation masks to neither the CNN evaluations nor the HU thresholding method, due to the lack of manual segmentations for all examined areas.

Reviewer 3

R3.1 I previously asked if the technique could be used for predicting tumor images? The authors answered as follows:
The answer to this question is definitely NO! We cannot say that CNNs can predict PET activity of tumors from CT scans without having any evidence that CT has any signal about the activity. We only aim to predict PET activity of BAT in the regions corresponding to BAT depots from CT scans since the earlier studies empirically show that CT carries such signal for BAT activity

However, I do not quite understand why “The answer to this question is definitely NO!”. Upon looking at the provided image, it appears that AI predicted not only FDG uptake in BAT but also non-BAT physiological uptakes across the entire slice. If the prediction is indeed based on changes in the CT values of fat, the question arises as to why this method cannot similarly apply to tumors, which also exhibit changes in CT values. Is it only due to the lack of oncology patients in the current cohorts?

Thanks for the follow-up question which allows us further clarification. If the underlying functional abnormality, e.g. tumor, has a signature in the CT images, neural networks-based methods have the potential to predict the corresponding PET activity from CT scans, provided they are trained with paired PET/CT datasets appropriate for the task. However, we have not tested for this nor we have a cohort consisting of oncology patients to test this hypothesis.

We believe that this is an important point to mention and clarify in the main manuscript. Therefore, we added the following paragraph in the Discussion section.

Can neural networks predict any functional activity in PET from CT scans? In this study, we utilized Convolutional Neural Networks (CNNs) to predict PET activity in brown adipose tissue (BAT) from CT scans, based on literature that suggests a correlation between CT's Hounsfield Units (HU) and PET's Standardized Uptake Values (SUV) within BAT regions, as previously discussed. The question may arise regarding the broader application of our method for predicting PET activity from CT scans in tissues other than BAT, including tumors. It is important to highlight the intrinsic differences between CT and PET imaging techniques; CT scans offer detailed structural information reflecting stable anatomical features, while PET images reveal dynamic metabolic and biochemical activities that can fluctuate significantly due to a range of factors, including health status and testing conditions. We would like to emphasize that our findings do not imply that our method is universally applicable across different tissues or conditions, given the fundamental differences between the imaging modalities. Our approach is specifically tailored to BAT, relying on distinct CT signatures of BAT activity, which have empirical backing. For the application of neural networks in predicting PET activity for other tissues, like tumors, identifying and thoroughly investigating comparable, specific CT signatures is essential.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have revised the manuscript and answered adequately to my comments. I have no further comment.

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for considering my comments. I believe that all my questions have been answered.